ZATOM-2: MULTITASK PRETRAINING ON ATOMISTIC DATA FOR GENERATIVE MODELING ACROSS DOMAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified atomistic modeling has the potential to accelerate discovery in chemistry, materials science, and biology by bridging data-rich chemical domains and data-scarce biological contexts. However, existing generative approaches to atomistic modeling remain highly specialized to scientific disciplines (chemistry vs. biology) or do not leverage both high-volume organic (molecule) and inorganic (material) data for general-purpose pretraining. To this end, we introduce Zatom-2, an atomistic generative model pretrained on approximately five million structures from the OMol25 and OMat24 electronic structure datasets. Zatom-2 features a multiscale Transformer architecture coupled with conditional flow matching that supports force conditioning and foundational pretraining tasks such as generation, structure prediction, and prediction of molecular and material energies and forces. Empirically, Zatom-2 achieves better molecular distribution fidelity than Zatom-1 and achieves strong performance on existing molecule and material generation benchmarks. Zatom-2 demonstrates the ability to control sample generation across low- and high-force regimes, and enhances protein generation in a low-data setting through joint generative-predictive pretraining and transfer learning, increasing protein backbone designability in a length extrapolation setting from 66.9% without pretraining to 75.1% after finetuning on 2,000 protein domains. Code and model weights will be made fully open source.

1 INTRODUCTION

Atomistic machine learning has made significant progress across chemistry, materials science, and structural biology, yet these domains have largely developed separate modeling frameworks. For small molecular and materials systems, generative models operate directly on atom identities and three-dimensional coordinates (Hoogeboom et al., 2022; Vonessen et al., 2026; Jiao et al., 2023; Zeni et al., 2025). In contrast, models like RFdiffusion3 and BoltzGen (Butcher et al., 2025; Stark et al., 2025) employ protein-specific tokenization to make all-atom generation tractable without a separate diffusion process over amino acid identities, while La-Proteina encodes sequence and side-chain information into learned per-residue latent variables (Geffner et al., 2026). These approaches are effective within their respective domains, but they do not explore training and finetuning a single model across molecules, periodic materials, and biomolecules. This separation is not intrinsic to the underlying physical systems. Molecules, materials, and proteins are all collections of interacting atoms in three-dimensional space, and many of the local environments that determine their structure and interactions recur across systems of varying size and type.

Recent large-scale quantum-chemical datasets make shared atomistic modeling increasingly practical. In particular, Open Molecules 2025 (OMol25) contains more than 100 million high-fidelity density functional theory calculations (Levine et al., 2025), and its diversity extends along two complementary axes. Chemically, OMol25 covers small organic molecules, biomolecules, metal complexes, electrolytes, varying charge and spin states, and a broad range of intra- and intermolecular interactions. Geometrically, it contains diverse conformers and reactive structures rather than restricting supervision to equilibrium geometries. Open Materials 2024 (OMat24) provides an analogous resource for periodic systems, comprising more than 110 million calculations over diverse

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.

inorganic compositions and configurations (Barroso-Luque et al., 2024). Crucially, both datasets provide energy and force labels in addition to atomic structures.

We argue that the diversity of OMol25 creates two exciting opportunities for generative modeling. First, its broad chemical coverage makes it possible to move beyond the narrow distributions that dominate existing molecular generation benchmarks and train models capable of directly generating chemically diverse and industrially relevant systems. Second, its compositional, conformational, and energy diversity makes it a natural pretraining corpus for transfer to settings where structural data are scarce. The latter is particularly compelling for biomolecular generation: although quantum-chemical datasets predominantly contain much smaller systems than proteins and DNA/RNA structures, they provide dense supervision over the local atomic environments and geometric interactions from which larger structures are composed. This motivates a central question of our work: *how should such large-scale atomistic data be used to learn representations that transfer effectively across both tasks and domains?*

To this end, we introduce **Zatom-2**, a unified atomistic architecture and pretraining framework for generative modeling across molecules, periodic materials and proteins. Uniquely, Zatom-2's pretraining curriculum includes the OMol25 (4M) and OMat24 (1M) datasets, and combines generation and structure prediction with auxiliary force and energy prediction objectives (see Figure 1). The latter objective is motivated by evidence that models trained to learn interatomic potentials (known as machine learning interatomic potentials, MLIPs) (Batatia et al., 2022; Wood et al., 2025; Neumann et al., 2024; Qu et al., 2026), have been demonstrated to learn meaningful, transferable representations of geometric systems (Wedig et al., 2025a; Li & Walsh, 2026; Didi, 2026), and we hypothesize that these auxiliary physics-grounded objectives help shape a generative model's latent representations in a way that is beneficial for downstream tasks. In particular, force and energy supervision encourages a backbone architecture to encode the local geometric structure of atomistic configurations influenced by the forces acting on the system. We further discuss related work in Appendix A.

Contributions. Based on these ideas, our main contributions in this work are as follows:

- We introduce Zatom-2, a *unified* model for generative modeling across atomistic domains, which supports molecule and material generation, structure prediction, and prediction of energies and forces by way of pretraining. Its architecture additionally supports protein generation, which we demonstrate in a finetuning task.
- Zatom-2: (**1**) achieves better molecular distribution learning than Zatom-1; (**2**) provides strong performance in molecule and material generation benchmarks; (**3**) introduces force conditioning to enable *control* over generated samples' relaxation state; and (**4**) demonstrates transfer of atomistic pretraining to protein generation, including strong extrapolation to sequence lengths beyond the training distribution.
- The Zatom-2 architecture features a unified `atom1` + `atom14` tokenization scheme for atomistic data, thereby enabling molecule, material, and all-atom protein generative modeling tasks within a single, standardized Transformer architecture which demonstrates performance advantages with data and modeling scaling.

2 PRELIMINARIES

2.1 ALL-ATOM REPRESENTATION AND MODELING

Diffusion models for small molecules and periodic materials often co-generate geometry and atom identity using separate corruption processes for coordinates and atom types (Vignac et al., 2023; Zeni et al., 2025; Joshi et al., 2025; Vonessen et al., 2026). Our formulation draws inspiration from the protein literature (Qu et al., 2025; Butcher et al., 2025), which models all-atom protein coordinates and predicts residue identities from learned geometric representations. To accommodate variation in the number of atoms per residue, Qu et al. (2025) introduces an `atom14` representation: each residue occupies 14 coordinate slots, with unused slots filled by virtual atoms. A protein with L residues is thus represented by $\mathbf{X}_0 \in \mathbb{R}^{L\times 14\times 3}$, whose shape is independent of residue identity. A sequence prediction head then recovers residue identities from the model's atom-level features, which enables coordinate diffusion without a separate diffusion process over amino acid types.

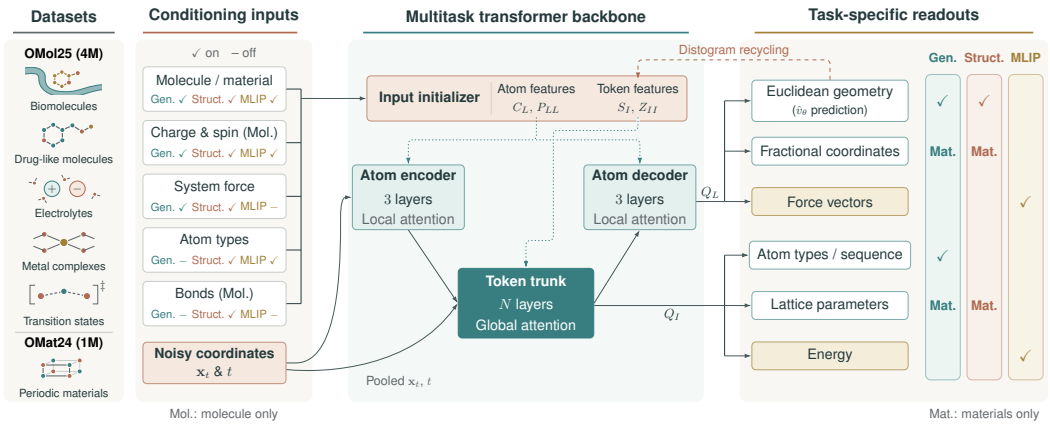


Figure 1: **Zatom-2 multitask pretraining across domains.** Zatom-2 is jointly trained on OMol25 (4M) and OMat24 (1M) as a generative model. Its input initializer receives task-dependent conditioning, summarized by task strips: checkmarks indicate enabled inputs and dashes indicate disabled inputs. MLIP task supervision (force and energy prediction) uses clean coordinates at $t = 1$ and known atom types, without bond or system force conditioning. The initializer provides atom (C_L, P_{LL}) and token (S_I, Z_{II}) conditioning through the dotted connections. Noisy coordinates $\mathbf{x}_t$ and flow time t enter the U-shaped atom encoder—token trunk—atom decoder. Atom features Q_L and token features Q_I feed task-specific readout heads. The recycling path converts the clean-coordinate estimate obtained from $\widehat{\mathbf{v}}_\theta$ into a token-level distogram, updating token-pair conditioning for a second forward pass. In this work, we experiment with $N = 9/18/36$ trunk layers.

Building on this principle, we pose the question, similarly to proteins, of whether molecule and material coordinates hold enough information to inform atom type prediction directly, and introduce an `atom1` representation for molecular systems, in which each atom constitutes a token with a three-dimensional coordinate. Like Feng et al. (2025), we generate atomic coordinates through diffusion and infer atom identities from representations learned by the denoising network. Distinctly, our formulation enables a shared all-atom generation framework for biological (e.g., protein) and chemical (e.g., molecule, material) data by allowing residue- and atom-level tokenization, respectively.

2.2 Conditional flow matching and sampling

Flow matching generative models (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) learn a continuous transport from a noise distribution at $t = 0$ to the data distribution at $t = 1$ through an ordinary differential equation (ODE). We adopt velocity prediction (Wang et al., 2025; Geffner et al., 2026), with a network $\widehat{\mathbf{v}}_\theta(\mathbf{x}_t, t)$ that predicts a velocity normalized by a scale σ_{data}. Given clean Cartesian coordinates $\mathbf{x}_1$ and a noise sample $\mathbf{x}_0 = \sigma_{\text{data}}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma_{\text{data}} = 16\,\text{Å}$, we define the interpolation path and normalized velocity target as

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \mathbf{v}_t = \frac{\mathbf{x}_1 - \mathbf{x}_0}{\sigma_{\text{data}}}. \tag{1}$$

The corresponding Cartesian velocity field is $\sigma_{\text{data}}\widehat{\mathbf{v}}_\theta(\mathbf{x}_t, t)$, yielding the clean endpoint estimate

$$\widehat{\mathbf{x}}_1 = \mathbf{x}_t + (1-t)\sigma_{\text{data}}\widehat{\mathbf{v}}_\theta(\mathbf{x}_t, t). \tag{2}$$

The model is trained via the l_2 regression objective $\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\widehat{\mathbf{v}}_\theta(\mathbf{x}_t, t) - \mathbf{v}_t\|_2^2\right]$.

For inference, let $\mathbf{y}_\sigma$ denote noisy Cartesian coordinates at noise level σ, corresponding to the additive corruption $\mathbf{y}_\sigma = \mathbf{x}_1 + \sigma\boldsymbol{\epsilon}$. We map these coordinates to the flow parameterization as

$$t = \frac{\sigma_{\text{data}}}{\sigma_{\text{data}} + \sigma}, \qquad \mathbf{x}_t = t\,\mathbf{y}_\sigma. \tag{3}$$

Since $t\sigma = (1-t)\sigma_{\text{data}}$, this transformation recovers the interpolation path in Equation 1. At each sampling step, we evaluate the velocity network at $(\mathbf{x}_t, t)$ and supply the sampler with the clean-coordinate estimate from Equation 2. This change of variables enables EDM-style sampling(Karras et al., 2022; Butcher et al., 2025; Williams et al., 2026) (n.b., which we observed to yield improved sample quality) while retaining a standard velocity-based flow matching training objective.

3 ZATOM-2

3.1 MULTISCALE TRANSFORMER BACKBONE

We formulate an architecture that supports multitask learning across generation, structure prediction, and property prediction tasks for molecules and periodic materials, as well as generation for proteins. To this end, we adopt our new multiscale tokenization scheme (Section 2.1) and define multitask input conditioning features, as well as bespoke output heads for each task and data domain.

Zatom-2 contains an input initializer and a U-shaped atom–token–atom Transformer (Figure 1), similar to the multiscale architectures of RFdiffusion3 and Emyx (Butcher et al., 2025; Williams et al., 2026); Appendix C details Zatom-2's forward pass and compares it to that of RFdiffusion3 and Emyx. The model learns atom representations Q_L for local geometric interactions and token representations Q_I for global information exchange, where L and I denote atoms and tokens, respectively.

The initializer embeds a series of features: domain labels, spins and charges of the system, system force labels (see Section 3.2), as well as fixed chemical identities, relative token positions and bonds for molecule structure prediction tasks. These embeddings are combined with coordinate and flow-time embeddings to initialize the atom and token representations and construct the conditioning features used throughout the denoising backbone (see Figure 1).

The atom encoder updates atom features through local attention over token-index neighborhoods and nearby atoms in space, then downcasts these features into token representations through cross-attention. The Transformer trunk applies global attention between tokens, before upcasting the updated token features back to the atom stream through cross-attention. The atom decoder further refines the fused features through local attention. All Transformer blocks use pair-biased attention (Jumper et al., 2021) and DiT-style adaptive normalization (Peebles & Xie, 2023).

Task-specific readouts. Dedicated heads map the shared representations to task-specific outputs. Atom features are used for normalized Cartesian velocity and fractional coordinate predictions, while token features are used for element identity and sequence type logits. Lattice lengths and angles are predicted from token features pooled within each system. An optional property module aggregates token contributions into a domain-normalized residual energy per atom and predicts normalized force vectors from atom features. For system b in electronic-structure domain d, the physical energy is reconstructed as

$$\widehat{E}_b = \sum_{\ell \in b} \varepsilon_{d,z_\ell} - n_b \left(\mu_d + s_d \widehat{e}_b\right), \tag{4}$$

where $\varepsilon_{d,z}$ are elemental reference energies estimated from the training split, μ_d and s_d normalize the residual energy per atom, and n_b is the atom count. Forces are normalized by a domain-specific RMS scale during training and restored to eV Å^{-1} at inference. The direct force readout supports computational efficiency and comparison with prior work. However, note that we enforce neither energy-gradient consistency nor exact rotational equivariance.

Training loss. To train Zatom-2, we minimize the weighted multitask objective

$$\begin{aligned}\mathcal{L} = \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{lDDT}}\mathcal{L}_{\text{lDDT}} \\ + g_{1\,\text{Å}}(t) \sum_{k \in \mathcal{K}} \lambda_k \mathcal{L}_k \\ + m_E \lambda_E \mathcal{L}_E + m_F \lambda_F \mathcal{L}_F + \mathbf{1}_{\text{MLIP}} \lambda_{\text{eq}} \mathcal{L}_{\text{eq}},\end{aligned} \tag{5}$$

where $\mathcal{K} = \{\text{element}, \text{sequence}, \text{fractional}, \text{lattice}\}$ indexes the time-gated training tasks, and $g_\delta(t) = \mathbf{1}[\sigma_{\text{data}}(1-t)/t < \delta]$ gates tasks for near-endpoint supervision. $\mathcal{L}_{\text{FM}}$ is the flow matching loss, $\mathcal{L}_{\text{lDDT}}$ is the smoothed lDDT loss (Mariani et al., 2013; Butcher et al., 2025), and $\mathcal{L}_E, \mathcal{L}_F$ are Huber losses. The energy/force gates (m_E, m_F) equal $(1, 1)$ for clean MLIP inputs, $(g_{1\,\text{Å}}, g_{0.25\,\text{Å}})$ for force-unconditioned generation examples, and $(0, 0)$ otherwise, preventing target-force leakage in force-conditioned sample generation. Both gates vanish unless input examples are uncropped and, for OMol25, have valid charge and spin annotations. Inspired by (Elhag et al., 2025), the MLIP-only latent equivariance loss $\mathcal{L}_{\text{eq}}$ trains a two-layer network to predict one rotated input view's pooled representation from another, conditioned on the two views' relative rotation. Empirically, we find that $\mathcal{L}_{\text{eq}}$ enables accurate atomic force prediction during joint generative-predictive pretraining.

3.2 Force conditioning

OMol25 and OMat24 contain substantial structural diversity, spanning high-force, rattled configurations to fully relaxed geometries (Appendix B). We hypothesize that explicitly informing the model of a system's force state can help it distinguish between geometrically distinct configurations that share the same discrete composition and implicitly inferred topology. To provide this information, we condition both generation and structure prediction tasks on the mean force norm of the system, using it as a global descriptor of the degree of structural relaxation:

$$s_F = \log_{10}\left(\frac{1}{L}\sum_{\ell=1}^{L}\|\mathbf{F}_\ell\|_2\right). \tag{6}$$

4 Experiments

We use the OMol25 4M subset (3,986,754 examples) (Levine et al., 2025) and the OMat24 1M subset (1,009,850 examples) (Barroso-Luque et al., 2024) for pretraining. For details on the datasets, including their composition and properties, we refer the reader to Appendix B and their original publications (Levine et al., 2025; Barroso-Luque et al., 2024).

The Experiments section is organized as follows. In Section 4.1, we present multiple pretraining configurations for OMol25 and OMat24, combining generation, structure prediction, and force and energy prediction tasks. To show the empirical effects of multitask and multi-domain training, we evaluate each configuration for generation and force conditioning quality. In Section 4.2, we perform a comparison study of Zatom-2 against established benchmarks, to offer a relative evaluation of the model's architecture. In Section 4.3, we conduct a scaling study to investigate the effects of model and data size on performance. Lastly, in Section 4.4, we explore the transferability of Zatom-2 to protein generation tasks, demonstrating its potential for broader applications in biomolecular design.

Metrics. We introduce force-consistency metrics to probe whether our training recipe enables Zatom-2 to capture the degree of rattling/relaxation of each generated sample, and, more broadly, the configurational diversity of OMol25 and OMat24. For each generated sample, we compute atomic forces using UMA-S-1p2 (Wood et al., 2025) and quantify agreement between the resulting mean force norm and the conditioning value using AUROC and MAE (Appendix D). We argue that forces provide a more relevant measure for the degree of rattling across OMol25 and OMat24 systems of varying sizes than unnormalized total energies. The rest of the metrics used in this section primarily assess generation quality, and are explained in Appendix D.

4.1 Atomistic pretraining on OMol25 & OMat24

We investigate a series of training configurations on OMol25, progressively augmenting the generation objective with structure prediction, OMat24 data, and force and energy supervision. We first train a small variant of Zatom-2 on OMol25 alone (70M parameters), with and without force conditioning (see Table 1). Compared to Zatom-1, Zatom-2 improves generation quality, especially as seen in lower AFD scores and mean force norms ($\overline{\|F\|}$) predicted by UMA (Wood et al., 2025), reflecting better training set distribution learning. Our method for force conditioning improves alignment to the dataset's average $\overline{\|F\|}$, and enables the model to generate force-consistent structures.

Table 1: **Pretraining configurations and generation quality.** Panel (a) shows multiple training configurations for OMol25 and OMat24 mixed training. Panel (b) reports their performance, where we also compare to 10,000 dataset samples and Zatom-1. All models are trained for 80 epochs. Forces are calculated using UMA (Wood et al., 2025) and metrics are reported in eV Å^{-1}. AUROC uses a conditioning mean force norm cutoff of 1 eV Å^{-1}. Structure prediction results are reported in Appendix Table 9.

a) Pretraining configurations

		Datasets		Tasks			
Model	Size	OMol25	OMat24	Generation	Structure prediction	Force & energy prediction	Force conditioning
Zatom-1 (released)	Large	✓	–	✓	–	–	–
Zatom-2 (I)	Small	✓	–	✓	–	–	–
Zatom-2 (II)	Small	✓	–	✓	–	–	✓
Zatom-2 (III)	Small	✓	–	✓	✓	–	✓
Zatom-2 (IV)	Base	✓	✓	✓	✓	–	✓
Zatom-2 (V)	Base	✓	✓	✓	✓	✓	✓
Zatom-2 (VI)–*final*	Large	✓	✓	✓	✓	✓	✓

b) Generation quality and conditioning adherence

	Sample quality						Force quality and conditioning adherence		
Model	RDKit Valid. ↑	Unique ↑	Diversity ↑	FP novelty ↑	AFD ↓	OMat24 Valid. ↑	$\overline{\|F\|}$ ↓	$\overline{\|F\|}$ MAE ↓	$\overline{\|F\|}$ AUROC ↑
Dataset	0.846	–	0.946	–	0.0025	0.791	1.311	–	–
Zatom-1 (released)	0.941	0.614	0.940	0.218	0.047	–	29.547	–	–
Zatom-2 (I)	0.840	0.923	0.937	0.349	0.016	–	1.154	0.827	0.586
Zatom-2 (II)	0.823	0.898	0.940	0.354	0.017	–	1.252	0.791	0.662
Zatom-2 (III)	0.846	0.885	0.939	0.368	0.019	–	1.627	0.722	0.730
Zatom-2 (IV)	0.836	0.884	0.938	0.354	0.018	0.727	1.595	0.740	0.713
Zatom-2 (V)	0.851	0.899	0.940	0.365	0.019	0.748	1.599	0.688	0.757
Zatom-2 (VI)–*final*	0.850	0.895	0.941	0.373	0.016	0.748	1.269	0.616	0.817

We then introduce structure prediction as an additional task. For OMol25, structure prediction conditions on atom types and molecular connectivity and predicts the corresponding geometry. Because connectivity could not be inferred reliably for all OMol25 structures, we restrict this task to a subset of 1,581,810 examples for which bonds could be assigned reliably (Appendix E.1). For OMat24, structure prediction is performed on the full dataset by conditioning on atom types and predicting both the atomic geometry and lattice parameters.

Adding OMol25 structure prediction (Zatom-2 (II) → Zatom-2 (III)) largely preserves generation quality, with a slight reduction in adherence to the dataset's average $\overline{\|F\|}$. We report results for the structure prediction task below, and in Appendix E. Incorporating OMat24 supervision and increasing the training budget (Zatom-2 (III) → Zatom-2 (IV), 140M parameters) partially restores this force consistency. Finally, we add force and energy prediction as auxiliary tasks and increase the model's capacity (Zatom-2 (IV) → Zatom-2 (V) → Zatom-2 (VI), 270M parameters), yielding our final configuration and the strongest overall performance across the evaluated metrics.

Force conditioning consistency. We evaluate how closely Zatom-2 generated samples adhere to their specified force conditioning first by measuring the absolute deviation between the mean force norm of each generated structure and its conditioning target using mean absolute error (MAE, see Appendix D.1). Second, we assess whether the model can controllably generate structures in low- versus high-force regimes of a given molecular system using AUROC. For the AUROC evaluation, we define dataset-specific force thresholds based on the corresponding OMol25 mean force norm distributions: $\tau = 1$ eV Å^{-1} for `geom_orca6`, and $\tau = 2.5$ eV Å^{-1} for `ani2x` (see Appendix Figure 4). Beyond evaluating whether Zatom-2 distinguishes between rattled and relaxed configurations, we use force consistency as the primary metric for structure prediction on OMol25.

Small molecule structure prediction, more commonly known as conformer generation, is typically assessed using precision and recall based on the root mean square deviation (RMSD) between sets of

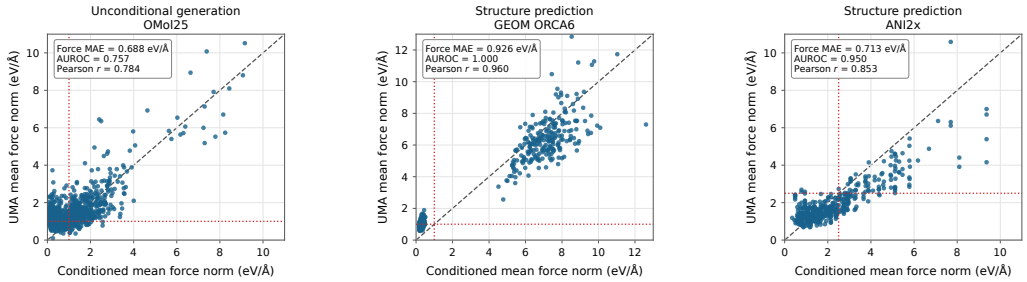


Figure 2: **Force-conditioning consistency.** Panel (a) compares the conditioned mean force norm with the mean force norm evaluated by UMA for unconditional OMol25 molecule generation. Panels (b) and (c) give the corresponding comparison for OMol25 structure prediction on the GEOM ORCA6 and ANI2x subsets, respectively. Red dotted lines mark the AUROC cutoff (1 eV Å^{-1} in (a)–(b), and 2.5 eV Å^{-1} in (c)).

generated and reference conformers for a given molecular graph (Liu et al., 2026). This evaluation is not directly applicable to OMol25: the dataset contains both equilibrium and non-equilibrium configurations and does not provide a canonical mapping from a molecular graph to a reference conformer ensemble (in fact, the dataset does not provide any molecular graphs or bond topologies). Moreover, conditioned on a molecular graph and target force magnitude, there is generally no unique ground-truth geometry, making RMSD to an individual reference structure an inappropriate measure of prediction quality. Force consistency instead evaluates whether a predicted geometry represents a physically compatible configuration under a specified force regime. Figure 2 shows that generated structures show an MAE ranging from 0.69 to 0.93 eVÅ^{-1}, and the model reliably distinguishes between low- and high-force regimes, achieving AUROC values between 0.76 and 1.00 across the three evaluated settings. Further results for structure prediction are presented in Appendix Table 9. This indicates that force conditioning provides reliable separation between high- and low-force regions, while precise force matching within each regime remains more challenging.

Energy and force prediction. By way of pretraining, Zatom-2 uniquely supports MLIP energy and force prediction for molecules and materials. We treat energy and force prediction as an auxiliary objective, motivated by the hypothesis that supervision on forces and energies encourages the model to better learn local geometric structure and improve representation quality, which previous methods have shown leads to improved generative performance (Yan et al., 2026). We report performance on MLIP prediction and tune generative / structure prediction / force & energy prediction task mixtures in Appendix E.2. As expected from a non-specialized model, MLIP accuracy remains below specialized potentials, however the task improves generative quality and especially enhances transfer learning, which is studied in Section 4.4. We further observe that an MLIP-pretrained model shows better separation of atomic embeddings in PCA space (see Appendix F.4) compared to a model which is not trained with force and energy supervision.

4.2 Generation on GEOM-Drugs and MP20

To contextualize Zatom-2's performance and enable comparison with prior architectures, we evaluate the model on three established generation benchmarks: joint QM9 molecule generation and MP20 material generation, and GEOM-Drugs molecule generation. As reported in Table 2 and Appendix Table 11, Zatom-2 achieves 95.72% validity on QM9, 96.90% validity on GEOM-Drugs, and a 4.92% yield of unrelaxed metastable, unique, and novel MP20 materials (MetaSUN) on LeMat-GenBench (Betala et al., 2025), improving or approaching state-of-the-art performance across these benchmarks. These results provide empirical support for our formulation in Sections 2.2 and 3.1, that atom identities and lattice parameters can be inferred from coordinate representations without separate generative trajectories over these modalities. We further evaluate Zatom-2 for structure prediction on GEOM-Drugs and MP20, and report comparisons to baselines in Appendix E.5 and E.6. Appendix F provides MolStar visualizations of generated molecules and materials.

Table 2: **Generation benchmarks on GEOM-Drugs and MP20.** Panel (a) reports unconditional GEOM-Drugs molecule generation metrics, and panel (b) reports MetaSUN yield for unrelaxed MP20 material generation.

(a) GEOM-Drugs molecule generation

Model	Valid ↑	Unique ↑	PB-valid↑	Valid & unique ↑	Valid & PB-valid ↑
EQGAT-diff	94.6	100.0	59.7	94.60	56.48
SemlaFlow	93.9	100.0	87.5	93.90	82.16
ADiT	95.3	100.0	85.3	95.30	81.29
TABASCO	97.6	99.03	91.6	96.65	89.40
Zatom-1	93.6	99.93	94.1	93.53	88.08
GEOM-Drugs (reference)	–	–	94.0	–	–
Zatom-2	96.90	99.61	91.47	96.52	88.63

(b) MP20 material generation (unrelaxed)

Model	MetaSUN ↑
ADiT (joint QM9+MP20)	1.0
Crystal-GFN	0.0
Crystalformer	3.1
LLaMat2-CIF	2.1
SymmCD	2.4
Zatom-1 (MP20-only, 80M)	0.20
Zatom-1 (joint QM9+MP20, 80M)	0.36
Zatom-1-L (joint QM9+MP20, 160M)	0.64
Zatom-2 (joint QM9+MP20)	4.92

4.3 Scaling model capacity and data coverage

We further investigate the effects of scaling Zatom-2 along its data and model parameter axes. Figure 3 shows how validation sample quality and force prediction performance evolve across training. At an equal number of epochs, across all four validation metrics, training on the full dataset consistently outperforms restricting each dataset domain to 500k examples. Parameter scaling provides an additional improvement under both data regimes, with the largest effect on OMat24 structure prediction. Similar, though smaller, gains are observed for molecular generation and force prediction fidelity. Overall, these results indicate that data coverage and model capacity provide complementary gains, and motivate further scaling to the full OMol25 (100M) and OMat24 (100M) datasets.

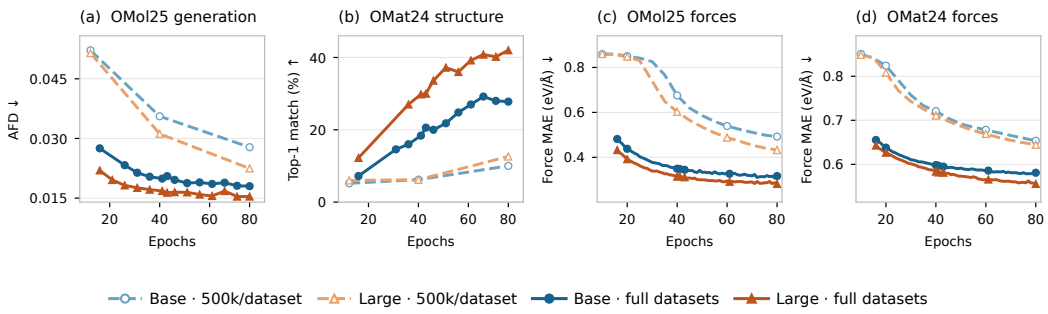


Figure 3: **Model and data scaling across training epochs.** We evaluate Zatom-2 models trained either on 500k examples per dataset or on the full OMol25 and OMat24 datasets. We report (a) OMol25 molecular generation quality measured by AFD, (b) OMat24 structure prediction measured by top-1 match rate, and force prediction MAE on (c) OMol25 and (d) OMat24. Training on the full dataset with more model parameters consistently improves validation performance across tasks.

4.4 Transfer to low-data protein generation

Finally, we study whether Zatom-2's atomistic pretraining transfers to a low-data protein generation setting. We curate a diverse subset of PDB structures from SCOPe (Appendix B.5), which we refer to as SCOPe-2k, and compare training from scratch against finetuning from three pretrained initializations: Zatom-2 (IV) (generation + structure prediction), Zatom-2 (V) (generation + structure prediction + force & energy prediction), and a variant of Zatom-2 (IV) without force conditioning. We also train RFdiffusion3 (Butcher et al., 2025) on the SCOPe-2k dataset. For each method, we select its checkpoint with the highest designability and novelty in a 96-backbone sweep (Appendix E.3).

We evaluate protein generation in-distribution, using sequence lengths represented in the training set (50–128 residues), as well as out-of-distribution (OOD), using sequences up to twice as long (129–256 residues). In Table 3, joint generative-predictive pretraining—combining generation, structure

Table 3: **SCOPe-2k protein generation and transfer. (a):** We evaluate 1,027 protein backbones for each configuration, 13 corresponding to each length between 50 and 128. **(b):** The same protocol is repeated for 129–256 residue lengths. FC denotes force-conditioning, and F/E pred. refers to a model trained with force and energy prediction as auxiliary tasks (Zatom (V)). The Zatom (V)-finetuned model substantially outperforms all configurations on the length extrapolation task. Metrics are explained in Appendix D.9.

Model / Zatom-2 initialization	Designability (%) ↑	Per-sequence success (%) ↑	Diversity (#clusters) ↑	Novelty (%) ↑
(a) In-distribution: 50–128 residues				
RFdiffusion3	**91.72**	**66.48**	447	12.74
None	88.02	61.23	395	10.62
Zatom-2 (IV): Generation + structure (w/o FC)	86.27	59.21	428	**16.82**
Zatom-2 (IV): Generation + structure	89.19	62.35	444	14.30
Zatom-2 (V): Generation + structure + F/E pred.	90.75	64.73	**453**	16.52
(b) Length extrapolation: 129–256 residues				
RFdiffusion3	55.86	29.54	462	89.16
None	66.89	34.91	473	90.36
Zatom-2 (IV): Generation + structure (w/o FC)	61.62	33.15	490	**96.04**
Zatom-2 (IV): Generation + structure	65.62	36.98	517	94.64
Zatom-2 (V): Generation + structure + F/E pred.	**75.10**	**46.50**	**621**	95.45

prediction, and force & energy prediction—has the strongest performance across most metrics when compared to other initializations for Zatom-2. In the in-distribution setting, it yields 453 distinct designable clusters, compared with 395 when training from scratch, with generation + structure pretraining offering similar performance improvements.

The benefits of pretraining are more pronounced in the OOD regime: Zatom-2 (V) substantially outperforms the other Zatom-2 initializations, improving designability from 66.89% to 75.10% compared to random initialization. Interestingly, pretraining is only beneficial to designability when force conditioning is enabled, supporting our hypothesis in Section 3.2 that force-aware pretraining is particularly beneficial when learning from data with high configurational diversity. Moreover, initializing from a vanilla generative model on OMol25 and OMat24 (Generation + structure (w/o FC)) even achieves worse performance than a model trained from scratch. All Zatom-2 configurations outperform RFdiffusion3 in the OOD setup, and in turn RFdiffusion3 outperforms Zatom-2 in-distribution.

Together, these results indicate that broad atomistic pretraining can transfer to data-limited generation tasks in biomolecular design. Interestingly, the strongest gains of pretraining arise when predictive objectives for forces and energies are incorporated alongside generation, providing evidence that joint generative-predictive training is a promising recipe for future atomistic generative models. In Appendix G, we provide a detailed discussion of this work's broader impacts and limitations.

5 Conclusion

In this work, we introduced Zatom-2, an atomistic generative model across molecules and materials, pretrained on the OMol25 (4M) and OMat24 (1M) datasets. We proposed a training recipe for generative models incorporating predictive objectives, and empirically found that force and energy supervision offers better separation of atomic embeddings in PCA space, boosts generative performance, and enables stronger transfer to protein generation tasks. We discuss the importance of force-guided generation on datasets with high configurational diversity, and propose a force consistency metric to assess the ability of the model to distinguish between rattled and relaxed configurations. Zatom-2 performs strongly across existing molecule and material generation benchmarks and achieves strong performance for in-distribution and out-of-distribution protein generation tasks through its generative-predictive pretraining recipe. The architecture also exhibits performance advantages with data and model scaling, motivating larger-scale pretraining efforts in future work.

Reproducibility Statement

To ensure that this work is reproducible, we have accompanied it with anonymized source code containing all documentation, data loading infrastructure, and training/inference materials necessary to reproduce the results presented in this manuscript. Additionally, Section 3 of the main text introduces the Zatom-2 architecture, and Appendix C provides additional model details, including an in-depth examination of Zatom-2's forward pass, its hyperparameters, and its required computing resources. Lastly, Appendix B outlines the datasets referenced in this work and how they were sourced/used, as well as what licenses apply to them.

AI Use Statement

In this work, we used generative AI tools to refine hypotheses, design or provide feedback on research methodology or experiments, implement methods and support qualitative data analysis. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, or clean and reformat datasets. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation do not apply to this work. As additional context, we used generative AI tools to create or edit figures or images, suggest experimental parameters, create or edit software code, draft parts of the research paper, brainstorm or source/search for information. We have reviewed all AI-assisted work. For instance, we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and LLM-generated code was verified and tested for correctness by at least 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://doi.org/10.1038/s41597-022-01288-4.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models, 2024. URL https://arxiv.org/abs/2410.12771.

Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YPpSngE-ZU.

Siddharth Betala, Samuel P. Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K. Joshi, Sékou-Oumar Kaba, Félix Therrien, Alex Hernandez-Garcia, Rocío Mercado, N. M. Anoop Krishnan, and Alexandre Duval. LeMat-GenBench: A unified evaluation framework for crystal generative models, 2025. URL https://arxiv.org/abs/2512.04562.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul T. Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo design of all-atom biomolecular interactions with RFdiffusion3. *bioRxiv*, 2025. doi: 10.1101/2025.09.18.676967. URL https://www.biorxiv.org/content/10.1101/2025.09.18.676967v2.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akin, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K. Fox, and Steven E. Brenner. SCOPe: Improvements to the structural classification of proteins—extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1):D553–D559, 2022. doi: 10.1093/nar/gkab1054.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hunter Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Po-Ssu J. Y. Leung, Timothy F. Huddy, Sam Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Jonathan M. Skelton, Kazuki Morita, and Aron Walsh. SMACT: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019. doi: 10.21105/joss.01361.

Kieran Didi. The unification of representation learning and generative modelling. Blog post, 2026. URL https://kdidi.netlify.app/blog/ml/2025-12-31-r4g/. Living survey; literature cutoff 2026-09-07.

Sathya Edamadaka, Soojung Yang, Ju Li, and Rafael Gómez-Bombarelli. Universally converging representations of matter across scientific foundation models, 2025. URL https://arxiv.org/abs/2512.03750.

Ahmed A Elhag, Arun Raja, Alex Morehead, Samuel M Blau, Hongtao Zhao, Christian Tyrchan, Eva Nittinger, Garrett M Morris, and Michael M Bronstein. Learning inter-atomic potentials without explicit equivariance. *arXiv preprint arXiv:2510.00027*, 2025.

Shikun Feng, Yuyan Ni, Yan Lu, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniGEM: A unified approach to generation and property prediction for molecules. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Lb91pXwZMR.

Eloy Félix, Andrew Dalke, Greg Landrum, and Roman Bushuiev. chembl/fpsim2: 0.7.3, January 2025. URL https://doi.org/10.5281/zenodo.14652272.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. In *International Conference on Learning Representations*, volume 2025, pp. 98803–98851, 2025.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic protein generation via partially latent flow matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RDerF20JYT.

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling. *arXiv preprint arXiv:2604.02270*, 2026. doi: 10.48550/arXiv.2604.02270. URL https://arxiv.org/abs/2604.02270.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, pp. 8867–8887, 2022.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – efficient 3D molecular generation with latent attention and equivariant flow matching. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=bee2G6pEh0.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *International Conference on Machine Learning*, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022. URL https://arxiv.org/abs/2206.00364.

Greg Landrum. RDKit: Open-source cheminformatics. *Release*, 1(1–79):4, 2013.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof T. Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3D molecule generation. In *International Conference on Learning Representations*, 2024.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models, 2025. URL https://arxiv.org/abs/2505.08762.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, 8:830–840, 2026. doi: 10.1038/s42256-026-01235-7. URL https://www.nature.com/articles/s42256-026-01235-7.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023. doi: 10.48550/arXiv.2301.12485. URL https://arxiv.org/abs/2301.12485.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Yunqing Liu, Yi Zhou, and Wenqi Fan. Geometric flow matching for molecular conformation generation via manifold decomposition. *arXiv preprint arXiv:2605.25577*, 2026. URL https://arxiv.org/abs/2605.25577.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating materials with riemannian flow matching. In *International Conference on Machine Learning*, 2024.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Tynan Perez, Samuel Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, Hrushikesh Sahasrabuddhe, Pietro Liò, Tommi Jaakkola, Rafael Gómez-Bombarelli, Rex Ying, N. Benjamin Erichson, and Michael W. Mahoney. Zatom-1: Towards a multimodal foundation model for 3D molecules and materials, 2026. URL https://arxiv.org/abs/2602.22251.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024. URL https://arxiv.org/abs/2410.22570.

Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3: 33, 2011. doi: 10.1186/1758-2946-3-33. URL https://pubmed.ncbi.nlm.nih.gov/21982300/.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

Hillary Pan, Alex M Ganose, Matthew Horton, Muratahan Aykol, Kristin A Persson, Nils ER Zimmermann, and Anubhav Jain. Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic chemistry*, 60(3):1590–1603, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

Tynan Perez and Rafael Gómez-Bombarelli. Distributional evaluations for atomistic generations: a set of methods and technical reports for evaluating atomistic generative models. https://github.com/TyJPerez/AtomisticGenEvals, 2026a.

Tynan Perez and Rafael Gómez-Bombarelli. Self-conditioned denoising for atomistic representation learning, 2026b. URL https://arxiv.org/abs/2603.17196.

Lucas Pinede, Soojung Yang, Juno Nam, and Rafael Gomez-Bombarelli. Unifying Force Prediction and Molecular Conformation Generation Through Representation Alignment. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. URL https://openreview.net/forum?id=yzkHGHvC74.

Eric Qu, Brandon M. Wood, Aditi S. Krishnapriyan, and Zachary Ward Ulissi. A recipe for scalable attention-based ML potentials: unlocking long-range accuracy with all-to-all node attention. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oZg9YF9dkz.

Wei Qu, Jiawei Guan, Rui Ma, Ke Zhai, Weikun Wu, and Haobo Wang. P(all-atom) is unlocking new path for protein design. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50786–50816. PMLR, 2025. URL https://proceedings.mlr.press/v267/qu25c.html.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K. Burley, Jaroslav Koča, and Alexander S. Rose. Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, 2021. doi: 10.1093/nar/gkab314. URL https://doi.org/10.1093/nar/gkab314.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9558–9568. PMLR, 2021. URL https://proceedings.mlr.press/v139/shi21b.html.

Anuroop Sriram, Benjamin K. Miller, Ricky T. Chen, and Brandon M. Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy John O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494. URL https://www.biorxiv.org/content/10.1101/2025.11.20.689494v1.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation, 2023. URL https://arxiv.org/abs/2302.09048.

Carlos Vonessen, Charles Harris, Miruna Cretu, and Pietro Liò. TABASCO: A fast, simplified model for molecular generation with improved physical quality. *Transactions on Machine Learning Research*, February 2026. URL https://openreview.net/forum?id=Kg6CSrbXl4.

Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista. SimpleFold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480*, 2025. URL https://arxiv.org/abs/2509.18480.

Steffen Wedig, Rokas Elijošius, Christoph Schran, and Lars Leon Schaaf. REM3DI: Learning smooth, chiral 3D molecular representations from equivariant atomistic foundation models. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025a. URL https://openreview.net/forum?id=jOmZsvXoK5.

Steffen Wedig, Rokas Elijošius, Christoph Schran, and Lars Leon Schaaf. REM3DI: Learning smooth, chiral 3D molecular representations from equivariant atomistic foundation models. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025b. URL https://openreview.net/forum?id=jOmZsvXoK5.

Nicholas J. Williams, Ward Haddadin, Matteo P. Ferla, Constantin Schneider, Nicholas B. Woodall, Ruby Sedgwick, Christian D. Madsen, Andrew L. Hopkins, Douglas E. V. Pires, and Edward O. Pyzer-Knapp. Emyx: Fast and efficient all-atom protein generation, 2026. URL https://arxiv.org/abs/2606.19377.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.23971.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=03RLpj-tc_.

Shaoheng Yan, Zian Li, Cai Zhou, Qiaojing Huang, Kai Liu, and Muhan Zhang. Toward better geometric representations for molecule generative models, 2026. URL https://arxiv.org/abs/2605.07693.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tYIMtogyee.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5. URL https://www.nature.com/articles/s41586-025-08628-5.

Chengqian Zhang, Yucheng Jin, Duo Zhang, Tiejun Li, and Han Wang. CrystalREPA: Transferring Physical Priors from Universal MLIPs to Crystal Generative Models, 2026. URL https://arxiv.org/abs/2605.08960.

Yang Zhang and Jeffrey Skolnick. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524.

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*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Zatom-2 Appendices

A Related Work . . . 18

A.1 Generative modeling of 3D molecules . . . 18

A.2 Generative modeling of periodic materials . . . 18

A.3 Representation learning and multitask atomistic prediction . . . 18

B Datasets . . . 19

B.1 OMol25 and OMat24 . . . 19

B.2 QM9 . . . 21

B.3 MP20 . . . 21

B.4 GEOM-Drugs . . . 21

B.5 SCOPe . . . 21

B.6 Dataset licenses . . . 21

C Additional Model Details . . . 23

C.1 Model forward pass . . . 23

C.2 Hyperparameters . . . 24

C.3 Compute resources . . . 25

D Evaluation Metrics . . . 25

D.1 Force self-consistency metrics . . . 26

D.2 Molecule generation metrics . . . 26

D.3 Material generation metrics . . . 27

D.4 QM9 and MP20 sample generation metrics . . . 27

D.5 GEOM-Drugs sample generation metrics . . . 28

D.6 Atomistic Fréchet Distance . . . 28

D.7 Structure prediction metrics . . . 29

D.8 Interatomic potential metrics . . . 29

D.9 Protein generation metrics . . . 29

E Additional Results . . . 30

E.1 OMol25 and OMat24 structure prediction . . . 30

E.2 MLIP performance before and after finetuning . . . 31

E.3 SCOPe-2k checkpoint trajectories 32
E.4 Comparison to generative baselines 33
E.5 GEOM-Drugs conformer generation benchmark. 34
E.6 MP20 material structure prediction 35

F Visualization 36

F.1 Generated molecules 36
F.2 Generated materials. 38
F.3 Generated proteins 40
F.4 Latent representations of OMol25 and OMat24 41

G Broader Impacts & Limitations 47

G.1 Broader impacts 47
G.2 Limitations 47

A Related Work

A.1 Generative modeling of 3D molecules

3D molecule generation methods commonly model atom identities and coordinates together. For example, equivariant diffusion models have done so while imposing Euclidean symmetry in their denoiser networks (Hoogeboom et al., 2022), and subsequent works have examined optimal architectures and loss functions supporting such equivariant molecular diffusion (Vignac et al., 2023; Le et al., 2024; Irwin et al., 2025). TABASCO instead demonstrated that a standard Transformer can support high-accuracy molecule generation (Vonessen et al., 2026), and Zatom-1 demonstrated the same capabilities for molecule-material generation (Morehead et al., 2026). Zatom-2 retains a Transformer backbone without exact rotation equivariance, uses rotational data augmentations, and extends its range of pretraining tasks to structure prediction, force-conditioned sample generation, and energy/force prediction. The model's QM9 and GEOM-Drugs results establish its competitiveness with well-known molecule generation baselines, while its OMol25 experiments test substantially broader chemical and conformational coverage and generative capabilities (Ramakrishnan et al., 2014; Levine et al., 2025).

A.2 Generative modeling of periodic materials

Materials generation additionally requires fractional coordinates and lattice geometry. Towards this end, CDVAE combines latent materials representations with diffusion-based sample generation (Xie et al., 2022). Similarly, DiffCSP performs joint equivariant diffusion of material modalities (Jiao et al., 2023), while FlowMM formulates materials generation in the framework of Riemannian flow matching (Miller et al., 2024). Uniquely, FlowLLM uses a language model to provide a base distribution for a materials generation flow matching model (Sriram et al., 2024). All-atom Diffusion Transformers and Zatom-1, however, shift toward supporting shared molecule-material generation capabilities, with the latter supporting prediction tasks as well (Joshi et al., 2025; Morehead et al., 2026). In this spirit, Zatom-2 supports domain-specific fractional coordinate and lattice geometry prediction while retaining a common Transformer trunk. The model's MP20 results demonstrate that its new `atom1` tokenizer enables competitive performance for materials generation without dedicated generation tracks for materials-specific input modalities, such as lattice parameters.

A.3 Representation learning and multitask atomistic prediction

Molecular representation learning has previously used paired 2D/3D views of atomistic data, masked modeling, and coordinate denoising to improve performance for downstream prediction tasks (Liu et al., 2022; Zhou et al., 2023; Zaidi et al., 2023). Analogously, Zatom-1 introduced a shared generative model for 3D molecules and materials and studied generative pretraining as a route to obtaining transferable atomistic representations (Morehead et al., 2026). Related efforts have coupled 3D molecule generation and property prediction within a single model framework (Feng et al., 2025).

Machine learning interatomic potential models (MLIPs) (Batatia et al., 2022; Neumann et al., 2024; Wood et al., 2025) are shown to learn rich representations of chemical domains, which are useful for downstream tasks (Wedig et al., 2025b) and converge in representation space as the improve in performance (Edamadaka et al., 2025; Li & Walsh, 2026). This poses the question whether they learn universal descriptors of local atomic geometries which can support generative modeling. This question has been explored using representation alignment (REPA) (Yu et al., 2025) to align the hidden representations of Boltzmann emulators (Pinede et al., 2025) and crystal generative models (Zhang et al., 2026) with those of pretrained MLIPs.

In contrast, Zatom-2 places generative flow matching, structure prediction, and energy/force prediction in a single mixture of pretraining tasks. This framing tests whether generative and structural objectives can coexist with an energy/force (MLIP) objective and whether their shared features remain useful during MLIP-heavy finetuning. Notably, Zatom-2 differs from a standalone universal potential such as UMA (Wood et al., 2025), which has the sole objective of atomistic property prediction.

Overall, Zatom-2 is positioned around a specific unification question: whether one scalable atomistic backbone can benefit from unified generative, structural, and property prediction capabilities across

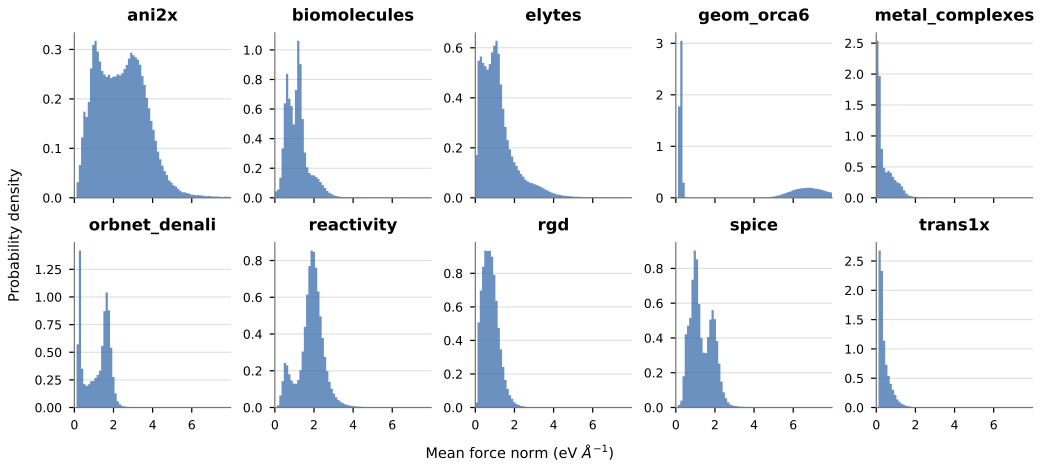


Figure 4: **Mean force norm distributions across OMol25-4M subsets.**

atomistic 3D systems. The experiments in Section 4 and throughout this work attempt to answer this question from a variety of angles.

B Datasets

B.1 OMol25 and OMat24

OMol25 is a molecular dataset with more than 100 million single-point density functional theory (DFT) calculations at a high level of theory (Levine et al., 2025). OMol25 contains 83 different atom types and combines small molecules, biomolecules, metal complexes, electrolytes, transition state geometries and reactive trajectories. Charge and spin multiplicity vary across the corpus, and a breakdown of the subsets used in the 4M training split (which we adopt in this work), together with their mean force norm distributions, is shown in Figure 4. The values were obtained by computing the mean of the Euclidean norms of the molecules' per-atom force vectors (i.e., the quantity inside the logarithm in Equation 6 of the main text).

Other than the composition diversity reflected in Figure 4, OMol25 contains per-system conformational diversity. For example, the GEOM-derived subset of OMol25 captures conformational variation within approximately 300,000 molecular families, with around 30% of each family's available conformers selected for DFT calculations. For its protein pocket–ligand subset, OMol25 performs molecular dynamics (MD) on extracted, cropped binding pockets and their ligands at 300 or 400 K, with restraints on the protein backbone. After filtering and subsampling each trajectory, the first and last retained frames are selected for DFT calculations, providing two configurations of the same pocket–ligand complex with variation in ligand conformation and relative ligand-pocket geometry. Similarly, for metal complexes, five configurations are randomly sampled from each 1–2 ps MLIP-based MD trajectory for DFT calculations. Detailed descriptions of OMol25 composition, rattling, and DFT calculations are presented in Levine et al. (2025).

In a complementary manner, OMat24 is a materials dataset containing approximately 118 million DFT calculations with energy and force labels (Barroso-Luque et al., 2024). Starting structures are taken from the Alexandria corpus and diversified with rattled Boltzmann samples, rattled relaxation trajectories, and ab initio MD. The dataset therefore includes both near-equilibrium and substantially distorted structures, with 89 different elements throughout. We use its public 1M training split (which contains 88 distinct elements) and, similarly to OMol25, report mean force norm distributions in Figure 5.

Figure 6 shows the total atom count distributions of both datasets. Atom counts include hydrogens.

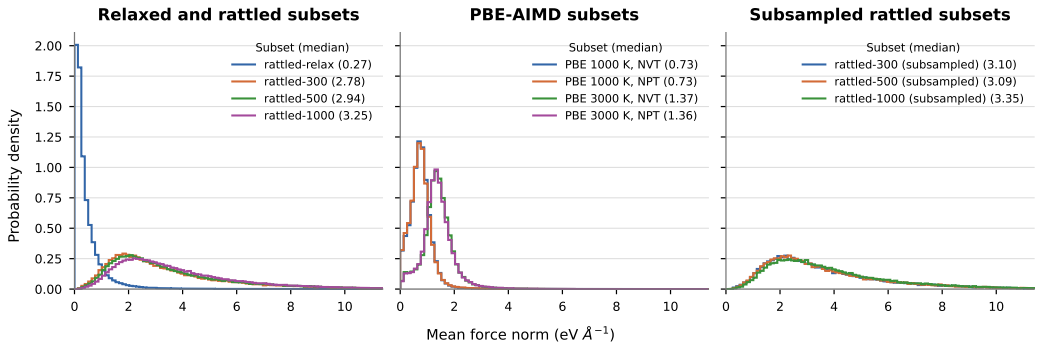


Figure 5: **Mean force norm distributions across OMat24-1M subsets.** Data subsets are grouped by generation protocol (rattled and then re-relaxed, rattled at various temperatures, or ab initio molecular dynamics trajectories; see Barroso-Luque et al. (2024) for more details). The legend shows subset medians in eV Å^{-1}.

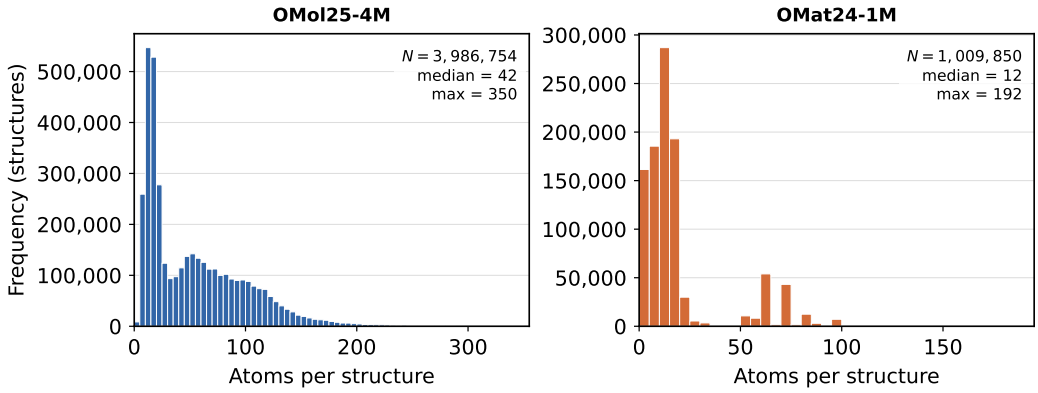


Figure 6: **Atom count distributions of OMol25-4M & OMat24-1M.** Hydrogens are included.

B.2 QM9

QM9 contains approximately 134,000 small organic molecules with quantum-chemically optimized geometries and molecular properties (Ramakrishnan et al., 2014). Each molecule has at most nine heavy atoms (C, N, O, or F), with hydrogens included. For joint molecule and material generation, we use Zatom-1's corresponding 100,000 QM9 training examples alongside its MP20 training subset described below, matching the corresponding training protocol of Zatom-1 (Morehead et al., 2026).

B.3 MP20

MP20 consists of 45,231 inorganic material structures from the Materials Project, with at most 20 atoms per unit cell and 89 elements across the dataset (Xie et al., 2022; Morehead et al., 2026). In this work, we use Zatom-1's corresponding 27,138 MP20 training structures and match its MP20 training conventions (Morehead et al., 2026).

For the structure prediction experiments in Appendix E.6, we also prepared a Crystalite partitioning of the dataset (Hadži Veljković et al., 2026). To do so, we processed Crystalite's published training/validation/testing CSV data splits separately with Niggli reduction enabled and primitive cell reduction and graph construction disabled.

B.4 GEOM-DRUGS

GEOM-Drugs provides energy-annotated conformer ensembles of drug-like molecules, covering larger and more structurally flexible systems than QM9 (Axelrod & Gómez-Bombarelli, 2022). We use it for unconditional molecule generation and a separate structure prediction (i.e., conformer generation) task. The latter uses ConfGF's processed training split (Shi et al., 2021) and the 200-molecule test split adopted by GO-Flow (Liu et al., 2026), containing 14,324 reference conformers.

B.5 SCOPE

SCOPe (Structural Classification of Proteins—extended) (Chandonia et al., 2022) takes experimentally determined structures largely from the PDB, splits proteins into structural domains (up to 256 residues), and organizes those domains hierarchically according to structural and evolutionary relationships (e.g., fold, family, and superfamily). We use the SCOPe ASTRAL-40 subset, which is obtained by clustering SCOPe domains at 40% sequence identity and selecting one representative from each cluster, ensuring that no two domains share more than 40% sequence identity. Genie (Lin & AlQuraishi, 2023) showed that using the short variant of SCOPe ASTRAL-40, containing only 3,936 domains of up to 128 residues after six domains were deleted because of parsing incompatibilities, is sufficient to train a well-balanced protein generative model. We call this the SCOPe-4k dataset.

We further sample selectively from the 488 SCOPe-4k fold groups to reduce the dataset size while preserving structural diversity:

1. Group domains by their SCOPe fold.
2. Randomize the order of the folds and the order of domains within each fold using a fixed random seed.
3. Cycle through the folds, taking one domain from each fold per cycle. If a fold runs out of domains, skip it.
4. Use the first 1,000 domains as the 1k dataset, the first 2,000 as the 2k dataset, and the first 4,000 as the 4k dataset.

B.6 DATASET LICENSES

Table 4 records the licensing terms for the datasets described in this appendix as well as the LeMat-Bulk reference dataset. These terms concern only the data; third-party software and pretrained models used in method evaluations retain their separate licenses.

Table 4: **Dataset licenses and access terms.** Dataset names link to the corresponding release or the provider's usage statement.

Dataset	Source data terms
OMol25	Creative Commons Attribution 4.0 (CC BY 4.0)
OMat24	CC BY 4.0
QM9	CC0 1.0 (original Figshare data deposit)
MP20	MIT for the processed ADiT release used by Zatom-1; underlying Materials Project data are CC BY 4.0
GEOM-Drugs	CC0 1.0 (Harvard Dataverse deposit)
SCOPe / ASTRAL	Provider states that all data are freely available to all users; no standard named license is specified
PDB source structures	CC0 1.0; attribution to structure authors is encouraged
LeMat-Bulk	CC BY 4.0, including attribution to upstream data providers

C Additional Model Details

C.1 Model forward pass

Algorithm 1 details Zatom-2's model forward pass and task readouts. Throughout this subsection, q, a are atom/token representations, c, s denote conditioning, and p, z are pair representations/biases. Subscripts 0 and $-$ denote initial values and the previous pass, respectively. D computes distograms of representative-token coordinates, W denotes a learned projection, and sg stops gradients. Enc, Trunk, and Dec are attention stacks; Down/Up are gated cross-attention between resolutions.

The feature dictionary f contains conditioning, atom-to-token and token-to-system maps, and fixed-coordinate masks. L, I, B index atom, token, and system levels, respectively: $\mathrm{MeanPool}_{L\to I}$ averages atoms within tokens, and $\mathrm{MeanPool}_{I\to B}$ averages tokens within systems. $\mathrm{Time}_{L/I}$ combines Fourier time features, RMSNorm, and a linear projection. Each Head includes its own adaptive normalization and output projection; the energy readout additionally sums domain-specific token contributions and divides by the system's atom count. The velocity, energy, and forces returned are, by construction, normalized; physical velocity is $\sigma_{\mathrm{data}}\hat{v}$, and property rescaling follows Section 3.1, namely Equation 4.

Algorithm 1 Zatom-2 forward pass with coordinate recycling and multitask readouts.

Require: Coordinates $x_t \in \mathbb{R}^{L\times 3}$, flow time t, features f, total passes $R \geq 1$ (default 2)

$(c_0, s_0, p, z_0) \leftarrow \mathrm{Initializer}(f)$ ▷ Static atom/token features and pair biases
$u \leftarrow x_t/\sigma_{\mathrm{data}}$; $q_0 \leftarrow \mathrm{RMSNorm}(c_0 + W_L u)$
$a_0 \leftarrow \mathrm{RMSNorm}(s_0 + \mathrm{MeanPool}_{L\to I}(W_I u))$
$(t_L, t_I) \leftarrow \mathrm{BroadcastTime}(t; f)$; set fixed atoms and fully fixed tokens to time 1
$c \leftarrow (q_0 + \mathrm{Time}_L(t_L))/2$; $s \leftarrow (a_0 + \mathrm{Time}_I(t_I))/2$
$G_L \leftarrow \mathrm{LocalNeighbors}(x_t, f; 128)$; $M_I \leftarrow \mathrm{SameSystemMask}(f)$
for $r = 1, \ldots, R$ **do**
  Evaluate this pass without gradients unless training and $r = R$
  $q \leftarrow q_0$; $a \leftarrow a_0$; $z \leftarrow z_0$ ▷ Restart node states on every pass
  if $r > 1$ **then**
    $z \leftarrow z_0 + W_D D(x_{\mathrm{rec}})$ ▷ 65-bin distogram of representative-token coordinates
  end if
  $q \leftarrow \mathrm{Enc}(q; c, p, G_L)$ ▷ Local atom DiT stack
  $a \leftarrow \mathrm{Down}(q, a; f)$ ▷ Tokens attend to their constituent atoms
  $a \leftarrow \mathrm{Trunk}(a; s, z, M_I)$ ▷ Global token DiT stack within each system
  $q \leftarrow \mathrm{Up}(q, a; f)$ ▷ Atoms attend to 24 channel groups of their token
  $q \leftarrow \mathrm{Dec}(q; c, p, G_L)$ ▷ Local atom DiT stack
  $\hat{v} \leftarrow \mathrm{Head}_{\mathrm{velocity}}(q, c)$; $\hat{v}[\text{fixed atoms}] \leftarrow 0$
  $\hat{x} \leftarrow x_t + (1 - t_L)\sigma_{\mathrm{data}}\hat{v}$ ▷ Endpoint estimate; fixed coordinates stay unchanged
  $\ell_{\mathrm{element}} \leftarrow \mathrm{Head}_{\mathrm{element}}(a, s)$; $\ell_{\mathrm{sequence}} \leftarrow \mathrm{Head}_{\mathrm{sequence}}(a, s)$
  $\hat{x}_{\mathrm{frac}} \leftarrow \mathrm{Head}_{\mathrm{fractional}}(q, c)$
  $g \leftarrow \mathrm{Down}_{I\to B}(a, g_0; f)$; $\bar{s} \leftarrow \mathrm{MeanPool}_{I\to B}(s)$ ▷ Learned system query g_0
  $(\hat{l}, \hat{\alpha}) \leftarrow \mathrm{Head}_{\mathrm{lattice}}(g, \bar{s})$ ▷ Lattice lengths and angles
  $\mathcal{O} \leftarrow \{\hat{x}, \hat{v}, \ell_{\mathrm{element}}, \ell_{\mathrm{sequence}}, \hat{x}_{\mathrm{frac}}, \hat{l}, \hat{\alpha}\}$
  if property heads are enabled **then**
    $\hat{e} \leftarrow \mathrm{Head}_{\mathrm{energy}}(a, s; f)$; $\hat{F} \leftarrow \mathrm{Head}_{\mathrm{force}}(q, c)$; $\mathcal{O} \leftarrow \mathcal{O} \cup \{\hat{e}, \hat{F}\}$
  end if
  $x_{\mathrm{rec}} \leftarrow \mathrm{sg}(\hat{x})$ ▷ Only coordinates feed the next pass
end for
return $\mathcal{O}$ ▷ Readouts from the final pass

To complement Algorithm 1, Table 5 compares one model forward pass (including recycling) in Zatom-2, RFdiffusion3's public source code (with standard memory mode enabled) (Butcher et al., 2025)[1] and Emyx's published pseudocode (Appendix J, Algorithms 2–17 of Williams et al., 2026, arXiv v1). The table lists graph construction first to align stages across models, while Algorithm 1 follows Zatom-2's implementation order. This reordering leaves the model's outputs unchanged.

[1] RFdiffusion3 source and configuration: `foundry/models/rfd3`, revision `b02eed6a6bdf`.

Table 5: **Aligned forward pass operations.** Enc, Trunk, and Dec are attention stacks; Down/Up are gated cross-attention between input resolutions. By definition, no recycling feedback is used in the first pass.

	RFdiffusion3	Emyx	Zatom-2
1	Build local atom neighborhoods (128 keys); use global token attention.	Build sparse atom/token graphs prioritizing sequence, bonds, ligand, and motif edges, then spatial neighbors.	Build local atom neighborhoods (128 keys, token-index/spatial neighbors); use global token attention.
2	Embed atom/token metadata; mix dense pair features with two triangle-free Pairformer blocks; pool atom pairs into token pairs.	Bottleneck-embed atom14 atom/token and global features; project relative positions and bonds to edge biases.	Bottleneck-embed atom1/atom14 metadata, domain, and charge/spin; residually-embed mean force norm; map bonds, relative positions, and any available fixed/reference geometry to head biases.
3	Add linear embeddings of $x_\sigma/\sqrt{\sigma^2+\sigma_{\text{data}}^2}$; condition on Fourier features of $\frac{1}{4}\log(\sigma/\sigma_{\text{data}})$.	Add Fourier coordinate embeddings; average geometry-aware conditioning with sinusoidal time features.	Add linear atom and mean-pooled linear token embeddings of x_t/σ_{data}; average each conditioning stream with Fourier features of t.
4	Cache $q_e = \text{Enc}(q_0; c, p)$ and $a_e = \text{Down}(q_e, a_0; s)$ *before* recycling.	Begin recycling before the atom encoder.	Begin recycling before the atom encoder.
5	Reset $q, a \leftarrow q_e, a_e$; concatenate z_0, noisy-input and recycled distograms; apply transitions and two triangle-free Pairformer blocks to obtain s, z.	Reset q, a to initial states plus projected, normalized $\text{sg}(q_-, a_-)$; set $z = z_0 + WD(\hat{x}_-)$ on sparse edges.	Reset $q, a \leftarrow q_0, a_0$; set $z = z_0 + WD(\hat{x}_-)$ for all token pairs. No node-state feedback.
6	Reuse cached encoder/downcast outputs.	$q \leftarrow \text{Enc}(q; c, p)$; $a \leftarrow \text{Down}(q, a)$.	$q \leftarrow \text{Enc}(q; c, p)$; $a \leftarrow \text{Down}(q, a)$.
7	$a \leftarrow \text{Trunk}(a; s, z)$, globally over tokens.	$a \leftarrow \text{Trunk}(a; s, z)$, along token edges.	$a \leftarrow \text{Trunk}(a; s, z)$, globally within each packed system.
8	For each decoder block: $q \leftarrow \text{Dec}_j(\text{Up}_j(q, a); c, p)$. Downcast detached states for sequence logits.	Upcast once using learned token copies; then $q \leftarrow \text{Dec}(q; c, p)$.	Upcast once using 24 channel groups per token; then $q \leftarrow \text{Dec}(q; c, p)$.
9	$u = W\text{RMSNorm}(q)$; $\hat{x} = c_{\text{skip}}x_\sigma + c_{\text{out}}u$. Predict sequence logits from the final downcast.	$v = \text{AdaLNHead}(q, c)$; $\hat{x} = x_t + (1-t)v$; zero motif velocities.	$\hat{v} = \text{AdaLNHead}(q, c)$; zero fixed-atom velocities; $\hat{x} = x_t + (1-t_L)\sigma_{\text{data}}\hat{v}$. Read out element/sequence, fractional coordinates, lattice, and optional energy/forces.
10	Recycle coordinates/distograms; repeat lines 5–9 (two passes by default).	Recycle detached nodes and endpoint geometry; repeat lines 5–9 ($N_r + 1 = 3$ inference passes).	Recycle endpoint geometry only; repeat lines 5–9 (two passes here).

C.2 HYPERPARAMETERS

Table 6 summarizes Zatom-2's architecture and OMol25/OMat24 pretraining settings. Joint pretraining samples the two domains (i.e., OMol25 molecules and OMat24 materials) equally with replacement, using ∼5 million draws per configured epoch for 80 epochs; an epoch is therefore a sampling budget rather than a full pass through each source dataset. Generation/structure prediction samples the two training tasks equally. Adding MLIP supervision changes the generation/structure prediction/MLIP task probabilities to 0.34/0.33/0.33. Force conditioning is supplied for 80% of eligible generation and structure prediction examples and disabled in the non-force-conditioned model ablations.

Table 6: **Key Zatom-2 hyperparameters.** S/base/L vary the token trunk depth while retaining the same feature widths. Token budgets are per GPU before diffusion batch/view augmentation; a smaller token budget is paired with twice as much gradient accumulation.

Hyperparameter	Setting
Approximate model size (S/base/L)	70M / 140M / 270M parameters
Token-trunk blocks (S/base/L)	9 / 18 / 36
Atom encoder / decoder blocks	3 / 3
Atom / token / token-pair widths	128 / 768 / 256
Atom / token attention heads	4 / 12
Attention	Local atom attention (128 keys); global token attention
Transformer dropout / drop-path rate	0.1 / 0.1
Denoiser passes	2 (one coordinate-derived recycling update)
Optimizer	Adam; constant learning rate 10^{-4}
Adam (β_1, β_2) / ϵ	$(0.9, 0.95)$ / 10^{-8}
Gradient norm clipping value / EMA decay rate	10 / 0.999
Training precision	`bf16-mixed` (default) / `32-true` (MLIP finetuning)
Packed token budget per step (S/base; L)	7,168; 3,584
Crop limit / augmented views per system	512 tokens / 2 views
Coordinate prior scale σ_{data}	16 Å
Training-time distribution	$0.98\,\text{Beta}(1.9, 1) + 0.02\,\text{Uniform}(0, 1)$, clipped to $[10^{-3}, 1 - 10^{-3}]$

The Euclidean coordinate loss has a weight of 4, with a smoothed lDDT loss coefficient of 0.25; atom element type and token sequence type cross-entropy losses each have weight 0.1. Joint generative/structure-MLIP pretraining uses energy/force/latent equivariance loss weights of 0.1/0.3/0.1 and a 10,000-step MLIP loss warmup. To balance gradient magnitudes between generative and predictive tasks, fractional coordinates, lattice length, and lattice angle weights are 0.25/0.05/0.05 with MLIP supervision and 10/1/1 otherwise. Notably, our SCOPe transfer learning experiments use a batch size of 4 input systems per GPU, four augmented views per system, and crops of at most 128 tokens and 1,792 atoms.

Our configuration for OMol25/OMat24 generated sample evaluation uses 200 sampling steps with force-only classifier-free guidance at a scale of 2. In contrast, our QM9/MP20 results use 100 steps without classifier-free guidance (n.b., to fairly compare these inference settings to those of Zatom-1) and a 3,584-token training budget per GPU. Both sampling configurations use EMA weights, two denoiser passes, an EDM schedule exponent of 7, a noise scale of 1.003, and a step scale of 1.5. Furthermore, their churn coefficients are 0.6 and 0.8, respectively.

C.3 Compute resources

Training and evaluation use a computing cluster with NVIDIA A100 GPUs containing 80 GBs of GPU memory. Distributed launchers configure four GPUs per node, for either four-node (16-GPU) jobs or sixteen-node (64-GPU) jobs. For pretraining, gradient accumulation preserves an approximate maximum of $64 \times 7{,}168 = 458{,}752$ tokens per optimizer update before the two augmented diffusion views, including when the per-GPU token budget is halved for Zatom-2-L's experiments.

Table 7 estimates training costs from training run metadata by multiplying elapsed run time by the number of GPUs used for each experiment.

D Evaluation Metrics

In this section, we describe the metrics defined in the `nucleus` and `zevals` code implementations underlying Zatom-2. Unless stated otherwise, lower error or distance is better.

Table 7: **Approximate 80GB A100 GPU hours by training configuration.** G/S/P denote generation, structure prediction, and MLIP prediction; FC denotes force conditioning. Epoch budgets count completed epochs; MLIP finetuning runs end at raw epoch 119. All finetuning costs exclude pretraining the corresponding initial model weights.

Model training configuration	Budget / endpoint	GPUs	GPU hours
Small, OMol25 G, no FC	80 epochs	16	1,830
Small, OMol25 G, FC	80 epochs	16	1,820
Small, OMol25 G+S, FC	80 epochs	64	2,280
Base, OMol25+OMat24 G+S, FC	80 epochs	64	2,740
Base, OMol25+OMat24 G+S+P, FC	80 epochs	64	2,850
Large, OMol25+OMat24 G+S+P, FC	80 epochs	64	4,090
Base, OMol25+OMat24 G+S, no FC	80 epochs	64	2,810
Base, fixed 500k per OMol25/OMat24 dataset, G+S+P	16 * 5 = 80 epochs	64	540
Large, fixed 500k per OMol25/OMat24 dataset, G+S+P	16 * 5 = 80 epochs	64	790
Base, QM9+MP20 G	10,000 epochs	16	2,380
Base, GEOM-Drugs G	80 epochs	64	970
Base, MLIP finetuning from G+S	40 epochs	64	1,520
Base, MLIP finetuning from G+S+P, Part 1	40 epochs	64	1,520
Base, MLIP finetuning from G+S+P, Part 2	+120 epochs	64	4,560
SCOPe, no pretraining	Step 25,000	16	210
SCOPe, G+S initialization, no FC	Step 125,000	16	270
SCOPe, G+S initialization, FC	Step 120,000	16	180
SCOPe, G+S+P initialization, FC	Step 130,000	16	320

D.1 Force self-consistency metrics

In this subsection, we define how we measure the consistency between a molecule or material sample's predicted forces and those used as conditioning for that sample.

- **Force self-consistency.** For a generated sample, we compute its mean atomic force norm, $N^{-1}\sum_{\ell=1}^{N}\|\mathbf{F}_\ell\|_2$, without relaxation. To predict a sample's forces, UMA-S-1p2 uses its OMol and OMat task heads for molecules and materials, respectively (Wood et al., 2025). We report the mean of this per-structure statistic in eV Å^{-1}. When running UMA on 1000 random structures from OMol25 and 1000 structures from OMat24, we get MAEs between calculated mean force norms and those extracted from the metadata of 0.00587 and 0.02462 eV Å^{-1}, respectively.
- **UMA target MAE.** This is the mean absolute deviation of UMA-S-1p2's predicted mean force norms and the requested conditioning value over samples successfully scored by UMA. As a reference for this force-norm proxy, we score 1,000 structures sampled uniformly without replacement from each of the OMol25 and OMat24 validation splits (seed 42), using their original geometries without relaxation.
- **UMA AUROC.** This measures the area under the receiver operating characteristic curve (AUROC) for distinguishing low- and high-force conditioning targets, using a dataset's ground-truth mean force norm values with cutoff $\geq \tau$ as the positive class. We use $\tau = 1$ eV Å^{-1} in all contexts, except for OMol25's ANI2x data subset, where $\tau = 2.5$ eV Å^{-1}, based on the subset distributions shown in Figure 4 and Figure 5.

D.2 Molecule generation metrics

In this subsection, we define the default metrics used to assess molecule sample quality, which differ slightly from those used in the main text's Section 4.2 for GEOM-Drugs/QM9 baseline experiments.

- **RDKit validity and uniqueness.** Validity is the fraction of generated molecule samples for which RDKit produces a valid SMILES representation of the sample's largest molecular fragment (Landrum, 2013). Uniqueness is the number of distinct valid SMILES divided by the number of valid samples.

- **Internal diversity.** For the set of unique valid SMILES, diversity is the mean pairwise Tanimoto distance, $1 - s_{ij}$, between radius-2, 2,048-bit Morgan fingerprints.
- **Novelty.** To calculate novelty, the experiments in the main text query each valid generated SMILES against a precomputed (OMol25) training set database of fingerprints generated by FPSim2 (Félix et al., 2025). In this setting, a method's corresponding novelty value is one minus the mean top-1 Morgan fingerprint Tanimoto similarity to the training set of all its generated molecule samples.
- **PoseBusters pass rate.** PoseBusters applies standardized loading, sanitization, connectivity, valence, geometry, clash, flatness, and internal energy tests (Buttenschoen et al., 2024). The inputs for each test are RDKit-valid molecules; a method's aggregate "PoseBusters-valid" rate is its fraction of RDKit-valid molecules passing every configured PoseBusters test.
- **Force diagnostics.** UMA force norms and force conditioning metrics follow Appendix D.1, using UMA's OMol head. Additionally, Table 1 of the main text reports UMA's OMol25 summary metrics.

D.3 Material generation metrics

In this subsection, we define the default metrics used to assess material sample quality, which differ slightly from those used in the main text's Section 4.2 for MP20 baseline experiments.

- **Compositional validity.** A material's composition is considered valid if SMACT finds that its oxidation states satisfy charge neutrality and the Pauling electronegativity test; note that single-element systems and all-metal alloys are accepted by the metric (Davies et al., 2019).
- **Structural and overall validity.** A material's structure is considered valid if it has a periodic cell with volume at least $0.1\,\text{Å}^3$ and no periodic interatomic distance below $0.5\,\text{Å}$. Overall validity of a material requires both compositional and structural validity and a successfully constructed CrystalNN fingerprint (Pan et al., 2021).
- **Uniqueness.** To calculate uniqueness, valid materials are grouped using Pymatgen's `StructureMatcher` utility with $(\text{ltol}, \text{stol}, \text{angle_tol}) = (0.3, 0.5, 10^\circ)$ (Ong et al., 2013). Uniqueness is the number of groups divided by the number of valid materials.
- **Internal diversity.** Each valid material is represented by its normalized mean atom-wise CrystalNN fingerprint. Diversity is then one minus the mean off-diagonal cosine similarity over all successfully fingerprinted generated materials.
- **Novelty.** To calculate novelty, one representative material from each unique valid group is compared with cached OMat24 training fingerprints. A material is considered novel if its cosine similarity to every reference fingerprint is below 0.99; novelty is reported over only successfully fingerprinted representatives.
- **Force diagnostics.** UMA force norms and force conditioning metrics follow Appendix D.1, using UMA's OMat head.

D.4 QM9 and MP20 sample generation metrics

The QM9 and MP20 benchmarks of Section 4.2 in the main text follow Appendix B of Zatom-1 (Morehead et al., 2026). Here, we describe only the differences from Appendices D.2 and D.3.

QM9. Table 11 reports individual PoseBusters pass rates for connectivity, bond angles, bond lengths, aromatic ring flatness, double bond flatness, internal energy, and internal steric clashes. Each constituent PoseBusters test result is the fraction of generated molecules passing that test among RDKit-valid molecules. Molecular validity follows Appendix D.2.

MP20. The LeMat-GenBench metrics in Appendix Table 12 differ from the material metrics in Appendix D.3 as follows (Betala et al., 2025):

- **Overall validity.** A valid material must pass charge balance, interatomic distance, coordination, and physical unit tests. These tests assess oxidation states and bond valences, radii-based separation, element-specific coordination, and density, lattice, format, and symmetry constraints.
- **Uniqueness.** The number of structurally distinct valid materials is divided by all attempted materials, rather than only valid materials.
- **Novelty.** A unique valid material is novel if it has no structural match in the (5M) LeMat-Bulk reference dataset. Reported novelty rates divide the number of such materials by all attempted samples; they do not use the default OMat24 fingerprint similarity criterion.
- **Metastable, unique, and novel (MetaSUN).** The fraction of all attempted materials that are valid, unique, novel, and satisfy $0 < E_{\text{hull}} \leq 0.1$ eV/atom, where E_{hull} is the mean energy above the hull across an MLIP ensemble of Orb, MACE, and UMA. Notably, this metastable test excludes stable materials with $E_{\text{hull}} \leq 0$, a portion of which can often be recovered via post hoc MLIP-based relaxation (n.b., which we exclude in this work to assess direct method output quality).

D.5 GEOM-Drugs sample generation metrics

GEOM-Drugs validity, uniqueness, and aggregate PoseBusters validity follow Appendix D.2; individual PoseBusters pass rates use the definition in Appendix D.4, following Zatom-1 (Morehead et al., 2026). The additional metric definitions needed for the results in Appendix E.4.1 are:

- **Valid & unique and Valid & PB-valid.** These joint (intersection) metrics divide the number of distinct valid SMILES and the number of RDKit-valid molecules passing all PoseBusters checks, respectively, by all attempted samples. They therefore include the effect of RDKit validity, whereas uniqueness and PoseBusters validity are conditional on RDKit validity.

D.6 Atomistic Fréchet Distance

Background. The Atomistic Fréchet Distance (AFD) is the Fréchet Inception Distance (FID) applied to 3D atomistic structures Perez & Gómez-Bombarelli (2026a). FID scores an image generator by the Fréchet distance between Gaussian fits to embeddings of real and generated images created with a pretrained model (Inception-v3) Heusel et al. (2017). AFD replaces the Inception network with a self-supervised pretrained embedding model for atomistic data and applies the same distance to sets of molecules or materials. AFD scores provide a label-free measure of distributional similarity between two sets of atomistic structures - a ground truth reference set and a generated set. Unlike other binary evaluation tests like molecule 'validity' or PoseBusters geometry checks, AFD scores are sensitive to both compositional and conformational diversity and better represent a generative model's ability to produce realistic and diverse atomistic structures. Lower AFD scores suggest generated data is less distinguishable from real data (lower is better). However, AFD does not certify that any single structure is valid and should be reported in combination with validity checks.

$$\text{AFD}^2 \;=\; \|\mu_r - \mu_g\|_2^2 \;+\; \text{Tr}\Big(\Sigma_r + \Sigma_g - 2\big(\Sigma_r\Sigma_g\big)^{1/2}\Big) \tag{7}$$

Here (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the reference and generated embeddings, and Equation 7 is the Fréchet distance between the two Gaussians. Embeddings are the 256-dimensional graph-level output of a frozen Conditional Equivariant Transformer (CT) Perez & Gómez-Bombarelli (2026b). The encoder defines the feature space and therefore what the score can detect. Notably, the CT embedding is $E(3)$-invariant and responds monotonically to coordinate noise and to atom mutation or removal. It is also insensitive to chirality and unit-cell multiplicity.

Choice of encoder. The encoder checkpoint chosen for AFD must match one's evaluation domain (e.g., molecules, materials) because cross-domain checkpoints are far less sensitive. AFD is not an absolute score and requires a ground-truth reference set to define its target distribution. Additionally, the Fréchet estimator is biased upward at finite N, so scores are comparable only at equal N with the same reference set and checkpoint. In this work, we use the pretrained CT checkpoint `ct-scd-omol25` (pretrained on OMol25-4M) to evaluate generated samples from the OMol25

distribution and the checkpoint `ct-scd-amp20` (pretrained on Alex-MP-20) to evaluate models trained on OMat24 (see Table 8).

Table 8: AFD baselines for encoder dataset pairs used in this work (finite-N floor): Each entry is the mean $\pm$ standard deviation of the AFD score between two non-overlapping ground-truth reference sets of size N, over 5 replicates (no structure is shared within or across replicates at a given N).

N	OMol25 `4m` / `ct-scd-omol25`	OMat24 `1m` / `ct-scd-amp20`
100	0.02321 ± 0.00021	0.0776 ± 0.0042
500	0.005461 ± 0.000078	0.01691 ± 0.00091
1,000	0.002527 ± 0.000032	0.008268 ± 0.00034
5,000	0.0000916 ± 0.0000047	0.001236 ± 0.000027

D.7 Structure prediction metrics

Molecules. For the OMol25 structure prediction evaluation in Table 9, each target has one reference structure and one generated candidate. We additionally report symmetry-aware best heavy-atom RMSD after rigid alignment, removing hydrogens and requiring the generated and reference molecular graphs to agree. Note that force consistency is the primary metric discussed in Section 4.1. We report the mean and median molecular RMSD in Å. In this work, molecule structure prediction evaluations use 500 held-out OMol25 test set targets unless otherwise specified. The GEOM-Drugs structure prediction evaluation instead follows Appendix E.5.

Materials. For structure prediction of periodic materials, we use Pymatgen's `StructureMatcher` utility with $(\mathrm{ltol}, \mathrm{stol}, \mathrm{angle_tol}) = (0.3, 0.5, 10^\circ)$ and volume rescaling enabled (Ong et al., 2013). Here, a match rate is the fraction of all prediction targets that match their predicted structures within the predefined geometric tolerances above. Normalized root mean square (RMS) displacement divides predicted-reference coordinate deviations by the characteristic length per atom, $(V/N)^{1/3}$, and is summarized only over matched targets; note that it is not an Angstrom RMSD per se. Failed conversions and matcher errors count as unmatched predictions.

D.8 Interatomic potential metrics

Energies and forces. For systems $b = 1, \ldots, B$ with n_b labeled atoms and ground-truth energy E_b and forces F_b, we report the MLIP metrics

$$\mathrm{MAE}_E = \frac{1}{B}\sum_b |\widehat{E}_b - E_b|, \qquad \mathrm{MAE}_{E/\mathrm{atom}} = \frac{1}{B}\sum_b \frac{|\widehat{E}_b - E_b|}{n_b}, \tag{8}$$

$$\mathrm{MAE}_F = \frac{1}{3\sum_b n_b}\sum_{b,\ell,k} |\widehat{F}_{b\ell k} - F_{b\ell k}|. \tag{9}$$

Total energy and energy-per-atom errors are in eV and eV/atom, respectively; force component errors are in eV Å^{-1}. Energy coverage is the fraction of systems with finite predictions and valid targets after applying elemental reference masks; force coverage is defined analogously over atoms with finite vector labels.

For reference, MLIP task results using these metrics are reported in Appendix E.2.

D.9 Protein generation metrics

The metrics used for Table 3 are defined below. For each generated backbone, we use SolubleMPNN (Dauparas et al., 2022) to design eight sequences and ESMFold2 (Candido et al., 2026) to refold each sequence. We apply no pLDDT confidence gate.

- **Self-consistency RMSD (scRMSD).** Each refold is aligned to its generated backbone by a Kabsch superposition of corresponding Cα atoms. Best-of-eight scRMSD is the minimum value for each backbone; means and medians are then taken across backbones.

- **Designability.** A protein backbone is considered designable when at least one of its eight refolds has Cα scRMSD below 1.5 Å (the best-of-eight criterion).
- **Per-sequence success rate.** For backbone i, let m_i be the number of its eight SolubleMPNN refolds with scRMSD below 1.5 Å. The reported rate is $N^{-1}\sum_{i=1}^{N} m_i/8$, the expected success of choosing one designed sequence uniformly per backbone.
- **Distinct designable clusters.** These are computed by restricting the symmetric TM-align matrix of all generated backbones to the best-of-eight designable subset, applying clustering at TM≥ 0.5, and counting the resulting clusters.
- **Novelty.** We select the best-of-eight designable backbones and compare every retained backbone against the canonical SCOPe-2k training set with TM-align (Zhang & Skolnick, 2005). For each query, we take the maximum TM-score normalized by query length. Novelty is defined as the number of backbones with score below 0.5, divided by the number of designable backbones.
- **Joint sequence–structure clusters.** For the checkpoint-sweep diagnostic in Appendix E.3, we again use the best-of-eight designable subset. A pair is similar only when its min-symmetrized TM-align score is at least 0.5 and its generated sequences have at least 10% identity with 70% bidirectional coverage. We apply complete-linkage clustering to this joint relation and count the resulting clusters, following the sequence-and-structure diversity motivation of Proteina and RFD3 (Geffner et al., 2025; Butcher et al., 2025).

E Additional Results

E.1 OMol25 and OMat24 structure prediction

To enable bond conditioning for the structure prediction task on OMol25, we infer bonds using OpenBabel (O'Boyle et al., 2011). Because OMol25 contains high-energy structures, reactive compounds and biomolecules, for which OpenBabel cannot correctly infer bonds, we filter out all compounds which contain disconnected fragments after bond inference, as a proxy for failure of inferring the correct bonds. This excludes 60.32% of OMol25 training examples, and we train on the remaining subset for structure prediction.

Table 9 reports molecule and material structure prediction results for the pretrained configurations in Table 1. Each target has one generated candidate and evaluation metrics are defined in Appendix D.7.

Table 9: **Molecule/material top-1 structure prediction after pretraining.** All models use the 80-epoch pretrained checkpoints described in Table 1. OMol25 RMSDs are reported in Å and forces are calculated with UMA. Compared to OMol25, OMat24 structure prediction is more sensitive to outliers, so we report median metrics for OMat24. OMol25 AUROC uses a target mean force norm cutoff of 1 eV Å^{-1}.

	OMol25				
Model	Mean RMSD ↓	Median RMSD ↓	Mean $\overline{\|F\|}$ ↓	Target MAE ↓	AUROC ↑
Zatom-2 (III)	1.739	1.714	2.271	0.725	0.798
Zatom-2 (IV)	1.708	1.675	2.138	0.787	0.787
Zatom-2 (V)	1.762	1.729	1.953	0.637	0.801
Zatom-2 (VI)	1.723	1.743	1.211	1.214	0.818

	OMat24			
Model	Match ↑	Median norm. RMS disp. ↓	Median $\overline{\|F\|}$ ↓	Target Median AE ↓
Zatom-2 (IV)	0.446	0.225	2.765	1.703
Zatom-2 (V)	0.294	0.349	32.458	30.553
Zatom-2 (VI)	0.460	0.321	18.867	17.380

E.2 MLIP PERFORMANCE BEFORE AND AFTER FINETUNING

Zatom-2 predicts energies and per-atom forces for both molecules and periodic materials with the same model used for generation and structure prediction. Table 10 compares pretrained and MLIP-finetuned versions of Zatom-2 for energy and force prediction. Here, evaluation uses 2.76 million OMol25 and 107,732 OMat24 validation examples, with fixed atom identities and geometries and no force conditioning. The finetuned Zatom-2 base models start from generation/structure pretraining (Zatom-2 (IV)) or joint generation/structure/MLIP pretraining (Zatom-2 (V)); Zatom-2 (VI) provides a reference point for how a larger pretrained model fairs for MLIP prediction.

Finetuning improves prediction, especially for molecules. After 40 epochs of finetuning, Zatom-2 (V)'s force MAE falls from 0.253 to 0.130 $\mathrm{eV\,\AA^{-1}}$ on OMol25 (49%) and from 0.637 to 0.587 $\mathrm{eV\,\AA^{-1}}$ on OMat24 (8%), with lower aggregate energy errors in both domains. Extended finetuning further reduces force MAE to 0.097 and 0.564 $\mathrm{eV\,\AA^{-1}}$, respectively. Starting from joint MLIP pretraining also yields lower aggregate errors than starting from generation/structure pretraining alone, suggesting that generative pretraining can still produce a rich model latent space but nonetheless one that is not as informative of atomistic properties as a latent space produced via MLIP supervision.

Generative capabilities are largely retained. Keeping generation and structure prediction examples in Zatom-2's finetuning data mixture preserves the model's capabilities for these tasks while improving MLIP accuracy. For Zatom-2 (V), OMol25 generation AFD improves from 0.019 to 0.016 and structure prediction RMSD changes from 1.762 to 1.729 Å; OMat24 generation validity and structure match rate remain comparable to their pretrained values. Nonetheless, retention is metric-dependent: for example, molecular structure prediction's force conditioning MAE worsens with extended finetuning even as UMA mean force norms and MLIP errors continue to decrease. Moreover, materials generation and structure prediction metrics degrade with MLIP finetuning, however, interestingly, the performance is in part recovered with more finetuning epochs (Zatom-2 (V) + FT (extended)).

Accuracy and physical consistency remain limiting. Zatom-2's energy and force errors remain substantially above specialized MLIP baselines such as eSEN, AllScAIP, and UMA (Morehead et al., 2026; Qu et al., 2026; Wood et al., 2025). These comparisons provide context rather than a controlled ranking, since training budgets and evaluation protocols partially differ between these methods (including our 1M OMat24 training subset versus AllScAIP's 100M OMat24 training set size). Zatom-2's improvements are also uneven across subsets of OMol25 and OMat24; for instance, biomolecular total-energy error increases slightly after initial finetuning, and rattled materials remain difficult. Lastly, Zatom-2's direct force prediction head enforces neither consistency between predicted energies and their gradients nor exact rotational equivariance with respect to the model's input geometries. These results demonstrate that Zatom-2 supports multitask energy and force prediction, but they currently do not establish Zatom-2 as a suitable method for high-accuracy energy-conserving molecular dynamics or reliable physical simulations.

Table 10: **MLIP accuracy and retention of generative capabilities.** "+ FT" denotes 40 epochs of finetuning with 80% of training steps allocated to MLIP prediction and 10% for generation and structure prediction each; here, "extended" means adding 120 additional MLIP-heavy finetuning epochs. Model names follow Table 1. Panel (a) reports generation and top-1 structure prediction results after finetuning, with forces calculated using UMA; corresponding pretrained results appear in Tables 1 and 9. For OMat24, we observe that generation and structure prediction are sensitive to outliers and we report median values. Panels (b)–(c) report validation MAEs (Appendix D.8); external baselines are published values with partially differing training and evaluation protocols.

(a) Generation and top-1 structure prediction retention

	OMol25								
	Generation				Structure prediction				
Model	AFD ↓	Mean $\overline{\|F\|}$ ↓	Target MAE ↓	AUROC ↑	Mean RMSD ↓	Median RMSD ↓	Mean $\overline{\|F\|}$ ↓	Target MAE ↓	AUROC ↑
Zatom-2 (IV) + FT	0.017	1.205	0.540	0.807	1.771	1.747	1.570	0.632	0.890
Zatom-2 (V) + FT	0.016	1.137	0.530	0.809	1.729	1.706	1.458	0.690	0.828
Zatom-2 (V) + FT (extended)	0.015	1.040	0.526	0.840	1.720	1.713	1.178	0.871	0.882

	OMat24							
	Generation				Structure prediction			
Model	Valid ↑	AFD ↓	Median $\overline{\|F\|}$ ↓	Target Median AE ↓	Match ↑	Median norm. RMS disp. ↓	Median $\overline{\|F\|}$ ↓	Target Median AE ↓
Zatom-2 (IV) + FT	0.812	1.114	11.167	9.651	0.406	0.269	10.834	8.790
Zatom-2 (V) + FT	0.759	1.182	38.826	36.866	0.308	0.354	32.128	30.098
Zatom-2 (V) + FT (extended)	0.758	1.177	25.783	24.169	0.382	0.310	20.843	19.019

(b) OMol25 MLIP prediction

Model	E MAE (eV) ↓	E/atom MAE (eV/atom) ↓	Force MAE (eV/Å) ↓	E coverage ↑
Zatom-1 (joint)	2.512780	0.034450	0.062990	–
eSEN-md-d. (4M)	0.077130	0.001320	0.006780	–
AllScAIP-md-ft-cons. (4M)	0.044400	–	0.007510	–
Zatom-2 (V)	1.366503	0.016426	0.252603	100%
Zatom-2 (VI)	1.253679	0.015344	0.228265	100%
Zatom-2 (IV) + FT	1.709557	0.018713	0.164550	100%
Zatom-2 (V) + FT	1.178403	0.013088	0.130012	100%
Zatom-2 (V) + FT (extended)	1.015206	0.011174	0.096865	100%

(c) OMat24 MLIP prediction

Model	E MAE (eV) ↓	E/atom MAE (eV/atom) ↓	Force MAE (eV/Å) ↓	E coverage ↑
UMA-L	–	0.009700	0.043500	–
AllScAIP-md-ft-cons.	–	0.010700	0.054300	–
Zatom-2 (V)	1.634724	0.100531	0.636802	99.9991%
Zatom-2 (VI)	1.637466	0.098387	0.617528	99.9991%
Zatom-2 (IV) + FT	1.618015	0.095861	0.604039	99.9991%
Zatom-2 (V) + FT	1.395262	0.089066	0.586901	99.9991%
Zatom-2 (V) + FT (extended)	1.338368	0.087149	0.563980	99.9991%

E.3 SCOPE-2K CHECKPOINT TRAJECTORIES

Because we train our protein generation models on a small dataset which is prone to overfitting, we evaluate intermediary model checkpoints during training and finetuning to select the checkpoint with the best designability / novelty tradeoff. We perform the study for all four initializations of Zatom-2 in Table 3 and for RFdiffusion3 (Butcher et al., 2025). Each checkpoint was evaluated using 96 unconditional backbones, eight SolubleMPNN sequences per backbone, and ESMFold2 refolding. The pretrained checkpoints begin at raw step 100k, so Figure 7 reports their local finetuning step after subtracting 100k, and scratch-trained Zatom-2 and RFD3 start at 0 steps. Metric definitions are given in Appendix D.9.

Based on convergence trends, we selected the following checkpoints for evaluation: Zatom-2 from scratch at `step-00025000`; Zatom-2 (IV) without force conditioning at `step-00125000`

(25k local finetuning steps); Zatom-2 (IV) with force conditioning at `step-00120000` (20k local steps); Zatom-2 (V) at `step-00130000` (30k local steps); and RFdiffusion3 at `step-00030000`.

The trajectories show faster convergence for pretrained models. The generation-and-structure-pretrained model exceeds 85% designability at 10k local steps, whereas Zatom-2 trained from scratch and RFdiffusion3 first exceed 85% designability 15k. The two models trained from scratch also deteriorate more sharply after they peak. The pretrained models are generally preserving high designability and novelty over the same late-training region. We interpret this pattern of models trained from scratch as evidence consistent with overfitting, whereas the pretrained models show more resilience against overfitting.

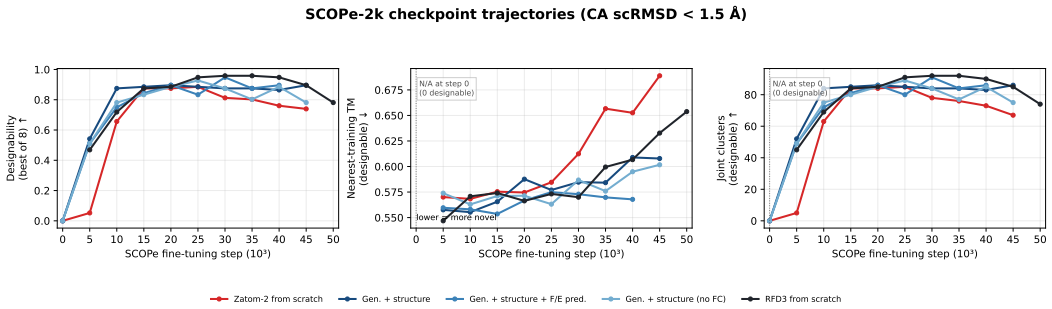


Figure 7: **SCOPe-2k convergence** From left to right, the panels show best-of-eight designability (CA scRMSD < 1.5 Å), the mean maximum exact, query-normalized TM-align score against the canonical SCOPe-2k training set among designable backbones (lower is more novel), and distinct joint sequence–structure clusters among designable backbones. Joint clusters require TM ≥ 0.5 together with at least 10% sequence identity and 70% coverage. Each point uses 96 generated backbones and eight refolds per backbone.

E.4 Comparison to generative baselines

E.4.1 Generation on QM9, GEOM-Drugs, and MP20

Table 11 reports results on QM9 generation for a model trained jointly on QM9 and MP20, and Table 12 contains full generation results for MP20. In Table 13 we report individual PoseBusters checks for GEOM-Drugs generation.

Table 11: **QM9 generation results.** Zatom-2 is trained jointly on QM9 and MP20 (Appendix B). Published baselines are from Table 2 of Morehead et al. (2026) (arXiv v4).

(a) Validity (%)

Model	Validity ↑
Equivariant Diffusion	91.90
Symphony	83.50
GeoLDM	93.80
ADiT (QM9-only)	92.19
ADiT (joint QM9+MP20)	94.45
Zatom-1 (QM9-only, 80M)	92.88
Zatom-1 (joint QM9+MP20, 80M)	94.94
Zatom-1-L (joint QM9+MP20, 160M)	95.26
Zatom-2 (joint QM9+MP20)	95.72

(b) PoseBusters sanity checks (% pass)

Test ↑	Symphony	Eq. Diff.	ADiT (joint)	Zatom-1 (joint, 80M)	Zatom-2 (joint)
Atoms connected	99.92	99.88	99.70	99.98	100.00
Bond angles	99.56	99.98	99.85	99.95	99.93
Bond lengths	98.72	100.00	99.41	99.97	99.93
Aromatic ring flat	100.00	100.00	100.00	100.00	100.00
Double bond flat	99.07	98.58	99.98	99.99	99.99
Internal energy	95.65	94.88	95.86	99.78	99.61
No steric clash	98.16	99.79	99.79	99.81	99.98

Table 12: **MP20 generation results (unrelaxed).**

Model	Overall valid ↑	Unique ↑	Novel ↑	MetaSUN ↑
ADiT (joint QM9+MP20)	90.6	87.8	26.0	1.0
Crystal-GFN	51.7	51.7	51.7	0.0
Crystalformer	69.9	69.4	31.8	3.1
LLaMat2-CIF	84.4	81.4	30.0	2.1
LLaMat3-CIF	15.4	15.2	10.5	0.2
SymmCD	73.4	73.0	47.0	2.4
Zatom-1 (MP20-only, 80M)	73.2	70.4	21.0	0.20
Zatom-1 (joint QM9+MP20, 80M)	88.5	84.4	8.1	0.36
Zatom-1-L (joint QM9+MP20, 160M)	95.0	90.4	3.7	0.64
Zatom-2 (joint QM9+MP20)	70.24	53.44	32.24	4.92

Table 13: **Individual PoseBusters checks for GEOM-Drugs generation.** Published baseline values are reported from Table 3 of Morehead et al. (2026) (arXiv v4). Results are percentages among RDKit-valid molecules.

Model	Atoms connected	Bond angles	Bond lengths	Aromatic ring flat	Double bond flat	Internal energy	No steric clash
EQGAT-diff	84.4	86.9	87.0	87.0	87.0	86.8	82.9
SemlaFlow	92.3	94.8	94.6	94.9	94.2	94.8	92.0
ADiT	93.0	92.3	92.5	95.4	95.3	91.3	91.8
TABASCO	99.9	99.2	99.4	100.0	99.8	99.4	94.3
Zatom-1	99.6	99.6	99.7	100.0	100.0	99.3	96.4
GEOM-Drugs (reference)	–	–	–	–	–	–	–
Zatom-2	99.18	99.51	99.17	100.00	99.97	99.31	94.34

E.5 GEOM-Drugs conformer generation benchmark

To further characterize Zatom-2's architecture, we evaluate it on the GEOM-Drugs conformer generation benchmark. While we use force consistency as a representative metric for structure prediction when conformer ensembles are unavailable for individual molecular graphs, this benchmark enables a direct assessment of Zatom-2's conformer prediction capabilities against established methods. We compare against GO-Flow (Liu et al., 2026), a specialized conformer prediction model, and test whether Zatom-2's structure generation sampler can recover a conformer ensemble when atom identities and molecular connectivity are fixed. Identical GEOM-Drugs training, validation, and test data splits are used for both Zatom-2 and GO-Flow.

Table 14 reports coverage (COV) and matching (MAT) metrics following Liu et al. (2026), using RDKit symmetry-aware RMSD after removing hydrogens. For each molecule, COV-R (recall) is the percentage of reference conformers with a generated conformer within 1.25 Å RMSD, and COV-P (precision) is the percentage of generated conformers with a reference conformer within the same threshold. MAT-R is the mean minimum RMSD from each reference conformer to the generated ensemble, and MAT-P reverses this direction. Higher COV and lower MAT indicate better performance. We report the mean of these per-molecule scores across the test set. Zatom-2 achieves higher precision than GO-Flow (Liu et al., 2026), but lower recall. Higher precision is valuable under a limited sampling budget: a larger fraction of generated candidates lies close to reference conformations; however, the model does not cover the entire conformational diversity of the reference ensemble as well.

Table 14: **GEOM-Drugs conformer generation evaluation.** Values are means across test molecules; COV is reported as a percentage and MAT as RMSD in Å. GO-Flow values are from Table 1 of Liu et al. (2026).

Model	COV-R mean ↑	MAT-R mean ↓	COV-P mean ↑	MAT-P mean ↓
GO-Flow	**94.82**	**0.7971**	70.13	1.1068
Zatom-2	86.09	0.8459	**83.07**	**0.9083**

E.6 MP20 MATERIAL STRUCTURE PREDICTION

We also benchmark Zatom-2 on periodic material structure prediction using the MP20 dataset. We compare against Crystalite (Hadži Veljković et al., 2026) and report Pymatgen's `StructureMatcher` match rate (MR) and normalized RMS displacement (RMSE) (Ong et al., 2013) across matched structures.

In this setting, Zatom-2 was conditioned on chemical composition, and we used the model to infer lattice geometry (lengths and angles) and fractional coordinates. To calculate match rates, we use PyMatGen's `StructureMatcher` with $(\text{ltol}, \text{stol}, \text{angle_tol}) = (0.3, 0.5, 10^\circ)$ and volume rescaling enabled (Ong et al., 2013). As shown in Table 15, without hyperparameter tuning, Zatom-2 approaches Crystalite's match rate (64.00% vs. 66.09%), but its higher RMSE (0.1156 vs. 0.0337) indicates lower geometric accuracy among matched structures.

Table 15: **MP20 structure prediction evaluation.** We report the Crystalite result from Hadži Veljković et al. (2026).

Model	MR (%) ↑	RMSE ↓
Crystalite	**66.09**	**0.0337**
Zatom-2	64.00	0.1156

F VISUALIZATION

In this section, we visualize Zatom-2's generated samples using MolStar (Mol*, version 5.11.0) (Sehnal et al., 2021). Figures 8–12 illustrate molecular geometries, periodic structures, and protein backbones. Appendix F.4 additionally examines the model's latent representations of reference OMol25 and OMat24 validation examples.

F.1 GENERATED MOLECULES

QM9. Figure 8 shows QM9 samples from the Zatom-2 checkpoint trained jointly on QM9 and MP20, corresponding to the results in Appendix E.4.1. We select the first nine distinct samples that yield connected, sanitizable neutral RDKit graphs from their saved coordinates. This selection filter is used only for this illustration.

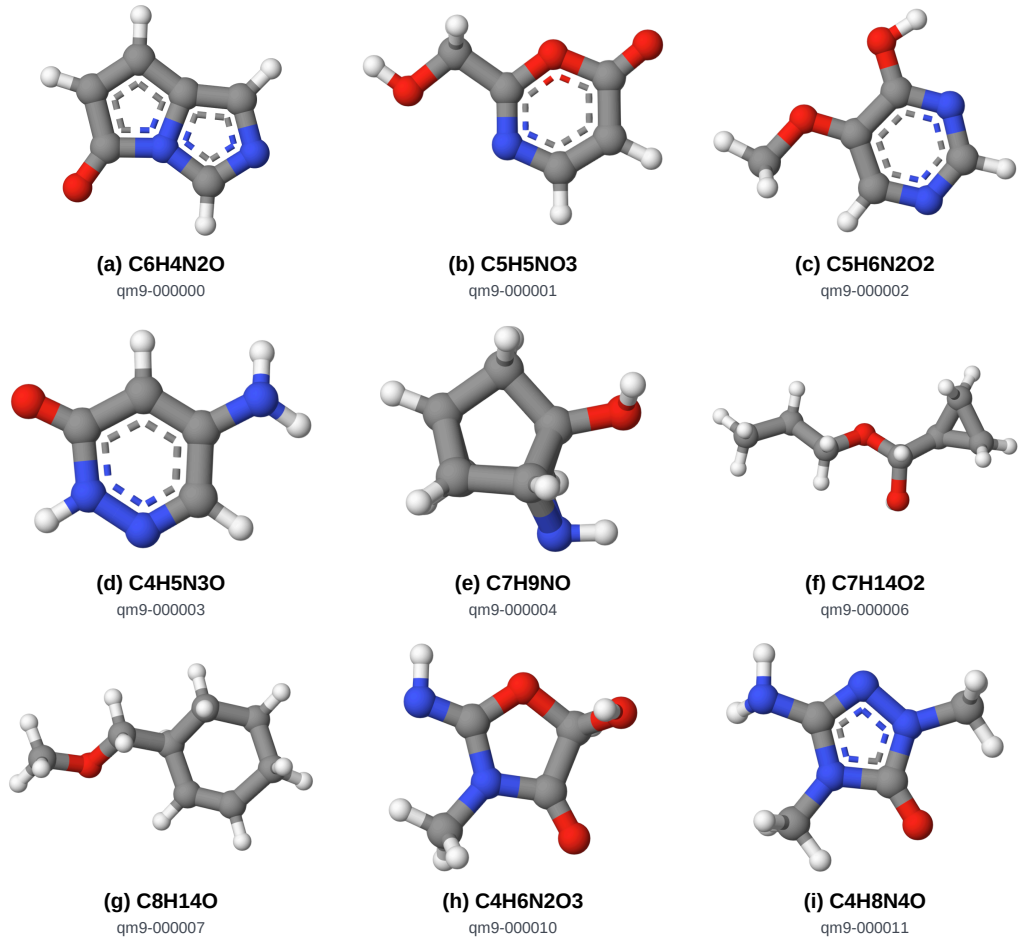


Figure 8: **Generated QM9 molecules.** MolStar ball-and-stick views of nine Zatom-2 molecules. Carbon is gray, hydrogen white, nitrogen blue, and oxygen red. Bonds are perceived by MolStar from the saved structures (note that coordinates are unchanged).

OMol25. Figure 9 shows *force-conditioned* samples from the Zatom-2 model pretrained via joint OMol25/OMat24 generation + structure + MLIP tasks. Here, we stratify by the OMol25 subset supplying each sample's atom count, charge, spin, and mean atomic force norm conditioning value (Levine et al., 2025). The twelve generated samples below cover all ten OMol25 subsets in a 1,000-sample batch of molecules drawn from Zatom-2. Original generated atom identities and coordinates are illustrated, without sanitization or post hoc optimization.

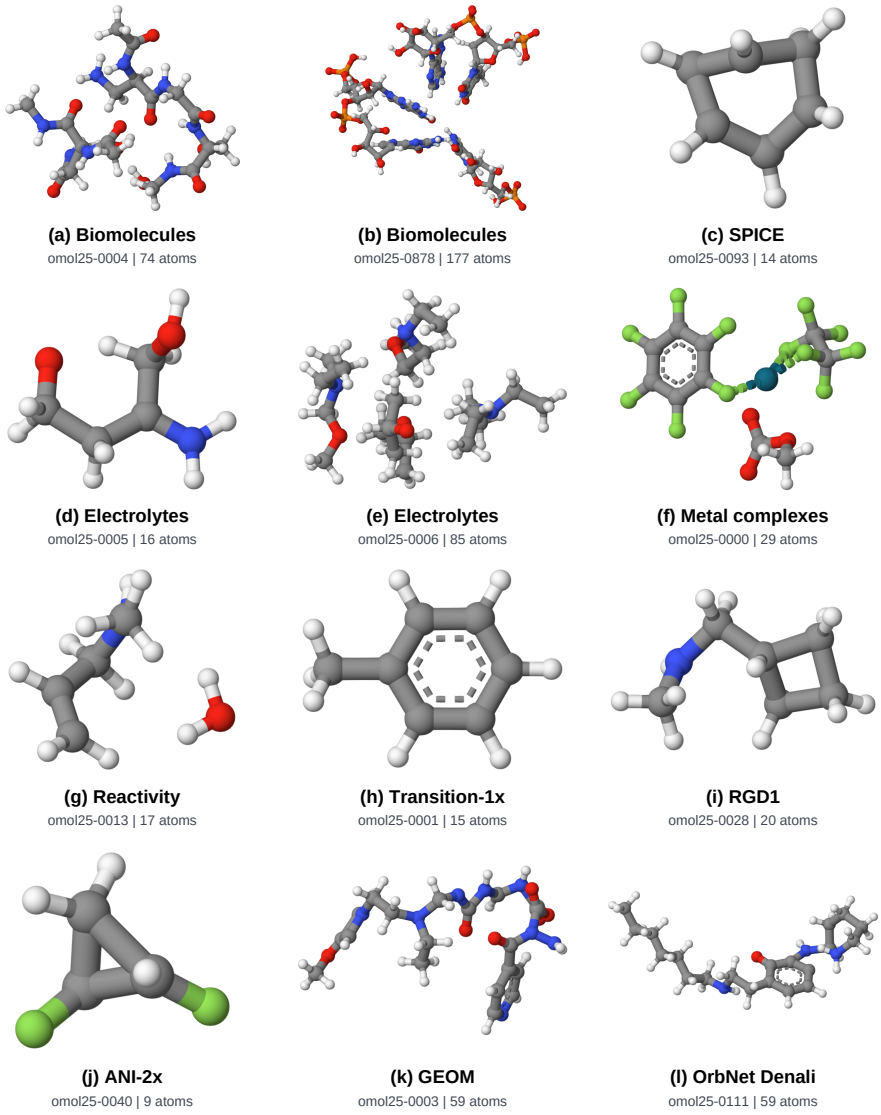


Figure 9: **Generated OMol25 molecules.** MolStar ball-and-stick views spanning biomolecules, electrolytes, metal complexes, reactivity, and community datasets (labels identify the source subset). Notably, Zatom-2's source category and chemical composition for each sample are not fixed during generation but rather are (re)produced. Note that Transition-1x and RGD1 supply metadata for known reaction path examples, but Zatom-2's generated samples for these subsets are not verified transition states. Atoms are colored by element type, and bonds are perceived by MolStar.

F.2 GENERATED MATERIALS

MP20. Figure 10 uses the same joint QM9 and MP20 Zatom-2 checkpoint as adopted in Appendix F.1. Here, we select the first nine distinct unit cell compositions considered valid according to LeMat-GenBench's evaluation protocol.

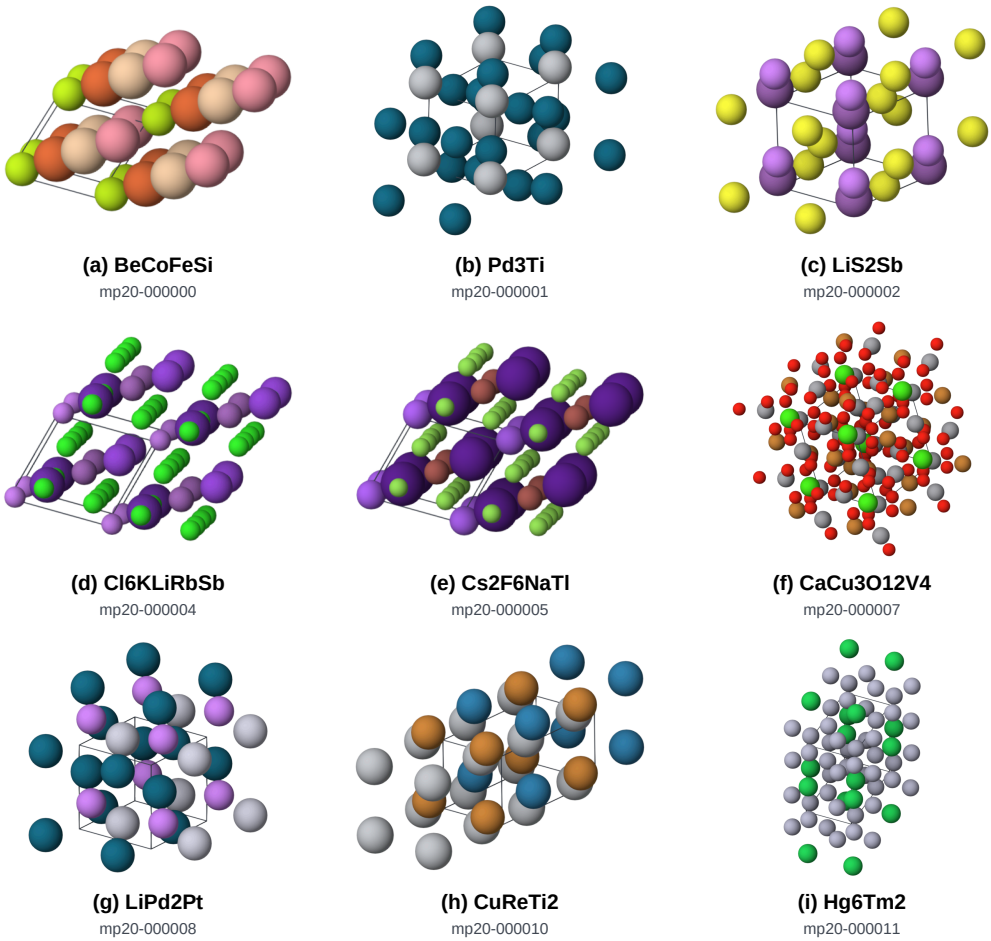


Figure 10: **Generated MP20 materials.** MolStar views of unrelaxed Zatom-2 materials, with atoms colored by element. Each panel displays a 2×2×2 periodic repetition of the saved cell; a gray outline marks one unit cell. Labels report the composition of a single saved cell (not a reduced formula) and its sample ID.

OMat24. Figure 11 shows *force-conditioned* generated materials from the same joint OMol25/OMat24 model as Figure 9. For the sake of legibility, we select the first nine distinct unit cell compositions with at most 40 atoms. Unlike the MP20 gallery in Figure 10, these examples are not filtered by LeMat-GenBench's validity (or MetaSUN) criteria.

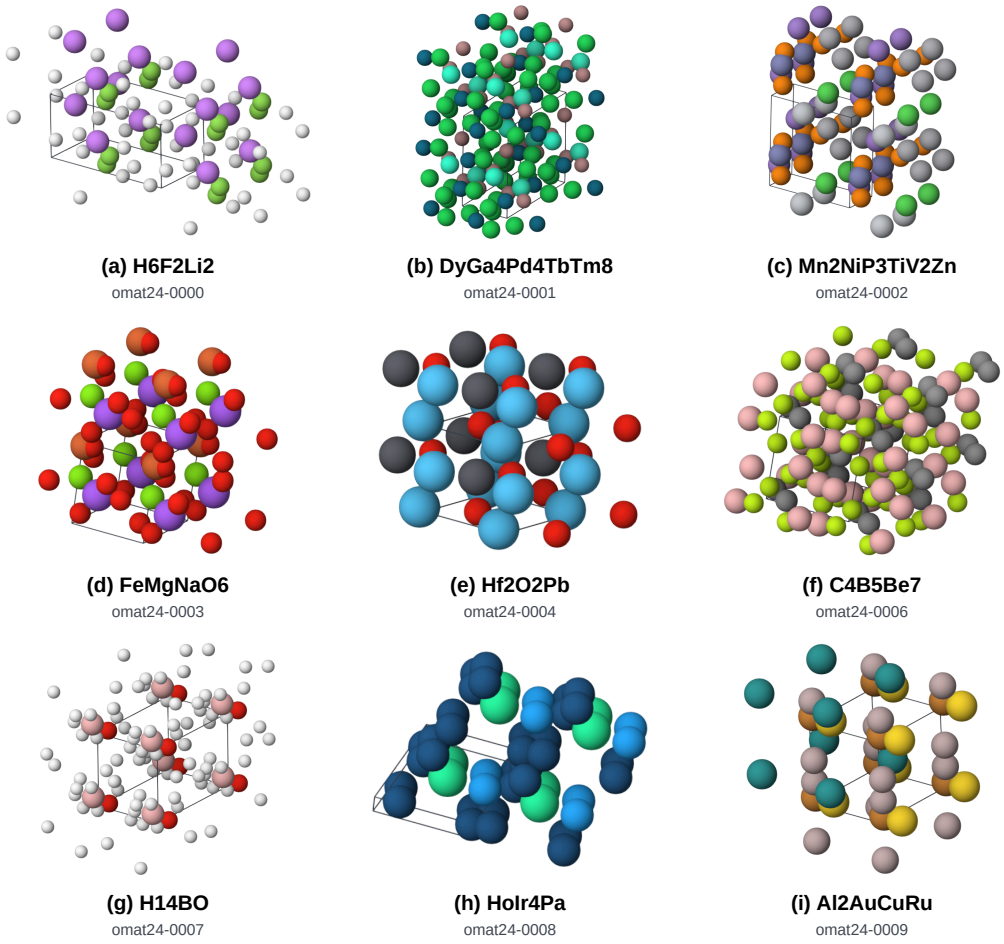


Figure 11: **Generated OMat24 materials.** MolStar views of nine force-conditioned Zatom-2 samples, with atoms colored by element. Each panel displays a $2 \times 2 \times 2$ periodic repetition of the generated cell, with one cell outlined in gray. Labels give the composition of the saved cell and an abbreviated sample ID. No relaxation, space group refinement, or symmetrization is applied.

F.3 GENERATED PROTEINS

SCOPe. Figure 12 shows generated backbones from the joint generation + structure + MLIP pretrained Zatom-2 model (with force conditioning) after 30,000 SCOPe-2k finetuning steps (see Section 4.4 of the main text). Samples come from Section 4.4's 1,027-backbone evaluation with 13 samples at each length from 50 to 128 residues. We select designable examples near lengths 60, 90, and 120 in three broad secondary-structure categories: helix-rich, mixed, and sheet-rich. Designability here uses the samples' precomputed best-of-eight Cα scRMSD threshold of 1.5 Å (Appendix D.9).

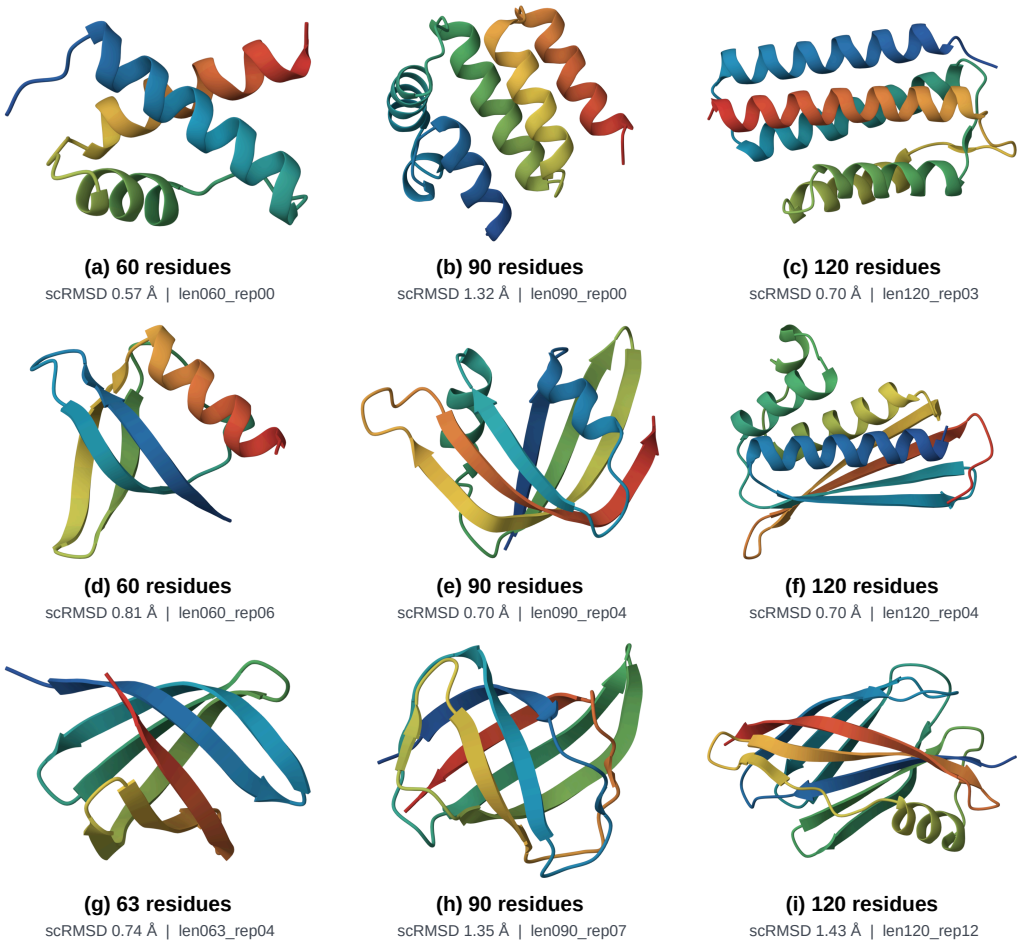


Figure 12: **Generated SCOPe-2k protein backbones.** MolStar cartoons colored from the N-terminus (blue) to the C-terminus (red). Rows show helix-rich, mixed, and sheet-rich examples, respectively. Labels display backbone length, best-of-eight Cα scRMSD in Å, and abbreviated sample IDs. The coordinates rendered for each sample are the generated backbones, not their ESMFold2 refolded structures.

F.4 LATENT REPRESENTATIONS OF OMOL25 AND OMAT24

System representations. To visualize how Zatom-2 organizes its molecular and material latent space with and without energy/force prediction during pretraining, we extract embeddings from the base-sized models, Zatom-2 (IV) and (V) of Table 1 in the main text, using their EMA weights and the same reference OMol25 and OMat24 validation examples. Notably, we evaluate both models under the clean-input conditions used for MLIP prediction, where we provide a clean flow endpoint $t = 1$ with known atom identities, Euclidean coordinates, and ground-truth charge/spin metadata and run one model forward pass. Here, force conditioning is disabled. We extract the final backbone representations before the task heads, allowing the same procedure to be used for model (IV), which has no energy/force prediction heads. Each embedding data point then averages the final 128-dimensional atom representations $\mathbf{q}_{s,i}$ (n.b., Q_L elsewhere in the text) over all N_s atoms of example s, including hydrogens:

$$\mathbf{h}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{q}_{s,i}. \tag{10}$$

Regarding data composition, we sample examples from these validation datasets without replacement, allowing 2–512 atoms total: 128 examples per OMol25 subset and 64 per OMat24 subset, totaling 768 molecules and 704 materials. Note that the "neutral organics" subset here pools the `ani2x`, `orbnet_denali`, and `geom_orca6` subsets into one subset for the sake of visualization. Importantly, both models receive identical examples and preprocessing. Figure 13 uses separate centered principal component analysis (PCA) fits for each of its panels, applying equal weight to each example as well as no channel standardization or vector normalization.

For joint PCA projections in panels (a,b) of Figure 13, most molecules and materials are separated visually in both models' latent spaces, while Zatom-2 *with* energy/force prediction notably shrinks the distance between molecules and materials and even partially overlaps a subset of these inputs. To better understand this observation, in panels (c,d) OMol25 subsets occupy partially distinct regions, whereas in panels (e,f) OMat24 subsets overlap substantially (n.b., these material subsets describe different sampling procedures rather than distinct material types). Comparing the two columns, energy/force prediction pretraining is associated with a greater concentration of variance in the first principal component: 95.4% versus 51.2% for OMol25 and 86.1% versus 44.3% for OMat24, with versus without energy/force prediction. In the next section, we investigate this phenomenon in more atomistic detail.

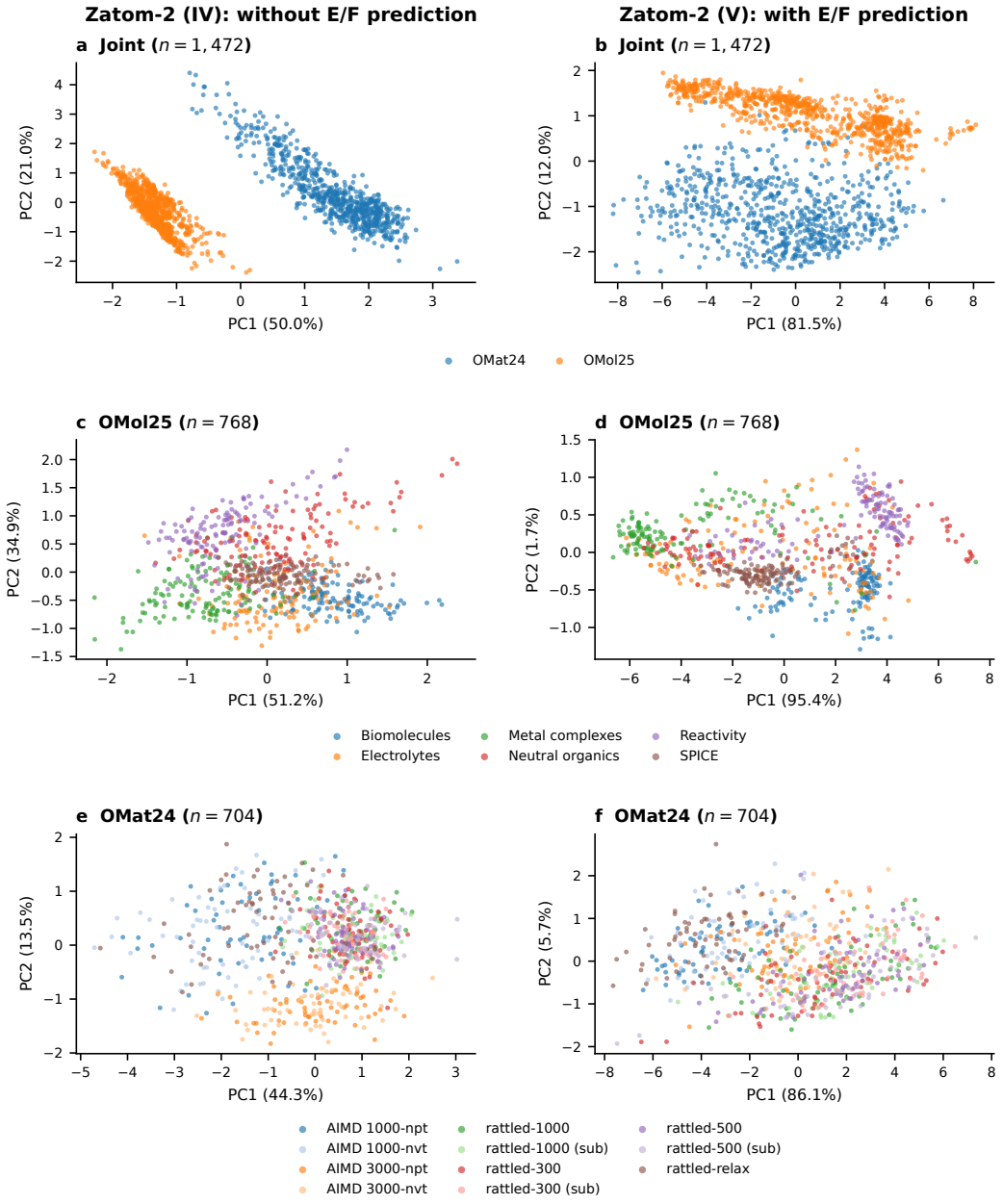


Figure 13: **Base Zatom-2 system embeddings across OMol25 and OMat24.** Left: Zatom-2 (IV), without energy/force prediction; right: Zatom-2 (V), with it. Each point represents the same validation example in both columns, mean-pooled over all atoms. Rows show results for the joint datasets (a,b), OMol25 subsets (c,d), and OMat24 subsets (e,f), with matching colors across checkpoints. Each panel has an independent PCA fit; axis percentages report variance. Note that axes are not aligned between models and that "(sub)" denotes subsampled OMat24 data.

Atom representations. To examine how energy/force prediction pretraining affects the organization of individual atom representations, we visualize the final backbone vectors $\mathbf{q}_{s,i}$ separately for each element, using the same models, validation examples, and MLIP clean-input conditions as above. Figures 14–16 show the per-element embeddings for OMol25 and OMat24 jointly, for OMol25 subsets, and for OMat24 subsets, respectively. Each comparison uses identical atoms across models and independent PCA fits, with element types selected by how common they are in the validation examples. Separating elements allows us to examine atomistic embedding organization while retaining variation that is averaged out in the system (e.g., per-molecule) representations.

In the joint dataset projections of Figure 14, molecular and material atom representations partially overlap, with the degree of overlap varying by element type and pretraining configuration. Energy/force prediction pretraining concentrates more variance in the first principal component, as observed for the system representations in Figure 14. However, most molecular and material atoms remain largely distinguishable in the full representation space with energy/force supervision, indicating that energy/force supervision may promote less obfuscation of atoms from different atomistic domains.

The clearest pretraining effect appears within OMol25 (Figure 15), where energy/force prediction pretraining produces more distinct subset-associated regions for H, C, N, and O. This organization also extends beyond the PCA projections: in the original 128-dimensional latent space, the fraction of ten nearest neighbors sharing an atom's subset label increases by 12.6–14.9% for these elements, considering only same-element neighbors from other validation examples and giving each query example equal weight.

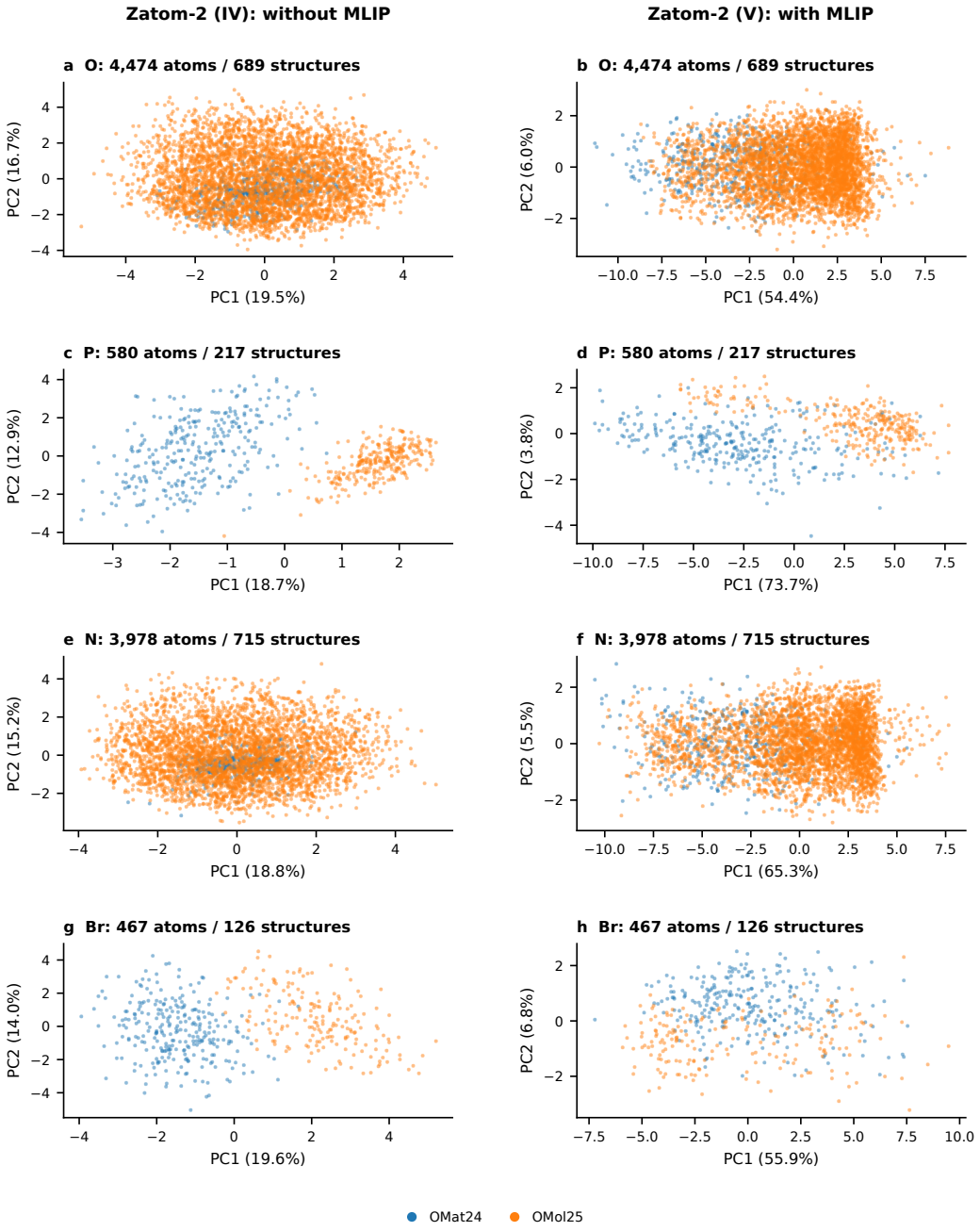


Figure 14: **Base Zatom-2 atom embeddings across OMol25 and OMat24.** Atom-level counterpart of Figure 13(a,b), separately for O, P, N, and Br, selected by shared structure coverage. Left: Zatom-2 (IV), without MLIP supervision; right: Zatom-2 (V), with it. Each point is one atom, colored by domain, with identical atoms in both columns. Titles give sampled atom and structure counts. Each panel fits centered PCA independently; axis percentages report variance, and axes are not aligned between models.

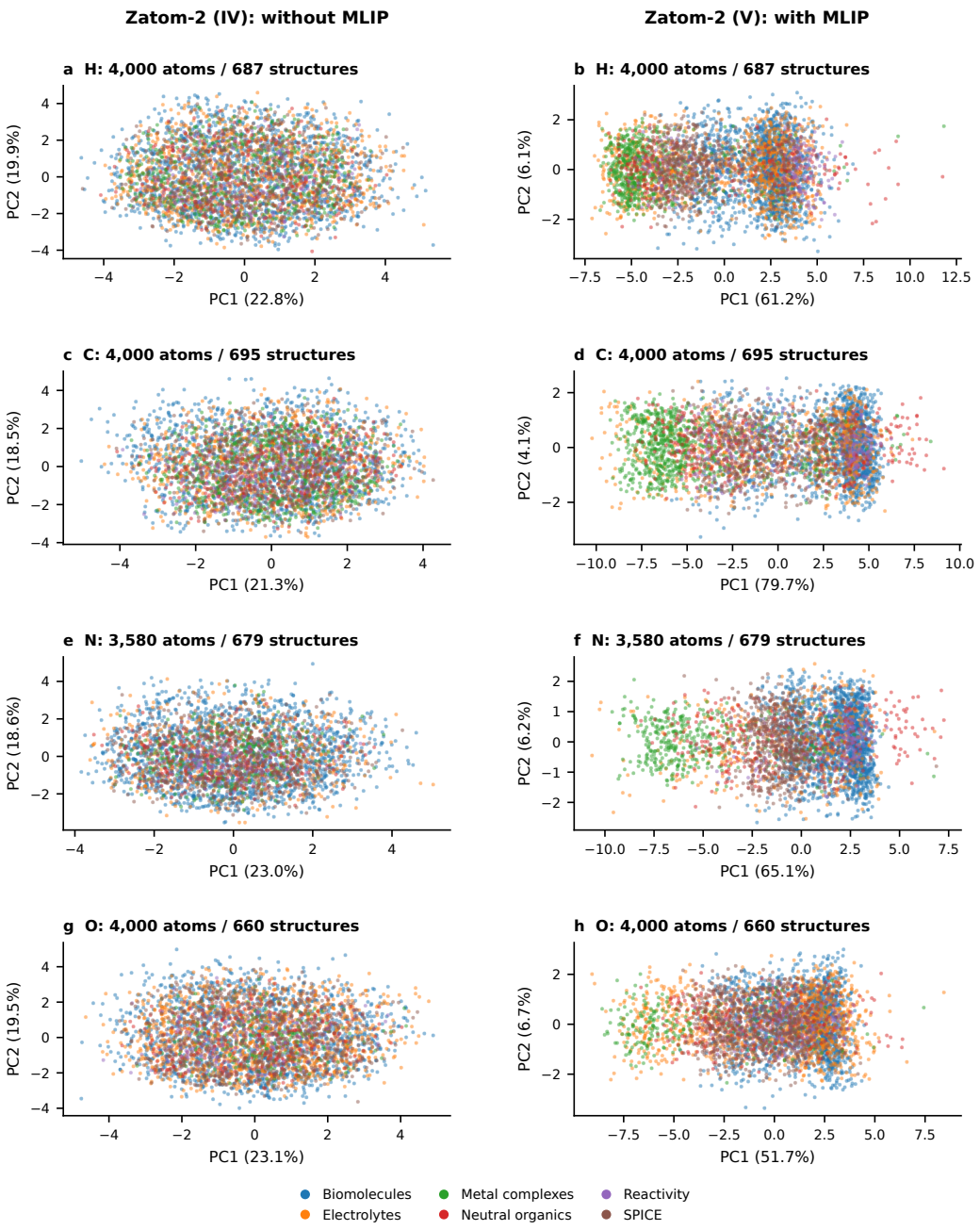


Figure 15: **Base Zatom-2 atom embeddings within OMol25.** Atom-level counterpart of Figure 13(c,d) for H, C, N, and O, the four elements present in the most sampled molecules. Left: without MLIP supervision; right: with it. Colors indicate the same OMol25 subsets as in Figure 13. Each point is an individual atom, with an independent PCA fit per panel.

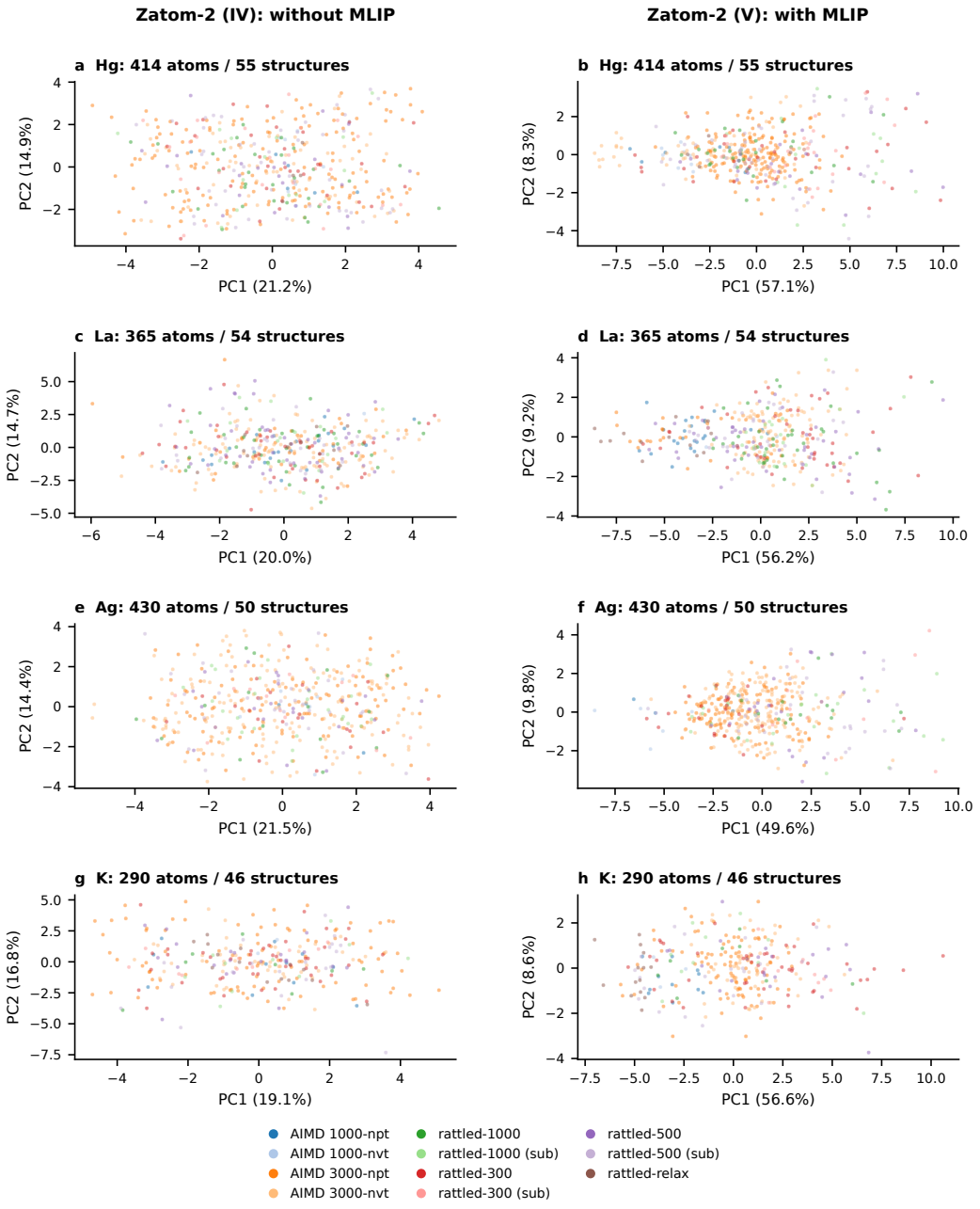


Figure 16: **Base Zatom-2 atom embeddings within OMat24.** Atom-level counterpart of Figure 13(e,f) for Hg, La, Ag, and K, selected by structure coverage in the sampled validation materials. Left: without MLIP supervision; right: with it. Each point is an atom, colored by the corresponding OMat24 subset; "(sub)" denotes a subsampled subset. Identical atoms appear in both columns, with independent PCA fits. MLIP supervision concentrates more variance in PC1.

G Broader Impacts & Limitations

G.1 Broader impacts

Shared atomistic representations could accelerate the discovery of medicines, catalysts, and energy materials and reduce the need to train separate models for each task. These benefits must be weighed against the energy cost of training and oracle evaluation, unequal access to computing resources, and biases in the chemical and structural coverage of the training data. Generative chemistry and protein design capabilities also carry dual-use risks, including the design of harmful compounds or biological agents. Responsible use calls for application-specific safety assessment, appropriate screening of proposed designs, and experimental validation before deployment.

G.2 Limitations

This work limits training to the 4M and 1M subsets of OMol25 and OMat24, respectively. The full datasets contain 100M examples each, and we leave training on the entire data for future work. Secondly, training with force conditioning enables Zatom-2 to distinguish clearly between low- and high-force regimes, but as the experiments in Section 4.1 highlight, within-regime force conditioning is not well captured by our force conditioning method on all subsets of the data. We leave improvements on force conditioning mechanisms, including conditioning on the per-atom force vector, to future work. Moreover, the UMA-based metric employed for force consistency is an error-prone machine learning method; however, we do wish to highlight that the MAEs between UMA-calculated mean force norms of OMol25 and OMat24 examples and their dataset mean force norm labels are 0.00587 and 0.02462 eV Å^{-1}, respectively (see Appendix D.1). We also note that our protein generation experiments are performed on a small, although diverse, dataset of proteins, and that future experiments can explore transfer to large-scale protein datasets via finetuning. Lastly, while we treat force and energy prediction as auxiliary tasks for generative training and do not intend to match specialized MLIPs' performance, results in Appendix E.2 leave room for improvement, which we leave for future work.