CHEMISTRY BEFORE COMPLEXITY: REPRESENTATION-FIRST MOLECULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular machine learning has increasingly focused on expressive architectures, yet classical fingerprints remain highly competitive with modern graph, language, and geometric models. We argue that *molecular representation remains a central bottleneck*: if chemically relevant information is not effectively exposed, additional model capacity may not recover it.

We introduce two chemically informed molecular fingerprints. **AtomGrid** encodes atomic-scale covalent, electrostatic, and dispersion interactions in a fixed coordinate system of chemical environments, while **GroupGrid** captures functional-scale organization through group multiplicity, topological separation, and charge complementarity. We study them together with ECFP, AtomPair, and MACCS as five complementary molecular views, evaluating each independently and combining only their prediction scores through a lightweight gated fusion model.

On ligand-based virtual screening with hard negatives, fixed fingerprints outperform substantially more complex learned encoders, with AtomGrid among the strongest standalone methods and the five-fingerprint fusion achieving the best overall performance. On molecular property prediction, the same fusion achieves the best aggregate performance across six OGB benchmarks, despite the proposed fingerprints being weaker standalone. These results suggest that future molecular learning should place greater emphasis on *chemistry-aware representation learning*, not only on increasing downstream model complexity.

1 INTRODUCTION

Molecular machine learning has become increasingly architecture driven. Graph neural networks, molecular language models, geometric networks, and large-scale pretraining have steadily increased the capacity of models to learn from molecular graphs, strings, and three-dimensional structures. Yet a striking empirical fact persists: *simple, carefully designed molecular fingerprints remain remarkably difficult to outperform*. Large-scale studies find classical fingerprints competitive with, and often stronger than, substantially more complex learned representations, particularly in the small- and moderate-data regimes common in molecular prediction (Jiang et al., 2021; Deng et al., 2023; Praski et al., 2025). Similar behavior appears in challenging structure–activity settings, including activity cliffs and cross-scaffold generalization (van Tilborg et al., 2022; Dabrowski-Tumanski et al., 2026).

This suggests that the bottleneck may lie less in model capacity than in the molecular information exposed to the learner. A sophisticated model cannot reliably exploit chemistry that its representation obscures, whereas a strong representation can make useful structure accessible to a simple predictor. Modern deep models largely learn which molecular features matter from data and training objectives; classical fingerprints instead encode explicit, task-independent chemical inductive biases, reducing how much structure must be rediscovered from limited supervision.

Extended-connectivity fingerprints provide the canonical example (Rogers & Hahn, 2010). ECFP captures recurring local atomic environments; AtomPair represents typed atoms across topological separations; and MACCS records predefined structural motifs. Their diversity highlights an important point: molecular behavior can depend on local substructures, long-range topology, functional organization, or physical interactions, and no single representation exposes all of this information equally well.

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.

We build on this observation and introduce two fixed molecular fingerprints that extend the representational space in complementary directions. **AtomGrid** captures atomic-scale organization through continuous covalent, electrostatic, and dispersion interactions between chemically defined environments. Its coordinates specify *which* environments interact, while its values specify *how strongly*. **GroupGrid** operates at the functional scale, representing relationships between chemically meaningful units through their multiplicity, topological separation, and charge complementarity. Unlike AtomGrid, GroupGrid is entirely two-dimensional and conformer free.

We study these representations alongside classical fingerprints **ECFP**, **AtomPair**, and **MACCS**, yielding five complementary views: local atomic environments (ECFP), topological atom-pair relationships (AtomPair), predefined structural motifs (MACCS), atomic-scale physical interactions (AtomGrid), and functional-group organization (GroupGrid). Each fingerprint is first evaluated independently with a lightweight predictor. We then ask: *how much can be gained by combining strong representations without introducing a high-capacity multimodal model?* Rather than concatenating high-dimensional descriptors, we fuse only five prediction scores with a lightweight gated model, so the main source of improvement is representation diversity rather than fusion capacity.

We evaluate this representation-first approach on two complementary settings: *ligand-based virtual screening* (LBVS) and *molecular property prediction* (MPP). On TopU-LBVS with hard negatives, fixed fingerprints outperform substantially more complex learned molecular encoders; AtomGrid is among the strongest standalone methods, and the five-fingerprint fusion achieves the best overall performance. On six OGB molecular property benchmarks, AtomGrid and GroupGrid are weaker individually, yet their complementary signal improves the full portfolio, whose gated fusion attains the best aggregate performance among the compared methods. This contrast is important: a useful representation need not be the strongest standalone predictor if it exposes information missing from other molecular views.

Taken together, these results support a representation-first view of molecular learning: *better performance does not necessarily require a more complex learner; it may require a better representation of the molecule*. We do not argue for replacing learned representations with fixed fingerprints. Rather, their continued strength suggests where molecular learning should move next: toward models that learn chemically informed representations, combining the adaptability of deep learning with inductive structure that exposes relevant molecular information instead of requiring it to be rediscovered from data alone. **Our contributions are**

- We introduce **AtomGrid**, a fingerprint of covalent, electrostatic, and dispersion interactions between chemically defined atomic environments.
- We introduce **GroupGrid**, a conformer-free representation of functional-group multiplicity, topological organization, and charge complementarity.
- We develop a **representation-first molecular learning framework** combining ECFP, AtomPair, MACCS, AtomGrid, and GroupGrid through lightweight score-level gated fusion.
- Across LBVS and MPP, we show that fixed fingerprints can rival or outperform strong learned molecular models and that complementary fingerprint fusion achieves the best aggregate performance, motivating greater emphasis on chemically informed representation learning.

2 Related Work

Molecular fingerprints. Extended-connectivity fingerprints (ECFP) encode the occurrence of local atomic environments (Rogers & Hahn, 2010); atom-pair fingerprints instead record feature pairs at given topological separations (Carhart et al., 1985); MACCS keys record a fixed vocabulary of predefined substructural motifs (Durant et al., 2002); and topological pharmacophore descriptors and extended reduced graphs (ErG) encode pharmacophore-type pairs, their distances, and scaffold-hopping similarity (Schneider et al., 1999; Stiefl et al., 2006). A separate lineage encodes continuous physical interaction rather than discrete substructure, including the Coulomb matrix of nuclear-charge/distance terms (Rupp et al., 2012).

AtomGrid and GroupGrid bridge these two lineages. AtomGrid organizes covalent, electrostatic, and dispersion interactions on a fixed chemical atom-type coordinate system, giving each tensor location a stable meaning across molecules. GroupGrid extends the relational idea of atom-pair and pharma-

cophore descriptors to functional groups, summarizing each type pair by multiplicity, topological separation, and charge complementarity in a permutation-invariant tensor. Related approaches convert graph or molecular structure into fixed-size local or image-like inputs for convolution (Niepert et al., 2016; Goh et al., 2017); our representations instead use chemically labeled coordinates with fixed cross-molecule semantics.

Deep learning for molecular representation. Molecular property prediction has increasingly been pursued through expressive learned models: message-passing graph neural networks (Gilmer et al., 2017; Yang et al., 2019; Xu et al., 2019), self-supervised graph pretraining (Wang et al., 2022), and three-dimensional geometric models that incorporate geometric information such as interatomic distances and angles (Schütt et al., 2017; Klicpera et al., 2020; Zhou et al., 2023). Yet large-scale comparisons repeatedly find fixed or simpler molecular representations competitive with substantially more complex graph models (Jiang et al., 2021; Deng et al., 2023; Praski et al., 2025; Boulougouri et al., 2024). This tension becomes particularly important when generalizing beyond familiar chemical space, including activity cliffs, cross-scaffold transfer, and structurally novel molecules (van Tilborg et al., 2022; Dabrowski-Tumanski et al., 2026; van Tilborg et al., 2026).

We take this literature as motivation rather than as a target to beat outright: rather than proposing another learned encoder, we ask whether fixed molecular representations can be extended in directions existing fingerprints leave relatively implicit—continuous atomic-scale physical interactions (AtomGrid) and functional-scale relational organization (GroupGrid).

Functional-group and multi-resolution models. A separate line of work brings functional-group and motif structure into the learned setting itself, including hierarchical motif-level graph networks (Zang et al., 2023), functional-group-aware pretraining (Nguyen et al., 2024; Li et al., 2023), and functional-group-level string encodings (Cheng et al., 2023). Most closely related, MMGX evaluates several reduced molecular graphs, atom-level, pharmacophore, junction-tree, and functional-group, within a shared message-passing architecture (Kengkanna & Ohue, 2024; Jin et al., 2018). Treating functional groups as first-class objects and comparing multiple molecular resolutions is therefore *not* our contribution; MMGX already establishes both. What distinguishes AtomGrid and GroupGrid is where the inductive bias is placed: MMGX pushes each reduced graph through a learned message-passing encoder that must still discover how atomic or functional relationships combine, whereas AtomGrid and GroupGrid are fixed tensors, computed once and read out by a lightweight, non-message-passing model.

Fusing molecular representations. Our fusion stage is closest in spirit to *data fusion* in ligand-based virtual screening, where rankings or scores from multiple similarity measures or representations are combined, typically by fixed rules such as sum, max, or weighted averages, and can improve retrieval effectiveness over individual constituent measures (Willett, 2006; 2013). That literature motivates the premise of our fusion experiment, that diverse fixed representations carry complementary evidence, while classical data-fusion approaches primarily use fixed combination rules operating on rankings or similarity scores. We instead treat fusion as a supervised learning problem over prediction scores, in the spirit of stacked generalization (Wolpert, 1992): a lightweight model learns how to weight and combine representations from data rather than by a fixed rule. We compare several such learned combiners, including an explicit quadratic stacker over pairwise score interactions, and find that a gated combination, which learns molecule-dependent trust in each representation rather than a fixed pairwise interaction term, performs best. Because the fusion model receives only the five scalar scores of our portfolio rather than raw descriptors, it remains close in spirit to classical rank-level data fusion while learning representation-specific weights end to end.

Benchmarks for LBVS and molecular property prediction. Virtual-screening benchmarks are commonly built around fingerprint-similarity retrieval and scaffold hopping (Riniker & Landrum, 2013; Stiefl et al., 2006), a setting in which local-fragment similarity is known to fail precisely when structurally dissimilar actives share functional organization, the same regime that defines activity cliffs (van Tilborg et al., 2022). We evaluate on TopU-LBVS (Kumar et al., 2026), which constructs hard decoys for exactly this reason. Property prediction is commonly evaluated on standardized suites such as MoleculeNet (Wu et al., 2018) and OGB (Hu et al., 2020), and increasingly on broad multi-model comparisons spanning fixed and learned representations (Deng et al., 2023; Praski et al., 2025); these comparisons motivate our evaluation of AtomGrid and GroupGrid against both families of baseline, on both tasks, rather than on a single benchmark.

3 METHODOLOGY

We take a representation-first view of molecular learning: before increasing predictor complexity, we ask what molecular information should be exposed to the learner. Classical fingerprints already encode chemically meaningful structure, including local environments, motifs, and topological relationships, before learning (Appendix A). We extend this idea with two new representations: **AtomGrid**, capturing atomic-scale physical interactions, and **GroupGrid**, capturing functional-scale organization.

Together with ECFP, AtomPair, and MACCS, they form five complementary molecular views spanning local environments, atom-pair topology, structural motifs, physical interactions, and functional organization. Each representation is first evaluated independently with a lightweight predictor; only then are the five scalar predictions combined by a small gated model. This design keeps the focus on improving molecular representation rather than increasing downstream model capacity.

3.1 FIVE COMPLEMENTARY MOLECULAR VIEWS

For molecule M, let $\Phi(M) = \{\phi_{\mathrm{ECFP}}, \phi_{\mathrm{AP}}, \phi_{\mathrm{MACCS}}, \phi_{\mathrm{Atom}}, \phi_{\mathrm{Group}}\}$. ECFP captures recurring local atomic environments, AtomPair encodes chemically typed atoms across graph distances, and MACCS records predefined structural motifs; we review these standard fingerprints in Appendix A.3.

These views remain incomplete in two important ways: they do not explicitly describe *physical interactions* between atomic environments or *how functional units are organized* across a molecule. AtomGrid and GroupGrid are designed to expose these two forms of structure directly.

We do not concatenate the five high-dimensional fingerprints. Each first produces an independent task prediction, and fusion operates only on the five scores. This deliberate bottleneck helps ensure that gains from fusion primarily reflect complementary molecular information rather than the capacity of a large multimodal encoder.

3.2 ATOMGRID: ATOMIC-SCALE PHYSICAL INTERACTION FINGERPRINT

Most conventional fingerprints describe which substructures or graph relationships occur. AtomGrid asks a different question: *Which chemical environments physically interact, and how strongly?*

Let $\mathcal{V}(M) = \{v_1, \ldots, v_n\}$ denote the atoms of molecule M. For each distinct pair (v_i, v_j), we construct three interaction channels.

Physical interaction channels. The covalent channel uses Wiberg bond order, $W_{ij}^{\mathrm{cov}} = \mathrm{WBO}(v_i, v_j)$, providing a continuous measure of electron sharing beyond binary bond connectivity. The electrostatic channel is $W_{ij}^{\mathrm{elec}} = q_i q_j / r_{ij}$, up to the corresponding physical constant, coupling molecular geometry with partial-charge distribution. The third channel is the pairwise dispersion contribution, $W_{ij}^{\mathrm{disp}} = E^{\mathrm{disp}}(v_i, v_j)$. We retain the three channels separately because they represent distinct physical mechanisms. A raw $n \times n$ interaction matrix is not yet a useful fingerprint: it is indexed by arbitrary atom order, so its coordinates have no consistent meaning across molecules. AtomGrid therefore replaces atom index with a *chemically aligned coordinate system*.

Chemically aligned coordinates. Each atom is assigned to one of K coarse chemical environments $\mathcal{A} = \{A_1, \ldots, A_K\}$, including major carbon, nitrogen, oxygen, sulfur, phosphorus, and halogen environments. The full typing scheme is given in Appendix B.3.

Each environment A_α receives a fixed axis budget b_α. Within a type, atoms are ordered by electrotopological-state (E-state) value, with canonical atom rank used only to break ties. The resulting global axis has fixed size $B = \sum_{\alpha=1}^{K} b_\alpha$ and is independent of the original atom ordering.

The three interaction matrices are mapped into this shared coordinate system, producing $X_{\mathrm{Atom}}(M) \in \mathbb{R}^{B \times B \times 3}$. A given region of the tensor therefore always refers to the same pair of chemical environments. AtomGrid separates

coordinate: which environments interact from *value: how strongly they interact*,

while remaining permutation invariant.

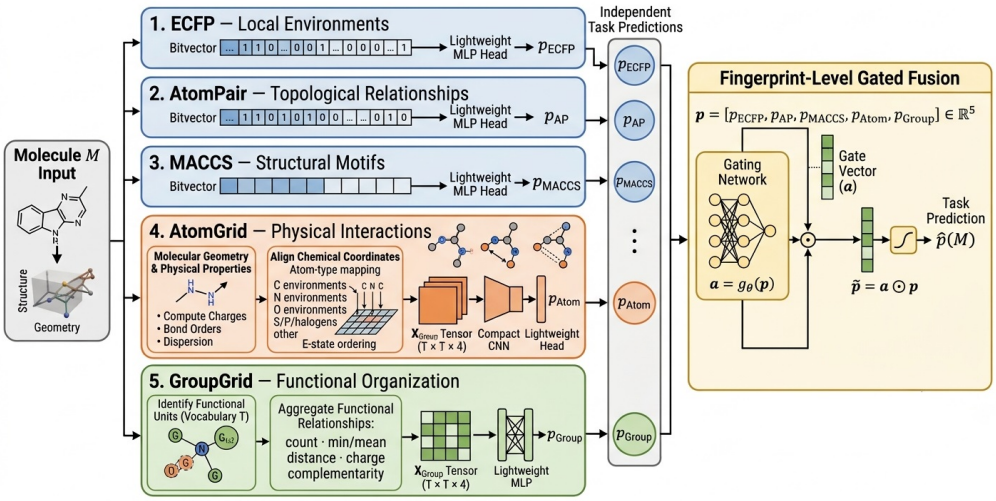


Figure 1: **Representation-first molecular learning framework.** Five complementary fingerprints encode local environments, topological relationships, structural motifs, physical interactions, and functional organization. Each produces an independent task score, and a lightweight gated model combines only these five scores for the final prediction.

Because zero may indicate either a true zero interaction or an unoccupied coordinate, we retain a binary support mask. An atom-type composition vector containing both counts and relative fractions is also retained. Support handling, normalization, axis budgets, and other implementation details are given in Appendix B.3.

AtomGrid readout. The fixed two-dimensional chemical layout allows a compact CNN to extract patterns across the three interaction channels, followed by a lightweight prediction head. Importantly, the CNN does not learn the molecular representation from raw atoms: chemical typing, coordinate alignment, and physical interaction channels are constructed before learning, and the CNN only learns how these predefined patterns relate to the task.

AtomGrid is more expensive to construct than conventional 2D fingerprints because it requires molecular geometry and additional interaction quantities, but these are computed once and cached. This motivates our second representation, which captures complementary structure while remaining fully two-dimensional.

3.3 GroupGrid: Functional-Scale Relational Fingerprint

AtomGrid operates at atomic resolution. GroupGrid asks a coarser question: *Which functional units are present, and how are they organized?*

This matters because molecules with similar local substructures can arrange their functional units differently, while distinct scaffolds may preserve similar functional organization. GroupGrid captures this relational structure directly and requires no molecular conformer.

Functional units and topology. We define a fixed vocabulary $\mathcal{T} = \{T_1, \ldots, T_T\}$ of coarse functional and pharmacophoric units, including aromatic and heteroaromatic rings, hydrophobic units, hydroxyls, ethers, carbonyls, carboxyl groups, esters, amines, amides, nitriles, sulfonyl-containing groups, charged centers, and halogens. For molecule M, let $\mathcal{G}(M) = \{g_1, \ldots, g_m\}$ denote the detected group instances. Overlapping patterns are resolved deterministically, with chemically specific groups assigned before more generic constituents; full details are given in Appendix B.3.

For group instances g_i and g_j, their topological separation is $d_{\text{top}}(g_i, g_j) = \min_{u \in g_i, v \in g_j} d_G(u, v)$, capturing relative placement along the molecular graph without conformer dependence.

Charge complementarity. Topology alone does not capture electronic character. For each group g_i, we define $Q_i^+ = \sum_{a \in g_i} \max(q_a, 0)$ and $Q_i^- = \sum_{a \in g_i} \max(-q_a, 0)$, and measure charge complementarity by $C(g_i, g_j) = Q_i^+ Q_j^- + Q_i^- Q_j^+$. This is not an electrostatic energy; it is a conformer-free measure of opposite-sign charge compatibility between functional units.

From group instances to a fixed fingerprint. Rather than ordering individual groups, GroupGrid aggregates all relationships for each functional-type pair (T_α, T_β): $x_{\alpha\beta} = \left[N_{\alpha\beta}, d_{\alpha\beta}^{\min}, d_{\alpha\beta}^{\text{mean}}, C_{\alpha\beta}\right]$, where $N_{\alpha\beta}$ is the number of group-instance pairs, $d_{\alpha\beta}^{\min}$ and $d_{\alpha\beta}^{\text{mean}}$ are their minimum and average topological separations, and $C_{\alpha\beta}$ is their average charge complementarity. Same-type entries use distinct unordered group instances.

Stacking these descriptors over the fixed vocabulary yields $X_{\text{Group}}(M) \in \mathbb{R}^{T\times T\times 4}$. Each cell therefore has a fixed chemical meaning across molecules; for example, the amine–aromatic cell always summarizes the relationship between amine and aromatic units. No arbitrary ordering, resizing, or learned pooling is needed, and the representation is permutation invariant.

We also retain a functional-group composition vector, separating *what functional units occur* from *how they are related*. GroupGrid is processed by a lightweight MLP, so the downstream model learns only how these pre-organized chemical relationships map to the task.

3.4 Fingerprint-Level Gated Fusion

Each representation is first evaluated independently. For $r \in \text{ECFP}, \text{AP}, \text{MACCS}, \text{Atom}, \text{Group}$, a standalone predictor outputs $p_r(M) = f_r(\phi_r(M))$. The predictor f_r is chosen according to the evaluation setting, with task-specific architectures and training details given in Appendix B.5

To test whether the representations provide complementary evidence, we fuse only their task scores rather than their high-dimensional descriptors:

$$p(M) = [p_{\text{ECFP}}, p_{\text{AP}}, p_{\text{MACCS}}, p_{\text{Atom}}, p_{\text{Group}}] \in \mathbb{R}^5.$$

This bottleneck prevents the fusion model from learning a new molecular representation from raw structure; it can only combine evidence supplied by the five precomputed views.

Gated representation weighting. Because different views may be useful for different molecules, a small gating network learns molecule-dependent contributions from the representation-level prediction scores. The resulting gated combination allows the contribution of each representation to vary across molecules while operating only on score-level information rather than on the underlying molecular descriptors. The task-specific gate formulations are given in Appendix B.5.

Training protocol. Fusion is trained on out-of-fold predictions from the base models using standard cross-fitted stacking, so each training molecule is scored only by models fit without it. The base predictors are then refit on the full training split to generate test predictions. We denote gated fusion of ECFP, AtomPair, and MACCS by **3FP-Gated**, and fusion of all five representations by **5FP-Gated**. Their comparison tests whether AtomGrid and GroupGrid add information beyond conventional fingerprints. Alternative fusion strategies, including averaging, MLP-based fusion, standardized-score fusion, and quadratic stacking, are reported in Appendix C.1.

4 Experiments

We evaluate our representation-first hypothesis in two complementary settings. First, TopU-LBVS (Section 4.1) tests ligand-based virtual screening under both random and hard negatives, directly stressing whether a representation captures activity-relevant chemical structure. Second, six OGB molecular property benchmarks (Section 4.2) test whether the same representations remain competitive with graph neural networks and pretrained molecular models in supervised prediction.

Across both settings, we compare ECFP, AtomPair, MACCS, AtomGrid, and GroupGrid individually and through score-level fusion against strong learned baselines. Ablations isolate the classifier head and fusion mechanism, with 5FP-Gated emerging as the default configuration. Our code is available at `https://anonymous.4open.science/r/MolGrid-DFC5`

4.1 Ligand-Based Virtual Screening

Benchmark and protocol. Our primary evaluation uses TopU-LBVS (Kumar et al., 2026), where the representation itself plays a particularly direct role: given known actives, the goal is to rank a screening library so that additional active compounds appear near the top. We evaluate seven targets spanning diverse protein classes—`aa2ar`, `aces`, `cp3a4`, `egfr`, `esr1`, `kcnh2`, and `thrb`—under

Table 1: **LBVS EF@1% results on TopU-LBVS**, TopU hard-negative split (left) and random split (right), each with average rank and average EF@1% across the seven targets. 3FP-Gated fuses ECFP, AtomPair, and MACCS; 5FP-Gated adds AtomGrid and GroupGrid. **Best**, **second**, and **third** scores in each column are marked. Full results with standard deviations are given in Table 10.

	TopU (hard negatives)									Random split								
Model	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb	**Rank**	**Avg.**	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb	**Rank**	**Avg.**
ECFP	5.4	**11.0**	13.9	6.2	5.8	**5.1**	**9.7**	**4.7**	**8.2**	25.4	28.5	16.4	**31.4**	24.0	19.7	29.8	6.7	25.0
Tanimoto	3.4	9.4	11.5	6.3	**7.4**	5.0	**8.1**	6.5	7.3	17.9	21.1	10.8	19.5	16.0	12.3	23.5	14.3	17.3
AtomPair	5.9	8.7	14.1	7.5	5.8	4.3	7.8	**5.0**	7.7	25.3	27.3	14.5	30.2	**24.7**	**20.1**	31.1	6.8	24.7
MACCS	4.3	**9.6**	12.2	**8.2**	6.5	5.0	6.0	5.4	7.4	19.1	29.9	11.7	25.6	24.4	17.8	29.2	10.1	22.5
GAT	1.8	2.6	8.3	2.2	4.8	1.9	1.6	15.1	3.3	11.3	16.8	7.0	18.2	15.2	8.4	19.4	16.3	13.8
GAT-FP	3.9	5.7	**15.4**	3.3	2.7	2.8	3.2	10.2	5.3	25.8	29.9	16.4	30.1	23.2	18.9	30.2	7.1	24.9
GIN	3.4	3.1	8.5	3.3	1.8	1.1	3.2	14.4	3.5	13.2	18.8	7.3	23.9	16.3	9.0	22.2	14.9	15.8
GIN-FP	4.1	5.4	13.8	4.3	4.2	2.7	1.9	10.6	5.2	25.5	29.2	15.7	30.3	23.5	18.7	**31.8**	7.1	25.0
GPS	1.5	2.0	5.6	1.3	1.8	2.2	3.2	15.9	2.5	12.4	14.2	4.5	14.1	9.8	9.2	13.0	16.6	11.0
GPS-FP	2.8	4.6	13.9	3.5	4.2	2.9	2.9	11.4	5.0	25.0	29.6	**16.7**	30.8	**24.7**	19.7	31.1	**5.3**	**25.4**
DMPNN	3.1	6.5	12.9	7.4	5.9	**9.1**	4.5	7.3	7.1	26.2	29.1	14.8	28.0	22.6	20.1	29.6	7.9	24.3
MolFormer	2.8	3.4	13.1	3.5	2.7	3.8	5.7	11.4	5.0	23.0	27.3	15.2	26.4	19.6	19.1	27.7	10.6	22.6
UniMol2	2.7	4.5	13.4	3.9	2.0	**7.1**	4.8	10.4	5.5	**27.9**	29.2	14.4	25.1	23.5	18.9	29.2	9.1	24.0
GroupGrid	**6.1**	5.0	13.5	**8.4**	4.8	3.0	3.7	7.5	6.3	22.4	**30.5**	16.0	27.8	**24.7**	**20.5**	**32.1**	**5.3**	24.9
AtomGrid	**7.8**	9.4	13.0	7.1	**6.8**	4.1	6.5	5.3	**7.8**	22.2	27.3	14.2	24.0	20.7	20.0	31.1	10.3	22.8
3FP-Fusion	3.9	5.1	**17.3**	6.5	4.2	3.4	2.5	8.9	6.1	**27.1**	**32.5**	**16.9**	**31.5**	**25.0**	19.3	30.2	**3.6**	**26.0**
5FP-Fusion	**7.7**	**11.7**	**14.3**	**8.0**	**6.7**	4.8	**8.0**	**3.0**	**8.8**	**29.7**	**33.8**	**16.9**	**33.0**	**27.5**	**21.4**	**32.8**	**1.1**	**27.9**

both TopU's hard-negative protocol and a conventional random-decoy split. The hard negatives are designed to be substantially more difficult to distinguish from actives than random decoys, reducing the usefulness of superficial physicochemical shortcuts and placing greater pressure on the molecular representation. We report EF@1%, which measures enrichment of actives within the top 1% of the ranked library; complete benchmark and metric definitions are given in Appendix A.1 and B.1.

We compare ECFP, AtomPair, MACCS, AtomGrid, and GroupGrid against fingerprint similarity, GAT, GIN, GPS, fingerprint-augmented variants of these GNNs, D-MPNN (Yang et al., 2019), MolFormer (Ross et al., 2022), and UniMol2. We additionally compare **3FP-Gated**, which fuses ECFP, AtomPair, and MACCS, with **5FP-Gated**, which adds AtomGrid and GroupGrid. Full per-target results and standard deviations are reported in Appendix D.1.

Fixed representations remain remarkably strong. The hard-negative setting produces the clearest evidence for our central claim. The six best average ranks are all obtained by fixed fingerprints or their fusion: 5FP-Gated, ECFP, AtomPair, AtomGrid, MACCS, and Tanimoto retrieval. The strongest learned baseline, D-MPNN, appears only after these six, while GAT, GIN, GPS, MolFormer, and UniMol2 rank substantially lower. Thus, increasing architectural sophistication does not automatically overcome the advantage of a chemically effective representation. Importantly, this pattern becomes *more*, not less, pronounced when easy random decoys are replaced by TopU hard negatives, precisely where superficial molecular cues are less useful.

The proposed fingerprints are useful representations on their own. AtomGrid is particularly strong in the hard-negative regime, reaching an average EF@1% of 7.8, third among all methods by average enrichment and ahead of every learned baseline. It also obtains the best result on `aa2ar`. GroupGrid, despite requiring only two-dimensional molecular information, achieves the best score of any method on `egfr`. These results are important because AtomGrid and GroupGrid are not useful merely as auxiliary features in an ensemble: the physical-interaction and functional-relational information they expose is itself predictive of bioactivity.

Complementary representations, rather than more fusion capacity, drive the largest gain. The comparison between 3FP-Gated and 5FP-Gated isolates the contribution of our two representations. On the hard-negative split, combining only ECFP, AtomPair, and MACCS yields an average EF@1% of 6.1 and average rank 8.9, providing little benefit over the strongest individual classical fingerprints. Adding AtomGrid and GroupGrid under the *same fusion mechanism* raises average EF@1% to 8.8 and improves average rank to 3.0, making 5FP-Gated the strongest method overall. The gain is therefore not explained simply by learning a better combination rule: it appears when the representation portfolio is expanded with chemically distinct physical and functional views.

The random-decoy setting shows the same conclusion from a complementary direction. Here many methods obtain much higher absolute enrichment because the screening problem is easier, yet 5FP-Gated remains strongest overall, with average EF@1% 27.9 and the best average rank, and achieves the top score on six of seven targets. Together, the two regimes support the same picture:

Table 2: **Molecular Property Prediction.** Performance comparison across OGB molecular property prediction datasets, together with the five fingerprints (evaluated standalone with MLP) and their gated score-level fusion. AvgAUC, AvgGap, and AvgRank are computed per method on the six datasets. See Table 11 for full results.

Family	Method # Compounds	SIDER 1427	ClinTox 1478	BACE 1513	BBBP 2039	TOX21 7831	HIV 41127	Avg. AUC↑	Avg. Gap↓	Avg. Rank↓
GNN	GCN	58.2	75.6	71.9	65.6	74.7	72.9	69.8	9.8	12.7
	GIN	57.7	76.1	72.4	65.8	75.6	76.8	70.7	8.9	10.7
	GSAGE	59.0	81.9	72.2	64.0	75.2	73.0	70.9	8.7	10.7
	GAT	59.7	84.1	72.9	66.3	**76.0**	75.4	72.4	7.2	7.8
	GPS	59.5	76.1	67.5	65.1	71.0	70.4	68.3	11.3	13.5
	SGFormer	58.3	78.1	66.2	66.1	73.0	72.2	69.0	10.6	12.7
Pretrained	MolBERT	64.0	**87.2**	81.9	**74.3**	**76.4**	77.9	**77.0**	**2.7**	**3.5**
	ChemBERTa	**69.2**	73.9	**82.6**	**72.1**	**78.9**	**78.7**	75.9	3.7	4.0
	MolFormer	**66.6**	**88.4**	**83.7**	**71.5**	75.0	**78.0**	**77.2**	**2.4**	**3.5**
	Uni-Mol	62.2	82.4	**82.6**	66.1	72.5	76.1	73.7	6.0	7.7
	GraphMVP	**67.3**	78.9	81.8	68.7	76.0	70.2	73.8	5.8	7.0
Fingerprint	ECFP	62.8	85.5	79.0	66.2	74.3	73.0	73.5	6.1	8.0
	AtomPair	60.6	66.7	80.8	70.0	69.5	68.2	69.3	10.3	11.2
	MACCS	64.1	86.1	73.9	69.8	70.5	70.2	72.4	7.2	8.8
	AtomGrid	60.2	59.5	75.5	63.7	67.1	71.0	66.2	13.4	13.3
	GroupGrid	60.0	78.5	70.1	59.4	64.1	61.2	65.5	14.1	14.5
	5FP-Fusion	64.3	**92.0**	**84.8**	70.0	74.9	**79.4**	**77.4**	**2.2**	**3.2**

strong molecular representations remain highly competitive with modern learned encoders, and the largest gains come from adding *complementary chemical information*, not simply from increasing the capacity of the downstream learner. Fusion-strategy ablations in Appendix C further confirm that the advantage of the five-fingerprint portfolio is not specific to a single combination rule.

4.2 Molecular Property Prediction

Benchmark and protocol. We next ask whether the representation-first picture observed in LBVS extends to standard supervised molecular prediction. We evaluate on six OGB molecular classification benchmarks (SIDER, ClinTox, BACE, BBBP, TOX21, and HIV) using the standard scaffold splits (Table 2). The datasets span approximately 1,400 to 41,000 molecules and cover toxicity, side effects, blood–brain-barrier penetration, and bioactivity endpoints. We report AUROC together with mean AUROC, mean gap to the best method, and mean rank across datasets; full dataset and metric details are provided in Appendix A.2 and B.2. We compare the five fingerprints against GCN, GIN, GraphSAGE, GAT, GPS, and SGFormer trained directly on the molecular graphs, as well as pretrained molecular models including MolBERT, ChemBERTa, MolFormer, Uni-Mol, and GraphMVP. This provides a complementary test to LBVS: rather than retrieval from known actives, MPP asks whether fixed chemically motivated representations remain useful when competing directly with representations learned end-to-end or through large-scale pretraining.

Classical fingerprints remain competitive with learned graph representations. Even in this supervised setting, greater architectural complexity does not guarantee a better molecular representation. ECFP alone reaches 73.5 mean AUROC, exceeding every GNN evaluated; the strongest, GAT, reaches 72.4, with MACCS matching it. Pretraining closes much of this gap (MolFormer and MolBERT reach 77.2 and 77.0) showing the value of additional molecular data. Still, simple fixed fingerprints remain remarkably strong relative to their model complexity.

Complementary representations outperform any single view. The full **5FP-Gated** model achieves the best aggregate performance among all compared methods: 77.4 mean AUROC, the smallest mean gap to the per-dataset best, and the best mean rank. It is the top-performing method on ClinTox, BACE, and HIV, while remaining competitive on the other three datasets. Its advantage over the strongest pretrained models is modest (77.4 versus 77.2 for MolFormer), but this is precisely the relevant comparison: five fixed molecular views combined only through their scalar predictions match or exceed models whose molecular representations are learned with substantially greater architectural and pretraining complexity.

Representation quality is not the same as standalone accuracy. Unlike in LBVS, AtomGrid and GroupGrid are relatively weak when used alone on MPP. Yet adding them to the classical fingerprints improves the final five-view model. This is an important part of our representation-first argument: a molecular view can be valuable not because it is universally the best standalone descriptor, but because it exposes information missing from stronger representations. AtomGrid contributes physical-interaction structure, while GroupGrid contributes functional organization; the gated model can use these signals selectively alongside ECFP, AtomPair, and MACCS.

Thus, LBVS and MPP reveal two complementary regimes. In LBVS, chemically designed fingerprints can directly outperform learned encoders; in MPP, pretrained models are much stronger, yet a lightweight fusion of complementary fixed representations remains competitive with the best of them and achieves the strongest aggregate result. Together, these findings reinforce the same conclusion: improving the molecular information presented to the learner can be as important as increasing downstream model capacity. Classifier-head and fusion ablations are reported in Appendix C.

4.3 Ablation studies

We study three design questions: fusion robustness (App. C.1), the individual contributions of AtomGrid and GroupGrid (App. C.2), and the standalone classifier head (App. C.3). Here, we summarize the first two, which most directly test our representation-first hypothesis.

Table 3: **Fusion strategy ablation**, average rank and average EF@1% across the seven TopU-LBVS targets, hard-negative and random splits.

	TopU (hard neg.)		Random split	
Method	**Rank↓**	**Avg.↑**	**Rank↓**	**Avg.↑**
3FP-average	6.9	6.1	6.6	26.1
3FP-prob-raw	7.6	5.9	6.9	26.0
3FP-prob-z	7.4	5.9	4.8	26.6
3FP-Quadratic	7.4	6.0	4.8	26.6
3FP-Gated	7.1	6.1	6.9	26.0
5FP-average	5.3	7.5	7.6	25.4
5FP-prob-raw	4.1	7.9	6.1	25.9
5FP-prob-z	4.6	8.4	5.1	26.6
5FP-Quadratic	2.4	**8.8**	4.1	27.1
5FP-Gated	**2.1**	**8.8**	**2.1**	**27.9**

Fusion robustness. We compare five score-level fusion mechanisms, from unweighted averaging to learned gating (Appendix B.4). On TopU hard negatives, expanding from the three classical fingerprints to all five representations improves performance under *every* fusion rule: average EF@1% rises from 5.9–6.1 for 3FP to 7.5–8.8 for 5FP. **Even simple 5FP averaging** (7.5) **exceeds the best 3FP fusion** (6.1), showing that the dominant gain comes from richer molecular information rather than fusion capacity alone. Among five-view models, 5FP-Gated gives the strongest overall results across LBVS and MPP, and we therefore use it as the default configuration. For further details, see Appendix C.1.

Table 4: **AtomGrid and GroupGrid contribution:** Individual contributions of the proposed fingerprints.

	TopU (hard neg.)		Random split	
Method	**Rank↓**	**Avg.↑**	**Rank↓**	**Avg.↑**
3FP-Gated	3.9	6.1	3.4	26.1
3FP+AtomGrid	2.7	8.2	2.9	26.7
3FP+GroupGrid	2.0	8.7	2.6	26.9
5FP-Gated	**1.4**	**8.8**	**1.1**	**27.9**

Contribution of AtomGrid and GroupGrid. To isolate the proposed representations from the fusion mechanism, we keep gated fusion fixed and compare 3FP-Gated with 3FP+AtomGrid, 3FP+GroupGrid, and the full 5FP-Gated model (Appendix C.1). On TopU hard negatives, average EF@1% increases from 6.13 for 3FP-Gated to 8.22 with AtomGrid and 8.72 with GroupGrid, reaching 8.8 when both are included; the corresponding random-split averages are 26.1, 26.7, 26.9, and 27.9. Thus, both representations contribute beyond the classical fingerprint portfolio, with their combination giving the strongest overall result. This controlled ablation supports the intended complementarity of atomic-scale physical interactions and functional-scale organization; full per-target results and discussion are provided in Appendix C.2.

5 Conclusion

We introduced **AtomGrid** and **GroupGrid**, two chemically informed molecular fingerprints that expose complementary atomic- and functional-scale structure beyond conventional representations. Together with ECFP, AtomPair, and MACCS, they form a diverse set of molecular views whose lightweight fusion performs strongly across both molecular property prediction and ligand-based virtual screening. More broadly, our results support a representation-first view of molecular learning: strong fixed fingerprints remain competitive with, and in some settings outperform, substantially more complex learned encoders. This suggests that improving how molecular information is represented can be as important as increasing model capacity.

The next step is not to replace learned representations with fixed fingerprints, but to learn *better representations*. Future work will develop deep models that incorporate stronger chemical inductive structure directly into representation learning, combining the adaptability of neural models with explicit atomic interactions, functional relationships, and multiscale molecular organization.

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 representation designs, methodological choices, experimental protocols, and scientific interpretations were developed and evaluated 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 scientific claims or determine the validity of the reported results. The authors take full responsibility for the final manuscript, code, experiments, and conclusions.

ETHICS STATEMENT

This work studies molecular representation learning for ligand-based virtual screening and molecular property prediction using publicly available research benchmarks. It does not involve human subjects, personally identifiable information, private clinical data, or collection of new biological data.

The proposed representations could ultimately be used in applications such as drug discovery, toxicity prediction, or molecular screening. Predictions from computational models should not be treated as experimental or clinical evidence without appropriate domain-specific validation. In particular, deployment in safety-critical biomedical or molecular-design settings would require independent experimental verification and appropriate safeguards. We are not aware of additional ethical concerns specific to the methodological and benchmark experiments presented here.

REPRODUCIBILITY STATEMENT

We provide implementation code and experiment scripts for constructing AtomGrid and GroupGrid, training the standalone fingerprint predictors, and performing score-level fusion. The release includes the preprocessing required for AtomGrid's chemical typing, E-state ordering, physical interaction channels, support handling, and normalization; the functional-group detection, topological-distance, and charge-complementarity calculations used by GroupGrid; and the training and evaluation code for the gated fusion model and its ablations.

The main paper describes the representation construction and learning framework, while Appendix B provides additional implementation details. Fusion is trained using out-of-fold base-model predictions to prevent leakage, and the same predefined benchmark splits and evaluation protocols are used across compared methods. Alternative fusion strategies and related ablations are reported in Appendix C.

Together, the released code, preprocessing procedures, fixed evaluation protocols, and appendix details are intended to enable reproduction of the standalone fingerprint results, fusion experiments, and reported ablations. Our code is available at https://anonymous.4open.science/r/MolGrid-DFC5

REFERENCES

Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, 2024.

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.

Austin H. Cheng, Andy Cai, Santiago Miret, Gustavo Malkomes, Mariano Phielipp, and Alán Aspuru-Guzik. Group SELFIES: A robust fragment-based molecular string representation. *Digital Discovery*, 2(3):748–758, 2023.

Pawel Dabrowski-Tumanski, Bartosz Topolski, Dariusz Plewczynski, and Tomasz Jetka. The geometry of activity cliffs: Representation dependence and multi-scale characterization of activity landscapes. *arXiv preprint arXiv:2605.30831*, 2026.

Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395, 2023.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42 (6):1273–1280, 2002.

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)*, pp. 1263–1272, 2017.

Garrett B. Goh, Charles Siegel, Abhinav Vishnu, Nathan O. Hodas, and Nathan Baker. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689*, 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.

Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):12, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2323–2332, 2018.

Apakorn Kengkanna and Masahito Ohue. Enhancing property and activity prediction and interpretation using multiple molecular graph representations with MMGX. *Communications Chemistry*, 7 (1):74, 2024.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020. Author later published as Johannes Gasteiger.

Surbhi Kumar, Yuhe Zhou, Varun Shiralkar, Niu Huang, and Baris Coskunuzer. Topu-lbvs: A realistic multi target benchmark for ligand based virtual screening. *arXiv preprint arXiv:2609.29740*, 2026.

Biaoshun Li, Mujie Lin, Tiegen Chen, and Ling Wang. FG-BERT: A generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics*, 24(6):bbad398, 2023.

Thao Nguyen, Kuan-Hao Huang, Ge Liu, Martin D. Burke, Ying Diao, and Heng Ji. FARM: Functional group-aware representations for small molecules. *arXiv preprint arXiv:2410.02082*, 2024.

Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2014–2023, 2016.

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

Sereina Riniker and Gregory A. Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26, 2013.

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

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.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012.

Gisbert Schneider, Werner Neidhart, Thomas Giller, and Gerard Schmid. "scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie International Edition*, 38(19):2894–2896, 1999.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Nikolaus Stiefl, Ian A Watson, Knut Baumann, and Andrea Zaliani. Erg: 2d pharmacophore descriptions for scaffold hopping. *Journal of chemical information and modeling*, 46(1):208–220, 2006.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23): 5938–5951, 2022.

Derek van Tilborg, Luke Rossen, and Francesca Grisoni. Molecular deep learning at the edge of chemical space. *Nature Machine Intelligence*, pp. 1–13, 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.

Peter Willett. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR & Combinatorial Science*, 25(12):1143–1152, 2006.

Peter Willett. Combination of similarity rankings using data fusion. *Journal of Chemical Information and Modeling*, 53(1):1–10, 2013.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

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.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.

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 Basics of Molecular Learning **13**

A.1 Ligand-based virtual screening (LBVS) 13

A.2 Molecular property prediction (MPP) 14

A.3 Classical Molecular Fingerprints 14

B Experimental Setup **15**

B.1 LBVS Dataset: TopU-LBVS-mini 16

B.2 MPP Dataset: OGBG-MOL 16

B.3 AtomGrid and GroupGrid: Implementation Details 16

B.4 Fusion Models and Training Protocol 17

B.5 Hyperparameters and Training Protocol 18

C Ablation Studies **18**

C.1 Fusion strategies 19

C.2 Individual Contribution of AtomGrid and GroupGrid 20

C.3 Classifier head: Random Forest vs. MLP 21

D Extended Results and Discussion **21**

D.1 LBVS Results 21

D.2 Molecular Property Prediction Results 22

A BASICS OF MOLECULAR LEARNING

This appendix gives a self-contained introduction to the two evaluation settings used in this work, ligand-based virtual screening (LBVS) and molecular property prediction (MPP), the metrics used to score them, and the three classical molecular fingerprints that form part of our representation portfolio. Readers already familiar with these settings can skip directly to Appendix B, which describes this paper's specific datasets, splits, and implementation choices.

A.1 LIGAND-BASED VIRTUAL SCREENING (LBVS)

Task. Ligand-based virtual screening is an early-stage drug discovery task: given one or more molecules already known to be active against a biological target (e.g. a protein implicated in a disease), the goal is to search a large pool of candidate molecules and rank them so that molecules likely to also be active against that target appear near the top. Unlike *structure-based* screening, which requires a 3D structure of the target protein, LBVS relies only on the 2D or 3D structure of the molecules themselves, comparing candidates to the known actives through a molecular representation and a similarity or learned scoring function. This makes it applicable even when no crystal structure of the target is available, and makes the quality of the molecular representation the central bottleneck of the task.

A key methodological concern in LBVS is the choice of *decoys* (presumed inactive molecules used to fill out the candidate pool during evaluation). If decoys are drawn without care, a model can appear to separate actives from decoys well simply by exploiting superficial physicochemical differences (molecular weight, simple substructure counts) rather than genuine activity-relevant structure; this is a widely recognized source of benchmark inflation in LBVS evaluation (Riniker & Landrum, 2013; van Tilborg et al., 2022). Property-matched, structurally challenging decoy sets are used to mitigate this.

Metric: Enrichment Factor at $k\%$ (EF@$k\%$). A standard LBVS metric that measures how much better than random a model's top-ranked predictions are. If a screened pool contains N molecules of which N_a are active, and a model's top-ranked $k\%$ of the pool (i.e. the top $\lceil (k/100)\, N \rceil$ molecules) contains n_a actives, then

$$\text{EF@}k\% = \frac{n_a / \lceil (k/100)\, N \rceil}{N_a / N}.$$

The numerator is the fraction of actives found among the top-ranked $k\%$ of candidates; the denominator is the fraction of actives one would expect to find in a random $k\%$ sample of the pool. An EF@$k\%$ of 1.0 indicates performance no better than random ranking, while an EF@$k\%$ of, say, 10 indicates that the top-ranked $k\%$ of the model's ranking is ten times enriched for actives relative to chance. Because the actives-to-pool ratio is typically small, EF@$k\%$ can attain large values when actives are concentrated near the top of the ranking; comparisons should therefore be made across methods on the same target and pool rather than interpreted against an absolute scale. Smaller k (e.g. EF@1%) emphasizes early-recognition performance relevant to practical screening budgets, while larger k (e.g. EF@10%) is more stable when the pool contains few actives.

Metric: Precision-Recall AUC (PR-AUC). The area under the precision-recall curve, obtained by plotting precision (the fraction of predicted actives that are truly active) against recall (the fraction of true actives recovered) as the classification threshold is varied. PR-AUC ranges from the pool's active prevalence N_a/N (the score of a random ranking) to 1.0 (perfect ranking), and, unlike ROC-AUC, is sensitive to class imbalance, making it well suited to the highly imbalanced actives-versus-decoys pools typical of LBVS.

A.2 Molecular property prediction (MPP)

Task. Molecular property prediction asks a model to predict a physicochemical, biological, or toxicological property of a molecule directly from its structure, without reference to any other molecule. This is typically framed as (multi-task) binary classification: given a molecule, predict whether it exhibits a given property (e.g. toxicity in a specific assay, blood-brain-barrier permeability, or a specific side effect). Unlike LBVS, which is a *retrieval* task defined relative to a set of known actives, MPP is a *supervised prediction* task defined per molecule, and therefore tests a representation's ability to support accurate downstream prediction rather than similarity-based ranking.

Metric: Area Under the ROC Curve (AUROC). The standard classification metric for these benchmarks. AUROC measures the probability that a model ranks a randomly chosen active (positive) molecule above a randomly chosen inactive (negative) molecule, averaged over all possible classification thresholds; equivalently, it is the area under the curve obtained by plotting the true positive rate against the false positive rate as the decision threshold is varied. AUROC ranges from 0 to 1: a value of 0.5 corresponds to random ranking, and 1.0 corresponds to perfect separation of actives from inactives. For multi-task datasets, AUROC is computed per task and averaged, following standard practice.

A.3 Classical Molecular Fingerprints

Our representation portfolio contains three established molecular fingerprints, ECFP, AtomPair, and MACCS, together with the proposed AtomGrid and GroupGrid fingerprints. These classical fingerprints are intentionally chosen because they encode substantially different notions of molecular structure: local atomic environments, topological atom relationships, and predefined chemical motifs, respectively. They therefore provide both strong standalone baselines and complementary inputs to the fingerprint-level fusion model described in the main text.

Extended-Connectivity Fingerprint. Extended-connectivity fingerprints (ECFPs), commonly implemented through the Morgan algorithm, encode the local neighborhoods surrounding each atom. Each atom is first assigned an initial identifier describing its local chemical environment. These identifiers are then iteratively updated by incorporating the identifiers and bond information of neighboring atoms. After a fixed number of iterations, the resulting local substructures are mapped into a fixed-dimensional fingerprint.

Conceptually, ECFP represents a molecule through the collection

$$\phi_{\mathrm{ECFP}}(M) = \{\text{local atomic environments occurring in } M\}\,.$$

Its principal inductive bias is therefore *locality*: molecular similarity is high when two molecules contain many similar circular atomic neighborhoods.

This local environment encoding has several advantages. It is deterministic, permutation invariant, inexpensive to compute, and strongly aligned with substructure-based reasoning in medicinal chemistry. At the same time, it does not explicitly describe how distant motifs are arranged relative to one another or how their physical interactions differ (Rogers & Hahn, 2010). These limitations motivate the complementary representations considered in the main text.

Atom-Pair Fingerprint. AtomPair fingerprints describe a molecule through pairs of chemically typed atoms together with their topological separation in the molecular graph. For atoms v_i and v_j, an atom-pair feature can be written schematically as

$$(\tau(v_i), d_G(v_i, v_j), \tau(v_j))\,,$$

where $\tau(v)$ denotes the chemical atom type and $d_G(v_i, v_j)$ is their shortest-path distance in the molecular graph.

The complete fingerprint aggregates such features over atom pairs in the molecule,

$$\phi_{\mathrm{AP}}(M) = \left\{\big(\tau(v_i), d_G(v_i, v_j), \tau(v_j)\big) : i < j\right\},$$

followed by the standard fixed-dimensional fingerprint construction (Carhart et al., 1985).

AtomPair therefore complements ECFP by replacing a primarily local neighborhood view with an explicitly *relational and topological* view. Two motifs that appear similarly under local fingerprints may nevertheless produce different AtomPair patterns when they occur at different molecular separations. AtomPair remains purely structural: it records atom identities and graph distance but does not explicitly encode continuous physical interactions or relationships among higher-level functional units.

MACCS Structural Keys. MACCS keys represent a molecule using a fixed dictionary of predefined structural patterns (Durant et al., 2002). Each coordinate indicates whether a particular chemical motif or property is present in the molecule. Thus,

$$\phi_{\mathrm{MACCS}}(M) = [\mathbf{1}\{P_1 \subseteq M\}, \ldots, \mathbf{1}\{P_D \subseteq M\}],$$

where P_k denotes the structural pattern associated with key k.

Unlike ECFP and AtomPair, whose features are generated systematically from local neighborhoods or atom pairs, MACCS uses an explicitly specified chemical vocabulary. Its inductive bias is therefore *motif presence*: the fingerprint directly records whether selected chemically meaningful patterns occur, but largely discards multiplicity, organization, and detailed interaction information.

The three conventional fingerprints consequently provide complementary structural views:

ECFP	:	local atomic environments,
AtomPair	:	topological atom relationships,
MACCS	:	predefined structural motifs.

Section 3 of the main text extends this portfolio in two additional directions, AtomGrid (atomic-scale physical interaction structure) and GroupGrid (functional-scale relational organization), motivating the five-fingerprint fusion studied in Section 3.4.

B Experimental Setup

This appendix describes the datasets, splits, and implementation details specific to this paper, building on the task and metric definitions given in Appendix A.

B.1 LBVS Dataset: TopU-LBVS-mini

We evaluate on TopU-LBVS-mini, a seven-target subset of TopU-LBVS (Kumar et al., 2026) spanning distinct protein classes: `aa2ar` (GPCR), `aces` (enzyme), `cp3a4` (cytochrome P450), `egfr` (kinase), `esr1` (nuclear receptor), `kcnh2` (ion channel), and `thrb` (protease), with one representative target selected from each protein class.

We evaluate two protocols following the TopU-LBVS benchmark. In Setting 1, models are trained on ChEMBL* after removing all target-specific TopU compounds to ensure zero train–test overlap. The remaining compounds are used to construct a 1:10 active-to-decoy training pool, of which 15% is reserved as a stratified validation set and the remaining 85% is used for training. The test set is the full TopU library for each target, constructed at a 1:40 active-to-decoy ratio.

In Setting 3, active memberships are preserved from Setting 1, while ChEMBL* and TopU inactives are pooled and randomly assigned to the training, validation, and test partitions. The active-to-decoy ratios remain 1:10 for training and validation and 1:40 for testing. EF@1% is the primary test metric, and PR-AUC is used for model selection, following the definitions in Appendix A.1.

B.2 MPP Dataset: OGBG-MOL

We evaluate on six datasets from the OGBG-MOL benchmark suite (Wu et al., 2018): BACE, BBBP, ClinTox, HIV, SIDER, and TOX21, using the standard OGB scaffold split, which partitions molecules by core structural skeleton rather than at random so that structurally distinct molecules appear in training and test sets. AUROC is the primary metric, reported per dataset together with mean AUROC, mean gap to the best-performing method, and mean rank, as defined in Appendix A.2.

Table 5: **Datasets.** Number of molecules, tasks, and mean share of positive labels across tasks for MoleculeNet datasets.

Dataset	# Molecules	# Tasks	% pos.
BACE	1,513	1	45.67%
BBBP	2,039	1	76.51%
ClinTox	1,477	2	50.61%
HIV	41,084	1	3.50%
SIDER	1,427	27	56.76%
Tox21	7,831	12	7.52%

B.3 AtomGrid and GroupGrid: Implementation Details

AtomGrid. Atoms are assigned to a fixed set of coarse chemical environments (aliphatic/aromatic/carbonyl carbon, amine/amide/aromatic nitrogen, hydroxyl-ether/carbonyl oxygen, sulfur, phosphorus, halogens, and a small number of additional coarse classes), each allocated a fixed axis budget. Within each type, atoms are ordered by E-state value, with canonical atom rank used to break ties.

For each seed, the per-type axis budgets are estimated from the training split only using the empirical 0.995 quantile with the "higher" quantile rule and a minimum budget of one, and are then frozen for validation and test molecules.

For each molecule, we generate an ETKDGv3 conformer with explicit hydrogens and optimize the geometry with GFN2-xTB. From the resulting electronic structure calculation, we obtain Wiberg bond orders, atomic partial charges, and pairwise D4 dispersion contributions. These quantities define the covalent, electrostatic, and dispersion channels, respectively, with the electrostatic interaction given by $q_i q_j / r_{ij}$. Hydrogens are included during conformer generation and the quantum-chemical calculation but are excluded when constructing the final heavy-atom AtomGrid. All calculations are performed in the gas phase.

The three interaction channels are mapped onto the shared chemically aligned coordinate system, together with a binary support mask that distinguishes unoccupied coordinates from zero-valued interactions and an atom-type composition vector. When multiple atom pairs map to the same grid

cell, covalent values are reduced by the maximum, while electrostatic and dispersion values are averaged. Self-interactions are excluded, and channel values are normalized using statistics computed from supported entries in the training split only.

GroupGrid. Functional-group instances are detected deterministically from the molecular graph over a fixed vocabulary (aromatic/heteroaromatic rings, aliphatic hydrophobic groups, hydroxyl/phenol, ether, carbonyl, carboxyl, ester, amine, amide, nitrile, sulfonyl/sulfonamide, charged centers, halogens, and an `other` category), with more chemically specific groups assigned before their generic constituents when patterns overlap. Atom-level charge features are computed using RDKit Gasteiger–Marsili partial charges, while topological distances are unweighted shortest-path distances on the heavy-atom molecular graph. Each functional-type pair is summarized by count, minimum and mean topological separation, and average charge complementarity, yielding a fixed tensor with no node ordering, resizing, or learned pooling, plus a support mask and a composition vector; unsupported cells are set to zero.

Downstream predictors. GroupGrid is processed by a lightweight MLP; AtomGrid is processed by the compact CNN described in Section 3.2. A separate support mask is retained to distinguish populated entries from padded or unsupported cells.

B.4 Fusion Models and Training Protocol

For each molecule M, the standalone fingerprint predictors produce

$$p(M) = [p_1(M), \ldots, p_R(M)] \in (0,1)^R.$$

We consider two representation portfolios: **3FP**, consisting of ECFP, AtomPair, and MACCS ($R = 3$), and **5FP**, which additionally includes AtomGrid and GroupGrid ($R = 5$). The main experiments use the gated five-fingerprint model, **5FP-Gated**; the alternatives below are evaluated in the fusion ablation.

Out-of-fold fusion protocol. All learned fusion models are trained on out-of-fold predictions from the standalone fingerprint models. The training set is partitioned into folds, and each held-out fold is scored by base predictors trained on the remaining folds. These predictions form the meta-training data for the fusion model. After fusion training, each base predictor is refit on the complete training split and used to generate validation and test scores. Thus, no fusion-training input is produced by a base model that was trained on the same molecule.

Unweighted averaging. The parameter-free baseline assigns equal weight to all representations:

$$\widehat{p}_{\mathrm{avg}}(M) = \frac{1}{R}\sum_{r=1}^{R} p_r(M).$$

MLP on raw probabilities. **FP-prob-raw** applies a lightweight MLP directly to the fingerprint score vector:

$$\widehat{p}_{\mathrm{raw}}(M) = f_{\mathrm{MLP}}(p(M)).$$

MLP on standardized logits. To account for differences in score scale and calibration, we transform each probability to a logit,

$$\ell_r = \log \frac{p_r + \epsilon}{1 - p_r + \epsilon},$$

and standardize it using statistics estimated from the out-of-fold training predictions:

$$z_r = \frac{\ell_r - \mu_r}{\sigma_r}.$$

FP-prob-z applies the same lightweight MLP to $z(M) = [z_1, \ldots, z_R]$. All normalization statistics are estimated from training predictions only and frozen at evaluation time.

Quadratic logistic stacking. **FP-Quadratic** explicitly models second-order interactions between fingerprint scores. From the standardized logits z, we construct

$$\psi(z) = \left[z_1, \ldots, z_R,\ z_1^2, \ldots, z_R^2,\ \{z_i z_j\}_{i<j}\right].$$

These features are passed to L2-regularized logistic regression: For $R = 5$, $\psi(z) \in \mathbb{R}^{20}$. These features are passed to L2-regularized logistic regression:

$$\widehat{p}_{\mathrm{quad}}(M) = \sigma\big(b + w^\top \psi(z(M))\big).$$

Gated fusion. **FP-Gated** uses a lightweight gating mechanism to assign molecule-dependent contributions to the representation-level scores. The gate operates only on score-level information, allowing the contribution of each representation to vary across molecules without accessing the underlying molecular descriptors.

B.5 Hyperparameters and Training Protocol

LBVS Hyperparameters All LBVS experiments use three seeds, $\{2026, 2027, 2028\}$, with hyperparameters selected using validation PR-AUC and the test set used only for final reporting. For AtomGrid and GroupGrid, we tune learning rate, weight decay, and dropout, using batch size 64 and at most 100 epochs with early stopping. The search space is $\{10^{-3}, 5\times10^{-4}, 10^{-4}\}$ for learning rate, $\{10^{-4}, 10^{-3}\}$ for weight decay, and $\{0.1, 0.2\}$ for dropout. For the classical fingerprint Random Forest baselines, AtomPair and MACCS are tuned over `max_features` $\in \{\mathrm{sqrt}, 0.1\}$ and `min_samples_leaf` $\in \{1, 2, 5\}$, with the number of trees selected using validation PR-AUC. For the fusion models, we tune learning rate over $\{10^{-3}, 5\times10^{-4}, 10^{-4}\}$ and weight decay over $\{10^{-4}, 10^{-3}\}$, with hyperparameters likewise selected using validation PR-AUC. All reported EF@1% values are computed only after model selection is fixed and are reported as mean $\pm$ standard deviation across the three seeds.

OGB Hyperparameters For each molecule we compute five representations: ECFP (Morgan, radius 2, 2048 bits), Atom-Pair (2048 bits), MACCS (167 keys), AtomGrid-2D and GroupGrid-2D (136 type-pair cells $\times$ 4 channels plus a 16-dimensional composition vector; distances clipped at 15 bonds). Fingerprint experts are L_2 logistic regressions ($C = 1$) on $\log(1 + x)$ counts; grid experts are two-layer MLPs (128 units, dropout 0.2, Adam 10^{-3}, 60 epochs). Out-of-fold (OOF) probabilities come from 5-fold stratified cross-fitting with folds shared across representations; the experts are then refit on the full training split to score validation and test. Logit statistics μ_r, σ_r are estimated on the OOF training probabilities and never refit on evaluation data.

A gate maps ten inputs—the standardised logits z_r and their deviations $d_r = |z_r - \bar{z}|$—to five *independent* sigmoids $a_r(x) \in [0, 1]$:

$$\ell_{\mathrm{final}} = \ell_{\mathrm{anchor}} + \lambda \tanh\Big(\textstyle\sum_r a_r(\ell_r - \bar{\ell})\Big), \qquad \ell_{\mathrm{anchor}} = \mathrm{logit}\Big(\tfrac{1}{5}\textstyle\sum_r p_r\Big). \tag{1}$$

The residual uses raw logit deviations, so representations contribute in proportion to actual disagreement rather than to $1/\sigma_r$. The gate's output layer is zero-initialised, so all gates start at 0.5 and, since $\sum_r(\ell_r - \bar{\ell}) = 0$, the model begins exactly at the equal-average anchor; the correction obeys $|\ell_{\mathrm{final}} - \ell_{\mathrm{anchor}}| \le \lambda$ throughout.

The gate is trained for 100 epochs (Adam, binary cross-entropy) with ROC-AUC checkpointed every 10 epochs. Hyperparameters (λ, hidden width, dropout, learning rate, weight decay, batch size) are chosen by 16-trial random search, selecting both configuration and epoch on validation ROC-AUC; test is used only for reporting. We use the official OGB scaffold split, train one gate per task, and average ROC-AUC over tasks. Each experiment is repeated five times with different model seeds over a shared base-probability table; we report mean $\pm$ standard deviation. The split is fixed, so this variability reflects model training and hyperparameter selection only, not split variance.

C Ablation Studies

This appendix isolates the main design choices behind our results. We first separate the effect of the *representation portfolio* from that of the *fusion mechanism*, comparing three- and five-fingerprint

Table 6: **Fusion-mechanism ablation on TopU-LBVS (hard-negative and random split)**, EF@1% (mean ± std), with average rank and average EF@1% across the seven targets. 3FP fuses ECFP, AtomPair, and MACCS; 5FP adds AtomGrid and GroupGrid. **Best**, **second-best**, and **third-best** results are marked.

TopU-LBVS Hard Negatives

Method	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb	Avg.Rank	Avg.EF1
3FP-average	3.1±0.8	6.0±1.5	**16.7±1.1**	5.6±0.4	3.9±0.5	**4.1±1.2**	3.5±0.6	6.9	6.1
3FP-prob-raw	3.4±0.9	5.4±0.5	**16.4±0.7**	5.9±0.7	3.3±1.4	3.8±1.2	3.2±0.6	7.6	5.9
3FP-prob-z	3.4±0.9	6.0±1.7	15.4±1.3	5.4±1.0	4.8±1.4	2.7±1.4	3.5±1.1	7.4	5.9
3FP-Quadratic	4.1±0.4	6.0±1.7	15.6±0.4	5.0±1.0	5.3±0.9	2.5±1.9	3.2±1.5	7.4	6.0
3FP-Gated	3.9±0.8	5.1±0.9	**17.3±0.0**	6.5±1.1	4.2±0.5	3.4±1.0	2.5±0.6	7.1	6.1
5FP-average	5.5±0.3	8.6±0.6	14.1±0.6	**7.9±0.3**	5.6±0.7	**4.7±0.8**	5.9±0.8	5.3	7.5
5FP-prob-raw	**6.6±0.4**	8.7±0.6	14.2±0.6	**8.0±0.3**	6.1±0.7	**4.8±0.8**	6.9±0.8	**4.1**	**7.9**
5FP-prob-z	**7.5±0.3**	**10.5±0.6**	14.1±0.6	**7.9±0.3**	**6.5±0.7**	**4.7±0.7**	**7.8±0.7**	4.6	**8.4**
5FP-Quadratic	**7.7±0.4**	**11.6±0.7**	14.9±0.6	**8.1±0.3**	**6.7±0.8**	**4.8±0.8**	**7.9±0.8**	**2.4**	**8.8**
5FP-Gated	**7.7±0.4**	**11.7±0.7**	14.3±0.6	**8.0±0.3**	**6.7±0.8**	**4.8±0.8**	**8.0±0.8**	**2.1**	**8.8**

TopU-LBVS Random Split

Method	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb	Avg.Rank	Avg.EF1
3FP-average	27.3±1.8	32.5±0.9	**16.9±0.4**	30.6±1.7	25.3±1.0	19.5±0.6	30.5±3.3	6.6	26.1
3FP-prob-raw	26.8±1.6	31.9±2.5	**17.3±0.6**	31.0±1.0	25.6±2.2	19.4±1.4	30.2±3.1	6.9	26.0
3FP-prob-z	**28.4±0.9**	32.2±0.5	**17.7±0.4**	**31.9±3.0**	24.4±2.2	20.6±1.5	31.1±2.8	**4.8**	**26.6**
3FP-Quadratic	**28.4±0.9**	**33.0±0.5**	**17.7±1.3**	**31.9±3.9**	24.1±2.7	20.0±0.7	30.8±3.1	**4.8**	**26.6**
3FP-Gated	27.1±1.3	32.5±0.9	**16.9±0.7**	31.5±0.8	25.0±1.5	19.3±1.6	30.2±3.1	6.9	26.0
5FP-average	25.3±1.6	31.1±1.3	13.4±2.0	30.1±1.3	25.3±1.8	**21.0±0.3**	31.4±1.1	7.6	25.4
5FP-prob-raw	25.1±0.9	32.0±0.5	13.8±1.5	30.7±1.7	25.5±0.5	**21.1±0.4**	**32.9±2.8**	6.1	25.9
5FP-prob-z	26.8±0.9	**32.7±1.8**	15.6±1.7	**31.6±2.0**	**26.9±1.4**	20.4±1.2	32.2±1.1	5.1	**26.6**
5FP-Quadratic	**29.1±1.3**	31.8±1.0	15.8±1.6	**31.6±2.9**	**27.9±0.9**	20.7±0.5	**33.1±1.1**	**4.1**	**27.1**
5FP-Gated	**29.7±1.4**	**33.8±1.3**	**16.9±1.1**	**33.0±2.0**	**27.5±0.5**	**21.4±0.4**	**32.8±0.9**	**2.1**	**27.9**

models across multiple score-level fusion strategies. We then hold gated fusion fixed and ablate AtomGrid and GroupGrid individually on TopU-LBVS to measure the information contributed by each proposed representation. Finally, on the six OGB molecular property benchmarks, we compare Random Forest and MLP heads to verify that the standalone fingerprint results are not driven by a particular classifier choice. Together, these experiments distinguish gains from richer chemical representations from gains due to downstream modeling choices.

C.1 Fusion strategies

Representation portfolio versus fusion mechanism. We compare the three classical fingerprints (**3FP**: ECFP, AtomPair, and MACCS) with the full portfolio (**5FP**: 3FP plus AtomGrid and GroupGrid) under the same five fusion rules (Appendix B.4). This separates two possible sources of improvement: a better way of combining predictions, or additional chemical information supplied by the representations.

Ligand-based virtual screening. The hard-negative split gives the clearest answer. Adding AtomGrid and GroupGrid improves performance under *every* fusion mechanism: average EF@1% increases from only 5.9–6.1 across all 3FP variants to 7.5–8.8 across the corresponding 5FP variants. Indeed, even unweighted 5FP averaging (7.5) substantially exceeds the best 3FP fusion (6.1). Thus, on the more challenging screening regime, the dominant gain comes from expanding the molecular representation portfolio rather than from increasing fusion complexity.

Within the five-view portfolio, learned fusion provides an additional gain. 5FP-Gated and 5FP-Quadratic both reach 8.8 average EF@1% on hard negatives, with gating achieving the best average rank (2.1 versus 2.4). On the easier random-decoy split, simple 3FP and 5FP combinations are largely comparable, but the advantage of the additional views emerges under adaptive fusion: 5FP-Gated reaches 27.9 average EF@1% and average rank 2.1, compared with 26.0 and 6.9 for 3FP-Gated. This suggests that AtomGrid and GroupGrid contribute complementary signals whose value is especially clear under hard negatives, while adaptive weighting helps exploit them when easier signals already make the classical fingerprints strong.

Table 7: **Fusion Strategy Ablation.** Five ways of combining the five MLP-head fingerprint scores, in increasing order of learned complexity: unweighted averaging (5FP-average); an MLP trained directly on the raw five-dimensional probability vector (5FP-prob-raw); concatenation of the five scores followed by an MLP (5FP-prob-z); the quadratic logistic stacker (5FP-Quadratic); and a gated combination (5FP-Gated).

Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	AUC↑	Gap↓	Rank↓
5FP-average	83.5	**70.2**	85.6	**78.4**	**63.7**	74.1	75.9	1.9	3.8
5FP-prob-raw	84.7	69.8	87.8	76.9	63.1	**74.6**	76.2	1.6	3.7
5FP-prob-z	**84.9**	69.3	**89.8**	**78.6**	**64.5**	74.1	**76.9**	**0.9**	**2.8**
5FP-Quadratic	**84.9**	**71.0**	**91.0**	76.6	62.4	**74.3**	**76.7**	**1.1**	**2.9**
5FP-Gated	**84.8**	**70.0**	**92.0**	**79.4**	**64.3**	**74.9**	**77.6**	**0.2**	**1.8**

Molecular property prediction. MPP provides a complementary test because AtomGrid and GroupGrid are relatively weak standalone predictors in this setting. Nevertheless, within the five-view portfolio, gated fusion is strongest on all three aggregate statistics, reaching mean AUROC 77.6, mean gap 0.2, and mean rank 1.8. Standardized-score fusion (76.9 AUROC) and quadratic stacking (76.7) are competitive, whereas unweighted averaging is weaker (75.9).

Taken together, the ablations reinforce the representation-first interpretation of the main results. On hard-negative LBVS, substantial gains appear simply by adding the two chemically distinct views, even with unchanged or parameter-free fusion. More adaptive fusion provides a further benefit, particularly for the heterogeneous five-view portfolio. We therefore use **5FP-Gated** as the default configuration, while attributing the primary gain to complementary molecular information rather than fusion capacity alone.

C.2 Individual Contribution of AtomGrid and GroupGrid

To isolate the contribution of each proposed fingerprint, we keep the gated fusion mechanism fixed and compare 3FP-Gated with two intermediate portfolios, 3FP+AtomGrid and 3FP+GroupGrid, and the full 5FP-Gated model. Table 8 reports results under both TopU hard-negative and random-decoy splits.

Both proposed fingerprints contribute complementary signal. On the TopU hard-negative split, adding either AtomGrid or GroupGrid improves over 3FP-Gated on six of seven targets, and 5FP-Gated is strongest on six of seven. The same overall pattern persists on the random split, although the individual gains are smaller and less uniform. Thus, the improvement of 5FP does not arise solely from one proposed representation: both atomic-scale physical interactions and functional-scale organization contribute information beyond the three classical fingerprints.

Table 8: **3FP+AtomGrid/GroupGrid ablation on TopU-LBVS**, EF@1% (mean ± std). GroupGrid+3FP and AtomGrid+3FP each add one proposed fingerprint to the classical 3FP portfolio (ECFP, AtomPair, MACCS), gated fusion in all rows. **Bold** marks the best result in each column.

TopU-LBVS Hard Negatives

Method	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb
3FP-Gated	3.9±0.8	5.1±0.9	17.3±0.0	6.5±1.1	4.2±0.5	3.4±1.0	2.5±0.6
GroupGrid+3FP	7.0±0.8	10.1±0.7	**18.4**±1.1	7.4±0.9	**6.7**±0.7	4.2±0.8	7.3±0.8
AtomGrid+3FP	5.9±0.4	9.4±0.9	17.7±1.1	7.3±0.7	6.3±0.7	4.3±0.5	6.7±1.1
5FP-Gated	**7.7**±0.4	**11.7**±0.7	14.3±0.6	**8.0**±0.3	**6.7**±0.8	**4.8**±0.8	**8.0**±0.8

TopU-LBVS Random Split

Method	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb
3FP-Gated	27.1±1.3	32.5±0.9	**16.9**±0.7	31.5±0.8	25.0±1.5	19.3±1.6	30.2±3.1
GroupGrid+3FP	28.8±0.9	32.7±1.3	16.2±1.0	32.9±1.3	25.0±0.9	21.0±0.3	31.8±2.1
AtomGrid+3FP	28.1±1.6	32.6±1.5	16.3±0.7	31.2±1.4	26.1±0.5	19.9±0.6	32.4±1.8
5FP-Gated	**29.7**±1.4	**33.8**±1.3	**16.9**±1.1	**33.0**±2.0	**27.5**±0.5	**21.4**±0.4	**32.8**±0.9

Fusion value need not follow standalone strength. Although AtomGrid is the stronger standalone fingerprint (Table 1), 3FP+GroupGrid outperforms 3FP+AtomGrid on six of seven TopU targets, with `kcnh2` as the exception. On the random split the difference is much smaller, with GroupGrid leading on four targets and AtomGrid on three. This illustrates an important aspect of our representation-first view: a representation can be valuable to the portfolio not because it is the strongest predictor by itself, but because it exposes information missing from the existing views. GroupGrid's functional-relational structure appears particularly complementary to the classical fingerprints under hard negatives.

`cp3a4` is the main exception. On `cp3a4`, adding AtomGrid or GroupGrid individually gives TopU EF@1% values of 17.7 and 18.4, compared with 17.3 for 3FP-Gated, whereas combining both lowers performance to 14.3. On the random split, neither individual addition improves over 3FP-Gated, and 5FP-Gated only matches it. This target therefore suggests a representation-specific interaction in which combining otherwise useful views does not yield a further gain. Understanding when complementary representations interfere rather than reinforce one another is an interesting direction for future work.

C.3 Classifier head: Random Forest vs. MLP

We compare Random Forest and MLP heads for each fingerprint when evaluated standalone on the six OGB molecular property benchmarks. The MLP improves mean AUROC for all five representations, with the largest gain for AtomPair (66.8 → 69.3) and smaller gains for ECFP (73.1 → 73.5), MACCS (72.2 → 72.4), AtomGrid (64.8 → 66.2), and GroupGrid (64.8 → 65.6). The same trend holds for average rank: every MLP variant ranks above its Random Forest counterpart.

These results indicate that the relative behavior of the fingerprints is not an artifact of a particular classical classifier. We therefore use the lightweight MLP as the default standalone head in the main experiments. Importantly, even the strongest standalone configuration, ECFP with an MLP, remains below the five-view gated fusion, reinforcing that the gains reported in the main text come from combining complementary representations rather than from simply choosing a stronger per-fingerprint classifier.

Table 9: **Classifier Head Ablation.** Each of the five fingerprints evaluated standalone with a Random Forest versus an MLP classifier head, on six OGBG-MOL datasets. Last three columns report averages (AUC, Gap, and Rank) computed per method over all ten methods shown.

Classifier	Fingerprint	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	AUC↑	Gap↓	Rank↓
Random Forest	ECFP	78.9	66.1	85.2	**72.7**	62.0	**73.8**	**73.1**	**1.6**	**3.7**
	AtomPair	**80.7**	69.1	59.6	62.5	60.1	68.9	66.8	7.9	5.8
	MACCS	73.8	**69.6**	**86.0**	70.1	**63.5**	70.4	72.2	2.5	4.2
	AtomGrid	75.4	63.1	57.5	70.9	57.0	65.1	64.8	9.9	7.7
	GroupGrid	67.7	58.8	80.0	61.1	57.6	63.7	64.8	9.9	9.0
MLP	ECFP	**79.1**	66.2	**85.5**	**73.0**	**62.8**	**74.3**	**73.5**	**1.2**	**2.7**
	AtomPair	**80.8**	**70.0**	66.7	68.2	60.6	69.5	69.3	5.4	4.3
	MACCS	73.9	**69.8**	**86.1**	70.2	**64.1**	**70.5**	**72.4**	**2.3**	**3.2**
	AtomGrid	75.5	63.7	59.5	**71.0**	60.2	67.1	66.2	8.5	6.2
	GroupGrid	70.1	59.4	78.5	61.3	60.0	64.1	65.6	9.2	8.3

D Extended Results and Discussion

This appendix reports the full per-target and per-dataset results underlying the summary statistics presented in the main text, together with standard deviations across runs and a discussion of the robustness and boundary cases these full results reveal. Section D.1 covers ligand-based virtual screening on TopU-LBVS; Section D.2 covers molecular property prediction on the six OGB benchmarks.

D.1 LBVS Results

Beyond the summary statistics discussed in Section 4.1, this full table with standard deviations reveals a stability pattern consistent with what we observe on MPP (Appendix D.2): 5FP-fusion combines the strongest mean performance with comparatively low run-to-run variance, while several

Table 10: LBVS results on TopU-LBVS across seven targets (mean ± std) for TopU (hard negatives) split and random split.

TopU-LBVS Hard Negatives

Model	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb
ECFP	5.4±0.4	**11.0±1.7**	13.9±0.5	6.2±1.2	5.8±1.1	**5.1±0.2**	**9.7±0.2**
Tanimoto	3.4±0.7	9.4±0.7	11.5±0.5	6.3±0.3	**7.4±0.4**	5.0±0.2	**8.1±0.3**
AtomPair	5.9±1.5	8.7±0.7	14.1±1.7	7.5±0.8	5.8±1.3	4.3±0.7	7.8±1.0
MACCS	4.3±0.7	**9.6±0.8**	12.2±1.1	**8.2±0.4**	6.5±0.5	5.0±0.1	6.0±0.8
GAT	1.8±0.4	2.6±0.9	**8.3±0.6**	2.2±0.4	4.8±1.4	1.9±0.6	1.6±1.5
GAT-FP	3.9±0.8	5.7±0.5	15.4±0.6	3.3±0.7	2.7±0.9	2.8±0.5	3.2±2.4
GIN	3.4±1.9	3.1±0.5	8.5±0.7	3.3±0.7	1.8±0.9	1.1±0.3	3.2±1.1
GIN-FP	4.1±2.5	5.4±1.8	13.8±2.3	4.3±1.0	4.2±1.4	2.7±0.4	1.9±1.0
GPS	1.5±2.0	2.0±0.5	5.6±2.7	1.3±0.7	1.8±0.9	2.2±0.4	3.2±1.1
GPS-FP	2.8±0.4	4.6±1.3	13.9±1.3	3.5±1.0	4.2±1.9	2.9±1.1	2.9±1.7
DMPNN	3.1±0.6	6.5±2.9	12.9±1.2	7.4±2.2	5.9±1.5	**9.1±3.4**	4.5±3.9
MOLFORMER	2.8±1.0	3.4±1.2	13.1±1.0	3.5±1.3	2.7±1.5	3.8±0.4	5.7±0.8
UniMol2	2.7±0.7	4.5±1.8	13.4±0.6	3.9±1.6	2.0±0.9	**7.1±2.1**	4.8±0.8
GroupGrid	**6.1±0.7**	5.0±0.8	13.5±0.5	**8.4±0.1**	4.8±1.5	3.0±1.3	3.7±1.9
AtomGrid	**7.8±0.6**	9.4±1.0	13.0±0.3	7.1±0.8	**6.8±0.8**	4.1±1.0	6.5±0.3
3FP-fusion	3.9±0.8	5.1±0.9	**17.3±0.0**	6.5±1.1	4.2±0.5	3.4±1.0	2.5±0.6
5FP-fusion	**7.7±0.4**	**11.7±0.7**	**14.3±0.6**	**8.0±0.3**	**6.7±0.8**	4.8±0.8	**8.0±0.8**

TopU-LBVS Random Split

Model	aa2ar	aces	cp3a4	egfr	esr1	kcnh2	thrb
ECFP	25.4±1.4	28.5±1.0	16.4±2.4	**31.4±1.0**	24.0±2.2	19.7±1.2	29.8±2.3
Tanimoto-NN	17.9±0.6	21.1±2.0	10.8±2.7	19.5±1.2	16.0±0.0	12.3±0.9	23.5±2.4
AtomPair	25.3±1.2	27.3±0.0	14.5±1.5	30.2±2.0	**24.7±1.9**	**20.1±2.2**	31.1±2.8
MACCS	19.1±1.2	29.9±2.3	11.7±1.0	25.6±5.4	24.4±0.5	17.8±0.8	29.2±2.0
GAT	11.3±3.1	16.8±2.0	7.0±1.1	18.2±2.3	15.2±1.8	8.4±0.6	19.4±1.5
GAT-FP	25.8±0.9	29.9±1.0	16.4±1.0	30.1±1.6	23.2±3.1	18.9±1.2	30.2±1.5
GIN	13.2±2.8	18.8±2.3	7.3±1.3	23.9±3.0	16.3±0.5	9.0±1.0	22.2±4.0
GIN-FP	25.5±1.3	29.2±1.3	15.7±1.0	30.3±1.4	23.5±2.2	18.7±1.9	**31.8±3.1**
GPS	12.4±1.5	14.2±0.5	4.5±1.7	14.1±1.6	9.8±1.5	9.2±1.7	13.0±3.3
GPS-FP	25.0±2.4	29.6±0.5	**16.7±1.6**	30.8±1.5	**24.7±1.9**	19.7±1.9	31.1±1.5
DMPNN	26.2±1.1	29.1±0.8	14.8±1.2	28.0±3.8	22.6±2.0	**20.1±0.6**	29.6±2.1
MOLFORMER	23.0±0.9	27.3±1.2	15.2±1.7	26.4±2.0	19.6±2.5	19.1±1.3	27.7±4.1
UniMol2	**27.9±0.2**	29.2±2.2	14.4±1.1	25.1±2.2	23.5±1.1	18.9±1.3	29.2±1.2
GroupGrid	22.4±1.5	**30.5±1.0**	16.0±1.3	27.8±3.2	**24.7±2.1**	**20.5±1.1**	**32.1±3.3**
AtomGrid (CNN)	22.2±1.0	27.3±1.0	14.2±1.7	24.0±2.3	20.7±1.0	20.0±0.6	31.1±3.1
3FP-fusion	**27.1±1.3**	**32.5±0.9**	**16.9±0.7**	**31.5±0.8**	**25.0±1.5**	19.3±1.6	30.2±3.1
5FP-fusion	**29.7±1.4**	**33.8±1.3**	**16.9±1.1**	**33.0±2.0**	**27.5±0.5**	**21.4±0.4**	**32.8±0.9**

learned baselines exhibit substantial variance on individual targets, most visibly DMPNN on `thrb` (4.5±3.9) and `kcnh2` (9.1±3.4) under hard negatives, and GAT-FP on the same target (3.2±2.4), where the standard deviation is comparable to or exceeds the mean itself. By contrast, under the TopU hard-negative split, 5FP-fusion's standard deviation does not exceed 0.8 EF@1% across the seven targets combinations we evaluate, indicating that its advantage over learned baselines is not an artifact of a favorable draw of seeds, and that its ranking relative to other methods would be stable under re-evaluation.

We also note, for completeness, that the one exception to the 5FP-over-3FP pattern established in Section 4.1 is visible directly in this table: on `cp3a4`, 3FP-fusion outperforms 5FP-fusion under both the hard-negative split (17.3±0.0 vs. 14.3±0.6) and, more narrowly, the random split (16.9±0.7 vs. 16.9±1.1, effectively tied). This is the only target of the seven where adding AtomGrid and GroupGrid does not help, and it holds under both decoy regimes, suggesting it reflects a genuine property of this target rather than split-specific noise; we leave a full investigation of what distinguishes `cp3a4` to future work.

D.2 Molecular Property Prediction Results

Beyond the summary statistics discussed in Section 4.2, this full-precision table surfaces two points worth discussing explicitly. First, variance is not uniform across method families: graph neural networks and pretrained models exhibit substantial run-to-run variance on several datasets, most

strikingly MolFormer on ClinTox (88.40 ± 15.90) and Uni-Mol on the same dataset (82.40 ± 13.50), where the standard deviation alone exceeds the entire margin by which MolFormer's mean trails 5FP+Gated (91.95 ± 0.52). A single unlucky run of MolFormer would therefore have produced a materially different ranking on this dataset, whereas 5FP+Gated's advantage on ClinTox holds robustly across seeds. More generally, 5FP+Gated combines the strongest mean performance with among the lowest variance of any method compared, on every dataset its standard deviation is at or below 0.52, an order of magnitude tighter than the pretrained and graph-based baselines, which is consistent with its architecture: the underlying fingerprints are deterministic, non-parametric descriptors, and the only stochastic components are the lightweight per-fingerprint heads and the fusion layer, rather than an end-to-end model trained from scratch on each dataset.

Second, we note that the five standalone fingerprints (ECFP, AtomPair, MACCS, AtomGrid, and GroupGrid) are reported with zero standard deviation across the five runs, reflecting their fully deterministic featurization and classifier fitting procedure under a fixed protocol, in contrast to the stochastic training of the GNN, pretrained, and fusion methods. This is expected given the closed-form nature of these baselines under our evaluation setup, but it also means the standalone fingerprint rows are not subject to the same run-to-run averaging as the other methods in this table, and comparisons of their exact standalone values against methods with nonzero variance should be read with this asymmetry in mind.

Table 11: Full results table for OGB datasets reporting mean AUROC ± standard deviation across 5 independent runs.

Method # Compounds	SIDER 1427	ClinTox 1478	BACE 1513	BBBP 2039	TOX21 7831	HIV 41127
GCN	58.20±2.59	75.56±3.12	71.90±4.04	65.60±1.43	74.68±1.09	72.90±1.12
GIN	57.73±3.46	76.10±3.00	72.43±4.36	65.76±1.91	75.63±1.94	76.75±1.36
GSAGE	58.95±1.97	81.90±5.47	72.21±4.04	64.00±2.16	75.17±1.94	73.05±1.47
GAT	59.67±2.34	84.07±4.52	72.89±4.00	66.29±3.06	**76.03±1.72**	75.41±3.55
GPS	59.55±1.53	76.14±4.68	67.47±3.86	65.06±1.52	70.97±0.43	70.44±1.08
SGFormer	58.32±0.80	78.07±1.26	66.24±4.78	66.08±1.31	72.97±0.45	72.21±1.33
MolBERT	64.00±2.20	**87.20±4.80**	81.90±1.70	**74.30±2.70**	**76.40±3.00**	77.90±1.30
ChemBERTa	**69.20±1.90**	73.90±1.40	**82.60±4.50**	**72.10±1.00**	**78.90±3.50**	**78.70±2.50**
MolFormer	**66.60±1.70**	**88.40±15.90**	**83.70±6.80**	**71.50±2.70**	75.00±1.80	**78.00±1.30**
Uni-Mol	62.20±2.60	82.40±13.50	**82.60±3.40**	66.10±3.10	72.50±5.80	76.10±0.10
GraphMVP	**67.30±2.00**	78.90±2.40	81.80±5.60	68.70±3.80	76.00±3.60	70.20±0.20
ECFP	61.99±0.00	85.21±0.00	78.91±0.00	66.12±0.00	73.80±0.00	72.69±0.00
AtomPair	60.06±0.00	59.63±0.00	80.68±0.00	69.05±0.00	68.89±0.00	62.50±0.00
MACCS	63.53±0.00	86.02±0.00	73.81±0.00	69.62±0.00	70.39±0.00	70.05±0.00
AtomGrid	57.04±0.00	57.46±0.00	75.41±0.00	63.13±0.00	65.06±0.00	70.93±0.00
GraphGrid	57.61±0.00	79.97±0.00	67.74±0.00	58.83±0.00	63.71±0.00	61.05±0.00
5FP+Gated (Ours)	64.30±0.05	**91.95±0.52**	**84.83±0.50**	69.95±0.13	74.88±0.16	**79.43±0.26**