SKELEMOL: MOLECULAR PROPERTY PREDICTION WITH SKELETAL IMAGES

Anonymous authors
Paper under double-blind review

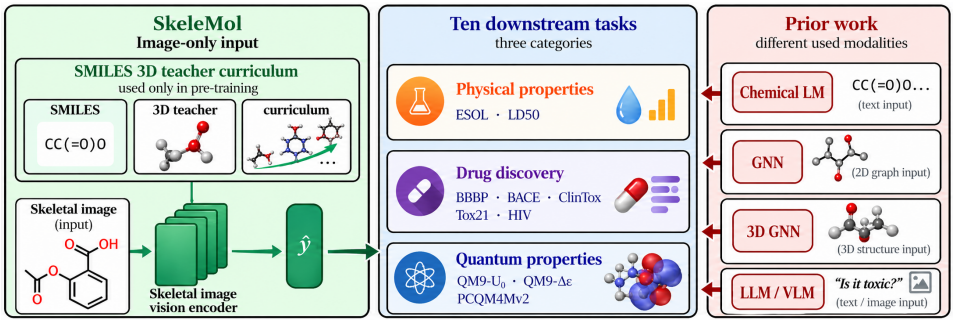


Figure 1: **Overview of SkeleMol.** *Left:* one vision encoder reads only the skeletal image at test time; SMILES, a 3D teacher and a curriculum are used only in pre-training. *Right:* prior work takes inputs from multiple modalities, each with its own kind of model.

ABSTRACT

Molecular property prediction is central to drug discovery and chemistry, as it lets candidate molecules be screened before synthesis and testing. It typically relies on specialized representations such as molecular graphs, SMILES strings, or 3D conformers, together with models pre-trained specifically on molecular data. Models for each representation are growing increasingly complex, and the field is fragmenting along input modalities. We investigate a different approach: representing molecules as skeletal images and adapting general-purpose vision encoders to molecular prediction. We introduce **SkeleMol**, an image-based framework that uses only a molecular drawing at inference time. It enables pre-trained vision encoders for molecular property prediction, replacing the molecule-specific architecture. During pre-training, we progressively inject chemical and geometric information through image-SMILES alignment, distillation from a 3D molecular model, and a chemistry-informed curriculum. Across 10 vision architectures and seven pre-training strategies, we find that the choice of molecular supervision has a substantially larger effect than the choice of vision backbone, with consistent strategy rankings across architectures. In particular, transferring 3D knowledge improves an image encoder despite the absence of conformers at inference, while curriculum learning further improves data efficiency. Across 10 benchmarks spanning physical, biological, and quantum properties, **SkeleMol** achieves state-of-the-art results on five tasks using 2 M pre-training molecules, while retaining a standard vision architecture and image-only inference.

1 INTRODUCTION

Molecular property prediction (MPP) maps a representation of a molecule to a physical, biological or quantum-chemical property, and the choice of representation fixes both the inductive bias of the model and the data on which it can be pre-trained. Message-passing networks operate on the molecular graph (Gilmer et al., 2017; Yang et al., 2019), 3D transformers on atomic coordinates from a generated conformer (Zhou et al., 2023), and chemical language models on SMILES strings (Edwards et al., 2022). Each learns its representation from molecular data alone, and the strongest string model is pre-trained on over a billion molecules (Ross et al., 2022). The skeletal image, the notation

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.

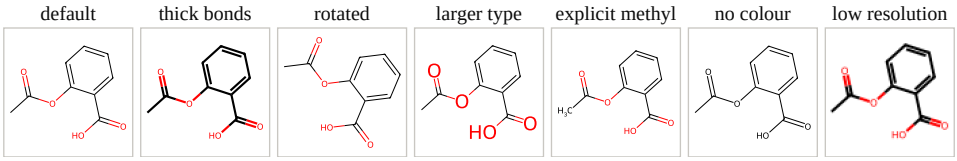


Figure 2: **One compound, several drawing conventions.** A chemist identifies every panel as the same molecule; an encoder trained on the defaults of one renderer must learn this invariance, which App. D.2 evaluates across several drawing variants.

in which chemists communicate structure, admits a different route: as an ordinary image, it can be encoded by vision backbones pre-trained on web-scale image collections (Radford et al., 2021; Zhai et al., 2023; Oquab et al., 2024), so that molecular pre-training adapts an existing visual representation instead of learning one from scratch. The image has nevertheless received far less attention as an input for MPP than the graph, the conformer or the string.

We study *image-only inference*: pre-training may use auxiliary signals, namely SMILES strings, text descriptions and embeddings from a 3D teacher, whereas at test time the model receives only the image x and predicts $\hat{y} = g(f_\theta(x))$ with an image encoder f_θ and a lightweight head g (Fig. 1). We introduce **SkeleMol** in this setting. The image encodes the 2D structure but not the 3D geometry on which many properties depend, and the most informative supervision, from 3D geometry and text, is unavailable at inference. Two questions follow: which pre-training signals transfer into f_θ, and how the missing geometry can be supplied during pre-training.

Our contributions are threefold. (i) A controlled study of image-only MPP over seven pre-training strategies and 10 architectures under a single downstream protocol, which shows that the pre-training strategy matters more than the architecture: every backbone in the strategy matrix ranks the strategies in the same order (Sec. 5). (ii) A training recipe that supplies the missing geometry. A frozen 3D teacher is distilled into f_θ during pre-training, and a curriculum over five chemistry-informed complexity tiers orders the image-SMILES stage, lowering regression errors with 42.6% fewer molecule-views, so the gain is attributable to ordering rather than to additional data (Secs. 3 and 4). (iii) An evaluation against graph, 3D and string models. A single image model is state of the art on five of the ten tasks with $550\times$ fewer pre-training molecules than the strongest string model it matches, the distilled student exceeds its 3D teacher on solubility and three of five classification tasks, and robustness analyses over splits, drawing style and augmentation support them.

The approach is simple and requires no molecule-specific architecture. For computer vision it adds a prediction problem that general-purpose image encoders address directly; for chemistry it provides a predictor that reads the notation chemists use; and for representation learning it shows that geometric knowledge transfers from a 3D teacher to a student that never observes a conformer.

2 Related Work

Graph and 3D molecular models. Message-passing networks (Gilmer et al., 2017) learn on the molecular graph, with later work adding directed message passing (Yang et al., 2019), attentive pooling (Xiong et al., 2020), self-supervised pre-training (Rong et al., 2020) and long-range coupling (Rampášek et al., 2022); Uni-Mol (Zhou et al., 2023; Lu et al., 2024) pre-trains a transformer on 3D conformers and requires a conformer at every inference. Self-supervised variants align 2D topology with 3D geometry (GraphMVP (Liu et al., 2022), 3D InfoMax (Stärk et al., 2022)), add a generative objective (MoleculeSDE (Liu et al., 2023)), blend atom relations (MOLEBLEND (Yu et al., 2024)) or model deep multimodal interactions (MOLINTERACT (Soga et al., 2025)); DeMol (Liu et al., 2026) pairs atom- and bond-centric graphs over 3.37 M PCQM4Mv2 molecules and reports the lowest quantum errors among the methods we compare.

Molecular language models. SMILES and SELFIES (Krenn et al., 2022) transformers, including ChemBERTa and ChemBERTa-2 (Chithrananda et al., 2020; Ahmad et al., 2022), MolT5 (Edwards et al., 2022), SELFormer (Yüksel et al., 2023), AdaMR (Ding et al., 2024) and MoLFormer (Ross et al., 2022), transfer masked or autoregressive pre-training to property prediction; tokenisation is compact and scalable but discards the 2D layout of the structure drawing.

ID	Strategy	Key Objective	Modalities	Dataset
S1	No Pre-training	Supervised fine-tune only	Image	–
S2	Image-Image (SimCLR)	NT-Xent on augmented image pairs	Image	MolTextNet-1M
S3	Image-SMILES	Sigmoid contrastive (Eq. 1)	Image + SMILES	MolTextNet-1M
S4	Image-Description	Sigmoid contrastive on NL descriptions	Image + Text	MolTextNet-1M
S5	GNN Distillation	S3, then hybrid loss (Eq. 2)	Image + SMILES, GNN	MolTextNet-1M, PCQM4Mv2†
S6	Curriculum	S5 with tier-ordered S3 stage (Sec. 4)	Image + SMILES, GNN	MolTextNet-1M, PCQM4Mv2†
S7	Joint SLIP	S2 + S3 simultaneously ($\mathcal{L}_{\text{SigLIP}} + \lambda\mathcal{L}_{\text{SimCLR}}$)	Image + SMILES	MolTextNet-1M

Table 1: **Pre-training strategies,** by increasing supervision. S5 and S6 run in two stages, image-SMILES on MolTextNet-1M, then distillation on PCQM4Mv2. †: uses Uni-Mol embeddings.

Vision and vision-language models. SimCLR (Chen et al., 2020) learns from images alone, without chemical grounding; CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2024), EVA-CLIP (Sun et al., 2023) and Long-CLIP (Zhang et al., 2024) supply transferable backbones, most of which can be paired with chemical text. MolVision (Adak et al., 2025) and general VLMs like GPT-4o (OpenAI, 2024) and Janus-Pro (Chen et al., 2025) answer prompted property queries, though quantitative chemistry remains hard for frontier models (Baranwal & Vyas, 2026), and all need inputs from multiple modalities or a large decoder at test time.

Image-based property prediction. Chemception (Goh et al., 2017) showed that a CNN trained on drawings alone matches expert-descriptor QSAR models, KekuleScope (Cortés-Ciriano & Bender, 2019) transferred ImageNet CNNs to potency prediction, and ImageMol (Zeng et al., 2022) scaled self-supervised pre-training to ten million images. These works establish that the image is predictive but leave open which supervision an image encoder requires and whether geometry absent from the image can be transferred to it. **SkeleMol** addresses both under one protocol: image-image contrastive pre-training outperforms only the no-pre-training baseline on every backbone; the largest gains come from cross-modal supervision, first from SMILES and then from a distilled 3D teacher; and a chemistry-ordered curriculum makes this supervision data-efficient.

Optical chemical structure recognition. DECIMER (Rajan et al., 2020), MolScribe (Qian et al., 2023) and MolNexTR (Chen et al., 2024) map a depiction back to a SMILES string or graph and randomise drawing style heavily to tolerate the conventions of patents and journals, which identifies the failure mode of a single-renderer pipeline and motivates our depiction-robustness analysis (Fig. 2, App. D.2). OCSR recovers the structure; we predict the property directly. Concurrent work injects bond adjacency into vision tokens but keeps a language decoder at inference (Wang et al., 2026), whereas **SkeleMol** keeps only an image encoder and a lightweight head.

Cross-modal distillation. Teacher-to-student transfer across modalities has been effective for audio-visual (Aytar et al., 2016), image-text (Kim et al., 2025) and camera-radar (Zhao et al., 2024) models; in molecular learning it usually compresses a graph model into a smaller one or, in GraphMVP and 3D InfoMax, moves 3D geometry into a 2D graph encoder. We instead distil a Uni-Mol teacher representation into an image encoder whose only test-time input is the image (Table 7).

3 METHODOLOGY

SkeleMol compares two design choices under one downstream protocol: the visual architecture and the molecular pre-training strategy. The curriculum changes only the order in which the image-SMILES corpus is presented (Sec. 4). Everything after pre-training is held fixed across the matrix: the renderer, the prediction head, the fine-tuning schedule and the three seeds. A difference between two cells is therefore attributable to the pre-training each received.

Visual architectures and input. The study spans two families (Table 8). Four vision-only encoders have no text tower: ResNet-18 and ResNet-50 are ConvNet controls at 12 M and 25 M parameters, while DINOv2 and DINOv3 are self-supervised ViT-B/14 backbones. Six vision-language encoders, SigLIP, SigLIP2, OpenCLIP, Long-CLIP, MetaCLIP and EVA-CLIP, pair a ViT-B/16 image tower with a text encoder. Eight of the ten share an 86 M-parameter ViT-B image tower, so comparing across them varies pre-training provenance rather than capacity. Because only the vision-language encoders have a text tower, the strategies that pair the image with SMILES or descriptions

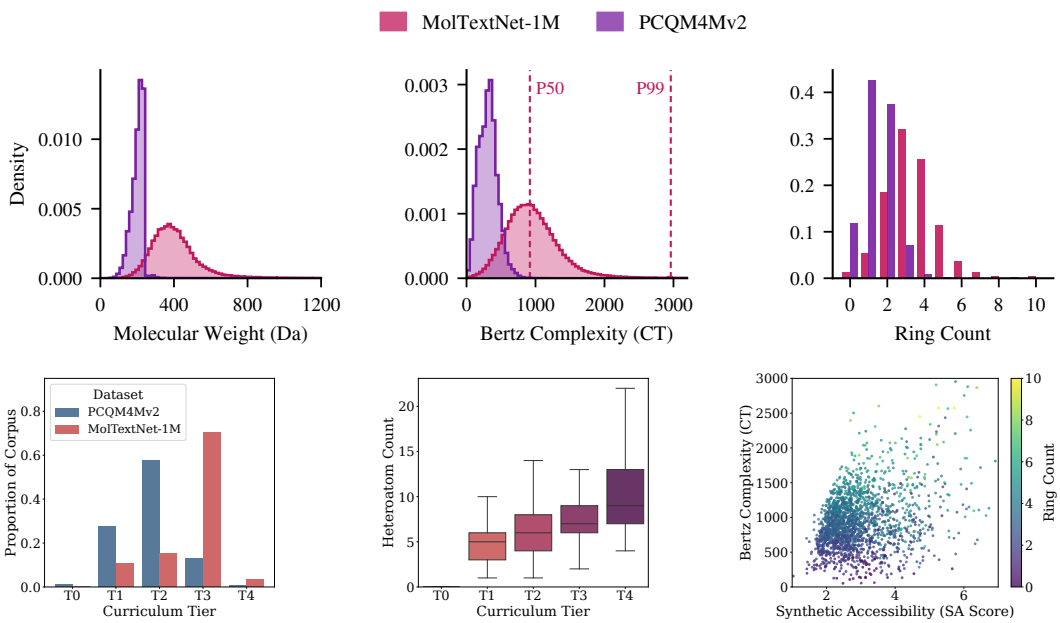


Figure 3: **Pre-training corpora and curriculum tiers.** Top: MolTextNet-1M is heavier, more complex and more ring-rich than PCQM4Mv2. Bottom: tier shares of both corpora, then heteroatom count by tier and SA score against Bertz complexity in MolTextNet-1M.

apply to them alone, whereas image-image pre-training applies to all ten. At inference only the image encoder f_θ and the lightweight head g are retained, so every configuration evaluates the same test-time predictor $\hat{y} = g(f_\theta(x))$ defined in Sec. 1.

RDKit renders each molecule as a 224×224 RGB skeletal image with default atom colouring, following the standard convention for chemical structure drawings. Because it is drawn from the molecular graph, the image encodes atoms, bonds, formal charges and stereo annotations but not 3D geometry, which S5 and S6 provide during pre-training. Pixel values are normalised per channel to $[-1, 1]$. The pooled output, the `[CLS]` token for ViTs or the global average pool for ResNets, is ℓ_2-normalised where the pre-training objective requires it.

Pre-training data and strategies. Two corpora of one million molecules each are used in sequence rather than merged. Every image-only and image-text stage uses MolTextNet-1M, which pairs drug-like molecules with their images and canonical SMILES. The teacher stage of S5 and S6 uses a PCQM4Mv2 subset, molecule-disjoint from the downstream PCQM4Mv2 task, with precomputed Uni-Mol embeddings and HOMO-LUMO targets (Table 1). Figure 4 shows examples from both corpora across the complexity tiers of Sec. 4. We specify seven strategies. Figure 5a summarises all seven schematically, and the tables report task-level results for S2 through S6, with S1 as the no-pre-training reference point and S7 included for completeness of the design space. They differ in objective, supervision, corpus and applicable architectures, so broad comparisons across strategies contrast complete training pipelines rather than isolated components.

S1 (No molecular pre-training). The encoder is initialised from ImageNet weights for ResNets or from its released checkpoint otherwise, and is fine-tuned directly on each downstream task. *S2 (Image-image contrastive).* Two augmented views of each image are aligned with the NT-Xent objective of SimCLR (Chen et al., 2020) at $\tau = 0.07$ (Eq. 5, App. B), using standard crops, flips, colour jitter and grayscale conversion; because these can alter chemically meaningful features of the image, App. D.4 also reports a crop-only variant, which improves BBBP AUC by 0.98 points.

S3 (Image-SMILES). For image i and text j, let $\hat{v}_i$ and $\hat{t}_j$ be normalised embeddings and $\ell_{ij} = 2\mathbf{1}[i = j] - 1$. SigLIP optimises

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{N^2} \sum_{i,j} \log \sigma\big(\ell_{ij} \left(e^\tau \, \hat{v}_i^\top \hat{t}_j + b\right)\big) \,, \tag{1}$$

where τ and b are learned. *S4 (Image-description).* S4 uses the same objective as S3 but replaces SMILES with descriptions that contain functional-group summaries and IUPAC names.

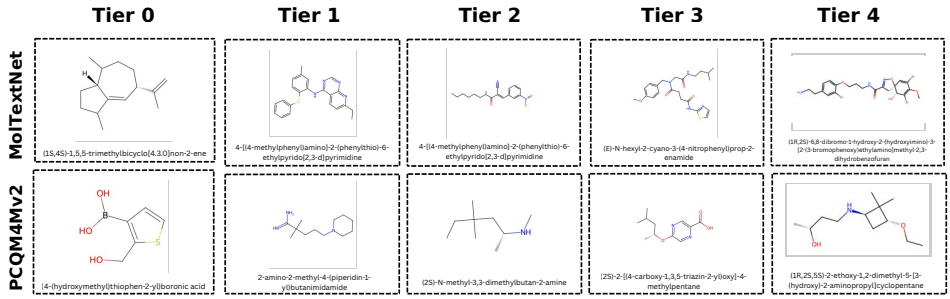


Figure 4: **Representative curriculum tiers.** Example molecules from each tier of both corpora.

S5 (GNN distillation). S5 continues from the S3 image-SMILES checkpoint with a teacher stage on PCQM4Mv2. A frozen Uni-Mol teacher produces $g_i \in \mathbb{R}^{512}$ from a 3D conformer. The image encoder produces $h_i = f_\theta(x_i)$, a trainable projection gives $q_i = \phi(g_i)$, and an auxiliary head $\mathcal{H}(h_i)$ predicts the HOMO-LUMO target y_i. With $\hat{h}_i = h_i/\|h_i\|_2$ and $\hat{q}_i = q_i/\|q_i\|_2$, the loss is

$$\mathcal{L}_{\text{Hybrid}} = \mathcal{L}_{\text{SigLIP}}(\hat{h}, \hat{q}) + \alpha\|q_i - h_i\|_2^2 + \beta\|\mathcal{H}(h_i) - y_i\|_2^2, \tag{2}$$

where $\alpha = 10$ and $\beta = 1$. The three terms act at different scales. The sigmoid term scores each image embedding against all projected teacher embeddings in the batch and thus constrains relative positions. The coordinate term regresses the student embedding onto the projected teacher embedding and carries the largest weight. Here ϕ is a linear map from the teacher's 512 dimensions to the student's 768, and $\mathcal{H}$ is a small MLP. The teacher, projection and auxiliary head are discarded after pre-training, so inference uses the image alone, as for every other strategy.

The embedding and gap terms are complementary rather than alternative forms of supervision, because they constrain the student embedding in subspaces of different dimension. With respect to the student embedding,

$$\nabla_{h_i} \beta\|\mathcal{H}(h_i) - y_i\|_2^2 = 2\beta\big(\mathcal{H}(h_i) - y_i\big)\nabla\mathcal{H}(h_i), \qquad \nabla_{h_i} \alpha\|q_i - h_i\|_2^2 = 2\alpha\,(h_i - q_i). \tag{3}$$

The first gradient is proportional to $\nabla\mathcal{H}(h_i)$, so for each molecule the gap term updates the embedding along a single direction of the 768-dimensional space and leaves unconstrained every perturbation that keeps $\mathcal{H}(h_i)$ fixed. The second gradient has no such restriction and acts in all 768 directions. The gap supplies one scalar per molecule and the teacher a 512-dimensional vector, so the gap term calibrates a single direction of the representation to a DFT quantity, while the relational structure used by the remaining properties is determined by the embedding terms. The teacher stage relies on this complementarity: the embedding terms determine the geometry of the representation, and the gap term aligns one of its directions with the HOMO-LUMO gap.

S6 (Curriculum). S6 runs the S5 recipe with its image-SMILES stage restricted, epoch by epoch, to the active structural-complexity tiers of Sec. 4; the teacher stage is unchanged. *S7 (Joint SLIP).* S7 combines S2 and S3: $\mathcal{L}_{\text{SLIP}} = \mathcal{L}_{\text{SigLIP}} + \lambda\mathcal{L}_{\text{SimCLR}}$ with $\lambda = 1$.

4 Chemistry-Informed Curriculum

Uniform sampling exposes the model to the full range of molecular structures from the first update, so a simple hydrocarbon and a densely fused polycyclic scaffold contribute equally to early gradients. Curriculum learning (Bengio et al., 2009) orders examples from simple to complex instead. We assign each molecule to one of five ordered tiers and expand the active pool as training proceeds. The tiers are proxies for structural complexity in the image. The two pre-training corpora occupy different chemical space (Fig. 3, top). MolTextNet-1M molecules are about twice as heavy as PCQM4Mv2 (mean molecular weight 421.2 vs 202.2 Da) and about three times as topologically complex (mean Bertz complexity 997.8 vs 313.1). Within MolTextNet-1M the Bertz distribution has a long upper tail, so uniform sampling spends much of early training on structures that the encoder cannot yet represent well. The curriculum addresses this non-uniformity.

Tier signals. The assignment rule uses five offline descriptors: heteroatom count n_{het}, the number and corpus commonness of detected functional-group types n_{FG}, aromatic substitution complexity

Tier	Name	Ordered assignment rule
T_0	Hydrocarbon	$n_{\text{het}} = 0$.
T_1	Common low-FG	$n_{\text{FG}} \leq 2$ and every detected functional-group type is among the six most common corpus types.
T_2	Moderate FG	$n_{\text{FG}} \leq 5$ and $S(m) \leq 4$.
T_3	Positional complexity	$S(m) > 4$, or $B(m)/n_{\text{atom}} > 50$ with $n_{\text{ring}} \geq 3$.
T_4	Residual complexity	Any molecule not assigned by the preceding rules.

Table 2: **Deterministic curriculum tiers.** Rules are evaluated from T_0 to T_4 and the first match is used. S: aromatic substitution complexity; B: RDKit Bertz complexity.

$S(m)$, Bertz complexity per atom $B(m)/n_{\text{atom}}$, and ring count n_{ring}. The rule is evaluated in order, so each molecule is placed in the first tier it matches (Table 2); Fig. 4 shows examples from each tier. The tier distribution reflects this corpus difference (Fig. 3, bottom): PCQM4Mv2 is dominated by lower tiers, consistent with its smaller quantum-chemistry molecules, while MolTextNet-1M carries a larger T_3 population of drug-like structures. Heteroatom count rises monotonically across the assigned tiers in both corpora, so the tier order corresponds to increasing chemical functionality.

Additive schedule. For a run of E epochs, let

$$K(e) = \min\left(4,\ 1 + \left\lfloor \frac{4e}{E} \right\rfloor\right), \qquad \mathcal{A}_e = \{T_0, \ldots, T_{K(e)}\}, \tag{4}$$

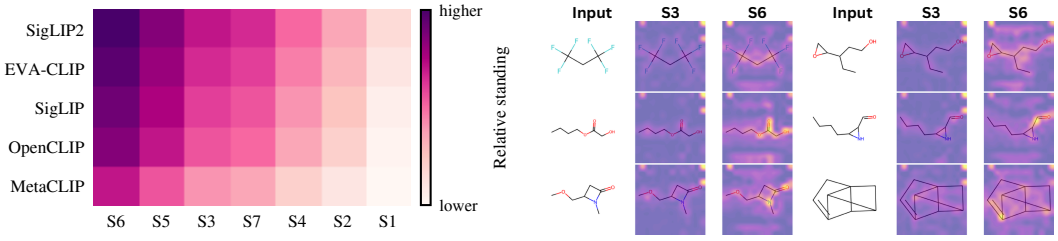
where $\mathcal{A}_e$ is the active tier set at epoch e. The epoch dataset is $\mathcal{D}_e = \{m \in \mathcal{D} : \text{tier}(m) \in \mathcal{A}_e\}$. For the 10-epoch experiments, T_0 and T_1 are active in epochs 0–2, T_2 is added in epochs 3–4, T_3 in epochs 5–7, and all tiers are active in epochs 8–9. Introducing tiers in this order exposes the encoder to simple bonding patterns before the most complex structures and reduces total exposure to 5.74 M molecule-views over the run, counting each molecule once per epoch in which its tier is active. The schedule applies only to the image-SMILES stage on MolTextNet-1M, whose objective and optimiser match the uniform run; the subsequent teacher stage is identical to that of S5.

5 Experiments and Analysis

Evaluation and implementation. The 10 tasks cover physical-property regression (ESOL and LD50), drug-discovery classification (BBBP, BACE, ClinTox, Tox21, HIV), and quantum-chemistry regression (QM9-U_0, QM9-$\Delta\varepsilon$, and PCQM4Mv2). We follow each benchmark's standard split and metric (Table 10 in App. C.3). Because these protocols differ, results are ranked per task and never averaged across tasks. Pre-training uses AdamW with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.98)$, 10 epochs, cosine decay, and 10% linear warmup. Downstream prediction uses a three-layer MLP ($d \to 512 \to 256 \to 1$) with LayerNorm, GELU, and dropout 0.2 between layers. The final two transformer blocks, or final two ResNet stages, are fine-tuned with the head for 50 epochs at a $10\times$ lower backbone learning rate. Literature baselines are reported under their original protocols.

All **SkeleMol** results are means over three seeds (42, 1337 and 2026). Regression targets are standardised with training-split statistics only; the epoch is selected on the validation split, and the test split is evaluated once. Table 3 reports representative configurations, Fig. 5a summarises the full strategy matrix, and App. C.2 gives every architecture under its best strategy. We distinguish two kinds of comparison. Controlled comparisons hold one factor fixed: the strategy matrix summarised in Fig. 5a, the two-backbone, two-strategy block of Table 3, and the fixed-backbone decomposition of Table 4. Best-configuration tables, Table 9 and Table 5, vary several factors at once; we interpret them per task and not as evidence for any single factor.

Pre-training strategy. Across the full matrix, molecular supervision produces larger and more consistent changes than switching among similarly sized backbones (Fig. 5a): every backbone ranks the strategies in the same order, from S6 at the top to image-image pre-training (S2) and no molecular pre-training (S1) at the bottom, and every strategy ranks the backbones in the same order. Table 3 quantifies the effect in task units with a two-by-two block: SigLIP2 and EVA-CLIP, each under S3 and S6. At a fixed backbone, moving from S3 to S6 reduces PCQM error by factors of 5.8 and 3.6 and $\Delta\varepsilon$ error by 32% for both, whereas replacing the backbone at a fixed strategy changes PCQM error by at most a factor of 1.9 and $\Delta\varepsilon$ error by at most 10%. On the drug-discovery columns both effects are smaller, since S3 already reaches 93.6 on BBBP and leaves little headroom. The



(a) (b)

Figure 5: **Effect of the pre-training strategy.** (a) Schematic of the strategy matrix (darker is better; task-level results in Table 3 and App. C.2): all backbones rank the strategies identically, and strategies vary far more than architectures. (b) Attention maps: S3 attends to atom labels and bond junctions, S6 to bonds, rings and functional groups.

			Physical (↓)		Quantum (↓)		Drug Discovery (AUC↑, %)				
ID	**Strategy**	**Arch.**	ESOL	LD50	$\Delta\varepsilon$	PCQM	BBBP	BACE	HIV	ClinTox	Tox21
S2	Img-Img	ResNet-50	1.330	0.580	0.0100	13.860	87.0	82.0	74.0	76.0	82.0
S2	Img-Img	DINOv3	0.774	0.630	0.0100	13.670	89.5	84.4	74.3	80.3	84.0
S3	Img-SMILES	SigLIP2	0.703	0.403	0.0102	0.549	<u>93.6</u>	<u>88.5</u>	<u>81.3</u>	81.0	**89.2**
S3	Img-SMILES	EVA-CLIP	0.752	0.460	0.0112	0.288	91.7	84.0	78.0	80.0	<u>86.0</u>
S4	Img-Desc.	Long-CLIP	<u>0.670</u>	0.460	0.0122	0.416	91.0	84.0	77.0	83.0	85.0
S6	**Curriculum**	**SigLIP2**	**0.627**	**0.369**	**0.00697**	<u>0.095</u>	**93.7**	**89.0**	**83.2**	**86.8**	77.6$^\diamond$
S6	**Curriculum**	**EVA-CLIP**	0.717	<u>0.384</u>	<u>0.0076</u>	**0.081**	92.0	87.8	79.0	<u>84.0</u>	–

Table 3: **Pre-training strategy comparison.** SigLIP2 and EVA-CLIP under S3 and S6 form a two-by-two block. ESOL: RMSE; LD50, $\Delta\varepsilon$, PCQM: MAE, $\Delta\varepsilon$ in Hartree, PCQM4Mv2 in eV; drug: ROC-AUC (%). $\diamond$: twelve-task scaffold mean (App. C.4). **Bold**: best. <u>Underline</u>: second.

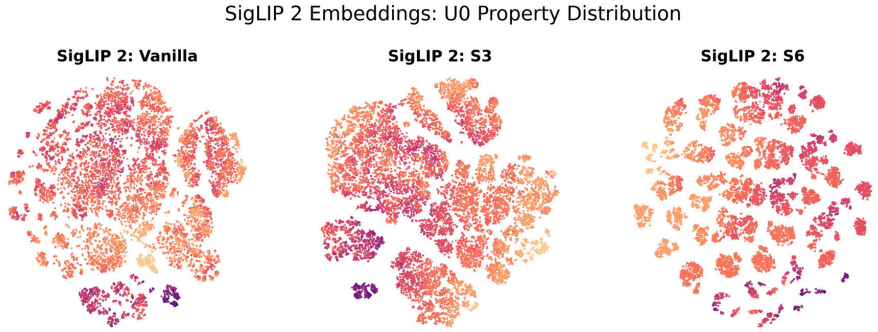


Figure 6: **QM9 validation embeddings.** t-SNE for the released SigLIP2, S3 and S6, coloured by U_0: S3 forms coarse clusters, S6 smaller groups with a smoother U_0 gradient.

step from image-image (S2) to image-SMILES training (S3) is the clearest: the strongest S2 row (DINOv3) reaches ESOL 0.774 and BBBP 89.5, against ESOL 0.703, BBBP 93.6 and HIV 81.3 for S3 with SigLIP2, because image-image contrastive learning aligns augmented views of the same image and receives no information about chemical identity.

Image-SMILES (S3) is also slightly stronger than image-description (S4) on most quantum and drug-discovery columns, for example QM9-$\Delta\varepsilon$ 0.0102 against 0.0122 and BBBP 93.6 against 91.0. Canonical SMILES encodes the complete molecular graph, whereas a description encodes it only partially, which limits what it can transfer. The teacher-supervised strategies S5 and S6 provide the largest improvements on quantum tasks; Table 4 decomposes them at a fixed backbone.

Distillation and curriculum. Table 4 decomposes the recipe at a fixed backbone. The S3 to S5 step adds the teacher stage; the S5 to S6 step changes only the order of the image-SMILES stage. The gradient decomposition in Eq. 3 predicts where the teacher stage acts: most strongly on the two gap tasks, PCQM and QM9-$\Delta\varepsilon$, where both terms bear on the property, and elsewhere to the extent that the teacher's geometry is informative. The results agree with this prediction. The teacher

Config	ESOL↓	LD50↓	$\Delta\varepsilon$↓	PCQM↓	BBBP↑
S3: Image-SMILES	0.703	0.403	0.0102	0.549	<u>93.6</u>
S5: GNN Distil., no curr.	<u>0.702</u>	<u>0.371</u>	<u>0.0080</u>	<u>0.1174</u>	<u>93.6</u>
S6: + Curriculum	**0.627**	**0.369**	**0.00697**	**0.095**	**93.7**
Δ S3→S5 (teacher stage)	-0.1%	-7.9%	-21.6%	-78.6%	$+0.0$
Δ S5→S6 (curriculum)	-10.7%	-0.5%	-12.9%	-19.1%	$+0.1$
Δ S3→S6 (combined)	-10.8%	-8.4%	-31.7%	-82.7%	$+0.1$

Table 4: **Curriculum ablation (SigLIP2), decomposed.** S3→S5 adds the teacher stage; S5→S6 reorders the image-SMILES stage. Three-seed mean; metrics, units and marks as in Table 3.

Category	Method	Pre-train molecules	Physical (↓)		Drug discovery (reported %, see caption)					Quantum (↓)		
			ESOL	LD50	BBBP	BACE	ClinTox	Tox21	HIV	QM9-$\Delta\varepsilon$	QM9-U_0	PCQM
3D GNN	Uni-Mol (Zhou et al., 2023)†	19 M	0.788	–	72.9	85.7	91.9	79.6	80.8	0.0047	0.304*	<u>0.070</u>
3D GNN	Uni-Mol2 (Ji et al., 2024)†	800 M	–	–	–	–	–	–	–	0.0035	–	–
Graph	DeMol (Liu et al., 2026)¶	<u>3.37 M</u>	–	–	75.1	**89.0**	<u>92.6</u>	<u>80.9</u>	81.2	**0.00099**	–	**0.060**
Graph SSL	GraphMVP (Liu et al., 2022)	–	–	–	68.5	76.8	79.0	74.5	74.8	0.00154†	0.00048†	–
Graph SSL	3D InfoMax (Stärk et al., 2022)	–	–	–	69.1	79.7	79.9	74.5	76.1	0.00155†	0.00049†	–
Graph SSL	MoleculeSDE (Liu et al., 2023)	–	–	–	71.8	79.5	87.0	76.8	78.8	0.00154†	0.00044†	–
Graph SSL	MOLEBLEND (Yu et al., 2024)	–	–	–	73.0	83.7	87.6	77.8	79.0	<u>0.00128</u>†	<u>0.00043</u>†	–
Graph SSL	MOLINTERACT (Soga et al., 2025)	–	–	–	68.5	79.1	88.4	77.3	79.5	0.00131†	**0.00028**†	–
1D string	MoLFormer-XL (Ross et al., 2022)	1.1 B	–	–	**93.7**	<u>88.2</u>	**94.8**	**84.7**	<u>82.2</u>	–	–	–
1D string	SELFormer (Yüksel et al., 2023)	**2 M**	–	–	90.2	83.2	–	65.3	68.1	–	–	–
1D string	ChemBERTa-2 (Ahmad et al., 2022)	77 M	–	–	79.9	79.3	60.1	–§	–	–	–	–
Read-across	GenRA (Helman et al., 2019)	–	–	0.580	–	–	–	–	–	–	–	–
LLM	GPT-4o (ICL)‡	–	0.980	0.870	77.0	56.0	59.0	42.0	82.0	8.380	17.943	0.680
VLM	Janus-Pro 7B (ICL)‡	–	0.610	0.720	68.0	78.0	83.0	69.0	92.0	8.530	17.211	0.620
VLM	BLIP-2 (LoRA)‡	–	1.070	0.490	93.0	86.0	89.0	99.0	92.0	4.920	22.120	1.990
Image-only inference	**SkeleMol** (SigLIP2, S6)	**2 M**	**0.627**	**0.369**	**93.7**	**89.0**	86.8	77.6◇	**83.2**	0.00697	0.499	0.095
Image-only inference	**SkeleMol** (EVA-CLIP, S6)	**2 M**	<u>0.717</u>	<u>0.384</u>	<u>92.0</u>	87.8	84.0	–	79.0	0.0076	0.647	0.081

Table 5: **Reported-source comparison.** Each source's own protocol (App. C.4). $\Delta\varepsilon$ and U_0 (raw total energy) in Hartree, PCQM in eV. $*$: our fine-tune of the released model on this split. †/‡: 3D conformer / LLM or VLM decoder at inference. ¶: from 26.8 meV. §: SR-p53 only (83.4). ◇: twelve-task scaffold mean. Bold/underline: best/second outside the LLM/VLM block.

stage, whose PCQM4Mv2 subset shares no molecule with the downstream PCQM task, accounts for most of the PCQM and QM9-$\Delta\varepsilon$ improvement; LD50, an acute-toxicity endpoint unrelated to the orbital gap, improves by 7.9%; and pairwise distances in the distilled encoder correlate with those in Uni-Mol's 3D embedding space at Spearman $\rho = 0.65$, against 0.24 for an untrained encoder (App. G.2), a geometry only the embedding terms constrain. The curriculum improves on S5.

Relative to S5, the curriculum in S6 reduces ESOL error by 10.7%, QM9-$\Delta\varepsilon$ by 12.9%, and PCQM by 19.1%, while drug-discovery AUC is essentially unchanged. This split is consistent with image-SMILES alignment already capturing the functional-group cues that drive pharmacological classification, so the curriculum mainly benefits tasks that depend on finer structural and electronic detail. In total, the S3 to S6 change reduces PCQM error by 82.7% and QM9-$\Delta\varepsilon$ error by 31.7%. Over 10 epochs, the image-SMILES stage of S6 processes 5.74 M molecule-views rather than S5's 10 M, so these gains are attributable to the ordering of the corpus rather than to additional data. S6 also produces embeddings in which U_0 varies more smoothly (Fig. 6; gap projections in App. G.1) and shifts attention toward bonds, aromatic rings and functional groups (Fig. 5b).

Reported-source comparisons. Table 5 compares input modalities, not individual models: each cited row is its source's best configuration under its own protocol, and the **SkeleMol** rows apply one strategy, S6, to off-the-shelf backbones. Drug results are ROC-AUC except for the LLM and VLM rows, which MolVision reports as Accuracy@0.5, so those blocks are not ranked against each other.

Against the molecule-specific baselines, a single model, SigLIP2 under S6, is state of the art on five of the ten tasks: ESOL, LD50, BBBP, BACE and HIV. It is best outright on ESOL (0.627 against Uni-Mol's 0.788), LD50 (0.369 against GenRA's 0.580) and HIV (83.2 against MoLFormer-XL's 82.2), and matches the best reported BBBP and BACE results, obtained by MoLFormer-XL and DeMol with $550\times$ and $1.7\times$ more pre-training molecules. The pre-training budgets behind these rows differ by three orders of magnitude: 1.1 B molecules for MoLFormer-XL and 77 M for ChemBERTa-2, against 2 M here. SELFormer, pre-trained on 2 M drug-like compounds, provides

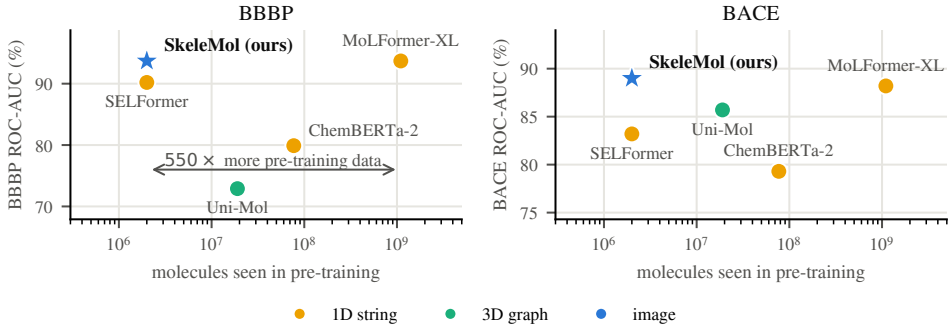


Figure 7: **Accuracy against pre-training corpus size. SkeleMol** lies at the upper left: the highest reported score from the smallest corpus. MoLFormer-XL matches it with 550× more molecules.

Method	Arch.	Params	FLOPs	Prep. ms	Success	ms/mol⋆
Uni-Mol†	3D GNN	48 M	52 G	34.17	89.6%	≥45
BLIP-2 (MolVision)‡	VLM	7.1 B	1.4 T	–	–	–
SkeleMol (ResNet-18)	CNN	12 M	1.8 G	6.65	100%	–
SkeleMol (ViT-B/16)	ViT	86 M	17.6 G	6.65	100%	**9.25**

Table 6: **Inference efficiency** on 10 k PCQM4Mv2 molecules. Prep.: SMILES to model input on one CPU core (RDKit rendering; ETKDGv3 with MMFF94 for Uni-Mol); Success: share completed. ⋆: SMILES to prediction, one CPU core plus one A6000. †, ‡: as in Table 5.

a budget-matched control, and at that budget the image model is ahead on all four shared tasks. A string representation is therefore not intrinsically stronger than an image. Figure 7 plots accuracy against corpus size. Relative to the five self-supervised graph methods, S6 is best on BBBP, BACE and HIV, with 93.7% on BBBP against at most 73.0% and 89.0% on BACE against 83.7%.

Student and teacher by property type. The student exceeds Uni-Mol on solubility, ESOL 0.627 against 0.788, and on three of the five classification tasks, BBBP, BACE and HIV, but not on the quantum targets; this split follows the physical determinants of each property. Solubility, permeability and activity are governed largely by functional groups, polarity and ring systems, all of which are determined by the 2D structure; ESOL was itself introduced as a linear model of four 2D descriptors (Delaney, 2004). The image depicts these features explicitly, and the S6 attention maps cover them (Fig. 5b). These tasks are also label-scarce, with about 1 500 training molecules for BBBP and BACE, a regime in which pre-training matters most (App. E.2). The HOMO-LUMO gap and U_0, in contrast, depend on the electronic structure at a specific 3D geometry, which the teacher observes through the conformer and the student receives only through distillation. Without computing a conformer, the student recovers enough of this geometry to remain among the leading methods on PCQM4Mv2 (App. C.1): both backbones, at 0.081 and 0.095 eV, are within a factor of 1.6 of the best 3D model, DeMol at 0.060 eV, whereas the prompted models' errors are over six times higher.

Inference cost. Table 6 compares cost from SMILES to prediction.

6 Conclusion

The skeletal image of a molecule is an effective input for predicting its properties. Reading only the image at test time, a standard vision encoder is state of the art on five of ten benchmarks, including both physical-property tasks, with 550× fewer pre-training molecules than the strongest chemical language model it matches. Two findings explain this. First, across ten encoders, pre-training matters more than architecture. Second, two pre-training signals contribute most. Distilling a 3D teacher and its HOMO-LUMO target gives the encoder geometric information that the image does not show, and ordering the image-SMILES corpus from simple to complex molecules lowers regression errors with 5.74 M instead of 10 M molecule-views. Because the model is a standard vision backbone, it benefits directly from advances in vision pre-training.

Limitations. Images omit 3D structure, which the teacher supplies only in training, so on quantum targets they are competitive rather than ahead; graphs can complement them.

AI USE STATEMENT

We used large language models (LLMs) as writing and coding aids in two ways. For the manuscript, they helped polish the text and identify and fix grammatical inconsistencies. For the experiments, they helped standardise the code. The authors reviewed every LLM-assisted change and take full responsibility for the content of this paper and its code.

ETHICS STATEMENT

This work involves no human subjects, no personally identifying information, and no new data collection. Every corpus and benchmark we use is public and is cited at first use; we use them under their original licences and keep each benchmark's published split, so the numbers we report are comparable to those of prior work rather than to a split of our own choosing. The molecules we render are drawn from existing screening and quantum-chemistry collections and carry no privacy concerns.

The application area does warrant a note. Property prediction for drug discovery is dual-use in principle, since a model that ranks candidates for a desirable property can be pointed at an undesirable one. The endpoints here are the standard physicochemical, toxicity and quantum benchmarks of the field, the model proposes no molecules and offers no synthesis route, and we release no capability that a published graph or 3D model does not already provide. We would also stress that benchmark accuracy is not evidence of fitness for clinical, regulatory or safety decisions; the reported errors are held-out estimates on curated public datasets and would need prospective validation before any such use.

REPRODUCIBILITY STATEMENT

The three training strategies are specified as pseudocode in Appendix B (Algorithms 1, 2 and 3); the curriculum's descriptor rules and tier thresholds are given in Table 2 and Appendix C.5. Section 3 describes the two pre-training corpora, and Section 5 the ten downstream tasks with their splits and metrics and the final hyper-parameters used for every run; all datasets are public benchmarks and are cited at first use. Every reported **SkeleMol** number is the mean over three seeds (42, 1337 and 2026), with per-seed values in Table 14 (Appendix D.3). The scaffold-split, depiction-robustness, seed-variance and augmentation-safety controls are collected in Appendix D, the cross-architecture and low-label controls in Appendix E, and the timing and sample-budget measurements in Appendix F. Upon publication we will release the source code and configuration files, the rendered-image indices and the curriculum tier labels derived over the existing corpora, and the pre-trained encoder checkpoints, all under a licence permitting free research use.

REFERENCES

Deepan Adak, Yogesh Singh 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.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards Chemical Foundation Models. In *ELLIS Machine Learning for Molecule Discovery Workshop*, 2022.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Aaditya Baranwal and Shruti Vyas. ChemPro: A Progressive Chemistry Benchmark for Large Language Models. *Artificial Intelligence Chemistry*, 4(1):100118, 2026. ISSN 2949-7477. doi: 10.1016/j.aichem.2026.100118.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, pp. 41–48. ACM, 2009.

Steven H. Bertz. The First General Index of Molecular Complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Yufan Chen, Ching Ting Leung, Yong Huang, Jianwei Sun, Hao Chen, and Hanyu Gao. MolNexTR: A Generalized Deep Learning Model for Molecular Image Recognition. *Journal of Cheminformatics*, 16:141, 2024.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. In *Machine Learning for Molecules Workshop at NeurIPS 2020*, 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.

John S. Delaney. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004.

Yan Ding, Hao Cheng, Ziliang Ye, Ruyi Feng, Wei Tian, Peng Xie, Juan Zhang, and Zhongze Gu. AdaMR: Adaptable Molecular Representation for Unified Pre-training Strategy. *arXiv preprint arXiv:2401.06166*, 2024.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 375–413, 2022. doi: 10.18653/v1/2022.emnlp-main.26.

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 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 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.

George Helman, Imran Shah, and Grace Patlewicz. Transitioning the Generalised Read-Across Approach (GenRA) to Quantitative Predictions: A Case Study Using Acute Oral Toxicity Data. *Computational Toxicology*, 12:100097, 2019.

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.

Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. COSMOS: Cross-Modality Self-Distillation for Vision Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14690–14700, 2025.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Falk, Flaviu Duber, Peter Harper, Matteo Jost, Rishikesh Magar, Kelsey Rose, et al. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588, 2022.

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.

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A Group Symmetric Stochastic Differential Equation Model for Molecule Multi-modal Pretraining. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Yunqing Liu, Yi Zhou, and Wenqi Fan. Enhancing molecular property predictions by learning from bond modelling and interactions. *arXiv preprint arXiv:2603.00568*, 2026.

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 (1):7104, 2024. doi: 10.1038/s41467-024-51321-w.

OpenAI. GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. *Journal of Chemical Information and Modeling*, 2023. doi: 10.1021/acs.jcim.2c01480.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. DECIMER: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12:65, 2020. doi: 10.1186/s13321-020-00469-w.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific Data*, 1:140022, 2014.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Lim, Guy Wolf, and Dominique Beaini. Recipe for a General, Powerful, Scalable Graph Transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 14501–14515, 2022.

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.

Patrick Soga, Zhenyu Lei, Camille Bilodeau, and Jundong Li. Deep Interactions for Multimodal Molecular Property Prediction. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, Lecture Notes in Computer Science, pp. 329–341. Springer, 2025. doi: 10.1007/978-981-96-8173-0_26.

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)*, pp. 20479–20502, 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.

Wenda Wang, Yihan Tong, Yuwei Hu, Xuchen Pan, Zhewei Wei, Yaliang Li, and Bolin Ding. MolSight: A Graph-Aware Vision-Language Model for Unified Chemical Image Understanding. *arXiv preprint arXiv:2607.01982*, 2026.

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.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59 (8):3370–3388, 2019.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal Molecular Pretraining via Modality Blending. In *International Conference on Learning Representations (ICLR)*, 2024.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. SELFormer: molecular representation learning via SELFIES language models. *Machine Learning: Science and Technology*, 4(2): 025035, 2023. doi: 10.1088/2632-2153/acdb30.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the Long-Text Capability of CLIP. In *European Conference on Computer Vision (ECCV)*, pp. 310–325, 2024.

Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15470–15480, 2024.

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.

Hao Zhu, Todd M. Martin, Lei Ye, Alexander Sedykh, Douglas M. Young, and Alexander Tropsha. Quantitative Structure–Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chemical Research in Toxicology*, 22(12):1913–1921, 2009.