MOLGRID: WHICH CHEMISTRY DO MOLECULAR MODELS READ?

Anonymous authors
Paper under double-blind review

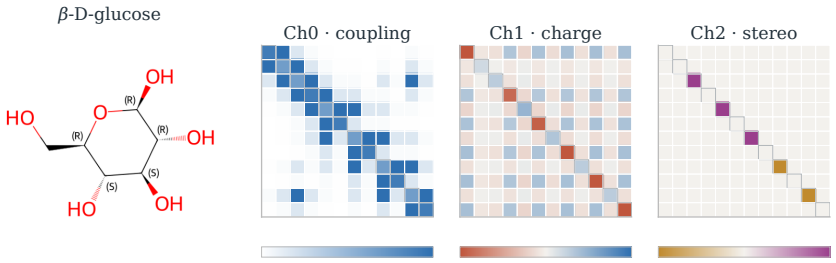


Figure 1: **MOLGRID represents a molecule as three symmetric matrices over its atom pairs.** For β-D-glucose (left), each channel (right) has one row and one column per heavy atom: the diagonal holds one value per atom and the off-diagonal one per atom pair. Channel 0 holds the element and the through-bond coupling between atoms, Channel 1 the partial charges, and Channel 2 the stereochemistry, with the five stereocentres coloured (R purple, S gold).

ABSTRACT

Molecular property prediction guides which molecules are made and tested, in drug discovery and beyond. Its models read graphs, SMILES strings or 3D structures, which leave relations between atoms, such as how strongly two atoms couple through bonds, for the network to infer. A good score does not tell us which of these relations a model actually relies on. We introduce **MOLGRID**, which states this pairwise chemistry explicitly as a symmetric image over atom pairs, with channels for through-bond coupling, partial charge and stereochemistry. Built from SMILES in milliseconds with no 3D input, it is read unchanged by pretrained vision backbones. Its diagonal holds one value per atom and its off-diagonal one per atom pair, the first two terms of a property's expansion over atoms, which predicts that additive properties read the atoms and delocalised ones the pairs. Because every value has a fixed address, we can shuffle either part or remove a channel and retrain a chemically pretrained backbone, against a control that only relabels the atoms. **MOLGRID** leads reproduced fingerprint, graph and SMILES-transformer baselines on five of seven MoleculeNet scaffold-split tasks each, and 3D Uni-Mol on solubility. Across nine properties, only the HOMO–LUMO gap reads the pairs, 5.7 percentage points beyond the control, while solvation reads the charges and permeability stereochemistry. Pretraining improves 25 of 28 backbone–task pairs without changing which property reads the pairs.

1 INTRODUCTION

Molecular property prediction spans many formats, from graph networks (Xu et al., 2019; Gilmer et al., 2017; Ying et al., 2021) and SMILES transformers (Chithrananda et al., 2020; Ross et al., 2022; Fabian et al., 2020; Cai et al., 2026) to 3D networks (Zhou et al., 2023; Ji et al., 2024) and vision-language and multimodal models (Li et al., 2025; Adak et al., 2025; Park & Lee, 2025). Their test-set accuracy does not show which input chemistry a model relies on.

Each of these formats describes a molecule through single atoms or tokens. The relations between atoms that a chemist reasons with are left for the network's layers to infer: how strongly two atoms

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.

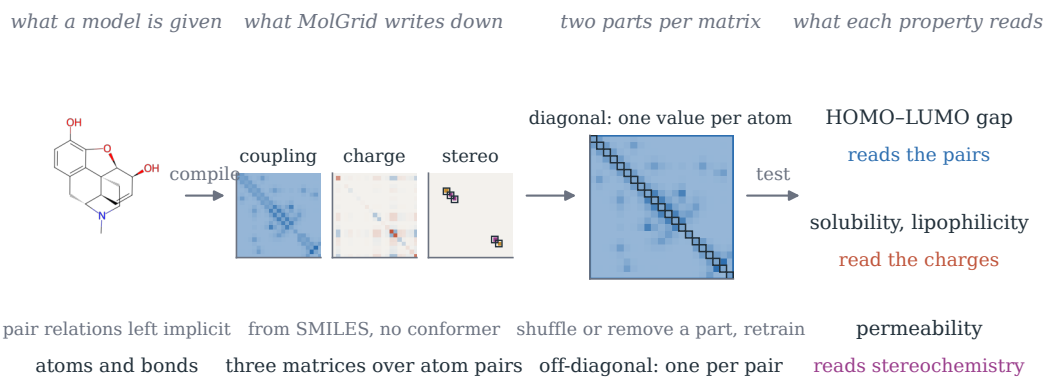


Figure 2: **Overview of MOLGRID and the test it enables.** A graph or SMILES model is given only atoms and bonds. MOLGRID computes three symmetric matrices over atom pairs (morphine shown), with one value per atom on each diagonal and one per atom pair off it. Retraining after shuffling or removing one part, against a relabelled copy, shows which part each property reads.

couple through the bonds between them, which differs between a saturated linker and a conjugated bridge of equal length, where it is polarised, and which centres are stereogenic.

This paper asks which chemistry a model reads from its input, and whether that can be tested by intervention, property by property. We hypothesise that much of a network's depth is spent inferring pairwise chemistry, that mainstream formats leave this chemistry implicit, and that stating it explicitly makes the question testable. We introduce MOLGRID to state it, use it to measure which chemistry each property depends on (Sec. 6.2), and show that it is also a competitive representation (Sec. 6.1).

MOLGRID is a fixed-size $3 \times N \times N$ symmetric tensor over atom pairs (Figures 1 and 2). Its three channels hold through-bond coupling, partial charge and stereochemistry, and it is computed from SMILES in milliseconds. In every channel the diagonal holds one value per atom and the off-diagonal one per atom pair, and we call these the *unary* (atom) and *binary* (pair) *registers*. Every value therefore has an address: its channel names the chemistry and its row and column name the atoms, so an intervention on a cell is an intervention on chemistry. Relabelling the atoms never moves a value between the two registers (Proposition 1), and the registers correspond to the first two terms of a property's expansion over atoms, which predicts which register each property depends on.

Because the chemistry is explicit, it can be changed one part at a time. We shuffle which atom pair holds each off-diagonal value, shuffle which atom holds each diagonal value, or remove a channel, and retrain the model each time, with ten seeds per experiment on the reported backbone and paired tests corrected for multiple comparisons. Relabelling the atoms is the matched control: the input is equally unfamiliar to the model, but every value stays with its atoms. We report four findings.

- ***Each property reads the register the decomposition predicts.*** Only the HOMO–LUMO gap reads the pairs, beyond the relabelling control in all five backbones, while all nine properties read the atoms, the charges matter most for solvation, and only permeability reads stereochemistry.
- ***Pretraining sharpens the reading without changing it.*** Chemical pretraining improves nearly every backbone and task, and the pretrained backbone still reads the pairs on the gap alone.
- ***Depth rebuilds pairwise chemistry only within its reach*** (Figure 3). Beyond it, a graph network recovers the coupling little better than raw atom features, and MOLGRID states it outright.
- ***The representation competes*** (Table 3). MOLGRID leads the reproduced D-MPNN, MolFormer-XL and ECFP4 baselines on five of seven tasks each, and the 3D model Uni-Mol on solubility.

2 RELATED WORK

Representation families each trade a chemistry property for a class of architecture. *Strings* tokenise SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2020) and leave the model to recover permutation invariance and pairwise relations from token order (Chithrananda et al., 2020; Ross et al., 2022;

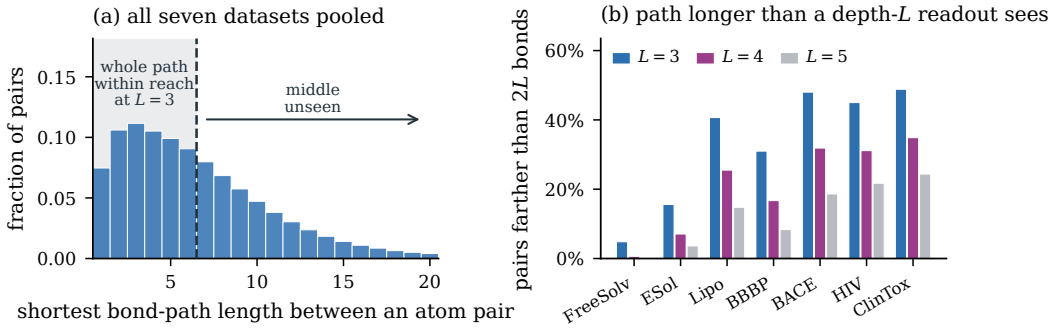


Figure 3: **Up to half of all atom pairs lie beyond the reach of a depth-three network.** A network of depth L sees the bond path between two atoms whole only if it is at most $2L$ bonds long. (a) Shortest bond-path lengths over heavy-atom pairs in the training sets, pooled, with the paths a depth-three network sees whole shaded. (b) The share beyond that reach, per dataset and depth.

Fabian et al., 2020). *Graphs* pass messages over topology (Gilmer et al., 2017; Yang et al., 2019; Xu et al., 2019; Kipf & Welling, 2017; Ying et al., 2021; Fang et al., 2022), and a pairwise readout sees a bond path whole only within twice the depth. Interatomic potentials raise the body order of each message so that expressivity no longer waits on depth (Batatia et al., 2022), and MOLGRID does the same for pairwise chemistry by stating it in the input. *3D point clouds* expose geometry to invariant and equivariant networks (Klicpera et al., 2020; Thölke & De Fabritiis, 2022; Zhou et al., 2023; Ji et al., 2024) at the cost of a conformer at inference. *Image renderings* apply vision backbones to depictions (Goh et al., 2017; Zeng et al., 2022), so pretraining must teach the model to decode a drawing. Cross-modal transfer instead learns an embedder into a pretrained model's input space (Lu et al., 2022; Shen et al., 2023), so an input feature carries no fixed meaning. In MOLGRID every cell has a fixed chemical meaning, and pretraining teaches the model to use it.

The closest prior work adds atom-pair information that is learned, geometric or partial. The Coulomb matrix (Rupp et al., 2012) is the best-known fixed-formula pair tensor, single-channel and built from 3D coordinates, and atom-pair fingerprints count typed pairs by their path length but pool them into an unordered bag (Carhart et al., 1985). Graph transformers learn pairwise biases from shortest-path distance and edge features, Graphormer averaging learned bond embeddings along a shortest path where Channel 0 multiplies fixed McConnell attenuations (Ying et al., 2021; Luo et al., 2023). Uni-Mol learns a pair track beside the atom track (Zhou et al., 2023), and Transformer-M and molecule attention transformers add distance and adjacency terms to self-attention (Luo et al., 2023; Maziarka et al., 2020; 2024). Closest of all, the dual-graph line adds architecture to learn bond-bond interactions (Liu et al., 2026), where we state the through-bond coupling as an input and change no architecture. The partial charges that line names as a next step are already one of our channels. None states a fixed, chemistry-typed value for every atom pair from SMILES alone.

Encoding is not use, and testing the difference needs an intervention. Probing shows what a representation encodes (Alain & Bengio, 2017; Hewitt & Liang, 2019) but not whether the model uses it (Belinkov, 2022). Causal use needs an intervention removing candidate information (Elazar et al., 2021; Fisher et al., 2019) and a model retrained on the altered input, as remove-and-retrain does for features (Hooker et al., 2019) and y-randomisation for labels in chemistry (Rücker et al., 2007). A test-time intervention can misstate a mechanism by pushing activations off the natural distribution (Grant et al., 2026). Accuracy alone cannot show what a representation adds, as simple baselines match or beat deep models (Deng et al., 2023; van Tilborg et al., 2022). Appendix A extends every comparison, covers modality transfer (Zhuo & Luo, 2026; Shen et al., 2025) and language models (Baranwal & Vyas, 2026) and asks if 3D geometry is needed.

3 WHY SHOULD A MOLECULE BE A PAIR TENSOR?

A molecular format is a choice about what to write down explicitly, and the four mainstream choices all write down first-order objects. SMILES states a token order but not adjacency. A 2D graph states the Hückel Hamiltonian while the couplings a chemist reasons with are functions of its inverse. A

Table 1: **Every channel survives relabelling, and the SMILES order keeps bonds next to the diagonal.** Each test molecule is relabelled at random, rebuilt and compared with the relabelled original. Channel 0 changes only where two equally short paths carry different bond types, by at most the largest change shown, and Channel 1 on one HIV zwitterion. The last two columns give the share of bonds within two cells of the diagonal in the SMILES order and under a random order.

		Unchanged under relabelling			Bonds within two cells	
Task	Molecules	Channel 0	largest Ch. 0 change	Channels 1, 2	SMILES order	random order
BBBP	199	78%	0.023	100.0%	81%	15%
BACE	147	77%	0.032	100.0%	79%	11%
ClinTox	146	84%	0.016	100.0%	77%	16%
HIV	1468	79%	0.072	99.9%	81%	14%
ESol	112	79%	0.032	100.0%	81%	21%
FreeSolv	64	91%	0.008	100.0%	78%	33%
Lipophilicity	420	91%	0.032	100.0%	81%	14%
QM9	–	–	–	–	78%	42%

3D structure states the through-space metric, and through-bond coupling is not a function of it. A depiction states a picture of a diagram, whose bonds and element symbols a network must first learn to read. Each leaves the pairwise chemistry for depth to rebuild.

What depth rebuilds has a closed form. How a perturbation at atom i is felt at atom j is governed not by the Hamiltonian H but by its resolvent, whose leading term for atoms joined by the shortest bond paths $\mathcal{P}_{ij}$ is

$$G(E) = (E\mathbb{I} - H)^{-1}, \qquad G_{ij}(E) \simeq \frac{1}{E-\varepsilon} \sum_{P \in \mathcal{P}_{ij}} \prod_{(u,v)\in P} \frac{\beta_{uv}}{E-\varepsilon}. \tag{1}$$

In Hückel theory H holds a site energy ε on its diagonal and a resonance integral β_{uv} on each bond, and when the bridge's levels lie far from the energy of interest the first term that connects i and j dominates (Löwdin, 1951). This is McConnell superexchange (McConnell, 1961), the standard model of through-bond communication in bridged systems (Beratan et al., 1991; Newton, 1991), decaying geometrically in the number of intervening bonds. Channel 0 writes its magnitude with bond-type attenuations along the shortest path, not G itself (Appendix C.1).

A network of finite depth rebuilds this term only within its reach. A depth-L message-passing network computes a polynomial of degree at most L in the graph operator, a truncated Neumann series for that inverse (Defferrard et al., 2016; Chien et al., 2021; He et al., 2021; Klicpera et al., 2019). A pairwise readout therefore sees a bond path whole only when the path is at most $2L$ bonds long, and Figure 3 shows how often real molecules exceed that. The middle of a longer path lies outside both receptive fields, and the readout can only guess it from the two ends. A graph network trained to depth three recovers the coupling across that middle segment at R^2 of 0.09, barely above the 0.04 that raw atom features give (Appendix E.1), while MOLGRID states the coupling directly.

Pairwise chemistry over n atoms is a symmetric function on atom pairs, so the minimal faithful container is a symmetric $n \times n$ array per property. A relabelling with permutation matrix P acts as $X \mapsto PXP^\top$ (Proposition 1). Under that action the diagonal transforms as a vector and the off-diagonal as a matrix, and no relabelling moves a value between them, the two orbits of relabelling on the cells (Maron et al., 2019). These are the *unary* and *binary registers*, holding what each atom is, by element, charge and chirality, and how each pair of atoms is related.

The chemistry gives that split independent meaning. A molecular property expanded over atoms is a many-body expansion,

$$P(\mathcal{M}) = \underbrace{\textstyle\sum_i p_1(i)}_{\text{unary register}} + \underbrace{\textstyle\sum_{i<j} p_2(i,j)}_{\text{binary register}} + \cdots, \tag{2}$$

whose first-order truncation is group contribution, the basis of atomic-contribution $\log P$ (Wildman & Crippen, 1999) and Joback-style estimation (Joback & Reid, 1987). A first-order term may depend on an atom's local environment, as group-contribution atom types and Gasteiger charges do, and still add up over atoms. Group contribution has predicted solubility for decades and never a frontier orbital energy, because delocalisation is not additive over atoms. A polyene's HOMO–LUMO gap, for instance, narrows toward a limit as the chain grows, which no sum of per-atom terms reproduces.

Table 2: **MOLGRID needs one RDKit call on a SMILES string, and a 3D conformer needs embedding and optimisation.** Build rate and single-core wall time per molecule, against the ETKDG embedding and MMFF optimisation a 3D model needs. ETKDG fails where embedding does not converge or MMFF lacks parameters, and MOLGRID only on unparseable SMILES and three molecules over 360 heavy atoms, past the budget. All 400 repeat builds were bit-identical.

Molecules	Input built	Succeeds	Median	99th pct.	Fails on
PubChem (20,000)	**MOLGRID** (SMILES only)	**99.95%**	**8.6** ms	200 ms	6 unparseable
	3D conformer (ETKDG+MMFF)	93.05%	119.5 ms	4549 ms	278 molecules
HIV test (4,112)	**MOLGRID** (SMILES only)	**99.88%**	**8.2** ms	126 ms	5 unparseable
	3D conformer (ETKDG+MMFF)	90.84%	97.5 ms	2477 ms	376 molecules

The split makes a testable prediction: a property additive over atoms reads the unary register, the first term of Eq. (2), and a delocalised property reads the binary register, the second. Permuting the off-diagonal tests the binary register alone, and zeroing a channel removes both.

Testing that prediction needs a model that reads a pair tensor, which sequence models and message passing do not read natively. Convolution and attention do, and unlike the second-order invariant graph network (Maron et al., 2019) they come with weights pretrained at scale.

MOLGRID puts the symmetry in the input and leaves the backbone free, the division of labour that canonicalisation methods use (Puny et al., 2022; Kaba et al., 2023; Mondal et al., 2023; Urbano et al., 2026). The input transforms equivariantly (Proposition 1), and the atom order enters only as the order of rows and columns. We take it from the SMILES string, a depth-first walk that keeps bonded atoms in nearby rows: about four bonds in five sit within two cells of the diagonal, two to seven times the share under a random order (Table 1). Convolution and patch attention exploit exactly this locality. The relabelling control of Sec. 5 removes it while keeping every chemistry-to-pair assignment, which makes it the matched control for the content-scramble.

4 MOLGRID: THREE CHANNELS, TWO REGISTERS

MOLGRID encodes a molecule of up to $N = 64$ heavy atoms as $X \in \mathbb{R}^{3\times N\times N}$, every cell (i, j) a relation between two atoms, the three channels stacking the three reads of Figure 2 and drawn in Figure 1. In each channel the diagonal is the unary register, carrying per-atom information as an atom's relation with itself, and the off-diagonal is the binary register.

Channel 0 carries through-bond electronic coupling and element identity. Its diagonal holds $X[0, i, i] = Z_i/Z_{\max}$, and its off-diagonal holds the leading term of the propagator of Eq. (1) in McConnell pathway form (McConnell, 1961; Beratan et al., 1991; Newton, 1991): a product, over the bonds of the shortest path between the two atoms, of a bond-type attenuation and the normalised bond order, worked through in Figure 7. A bonded pair carries its attenuated bond order, so the channel contains a bond-weighted adjacency as its first shell. The exact cell is in Appendix C.1.

Channel 1 carries charge, because two isomers with identical skeletons behave differently when their charge distributions diverge. The Gasteiger algorithm (Gasteiger & Marsili, 1980; Sanderson, 1951) gives a millisecond, well-validated proxy: per-atom charges q_i sit on the diagonal and the signed product q_iq_j off it, so a single sign flip marks a polarised pair. Channel 2 carries stereochemistry, because stereoisomers with identical Lewis structures can act differently on a biological target. The Cahn–Ingold–Prelog code sits on its diagonal ($+1$ for R, -1 for S, 0 for undefined) and bond-stereo (E/Z) tags off it, so chirality enters without 3D coordinates. Stereochemistry holds the third slot on the evidence: two fourth channels and five other candidates for that slot, tested under a rule declared in advance, never beat the three-channel set on every task (Appendix C.2).

Two hard properties follow from the encoding's cell-wise construction. Every cell has the form

$$X[c, i, j] = f_c\big(\phi(i),\, \phi(j),\, \rho(i, j)\big), \tag{3}$$

where $\phi(i)$ collects the intrinsic attributes of atom i and $\rho(i, j)$ is an intrinsic, graph-isomorphism-invariant relation of the *unordered* pair, reducing to the per-atom term on the diagonal. Both travel with the atoms rather than their labels, which is all the two symmetries need.

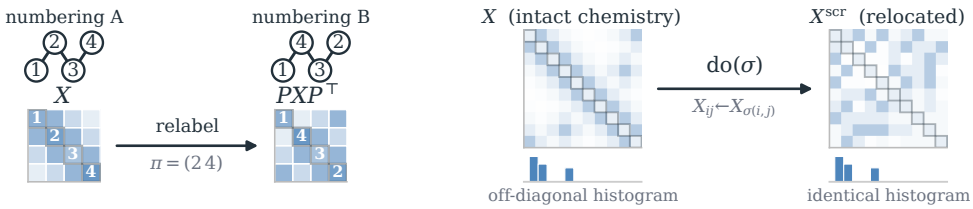


Figure 4: **The control keeps each value with its atoms, and the intervention moves it.** Left, the relabelling control: two numberings of one molecule give grids related by $X \mapsto PXP^\top$. Right, the content-scramble on Channel 0: the off-diagonal values move to other atom pairs, while the diagonal, the value histogram (below) and the symmetry stay as they were.

Proposition 1 (Relabelling acts register by register)**.** *For every atom relabelling $\pi \in S_n$ and every channel c whose cells have the form of Eq.* (3) *with f_c symmetric in its two atoms,*

$$X_\pi[c, i, j] = X[c, \pi(i), \pi(j)], \qquad X[c, i, j] = X[c, j, i], \tag{4}$$

exactly, with no learned parameter. A relabelling therefore moves diagonal cells only onto diagonal cells and off-diagonal cells only onto off-diagonal cells. Every relabelling-invariant linear readout of a channel has the form $w_1 \sum_i X[c, i, i] + w_2 \sum_{i \neq j} X[c, i, j]$ (Maron et al., 2019). Channels 1 and 2 have the cell form, and Channel 0 has it when equally short paths are summed.

Appendix C.1 proves it and bounds the single-path tie-break of our builder. Any operator T respecting row–column permutation, $T(PXP^\top) = PT(X)P^\top$, composed with a permutation-invariant readout, is *exactly* invariant to atom relabelling. A convolutional or attention backbone is not such an operator. It reads the atoms in SMILES order and brings pretrained weights.

A representation must be producible for any new molecule. MOLGRID builds from the SMILES string alone for 99.95% of 20,000 PubChem molecules outside every benchmark, at a median of 8.6 ms, fourteen times faster than a force-field 3D structure and bit for bit deterministic (Table 2).

The *presentation* is how the stored tensor reaches the backbone. *Fill* crops the molecule's block to the 224×224 input, so the molecule fills the field of view at any size, and *padded* resamples the whole canvas with the block at a random offset, so the backbone sees its scale (Appendix C.1). Pretraining uses padded, and validation selects per task at finetune.

5 METHODS

Our experiments use three methods: the chemical pretraining recipe, the finetuning recipe, and the two interventions that test the prediction of Sec. 3. MOLGRID is not a natural image, so we continue pretraining on it before any downstream task, and the reported backbone is DINOv2-B (Oquab et al., 2024). The corpus is 3.56M molecules of PCQM4Mv2, 3.31M of them with DFT labels (Nakata & Shimazaki, 2017; Hu et al., 2021), compiled to MOLGRID on the padded canvas. The backbone trains for six epochs at batch 512 under a complexity-ordered curriculum and three objectives. *Teacher alignment* aligns the pooled feature with a frozen Uni-Mol2 embedding (Ji et al., 2024) of the same molecule's 3D structure. *Property prediction* regresses the corpus's DFT HOMO–LUMO gap, and *functional-group supervision* predicts 85 RDKit substructure indicators. Teacher alignment brings in 3D knowledge as privileged training information (Liu et al., 2022; Stärk et al., 2022), with no 3D input at finetuning or inference. Appendices D.1 and D.2 give the loss weights and schedule.

Finetuning trains a linear task head on the pooled feature together with the backbone, for up to 30 epochs at batch 64 with early stopping. Four choices make up the recipe, each tested one variable at a time on all seven tasks before it entered (Appendix D.3). A linear probe precedes finetuning (Kumar et al., 2022), and an exponential moving average of the weights is evaluated (Morales-Brotons et al., 2024). Two auxiliary heads predict 85 functional-group indicators and 12 RDKit descriptors during training only, so the fingerprint is a prediction target and never an input.

When a vision backbone predicts well from MOLGRID, does it read the chemistry we wrote into the channels, or the tensor's shape, sparsity and value statistics? Accuracy alone cannot separate the two (Geirhos et al., 2020; Lapuschkin et al., 2019), so we intervene on one register at a time.

Table 3: **MOLGRID leads each reproduced 2D baseline on five of seven scaffold tasks, with no conformer at inference.** Test mean$_{\pm sd}$ over training seeds on one shared scaffold split. ECFP4-RF and D-MPNN (Chemprop's defaults) are trained from scratch, and MolFormer-XL and Uni-Mol start from their released checkpoints (Appendix H.1). MOLGRID is selected on validation from two declared configurations and scored on the molecules within its 64-atom budget. Bold marks the best mean and underlining the second. The last row counts the tasks on which MOLGRID leads, ties or trails each baseline by more than the pooled seed spread, with the total under MOLGRID. †Uni-Mol cannot build 3D coordinates for a few BBBP, ClinTox, HIV and Lipophilicity molecules, so it scores fewer of them, and its HIV run has one seed. The last column is the best configuration we ran, chosen with the test score in view, and its record totals the same comparisons as an upper bound on the modality rather than a claim, since each task may pick a different configuration.

Task	Metric	ECFP-RF	D-MPNN	MolFormer-XL	Uni-Mol (3D)	**MolGrid**	best config.
FreeSolv	RMSE ↓	$5.339_{\pm 0.028}$	$\underline{2.088}_{\pm 0.044}$	$2.778_{\pm 0.110}$	$\mathbf{1.980}_{\pm 0.086}$	$2.237_{\pm 0.162}$	$1.913_{\pm 0.046}$
ESol	RMSE ↓	$1.792_{\pm 0.008}$	$0.970_{\pm 0.012}$	$0.905_{\pm 0.017}$	$\underline{0.876}_{\pm 0.005}$	$\mathbf{0.848}_{\pm 0.020}$	$0.830_{\pm 0.030}$
Lipo	RMSE ↓	$0.928_{\pm 0.002}$	$\mathbf{0.625}_{\pm 0.019}$	$0.727_{\pm 0.009}$	$\underline{0.632}_{\pm 0.003}$	$0.687_{\pm 0.003}$	$0.657_{\pm 0.013}$
BBBP	AUROC ↑	$0.691_{\pm 0.009}$	$0.663_{\pm 0.027}$	$\underline{0.719}_{\pm 0.009}$	$0.705_{\pm 0.008}$	$\mathbf{0.731}_{\pm 0.013}$	$0.740_{\pm 0.005}$
BACE	AUROC ↑	$\mathbf{0.870}_{\pm 0.004}$	$0.805_{\pm 0.008}$	$0.813_{\pm 0.016}$	$\underline{0.848}_{\pm 0.007}$	$0.839_{\pm 0.010}$	$0.853_{\pm 0.008}$
HIV	AUROC ↑	$0.753_{\pm 0.004}$	$0.724_{\pm 0.018}$	$\mathbf{0.790}_{\pm 0.007}$	–	$\underline{0.754}_{\pm 0.010}$	$0.761_{\pm 0.010}$
ClinTox	AUROC ↑	$0.705_{\pm 0.014}$	$0.854_{\pm 0.007}$	$\mathbf{0.973}_{\pm 0.014}$	$0.713_{\pm 0.066}$	$\underline{0.906}_{\pm 0.016}$	$0.923_{\pm 0.003}$
MOLGRID leads / ties / trails		5 / 1 / 1	5 / 0 / 2	5 / 0 / 2	3 / 1 / 2†	18 / 2 / 7	20 / 2 / 5

Adapting permutation feature importance (Fisher et al., 2019) and amnesic counterfactuals (Elazar et al., 2021) to a structured input (Figure 4, Appendix E.2): let π permute the unordered atom-pair set of $X \in \mathbb{R}^{C\times n\times n}$. The *content-scramble* moves which pair carries which chemistry, applying the same π to the off-diagonal cells of every channel and leaving the diagonal unchanged:

$$\tilde{X}[c,i,j] = \tilde{X}[c,j,i] = X[c,\pi(i,j)] \;\; (i \neq j), \qquad \tilde{X}[c,i,i] = X[c,i,i]. \tag{5}$$

It keeps the diagonal, each channel's value histogram and the symmetry, and destroys only which two atoms a coupling, a charge product or a stereo flag belongs to. The *channel ablation* is complementary: it zeroes one channel, diagonal included, so it removes both registers where the scramble tests only the binary one. A per-channel variant permutes one channel alone. The *diagonal scramble* permutes which atom holds which per-atom triple and leaves every off-diagonal cell in place, so the unary register is tested by a permutation as the binary one is. The *relabelling control* is the same molecule with its atoms renumbered, rows and columns permuted together, $X \mapsto PXP^\top$, so every chemistry-to-pair assignment survives and only the order of the atoms changes. Because we retrain on the altered input rather than perturb a trained model, each model is evaluated on the kind of input it was fitted to. The test-time divergence that can misstate a mechanism (Grant et al., 2026) therefore does not arise. In each backbone's own feature space the relabelled input lies about as far from the intact one as the scrambled input does (Appendix E.2). The scramble's cost beyond the relabelling, paired by seed, is therefore the value of the chemistry-to-pair assignment. We call it the *pair excess*, as a probe's selectivity is its accuracy beyond a control task (Hewitt & Liang, 2019). A property *reads* a part of the tensor when losing that part raises its test error in a paired test over seeds that survives the correction across the map. For shuffled pairs the cost is the pair excess, not the raw scramble cost.

6 EXPERIMENTS

Each main task uses one fixed Bemis–Murcko scaffold split (Bemis & Murcko, 1996) at $80/10/10$ with the metric MoleculeNet recommends (Wu et al., 2018), and every number is a test mean with its standard deviation over at least three training seeds. The reported MOLGRID number is selected on validation, and Table 3 adds the best configuration we ran as an upper bound. Baselines are reproduced on the same partitions and task definitions and score the full test sets. The 64-atom budget leaves 2 to 93 of the largest test molecules out of MOLGRID's classification sets, and five HIV SMILES do not parse (Appendix H.2). A lead counts when it clears the pooled seed standard deviation $\sqrt{(\sigma_a^2+\sigma_b^2)/2}$, a descriptive margin, and every causal claim is a paired test over seeds, corrected across the map. An arm and its control differ in one variable and run on one GPU class.

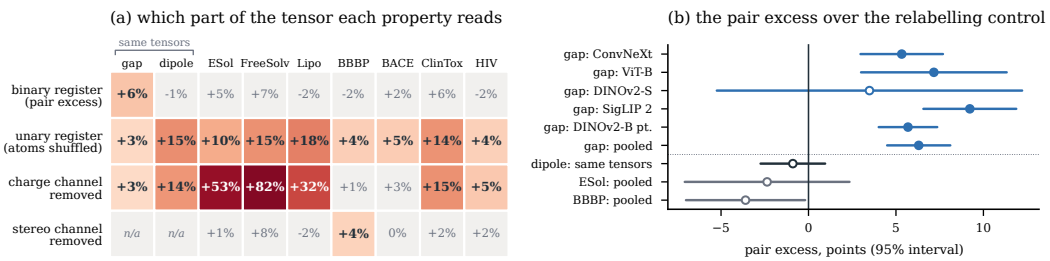


Figure 5: **The causal map: each property reads the part of the tensor the decomposition predicts.** (a) Relative test-error increase after each intervention on the reported DINOv2-B at ten seeds, coloured where a paired test over seeds survives a Benjamini–Hochberg correction at $q = 0.05$ over the 34 testable cells, which removes none of the cells that pass uncorrected. The binary row is the pair excess, the gap and the dipole share the same tensors, and the QM9 stereo cells are marked n/a because its SMILES carry no stereo tags. (b) The pair excess with its 95% interval, filled where it is positive with $p \leq 0.05$, and pooled rows combining the four ImageNet families.

6.1 Where does MolGrid stand against reproduced baselines?

Selected on validation, MolGrid leads each reproduced 2D baseline on five of seven tasks (Table 3). The baselines are the D-MPNN graph network (Yang et al., 2019), MolFormer-XL and the ECFP4 random forest, and every lead clears the pooled seed spread. MolGrid also leads the 3D model Uni-Mol on aqueous solubility (ESol), where the two score the same molecules. By mean, MolGrid is the only method in the top three on all seven tasks. Across the four baselines it leads 18 of the 27 comparisons, and the best configuration we ran leads 20. Where a comparator leads, the shortfall is concentrated, on Lipophilicity in the molecules larger than those in the pretraining corpus and on FreeSolv in two of its 64 test molecules, which carry half of it (Appendix G.4).

The standing belongs to the representation rather than to one backbone. Inside MolSight's published image pipeline, replacing its rendered drawing with MolGrid stage for stage lowers the QM9 gap error by 17% (Appendix G.3). Twelve backbones from 94K to 93M parameters share the seven task wins on MolGrid, six of them taking at least one (Appendix F). The locked recipe also carries to four further encoders, three text-trained transformers and one convolutional masked autoencoder, each leading ECFP4 on four or five of the seven tasks (Appendix F.2). On the four multi-label sets MolGrid leads the reproduced fingerprint bar on three (Appendix G.1).

6.2 Which register does each property read?

Only the HOMO–LUMO gap pays for losing which atom pair carries which value, and every property pays for losing which atom carries which value. The interventions of Sec. 5 establish the first half on the QM9 gap, ESol and blood–brain-barrier permeability (BBBP) in five backbones of four families, four from ImageNet initialisation and the reported DINOv2-B after chemical pretraining. The causal map of Figure 5 carries every intervention to nine properties on the reported backbone.

The relabelling control answers the standing objection that an intervention shows the model an unfamiliar input, and the dipole answers it by design. It trains on the gap's own intervened tensors, so its inputs are exactly as unfamiliar, yet it pays no pair excess, and unfamiliarity cannot produce a cost that only one target pays. The gap's pair excess is positive in all five backbones, with its interval above zero in four and in the pooled estimate (Figure 5b). On the reported backbone it is 5.7 points (95% interval 4.0 to 7.4), positive on all ten paired seeds, and unchanged without the QM9 test molecules whose skeleton appears in the pretraining corpus (Appendix H.2).

The binary register carries the gap, and permuting through-bond coupling alone carries the largest single-channel share of it in each of the three families tested (Table 5). Solvation and permeability pay no pair excess, and pooled over the four ImageNet families their intervals exclude one as large as the gap's (Figure 5b, Table 6), so neither reads the pairs.

The unary register carries what composition solves. Removing the charge channel raises ESol error by about half in every family, several times what it costs the gap. Solvation reads those charges atom

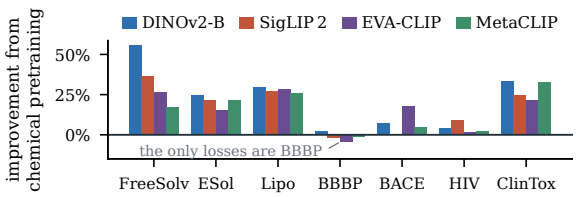


Figure 6: **Chemical pretraining on MOLGRID improves nearly every backbone and task.** Improvement from pretraining for four backbones on seven tasks, all else held fixed. The only losses are small ones on BBBP. Per-cell values are in Table 11.

by atom: keeping only the per-atom charges recovers most of the channel's value on ESol and nearly all of it on FreeSolv and Lipophilicity, where the pair products alone recover far less (Table 8). An additive property needs its atoms' charges, not their pairs. Solvation does without the pairs and the gap largely without the charges, the double dissociation that the decomposition predicts.

Across the nine properties of the causal map (Figure 5a), the charges matter most for the three solvation tasks, and only BBBP reads stereochemistry. On the gap's molecules the dipole, a sum over atoms of charge times position, reads the charges but not the pairs, so two quantum targets of the same molecules split by additivity. **Across five backbones the gap reads the pairs, solvation the charges and permeability neither, so the pattern does not depend on the backbone.**

6.3 WHAT DOES CHEMICAL PRETRAINING CHANGE?

Each of four backbones is finetuned twice, once from ImageNet initialisation and once after the chemical pretraining of Sec. 5, with all else fixed. Pretraining helps on 25 of 28 backbone–task pairs and clears the pooled seed spread on 22 (Figure 6). The gain is large on the regression tasks, where four encoders that begin far apart converge on nearly one value. The pretrained DINOv2-B reads the pairs where the four ImageNet-initialised families do, on the gap alone (Table 6). A matched pair shows it directly: SigLIP 2 pays nothing beyond the control for shuffled pairs on ESol before or after pretraining, while pretraining lowers its ESol error by 14%.

7 CONCLUSIONS

We asked which chemistry a molecular property model reads, and made the question testable with MOLGRID, which states pairwise chemistry explicitly. Retraining after changing one part at a time, against a control that only relabels the atoms, supports four conclusions.

- **Each property reads the chemistry its physics predicts.** Only the HOMO–LUMO gap, a delocalised property, depends on which atom pair holds which coupling. Shuffling the pairs raises its error 5.7 percentage points more than relabelling the atoms does, an excess positive on all ten seeds. Every property reads the per-atom values. Solvation and lipophilicity depend most on the partial charges, read atom by atom, and only permeability reads stereochemistry.
- **Stating pairwise chemistry matters beyond the reach of depth.** A graph network of depth three recovers the coupling across the middle of a long bond path barely better than raw atom features (R^2 of 0.09 against 0.04), yet up to half of the atom pairs in a dataset lie beyond that reach. MOLGRID states the coupling in every cell.
- **Pretraining improves accuracy without changing what is read.** Chemical pretraining improves 25 of 28 backbone–task pairs, with large gains on the regression tasks, and the pretrained backbone still reads the pairs for the gap alone, as the four ImageNet-initialised backbones do.
- **The representation competes.** Without 3D input, MOLGRID leads reproduced D-MPNN, MolFormer-XL and ECFP4 baselines on five of seven tasks each, and 3D Uni-Mol on solubility. By mean it is the only method in the top three on all seven tasks.

These results give a design rule: a delocalised property needs its pairwise chemistry stated, while an additive one is served by per-atom descriptors, as group contribution has long served solubility. **Once the chemistry is stated explicitly, what a model reads can be measured rather than guessed.**

Limitations. We did not explore fusion of MOLGRID with graph and 3D models, or molecules beyond $N = 64$ heavy atoms, a budget that covers nearly all benchmark molecules. Larger grids need no change to the method, since the backbones resample every grid to 224×224 pixels.

REPRODUCIBILITY STATEMENT

We release the MOLGRID dataset builders, the content-scramble intervention, the pretraining and finetuning recipes as frozen configuration files, and the regression suite that pins the invariants the claims rest on. Together these reproduce every result in the paper. Sec. 4 specifies the representation in full, Appendix C.1 gives the exact per-channel formulas, the padding convention and the two presentations, and Proposition 1 states, and Appendix C.1 proves, the equivariance and diagonal symmetry. Sec. 5 defines the causal content-scramble, and Appendix E.2 verifies its invariants on the shipped arrays. The released analysis scripts reproduce the depth probe of Appendix E.1 and the learner-agnostic probe of Appendix G.2. All datasets are public: MoleculeNet under a uniform Bemis–Murcko scaffold split with the evaluated split sizes stated (Appendix H.2), QM9 with its shipped DFT-optimised geometry, and PCQM4Mv2 for pretraining. Appendix D.1 states the pretraining objectives in full, and Appendix D.2 lists the as-run hyperparameters of both stages. Every reported benchmark number is a mean over at least three seeds with its seed standard deviation, or a 95% interval where a caption says so, and the reported configuration is selected on validation. We save the per-molecule predictions of the reported MOLGRID configurations, and of the reproduced baselines where their pipelines allow, so that paired analyses can be recomputed, and we disclose the compute budget in Appendix H.3. The AI use statement below describes the role of language models.

ETHICS STATEMENT

This work introduces a representation and an analysis methodology for molecular property prediction. It involves no human subjects and no personal or private data, and it uses only public benchmark datasets (MoleculeNet, QM9, PCQM4Mv2) under their published licenses. Molecular property prediction is dual-use. A model that predicts toxicity or poor absorption to steer discovery toward safer compounds could in principle be applied to harmful molecules. We view this as the baseline risk of the field rather than a risk introduced here. Our contribution is a representation evaluated on established public benchmarks, not a generative or optimisation system for the design of new molecules, and the released artifacts are smaller than, and comparable in capability to, existing public chemistry backbones. We disclose the pretraining compute (Appendix H.3) for transparency about environmental cost and declare no competing interests. Appendix I gives a fuller discussion.

AI USE STATEMENT

We used large language models as assistive tools in two ways. In preparing the manuscript, they helped polish the text and identify and correct grammatical inconsistencies. In the experiments, they helped standardise the code. The authors take full responsibility for the content of this paper.

REFERENCES

Deepan Adak, Yogesh S. Rawat, and Shruti Vyas. MolVision: Molecular property prediction with vision language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations (ICLR)*, 2021.

Aaditya Baranwal and Shruti Vyas. ChemPro: A progressive chemistry benchmark for large language models. *Artificial Intelligence Chemistry*, 4(1):100118, 2026. doi: 10.1016/j.aichem.2026.100118.

Aaditya Baranwal, Akshaj Gupta, Yogesh S. Rawat, and Shruti Vyas. MolSight: Molecular property prediction with images. *arXiv preprint arXiv:2605.10157*, 2026.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

David N. Beratan, Jonathan N. Betts, and José N. Onuchic. Protein electron transfer rates set by the bridging secondary and tertiary structure. *Science*, 252(5010):1285–1288, 1991. doi: 10.1126/science.1656523.

Feiyang Cai, Katelin Zacour, Tianyu Zhu, Tzuen-Rong Tzeng, Yongping Duan, Ling Liu, Srikanth Pilla, Gang Li, and Feng Luo. ChemFM as a scaling law guided foundation model pre-trained on informative chemicals. *Communications Chemistry*, 9:3, 2026. doi: 10.1038/s42004-025-01793-8.

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. doi: 10.1021/ci00046a002.

Bryan Cheng, Austin Jin, and Jasper Zhang. When three-dimensional conformer ensembles improve molecular property prediction beyond two-dimensional fingerprints: A systematic study. *arXiv preprint arXiv:2606.08825*, 2026.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations (ICLR)*, 2021.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Isidro Cortés-Ciriano and Andreas Bender. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *Journal of Cheminformatics*, 11:41, 2019. doi: 10.1186/s13321-019-0364-5.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

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*, 14:6395, 2023. doi: 10.1038/s41467-023-41948-6.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Carl Edwards, Chi Han, Gawon Lee, Thao Nguyen, Sara Szymkuć, Chetan Kumar Prasad, Bowen Jin, Jiawei Han, Ying Diao, Ge Liu, Hao Peng, Bartosz A. Grzybowski, Martin D. Burke, and Heng Ji. mCLM: A modular chemical language model that generates functional and makeable molecules. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.12565.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics (TACL)*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022. doi: 10.1038/s42256-021-00438-4.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. doi: 10.1016/j.imavis.2024.105171.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research (JMLR)*, 20(177):1–81, 2019.

Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 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.

Garrett B. Goh, Charles Siegel, Abhinav Vishnu, Nathan O. Hodas, and Nathan Baker. How much chemistry does a deep neural network need to know to make accurate predictions? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Yonatan Harnik, Hadas Shalit Peleg, Amit H. Bermano, and Anat Milo. Data efficient molecular image representation learning using foundation models. *Chemical Science*, 16(24):10833–10841, 2025. doi: 10.1039/d5sc00907c.

Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hamza Hentabli, Billel Bengherbia, Faisal Saeed, Naomie Salim, Ibtehal Nafea, Abdelmoughni Toubal, and Maged Nasser. Convolutional neural network model based on 2D fingerprint for bioactivity prediction. *International Journal of Molecular Sciences*, 23(21):13230, 2022. doi: 10.3390/ijms232113230.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. Triplet interaction improves graph transformers: Accurate molecular graph learning with triplet graph transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Uni-Mol2: Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kevin G. Joback and Robert C. Reid. Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications*, 57(1-6):233–243, 1987. doi: 10.1080/00986448708960487.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning (ICML)*, 2023.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Greg Landrum et al. RDKit: Open-source cheminformatics, 2025. `http://www.rdkit.org`.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019. doi: 10.1038/s41467-019-08987-4.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, et al. ChemVLM: Exploring the power of multimodal large language models in chemistry area. In *AAAI*, 2025.

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

Yunqing Liu, Yi Zhou, and Wenqi Fan. Enhancing molecular property predictions by learning from bond modelling and interactions. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.00568.

Per-Olov Löwdin. A note on the quantum-mechanical perturbation theory. *The Journal of Chemical Physics*, 19(11):1396–1401, 1951. doi: 10.1063/1.1748067.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7628–7636, 2022. doi: 10.1609/aaai.v36i7.20729.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. *Nature Communications*, 15: 7104, 2024. doi: 10.1038/s41467-024-51321-w.

Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2D & 3D molecular data. In *International Conference on Learning Representations (ICLR)*, 2023.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. Relative molecule self-attention transformer. *Journal of Cheminformatics*, 16:3, 2024. doi: 10.1186/s13321-023-00789-7.

Harden M. McConnell. Intramolecular charge transfer in aromatic free radicals. *The Journal of Chemical Physics*, 35(2):508–515, 1961. doi: 10.1063/1.1731961.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pretrained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024.

Maho Nakata and Tomomi Shimazaki. PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57 (6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083.

Marshall D. Newton. Quantum chemical probes of electron-transfer kinetics: The nature of donor-acceptor interactions. *Chemical Reviews*, 91(5):767–792, 1991. doi: 10.1021/cr00005a007.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. Rethinking CNN models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

Junwoo Park and Sujee Lee. MoltiTox: A multimodal fusion model for molecular toxicity prediction. *Frontiers in Toxicology*, 7:1720651, 2025. doi: 10.3389/ftox.2025.1720651.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations (ICLR)*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

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

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

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

Christoph Rücker, Gerta Rücker, and Markus Meringer. y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, 47(6):2345–2357, 2007. doi: 10.1021/ci700157b.

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. doi: 10.1103/PhysRevLett.108.058301.

R. T. Sanderson. An interpretation of bond lengths and a classification of bonds. *Science*, 114(2973): 670–672, 1951. doi: 10.1126/science.114.2973.670.

Alok Sharma, Edwin Vans, Daichi Shigemizu, Keith A. Boroevich, and Tatsuhiko Tsunoda. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports*, 9:11399, 2019. doi: 10.1038/s41598-019-47765-6.

Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: Align then refine. In *International Conference on Machine Learning (ICML)*, 2023.

Shuaike Shen, Jiaqing Xie, Zhuo Yang, Antong Zhang, Shuzhou Sun, Ben Gao, Tianfan Fu, Biqing Qi, and Yuqiang Li. SpecMol: A spectroscopy-grounded foundation model for multi-task molecular learning. *arXiv preprint arXiv:2509.21861*, 2025.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves GNNs for molecular property prediction. In *International Conference on Machine Learning (ICML)*, 2022.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations (ICLR)*, 2022.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Alonso Urbano, David W. Romero, Max Zimmer, and Sebastian Pokutta. RECON: Robust symmetry discovery via explicit canonical orientation normalization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.13289.

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. doi: 10.1021/acs.jcim.2c01073.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology*, 13(1):e1005324, 2017. doi: 10.1371/journal.pcbi.1005324.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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. doi: 10.1039/C7SC02664A.

Hongxin Xiang, Li Zeng, Linlin Hou, Kenli Li, Zhimin Fu, Yunguang Qiu, Ruth Nussinov, Jianying Hu, Michal Rosen-Zvi, Xiangxiang Zeng, and Feixiong Cheng. A molecular video-derived foundation model for scientific drug discovery. *Nature Communications*, 15(1):9696, 2024. doi: 10.1038/s41467-024-53742-z.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024a.

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

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3D: A benchmark for predicting 3D geometries from molecular graphs. *arXiv preprint arXiv:2110.01717*, 2021.

Zhao Xu, Yaochen Xie, Youzhi Luo, Xuan Zhang, Xinyi Xu, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. 3D molecular geometry analysis with 2D graphs. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 343–351, 2024b. doi: 10.1137/1.9781611978032.39.

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. doi: 10.1021/acs.jcim.9b00237.

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

Atsushi Yoshimori. Prediction of molecular properties using molecular topographic map. *Molecules*, 26(15):4475, 2021. doi: 10.3390/molecules26154475.

Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022. doi: 10.1038/s42256-022-00557-6.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, 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.

Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2D and 3D pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2626–2636, 2022. doi: 10.1145/3534678.3539368.

Yitan Zhu, Thomas Brettin, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, Yvonne A. Evrard, James H. Doroshow, and Rick L. Stevens. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific Reports*, 11:11325, 2021. doi: 10.1038/s41598-021-90923-y.

Wei Zhuo and Siqiang Luo. Modality-free graph in-context alignment. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.13434.