ZATOM-1: A MULTIMODAL FLOW FOUNDATION MODEL FOR 3D MOLECULES AND MATERIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

General-purpose 3D chemical modeling encompasses molecules and materials, requiring both generative and predictive capabilities. However, most existing AI approaches are optimized for a single domain (molecules or materials) and a single task (generation or prediction), which limits representation sharing and transfer. We introduce ZATOM-1, the first foundation model that unifies generative and predictive learning of 3D molecules and materials. ZATOM-1 is a Transformer trained with a multimodal flow matching objective that jointly models discrete atom types and continuous 3D geometries. This approach supports scalable pretraining with predictable gains as model capacity increases, while enabling fast and stable sampling. We use joint generative pretraining as a universal initialization for downstream multi-task prediction of properties, energies, and forces. Empirically, ZATOM-1 matches or outperforms specialized baselines on both generative and predictive benchmarks, while reducing the generative inference time by more than an order of magnitude. Our experiments demonstrate positive predictive transfer between chemical domains from joint generative pretraining: modeling materials during pretraining improves molecular property prediction accuracy.

1 INTRODUCTION

Life, materials, and many modern functions, ranging from small molecule therapeutics to consumer electronic devices, are driven by chemistry. However, navigating the design space of chemical systems for high-impact downstream applications is notoriously time-consuming and error-prone (Blanco-Gonzalez et al., 2023; Venkatasubramanian, 2019; Wang et al., 2023). As such, developing our understanding of chemical systems with artificial intelligence (AI) methods represents a promising area of investment for deep learning research, given that these computational techniques have led to significant breakthroughs across related scientific fields in recent years (Jumper et al., 2021; Abramson et al., 2024).

Recently, deep learning has begun to unify the generation of new 3D molecules and materials (Joshi et al., 2025), and several specialized prediction models currently exist for advanced molecule and material property prediction (Liao & Smidt, 2023; Qu & Krishnapriyan, 2024). Nonetheless, no prior works offer a unified model capable of both generative modeling and representation learning of 3D molecules and materials. To build such a unified model, several design choices must be addressed, including how to tokenize its inputs as well as what and how to learn using these tokens.

Inspired by previous research on scientific foundation models (Subramanian et al., 2023; Abramson et al., 2024), we argue for the importance of tokenizing atomistic systems at the fundamental level of atomic granularity: 3D atom coordinates and element types. The next question is what to learn with these tokens. Previous works have shown that image and video diffusion models, employing standardized deep learning architectures, can learn rich embeddings for downstream prediction tasks in a self-supervised manner (Li et al., 2023; Vélez et al., 2025). In addition, scientific machine learning (SciML) is increasingly adopting standardized deep learning architectures to train highly scalable models (Abramson et al., 2024; Kreiman et al., 2025; Wadell et al., 2025). As such, it is clear that generative modeling, due to its self-supervised (i.e., label-free) nature, has the potential to serve as a powerful and flexible pretraining method for atomic foundation models.

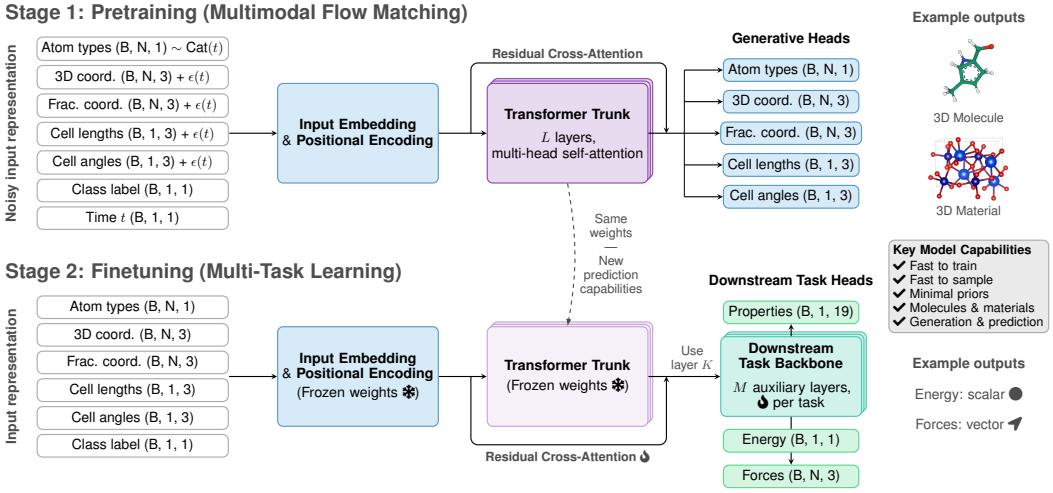


Figure 1: **Architecture of the ZATOM-1 foundation model for 3D molecules and materials.** A unified Transformer trunk processes (noisy) atomic, structural, and conditional inputs. The model features two training stages: (1) multimodal flow pretraining using the final trunk layer L's representations to create new molecules or materials; and (2) multi-task finetuning of additional downstream task backbones and heads that use a specified trunk layer K's representations to predict properties, energies, and forces.

In this spirit, we propose a question: *How much chemical knowledge can we represent with a 3D generative model?*

ZATOM-1, our answer to this question, is illustrated in Figure 1. It is a unified flow matching foundation model for 3D molecules and materials, based on the following ideas:

1. *Ambient all-atom generative modeling.* We represent 3D molecules and materials in $\mathbb{R}^3$ and perform generative flow matching directly within this Euclidean space. Importantly, with this approach, no pretrained autoencoders are necessary, e.g., for latent diffusion, delivering notable speed-ups in model training and inference.
2. *Representation learning using generative models.* We frame generative modeling of 3D molecules and materials as an ideal pretraining task for chemical representation learning (Briq et al., 2025) using modern, standardized Transformers (Vaswani et al., 2017).

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

- We introduce ZATOM-1, a first-of-its-kind foundation model for diverse chemical tasks, including *de novo* generation of periodic materials and non-periodic 3D molecules, multi-task (and for materials zero-shot) property prediction, and energy and force modeling.
- ZATOM-1: (**1**) achieves competitive results for 3D materials generation on the LeMat-GenBench benchmark using the MP20 materials dataset; (**2**) exceeds state-of-the-art for 3D molecule generation on the QM9 and GEOM-Drugs molecular datasets and achieves state-of-the-art multi-task QM9 property prediction; (**3**) provides a 3x reduction in GPU training hours compared to a latent diffusion baseline; and (**4**) demonstrates broad, positive transfer learning through cross-domain generative modeling.
- ZATOM-1's architecture features a standardized Transformer with Query-Key Normalization (Wortsman et al., 2024), Flash Attention (Dao, 2024), and SwiGLU feed-forward networks (Shazeer, 2020), yielding predictably improving performance through parameter scaling. This also allows it to deliver a **12.5x** speed-up over a 500M parameter latent diffusion baseline, where ZATOM-1 (300M) can generate 10,000 samples in under 4 minutes with a single NVIDIA A100 GPU.

Overall, this work represents an advancement in chemical representation learning by introducing the first broadly applicable foundation model for 3D molecules and materials.

2 ZATOM-1

ZATOM-1's architecture is trained in two stages: (1) generative pretraining for 3D molecule and material generation; and (2) predictive finetuning for energy, force, and property prediction tasks. For generative pretraining, we use conditional flow matching, a scalable training technique for normalizing flows (Lipman et al., 2023). As a non-autoregressive generative modeling paradigm, conditional flow matching trains a neural network to approximate a time-conditioned velocity field required to transport samples from a source distribution to a target distribution using ordinary differential equation (ODE) integration. Importantly, time integration is only required during sampling, not model training, offering scalable training speeds as a result.

To efficiently model the multimodal nature of 3D atomic systems, we use multimodal flow matching over a (discrete) atom type modality and four (continuous) geometric modalities, yielding a joint distribution of these modalities for generative modeling. Previous works for 3D materials generation have attempted multimodal flow matching over similar modality structures (Miller et al., 2024). However, we argue that reliance on sparse graph neural network architectures, as well as hand-crafted generative priors, has hindered performance and training/inference speed for generative modeling in this domain (Krishnapriyan et al., 2021; Qu & Krishnapriyan, 2024). Instead, we adopt a standardized Transformer architecture and Gaussian multimodal flow matching for scalable joint modeling of 3D molecules and materials. We discuss related works in Appendix A.

2.1 PROBLEM FORMULATION

We aim to learn a joint generative model of 3D molecules and materials and to obtain rich neural embeddings of both data domains. Towards this end, we represent a 3D molecule or material with N atoms using five unified modalities:

$$\begin{aligned}
\text{Atom types } \boldsymbol{A} = \{a_i\}_{i=1}^N &\in \mathbb{Z}^{1\times N}, \\
\text{3D coordinates } \boldsymbol{X} = \{x_i\}_{i=1}^N &\in \mathbb{R}^{3\times N}, \\
\text{Fractional coordinates } \boldsymbol{F} = \{f_i\}_{i=1}^N &\in [0,1)^{3\times N}, \\
\text{Unit cell/lattice lengths } \boldsymbol{L}_{\text{len}} = \{l_{\text{len}}\} &\in \mathbb{R}^{3\times 1}, \\
\text{Unit cell/lattice angles } \boldsymbol{L}_{\text{ang}} = \{l_{\text{ang}}\} &\in \mathbb{R}^{3\times 1}.
\end{aligned}$$

Atom types $\boldsymbol{A}$ are a discrete modality defined over the set of integers $\mathbb{Z}$, while 3D coordinates $\boldsymbol{X}$ are a continuous modality of $\mathbb{R}^3$ expressed in Angstroms. Fractional coordinates $\boldsymbol{F} = \boldsymbol{L}^{-1}\boldsymbol{X}$ reside in the range $[0, 1)$, and rotation-invariant lattice side lengths $\{l_{\text{len}}\} = \{a, b, c\} \in \mathbb{R}^{3\times 1}$ and the three internal angles between them $\{l_{\text{ang}}\} = \{\alpha, \beta, \gamma\} \in [60^\circ, 120^\circ]^{3\times 1}$ parametrize a parallelepiped $\boldsymbol{L} \in \mathbb{R}^{3\times 3}$ (Grosse-Kunstleve et al., 2004) characterizing the repeating unit cell (Miller et al., 2024). For the sake of numerical stability, during training and sampling, lattice lengths are normalized by $\sqrt[3]{N}$, and lattice angles are converted from degrees to radians. Further, during training and sampling, for non-periodic molecules the lattice lengths, angles, and fractional coordinate inputs are masked to null values $\varnothing$. For periodic materials 3D coordinate inputs are masked to null values $\varnothing$ to enforce materials rotation invariance by default (with training losses masked accordingly).

2.2 NETWORK ARCHITECTURE

We use a standard Transformer architecture, titled the **Trunk-based Flow Transformer (TFT)**, to learn a shared latent representation of each molecular and material modality. For a given 3D atomic system $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{F}, \boldsymbol{L}_{\text{len}}, \boldsymbol{L}_{\text{ang}})$ with a periodic (material) or non-periodic (molecule) class label c and (optional) additive noise, a TFT model $\mathcal{T}$ with L Transformer trunk layers learns latent representations $\boldsymbol{Z} = \{z_i\}_{i=1}^N \in \mathbb{R}^{d\times N}$ from each molecular and material modality to predict denoised modalities $\boldsymbol{S} = \left(\boldsymbol{A}', \boldsymbol{X}', \boldsymbol{F}', \boldsymbol{L}'_{\text{len}}, \boldsymbol{L}'_{\text{ang}}\right)$ and auxiliary properties $\boldsymbol{P} = \left(\boldsymbol{P}^{\text{props'}}, \boldsymbol{P}^{\text{energy'}}, \boldsymbol{P}^{\text{forces'}}\right) = \left(p^{\text{props'}} \in \mathbb{R}^{19},\ p^{\text{energy'}} \in \mathbb{R}^1,\ \{p_i^{\text{forces'}}\}_{i=1}^N \in \mathbb{R}^{3\times N}\right)$:

$$\boldsymbol{S}, \boldsymbol{P} \ = \ \mathcal{T}\left(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{F}, \boldsymbol{L}_{\text{len}}, \boldsymbol{L}_{\text{ang}}, c\right), \tag{1}$$

where class label dropout with a 10% drop probability is applied during training to support (optional) classifier-free guidance during inference.

Algorithm 1 Pseudocode for TFT model $\mathcal{T}$

Input: Noisy 3D atoms $(\{a_i\}, \{x_i\}, \{f_i\}, l_{\text{len}}, l_{\text{ang}})$, class c, time t
Output: Denoised 3D atoms $(\{a'_i\}, \{x'_i\}, \{f'_i\}, l'_{\text{len}}, l'_{\text{ang}})$, predictions $\{p^{\text{props}'}, p^{\text{energy}'}, p_i^{\text{forces}'}\}$

Lin^{B}, Lin^{NB} = Linear(bias=True), Linear(bias=False)
Mask inputs and project to d_{model}
$x_i,\ (f_i, l_{\text{len}}, l_{\text{ang}}) = x_i \cdot \mathbb{I}_{\text{molecule}},\ (f_i, l_{\text{len}}, l_{\text{ang}}) \cdot \mathbb{I}_{\text{material}}$
$h_i = \text{Embed}(a_i) + \text{Embed}(c) + \text{Embed}(t) + \sum_{v \in \{x_i, f_i, l_{\text{len}}, l_{\text{ang}}\}} \text{Lin}_v^{\text{NB}}(v)$ $\quad h_i \in \mathbb{R}^{d_{\text{model}}}$
Apply encoder trunk (L layer transformer)
$\{z_i, z_i^{\text{K}}\} = \text{TransformerEncoder}^{L}(\{h_i\})$
Apply residual cross-attention to input embeddings
$\{h_i^a, h_i^x, h_i^f, h_i^{l_{\text{len}}}, h_i^{l_{\text{ang}}}\} = \text{TransformerDecoderBlock}(\{z_i, h_i\})$
Denoise inputs
$a'_i = \arg\max(\text{Lin}^{\text{B}}(\text{SiLU}(\text{Lin}^{\text{B}}(h_i^a))))$ $\quad a'_i \in \mathbb{Z}$
$x'_i = \text{Lin}^{\text{NB}}(\text{LayerNorm}(h_i^x))$ $\quad x'_i \in \mathbb{R}^3$
$f'_i = \text{Lin}^{\text{NB}}(\text{LayerNorm}(h_i^f))$ $\quad f'_i \in \mathbb{R}^3$
$l'_{\text{len}} = \text{Lin}^{\text{NB}}\left(\text{LayerNorm}\left(\frac{1}{N}\sum_{i=1}^{N} h_i^{l_{\text{len}}}\right)\right)$ $\quad l' \in \mathbb{R}^3$
$l'_{\text{ang}} = \text{Lin}^{\text{NB}}\left(\text{LayerNorm}\left(\frac{1}{N}\sum_{i=1}^{N} h_i^{l_{\text{ang}}}\right)\right)$ $\quad l' \in \mathbb{R}^3$
Predict outputs (cross-attention + M layer transformer)
$\{h_i^{\text{props}}, h_i^{\text{energy}}, h_i^{\text{forces}}\} = \text{TransformerDecoderBlock}_y(\{z_i^{\text{K}}, h_i\})$, for $y \in \{\text{props}, \text{energy}, \text{forces}\}$
Update $\{h_i^y\} \leftarrow \text{TransformerEncoder}_y^{M}(\{h_i^y\})$
$p^{\text{props}'} = \text{Lin}^{\text{B}}\left(\frac{1}{N}\sum_{i=1}^{N} h_i^{\text{props}}\right)$ $\quad p \in \mathbb{R}^{19}$
$p^{\text{energy}'} = \text{Lin}^{\text{B}}\left(\frac{1}{N}\sum_{i=1}^{N} h_i^{\text{energy}}\right)$ $\quad p \in \mathbb{R}^{1}$
$p_i^{\text{forces}'} = \text{Lin}^{\text{NB}}\left(h_i^{\text{forces}}\right)$ $\quad p \in \mathbb{R}^{3}$

Algorithm 2 Pseudocode for multimodal sampling

Input: Denoising model $\mathcal{T}$, number of atoms N, sampling steps $\{t_i\}_{i=0}^{S}$
Output: Sampled 3D atoms $(\mathbf{a}, \mathbf{x}, \mathbf{f}, \mathbf{l}_{\text{len}}, \mathbf{l}_{\text{ang}})$

def DiscreteFlowStep($\mathbf{a}_t, \mathbf{p}'_{t,a}, t, \Delta t$):
  # NOTE: All indexed ops are vectorized
  $\mathbf{r}_t(i, \cdot) \leftarrow \frac{\Delta t}{1-t}\mathbf{p}'_{t,a}(i)$ (Campbell et al., 2024)
  $\mathbf{r}_t(i, \mathbf{a}_t(i)) \leftarrow -\sum_{j \neq \mathbf{a}_t(i)} \mathbf{r}_t(i, j)$
  $\mathbf{p}_{t+\Delta t}(i, j) \leftarrow 1_{\mathbf{a}_t(i)=j} + \mathbf{r}_t(i, j)$
  $\mathbf{a}_t(i) \leftarrow \text{Categorical}(\mathbf{p}_{t+\Delta t}(i, \cdot))$
  return $\mathbf{a}_t$

def EuclideanStep($\mathbf{z}_t, \mathbf{z}'_t, t, \Delta t, g(\cdot), \gamma_g$):
  $\mathbf{v}_t \leftarrow \frac{1}{1-t}(\mathbf{z}'_t - \mathbf{z}_t)$ (Geffner et al., 2025)
  $\mathbf{s}_t \leftarrow g(t)\frac{t\mathbf{v}_t - \mathbf{z}_t}{1-t}$
  $d\mathbf{W}_t \leftarrow \sqrt{2\,\gamma_g\, g(t)}\,\mathcal{N}(0, I)$
  $\mathbf{z}_t \leftarrow (\mathbf{v}_t + \mathbf{s}_t + d\mathbf{W}_t)\Delta t$
  return $\mathbf{z}_t$

Initialize all modalities with random noise
$\mathbf{a}_t \leftarrow \text{Categorical}(\delta(\frac{1}{\#\text{ atom types}}))$
Initialize $\mathbf{z}_t \leftarrow \mathcal{N}(0, I)$ for each continuous modality $\mathbf{z} \in \{\mathbf{x}, \mathbf{f}, \mathbf{l}_{\text{len}}, \mathbf{l}_{\text{ang}}\}$
# Iteratively denoise over sampling steps
for $i = 1$ to S **do**
  $\Delta t \leftarrow t_i - t_{i-1}$
  # Predict denoised endpoints using the model $\mathcal{T}$
  $(\mathbf{p}'_{t,a}, \mathbf{x}'_t, \mathbf{f}'_t, \mathbf{l}'_{t,\text{len}}, \mathbf{l}'_{t,\text{ang}}) \leftarrow \mathcal{T}(\mathbf{a}_t, \mathbf{x}_t, \mathbf{f}_t, \mathbf{l}_{t,\text{len}}, \mathbf{l}_{t,\text{ang}}, t_i)$
  # Perform one sampling step for each modality
  $\mathbf{a}_t \leftarrow$ DiscreteFlowStep($\mathbf{a}_t, \mathbf{p}'_{t,a}, t_i, \Delta t$)
  Update $\mathbf{z}_t \leftarrow$ EuclideanStep($\mathbf{z}_t, \mathbf{z}'_t, t_i, \Delta t$) for each $\mathbf{z}$
end for
# Return final denoised sample
return $(\mathbf{a}, \mathbf{x}, \mathbf{f}, \mathbf{l}_{\text{len}}, \mathbf{l}_{\text{ang}})$

Algorithm 1 provides pseudocode for the forward pass of a TFT model with L trunk layers, where layer $K \leq L$ is used for auxiliary task representations. Using random data augmentations during training, the model learns the rotational and periodic symmetries of molecules and materials, respectively (Abramson et al., 2024; Joshi et al., 2025). Additionally, inspired by the Platonic Transformer (Islam et al., 2025), we experiment with an O(3)-equivariant version of ZATOM-1 (PLATOM-1) featuring a TFT model variant titled the **Trunk-based Flow Platoformer (TFP)**, which we describe in more detail in Appendix B.

2.3 STAGE 1: MULTIMODAL FLOW PRETRAINING

At a high level, given a pair of data samples $(\mathbf{x}_0, \mathbf{x}_1)$ from their respective source and target distributions, flow matching (FM) offers a scalable technique for training generative models. Namely, using FM, one can train a neural network $v_\theta(\mathbf{x}_t, t)$ to sample data from the target distribution by matching the expected velocities transporting samples from the (often simpler) source to the target distribution:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)}[\|v_\theta(\mathbf{x}_t, t) - u_t\|_2^2], \tag{2}$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ and $u_t = \mathbf{x}_1 - \mathbf{x}_0$.

In this work, to optimize continuous modalities, we use Euclidean diffusion/conditional flow matching (CFM) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Tong et al., 2024), leveraging the equivalence between these two methods (Gao et al., 2024). For optimizing atom types, we employ Discrete CFM, which is parameterized within the framework of Discrete Flow Models (Campbell et al., 2024).

Specifically, consider a molecule or material comprised of N atoms, characterized by zero-centered ground-truth 3D coordinates $\boldsymbol{X}$, fractional coordinates $\boldsymbol{F}$, lattice lengths $\boldsymbol{L}_{\text{len}}$, angles $\boldsymbol{L}_{\text{ang}}$, and atom types $\boldsymbol{A}$. The continuous 3D coordinates and fractional coordinates, along with lattice lengths and angles, undergo partial noise perturbation represented as $\boldsymbol{X}_t = t \cdot \boldsymbol{X} + (1-t) \cdot \epsilon$, where the noise ϵ is drawn from a Gaussian distribution $\mathcal{N}(0, I)$. For the discrete modality, the noisy atom types $\boldsymbol{A}_t$ are derived by interpolating between atom type probabilities and sampling from a categorical distribution, expressed as $\boldsymbol{A}_t \sim \text{Cat}(t \cdot \delta(\boldsymbol{A}) + (1-t) \cdot \delta(\frac{1}{\#\text{ atom types}}))$, where $\delta(\cdot)$ generates a one-hot encoding

(Campbell et al., 2024). During the training phase, the model receives $\left(\boldsymbol{X}_t, \boldsymbol{F}_t, \boldsymbol{L}_{t_{\text{len}}}, \boldsymbol{L}_{t_{\text{ang}}}, \boldsymbol{A}_t\right)$ as input and learns to predict the terminal points of the trajectory.

Flow matching losses. The objective function for each continuous modality is given by

$$L_{\text{metric}}(\boldsymbol{X}) = \mathbb{E}_{\epsilon,t}\left[\frac{1}{N}\|\boldsymbol{X}' - \boldsymbol{X}\|_2^2\right]. \tag{3}$$

For the discrete atom types, the objective is defined using cross-entropy:

$$L_{\text{discrete}}(\boldsymbol{A}) = \mathbb{E}_t\left[-\sum\nolimits_i a_i \log(a_i')\right]. \tag{4}$$

These objectives are consolidated into a multi-objective framework, employing a weighting factor $\lambda_{\text{discrete}} \in [0, 1]$, expressed as

$$\begin{aligned} L_{\text{total}} = {} & L_{\text{metric}}(\boldsymbol{X}) + L_{\text{metric}}(\boldsymbol{F}) \\ & + L_{\text{metric}}(\boldsymbol{L}_{\text{len}}) + L_{\text{metric}}(\boldsymbol{L}_{\text{ang}}) \\ & + \lambda_{\text{discrete}} \cdot L_{\text{discrete}}(\boldsymbol{A}). \end{aligned} \tag{5}$$

During training, we set $\lambda_{\text{discrete}} = 0.1$ and sample from $t \sim \text{Beta}(\alpha_t, 1)$, where $\alpha_t = 1.8$, in line with prior studies (Geffner et al., 2025; Vonessen et al., 2025). Further, to learn data symmetries and accelerate convergence during training, we randomly rotate and (for periodic materials) translate multiple copies of each input example, which are noised to different levels according to batch-wise t (Abramson et al., 2024). As $t \to 1$, the model resembles an identity function, since we adopt an endpoint formulation for flow matching due to its empirical performance for scientific applications (Stark et al., 2024). To enable the model to effectively learn accurate atom placements even when $t \to 1$ causes the losses to approach zero, based on the sampled time t, we scale the loss with $\beta(t) \cdot L_{\text{total}}(\boldsymbol{X}_t, \boldsymbol{F}_t, \boldsymbol{L}_{t_{\text{len}}}, \boldsymbol{L}_{t_{\text{ang}}}, \boldsymbol{A}_t)$ via:

$$\beta(t) = \min\left\{100,\ \tfrac{1}{(1-t)^2}\right\}. \tag{6}$$

Multimodal sampling. Algorithm 2 describes how we perform unconditional sampling of each continuous and discrete modality to generate 3D molecules and materials. Concisely, we sample random atom types or Gaussian noise for discrete and continuous modalities, respectively, and iteratively denoise each using a series of Euler steps operating on the multimodal predictions of the TFT model $\mathcal{T}$. Notably, other ODE solvers can be used, though we find that multimodal Euler steps provide strong empirical performance without considerable hyperparameter tuning.

2.4 Stage 2: Finetuning

As illustrated in Figure 1, we treat conditional flow matching of the modalities $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{F}, \boldsymbol{L}_{\text{len}}, \boldsymbol{L}_{\text{ang}})$ as a self-supervised pretraining task akin to bidirectional modeling used to pretrain non-autoregressive BERT-style models (Devlin et al., 2019) for enhanced representation learning (Zhang et al., 2025) and *label-free* transfer learning. Once the model weights listed up to Line 10 of Algorithm 1 have been (pre)trained, for finetuning, we freeze these weights and enable training of the M-layer auxiliary Transformer trunk weights following Line 10 using trunk layer K's frozen embeddings. By default, $K = L$, which we ablate in Section 4.

Predictive Losses. For property prediction, we supervise the task's embeddings $p_b^{\text{props'}}$ using the mean absolute error (MAE) over each batch element b as follows:

$$L_{\text{MAE}}(\boldsymbol{P}^{\text{props'}}) = \frac{1}{B}\sum\nolimits_b \left|p_b^{\text{props'}} - p_b^{\text{props}}\right|, \tag{7}$$

where B is the number of samples in a given mini-batch, $p_b^{\text{props'}}$ denotes the predicted property for the b-th sample, and p_b^{props} denotes the true property for the same sample. Similarly, energy embeddings $p_b^{\text{energy'}}$ are supervised using the mean squared error (MSE) over each batch element:

$$L_{\text{MSE}}(\boldsymbol{P}^{\text{energy'}}) = \frac{1}{B}\sum\nolimits_b \left(p_b^{\text{energy'}} - p_b^{\text{energy}}\right)^2. \tag{8}$$

Lastly, force embeddings $p_b^{\text{forces'}}$ are also learned using a batch-wise MSE loss, with the loss weight λ_{forces}:

$$L_{\text{MSE}}(\boldsymbol{P}^{\text{forces'}}) = \mathbb{E}\left[\frac{1}{N}\|\boldsymbol{P}^{\text{forces'}} - \boldsymbol{P}^{\text{forces}}\|_2^2\right]. \tag{9}$$

3 EXPERIMENTAL SETUP

Datasets. Following Joshi et al. (2025), we train models on periodic materials from MP20 and non-periodic molecules from QM9 for our generative pretraining experiments. MP20 (Xie et al., 2022) is a collection of 45,231 metastable crystal structures from the Materials Project (Jain et al., 2013), with up to 20 atoms in each material's unit cell spanning 89 different atom/element types. QM9 (Wu et al., 2018) is comprised of 130,000 stable small molecules with up to 9 heavy atoms (C, N, O, F) along with hydrogens. To ensure fair comparisons, we split the data according to prior work (Xie et al., 2022; Hoogeboom et al., 2022). Additionally, we report generative pretraining results for GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022), a collection of 430,000 large molecules containing up to 180 atoms, following the benchmarking setup of Vignac et al. (2023). Lastly, in Appendix E, we employ the QMOF dataset of 14,000 metal-organic framework (MOF) structures (Rosen et al., 2021) after removing structures containing more than 150 atoms, following Joshi et al. (2025).

For predictive finetuning, we use the QM9 and Matbench datasets, the latter supporting several fundamental property prediction tasks in materials science (Dunn et al., 2020). Further, in Appendix E, we adopt the OMol25 (4M) (Levine et al., 2025) and MPtrj (Deng et al., 2023) datasets for molecule and materials energy and force prediction, respectively. Data splits follow prior work for fair evaluation (Liao & Smidt, 2023; Dunn et al., 2020; Levine et al., 2025; Deng et al., 2023).

Training and hyperparameters. Our generative pretraining procedure for each model is described in Tables 7 and 8 of Appendix C. Predictive finetuning is outlined in Table 9 of Appendix C, during which we finetune the weights dedicated to a particular task individually to maximize training batch throughput and minimize time to convergence.

Key inference hyperparameters include the white noise scale γ_g of each modality (following Geffner et al. (2025)) and the number of integration steps T. Through an extensive hyperparameter sweep illustrated in Figure 5 of Appendix C, we find that $T = 50$ or 100 with $\gamma_g = 0.01$ generally works well for molecule and material generation, except for QM9-sized small molecules, which benefit from adding additional white noise (e.g., 50) to non-periodic atom positions to improve sample uniqueness/diversity. Like Vonessen et al. (2025), $g(t) = \frac{1}{t+0.01}$ for noise scaling by default.

Evaluation metrics. Following previous work (Hoogeboom et al., 2022; Joshi et al., 2025), for generative pretraining, we evaluate ZATOM-1's ability to generate valid and realistic molecules and materials by sampling 10,000 molecules and 2,500 materials and computing validity, uniqueness, and PoseBusters sanity checks for molecules (Buttenschoen et al., 2024) and calculating validity, uniqueness, novelty, energy, and stability metrics using LeMat-GenBench (Duval et al., 2025). These generative metrics are described in further detail in Appendix D. For predictive finetuning, we assess ZATOM-1's MAE for property, energy, and force prediction tasks, scaled/unit-converted according to literature benchmarking conventions.

Baselines. For generative pretraining, we compare ZATOM-1 trained jointly on MP20 and QM9 to material-only and molecule-only variants, as well as state-of-the-art baselines for both data sources trained in one or two-stage configurations. Generated samples are shown in Appendix F.

Our experiments for materials generation using LeMat-GenBench include comparisons to: (1) six equivariant/invariant generative models adopting handcrafted multimodal product manifolds: MatterGen (Zeni et al., 2025), Crystal-GFN (Hernández-García et al., 2023), DiffCSP (Jiao et al., 2023), DiffCSP++ (Jiao et al., 2024), SymmCD (Levy et al., 2025), and FlowMM (Miller et al., 2024); (2) ADiT (Joshi et al., 2025), a non-equivariant molecule & material flow matching model that learns symmetries from data; and (3) four language model-based generative methods for materials: PLaID++ (Xu et al., 2025), WyFormer (Kazeev et al., 2025), Crystalformer (Cao et al., 2025), and LLaMat (Mishra et al., 2024).

Our molecule generation experiments compare ZATOM-1 to: (1) four generative methods operating on an equivariant multimodal product manifold: Equivariant Diffusion (Hoogeboom et al., 2022), Symphony (Daigavane et al., 2024), EQGAT-diff (Le et al., 2024), and SemlaFlow (Irwin et al., 2025); (2) two latent diffusion/flow matching generative models: GeoLDM (Xu et al., 2023) and ADiT (Joshi et al., 2025); and (3) non-equivariant TABASCO (Vonessen et al., 2025) which learns symmetries from data.

Table 1: **Crystal generation results on LeMat-GenBench with MP20 training.** Core model evaluation metrics are calculated for 2,500 sampled crystals using an *MLIP ensemble* with *LeMat-Bulk* as the reference set. Results are organized into three categories separated by horizontal lines: (1) for models that incorporate *relaxation* as part of their generation pipeline (MatterGen, PLaID++, etc.); (2) for other baseline generative models; and (3) for our proposed generative models. Best values are shown in **bold**, second-best values are underlined. Notably, jointly trained ZATOM-1 (80M) achieves its results using ∼400 GPU training hours, 3x faster than jointly trained ADiT (180M).

Model	# Params	Valid% ↑	Unique% ↑	Novel% ↑	Energy-based			Stability		Metastability	
					E_f (Std) ↓	E_{hull} (Std) ↓	RMSD (Std) ↓	Stable% ↑	SUN% ↑	Metastable% ↑	MSUN% ↑
With relaxation											
MatterGen	47M	95.7	**95.1**	**70.5**	-0.70 ± 0.79	0.18 ± 0.18	0.39 ± 0.50	2.0	0.2 (6)	33.4	**15.0**
PLaID++	7B	**96.0**	77.8	24.2	-0.50 ± 0.44	0.09 ± 0.16	$\mathbf{0.13 \pm 0.29}$	**12.4**	**1.0 (26)**	60.7	7.6
WyFormer	8B	93.4	93.0	66.4	-0.43 ± 0.95	0.50 ± 0.51	0.81 ± 0.98	0.5	0.1 (3)	15.7	1.9
WyFormer-DFT	8B	95.2	95.0	66.4	-0.67 ± 0.91	0.27 ± 0.36	0.42 ± 0.60	3.7	0.4 (9)	24.8	7.8
Without relaxation											
Jointly trained ADiT	180M	90.6	87.8	26.0	-0.73 ± 0.93	0.33 ± 0.45	0.38 ± 0.40	0.4	0.0 (0)	36.5	1.0
Crystal-GFN	1M	51.7	51.7	51.7	$\mathbf{-1.30 \pm 2.63}$	2.09 ± 2.38	1.87 ± 0.97	0.0	0.0 (0)	0.0	0.0
Crystalformer	5M	69.9	69.4	31.8	-0.17 ± 1.48	0.70 ± 1.33	0.66 ± 1.05	1.4	0.0 (0)	28.8	3.1
DiffCSP	12M	95.7	94.8	66.2	-0.64 ± 0.96	0.28 ± 0.56	0.59 ± 0.63	2.3	0.1 (3)	29.8	8.5
DiffCSP++	12M	95.3	**95.1**	62.0	-0.52 ± 0.87	0.41 ± 0.44	0.69 ± 0.82	1.0	0.2 (4)	26.4	5.0
LLaMat2-CIF	7B	84.4	81.4	30.0	-0.47 ± 1.01	0.44 ± 0.71	0.54 ± 0.71	0.7	0.1 (3)	34.7	2.1
LLaMat3-CIF	8B	15.4	15.2	10.5	0.74 ± 2.69	1.71 ± 2.48	1.01 ± 1.10	0.1	0.0 (0)	2.1	0.2
SymmCD	12M	73.4	73.0	47.0	-0.02 ± 1.22	0.88 ± 1.07	0.87 ± 1.09	1.4	0.1 (2)	18.6	2.4
Ours: Without relaxation											
MP20-only ZATOM-1	80M	73.2	70.4	21.0	-0.53 ± 0.97	0.44 ± 0.64	0.42 ± 0.49	0.6	0.00 (0)	32.7	0.20 (5)
Jointly trained ZATOM-1	80M	88.5	84.4	8.1	-0.84 ± 0.89	0.19 ± 0.37	0.21 ± 0.27	2.7	0.04 (1)	51.8	0.36 (9)
Jointly trained ZATOM-1-L	160M	95.0	90.4	3.7	-0.87 ± 0.86	$\mathbf{0.09 \pm 0.21}$	0.14 ± 0.20	3.1	0.04 (1)	**71.5**	0.64 (16)
Jointly trained ZATOM-1-XL	300M	94.9	90.1	4.8	-0.91 ± 0.87	0.11 ± 0.23	0.16 ± 0.19	3.3	0.00 (0)	64.3	0.36 (9)

Table 2: **Molecule generation results on QM9.** We report (a) validity and uniqueness rates as well as (b) % pass rates on 7 sanity checks from PoseBusters for 10,000 sampled molecules. All models explicitly generate hydrogen atoms. Notably, jointly trained ZATOM-1 (80M) achieves its results using ∼400 GPU training hours compared to jointly trained ADiT (180M), which requires ∼1,200 GPU pretraining + finetuning hours.

(a) **Validity results**

Model	# Params	Validity (%) ↑	Unique (%) ↑
One-stage training			
Equivariant Diffusion	20M	91.90	98.69
Symphony	2M	83.50	97.98
Two-stage training			
GeoLDM	20M	93.80	98.82
QM9-only ADiT	180M	92.19	97.90
Jointly trained ADiT	180M	94.45	97.82
Ours: One-stage training			
QM9-only ZATOM-1	80M	92.88	97.71
Jointly trained ZATOM-1	80M	94.94	97.16
Jointly trained ZATOM-1-L	160M	95.26	96.84
Jointly trained ZATOM-1-XL	300M	95.19	97.10

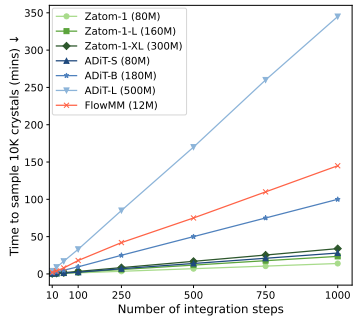
(b) **PoseBusters results**

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

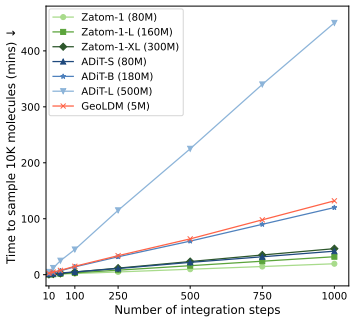
For predictive finetuning, we compare ZATOM-1 to: (1) 11 highly-optimized property prediction methods: DimeNet++ (Gasteiger et al., 2020), EGNN (Satorras et al., 2021), PaiNN (Schütt et al., 2021), TorchMD-NET (Thölke & Fabritiis, 2022), SphereNet (Liu et al., 2022b), SEGNN (Brandstetter et al., 2022), EQGAT (Le et al., 2022), Equiformer (Liao & Smidt, 2023), EquiformerV2 (Liao et al., 2024), PΘNITA (Bekkers et al., 2024), and the Platonic Transformer (Islam et al., 2025); (2) an unoptimized single-task property prediction method: AIM-STL (Minot & Schneider, 2025); and (3) two unoptimized multi-task property prediction methods: AIM-MTL-Scalar and AIM-MTL-Matrix (Minot & Schneider, 2025). Lastly, our energy and force prediction experiments for molecules and materials compare ZATOM-1 to the equivariant eSEN (Fu et al., 2025) and Orb-v1 (Neumann et al., 2024) models, respectively.

4 RESULTS

State-of-the-art molecule (crystal) generation. Tables 1 and 2 show that ZATOM-1 benefits from positive transfer learning between chemical domains compared to MP20-only or QM9-only training. In particular, Table 1's LeMat-GenBench results demonstrate that ZATOM-1 can generate highly stable, unique, and novel (SUN) materials after rigorous sample evaluation, whereas previous flow matching-based methods, such as ADiT, cannot. Nonetheless, as Duval et al. (2025) have suggested, ZATOM-1's novelty and SUN rates and model scaling benefits could likely be improved by training on a larger set of diverse 3D materials, such as (5M) LeMat-Bulk (Siron et al., 2025), in future



(a) Crystals – MP20



(b) Molecules – QM9

Figure 2: **ZATOM-1 is significantly faster than equivariant and latent diffusion models.** We plot the number of integration steps for ZATOM-1, equivariant diffusion FlowMM, and latent diffusion ADiT vs. time to generate 10,000 samples on a single NVIDIA GPU. ZATOM-1 scales significantly better with the number of integration steps compared to equivariant and latent diffusion models.

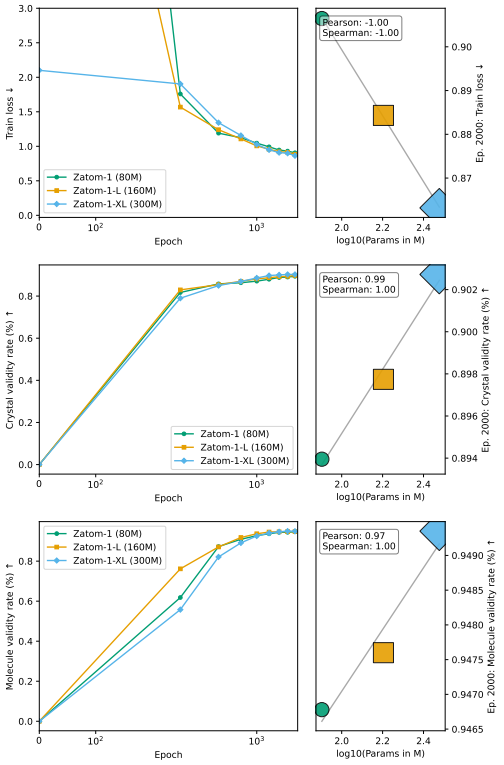


Figure 3: **Scaling up ZATOM-1 predictably improves performance.** We show the effect of increasing the number of ZATOM-1 model parameters on the training loss and generation validity rates. **Left:** training loss and validity rates vs. epochs. **Right:** correlation plots for training loss and validity rates at epoch 2,000 vs. ZATOM-1 parameters (in Millions).

work. Table 2's QM9 results highlight that ∼99% of ZATOM-1's generated molecules pass all of the PoseBusters sanity checks, while ∼95% of ADiT's molecules satisfy each check, establishing a new state-of-the-art. In line with Table 2's results, Table 3 shows that ZATOM-1 is the first generative model to reach a (maximum-possible) PoseBusters validity rate of 94% for larger molecules in GEOM-Drugs (Vonessen et al., 2025), demonstrating the strength of ZATOM-1's molecule & material pretraining for generating large, drug-like molecules. Important to note is that, due to QM9 (MP20)'s small size, for molecule (material) generation, increasing model size exchanges the model's rate of sample uniqueness (novelty) for increased sample validity (stability). For interested readers, we report results for O(3)-equivariant PLATOM-1 in Appendix B.

ZATOM-1 is faster than baselines. For both QM9 and MP20 sample generation, Figure 2 illustrates that ZATOM-1 offers an order of magnitude faster inference speeds than ADiT, a state-of-the-art baseline, and Table 3 shows that ZATOM-1's speeds scale better for GEOM-Drugs sampling.

ZATOM-1 improves with scale. Figure 3 highlights that ZATOM-1's training and validation metrics improve predictably as the model's size increases, suggesting that further performance improvements may be obtained with parallel data and model scaling (Kaplan et al., 2020).

Chemical transfer learning improves predictions. Table 4 demonstrates that ZATOM-1 achieves state-of-the-art *multi-task* property prediction performance for QM9. For the majority of QM9's target properties, these results further mark the *first* instance of positive predictive transfer learning and zero-shot generalization via joint generative pretraining in SciML, indicating that the final trunk

Table 3: **Molecule generation results on GEOM-DRUGS.** *Left:* Validity, uniqueness, and % pass rates on PoseBusters for 10,000 sampled molecules. ZATOM-1 matches or exceeds state-of-the-art equivariant/latent diffusion baselines and generates more PoseBusters-valid samples than the recent baseline TABASCO. *Right:* We plot the number of integration steps for ZATOM-1, (latent diffusion) ADiT, and (equivariant) SemlaFlow vs. time to generate 10,000 molecules on a single NVIDIA GPU. ZATOM-1 scales competitively with the number of integration steps compared to SemlaFlow/ADiT.

Metric (% pass) ↑	EQGAT-diff	SemlaFlow	ADiT	TABASCO	ZATOM-1	GEOM-Drugs
Validity	94.6	93.9	95.3	97.6	93.6	-
Uniqueness	100.0	100.0	100.0	99.03	99.93	-
Atoms connected	84.4	92.3	93.0	99.9	99.6	-
Bond angles	86.9	94.8	92.3	99.2	99.6	-
Bond lengths	87.0	94.6	92.5	99.4	99.7	-
Aromatic ring flat	87.0	94.9	95.4	100.0	100.0	-
Double bond flat	87.0	94.2	95.3	99.8	100.0	-
Internal energy	86.8	94.8	91.3	99.4	99.3	-
No steric clash	82.9	92.0	91.8	94.3	96.4	-
PoseBusters valid	59.7	87.5	85.3	91.6	94.1	94.0

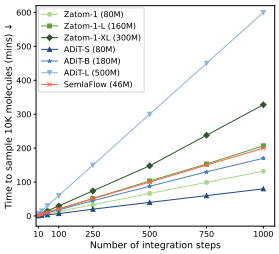


Table 4: **Property prediction for 3D molecules (QM9) / materials (Matbench, zero-shot).** Results: *test* mean absolute errors (std. dev. over three runs with random seeds); †: using different data partitions; /: a joint molecule & material property prediction task. Best (second-best) results for each (optimized or unoptimized) category are in **bold** (underlined). Notably, ZATOM-1 (80M ❄ + 20M 🔥) and AIM-MTL are the only multi-task learning (MTL) methods represented in these results, which they achieved with only ∼100 GPU hours of additional finetuning for molecule property prediction. In contrast, single-task learning (STL) baselines such as EquiformerV2 (11M) traditionally require ∼150 GPU training hours *for each task*, due to their extensive hyperparameter optimization per task.

Model / Metrics ↓	# Params	α (a_0^3)	$\Delta\varepsilon$ (meV)	$\varepsilon_{\text{HOMO}}$ (meV)	$\varepsilon_{\text{LUMO}}$ (meV)	μ (D)	C_ν (cal/mol K)	G^{ATOM} (meV)	H^{ATOM} (meV)	R^2 (a_0^2)	U^{ATOM} (meV)	U_0^{ATOM} (meV)	$ZPVE$ (meV)
Optimized single-task learning													
DimeNet++†	5M	.044	33.0 / **199**	25	20	.030	.023	8.00	7.00	.331	6.00	6.00	1.21
EGNN†	1M	.071	48.0 / —	29	25	.029	.031	12.0	12.0	.106	12.0	11.0	1.55
PaiNN	1M	.045	46.0 / —	28	20	.012	.024	**7.35**	**5.98**	.066	**5.83**	**5.85**	1.28
TorchMD-NET	7M	.059	36.0 / —	20	18	.011	.026	7.62	6.16	**.033**	6.38	6.15	1.84
SphereNet	2M	.046	32.0 / —	23	18	.026	**.021**	8.00	6.00	.292	7.00	6.00	**1.12**
SEGNN†	1M	.060	42.0 / —	24	21	.023	.031	15.0	16.0	.660	13.0	15.0	1.62
EQGAT	-	.053	32.0 / —	20	16	.011	.024	23.0	24.0	.382	25.0	25.0	2.00
Equiformer	4M	.046	30.0 / —	15	14	.011	.023	7.63	6.63	.251	6.74	6.59	1.26
EquiformerV2	11M	.050	**29.0** / —	**14**	**13**	.010	.023	7.57	6.22	.186	6.49	6.17	1.47
PΘNITA	-	**.038**	30.4 / —	16	15	.012	.024	8.63	8.04	.235	8.67	8.31	1.29
Platonic Transformer	-	.049	37.4 / —	22	17	**.010**	.024	12.0	12.0	.222	11.9	13.0	1.30
Unoptimized single-task learning													
AIM-STL†	-	.181	— / —	61	54	**.067**	.072	66.2	63.9	**.503**	64.2	58.8	4.54
Unoptimized multi-task learning													
AIM-MTL-Scalar†	-	.268	— / —	59	71	.089	.103	113	112	4.18	112	112	9.82
AIM-MTL-Matrix†	-	.251	— / —	61	72	.088	.103	112	118	4.07	116	116	12.2
Non-pretrained ZATOM-1 (OURS)	80M ❄ + 20M 🔥	.140±.001	72.3±0.1 / 3358±6	54±0.0	49±0.0	.157±.001	.064±.000	45.5±0.2	44.7±0.2	4.72±0.02	45.2±0.2	45.8±0.2	2.89±0.01
QM9-pretrained ZATOM-1 (OURS)	80M ❄ + 20M 🔥	.095±.001	49.5±0.3 / 5297±29	36±0.2	34±0.1	.100±.000	.044±.000	**23.1±0.2**	23.1±0.2	3.63±0.01	**23.5±0.2**	**22.6±0.2**	2.04±0.01
Jointly pretrained ZATOM-1, $K = L/2$ (OURS)	80M ❄ + 20M 🔥	.099±.001	51.2±0.3 / **2640±8**	37±0.1	34±0.2	.121±.001	.044±.000	24.1±0.1	23.3±0.2	3.96±0.01	**23.5±0.2**	23.4±0.1	2.01±0.01
Jointly pretrained ZATOM-1-XL (OURS)	300M ❄ + 20M 🔥	.104±.000	53.1±0.1 / 3478±14	38±0.1	37±0.1	.110±.000	.046±.000	26.6±0.0	26.2±0.1	3.75±0.02	29.2±0.0	26.6±0.1	2.14±0.01
Jointly pretrained ZATOM-1 (OURS)	80M ❄ + 20M 🔥	**.091±.001**	**46.2±0.2** / 3891±20	**34±0.1**	**32±0.2**	.090±.001	**.041±.000**	23.3±0.4	**23.0±0.4**	3.29±0.01	24.0±0.4	23.3±0.4	**1.98±0.01**

layer of ZATOM-1 excels in 3D molecular tasks ($K = L$), while its middle trunk layers are better suited for (zero-shot) 3D material tasks ($K = L/2$). Interestingly, increasing the size of ZATOM-1's pretrained trunk does not guarantee improvements for property prediction, which may be due to the modest size of ZATOM-1's pretraining datasets.

Additional results. We refer interested readers to Appendix E for the results of additional experiments, including those for metal-organic framework generation, materials finetuning for property prediction, and energy and force prediction towards machine learning-based interatomic potentials (MLIPs) (Wood et al., 2025). Notably, we demonstrate promising performance for materials force prediction compared to Orb-v1 and further find that ZATOM-1 can effectively balance (supervised) molecular and material property prediction with additional cross-domain finetuning.

5 CONCLUSION

In this work, we introduced ZATOM-1, the first general-purpose foundation model for 3D molecules and materials. Empirical results show that ZATOM-1 achieves state-of-the-art molecule generation and competitive materials generation, with fast training and inference speeds. Additionally, it demonstrates the advantages of generative pretraining for downstream prediction tasks, marking SciML's first instance of positive predictive transfer learning through joint generative pretraining.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 2025. ISSN 0021-9606. doi: 10.1063/5.0297006.

Erik Bekkers, Maxime W Lafarge, Mitko Veta, Koen A.J. Eppenhof, Josien P.W. Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

Erik J Bekkers, Sharvaree Vadgama, Rob Hesselink, Putri A Van der Linden, and David W. Romero. Fast, expressive $\mathrm{SE}(n)$ equivariant networks through weight-sharing in position-orientation space. In *The Twelfth International Conference on Learning Representations*, 2024.

Alexandre Blanco-Gonzalez, Alfonso Cabezon, Alejandro Seco-Gonzalez, Daniel Conde-Torres, Paula Antelo-Riveiro, Angel Pineiro, and Rebeca Garcia-Fandino. The role of ai in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6):891, 2023.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*, 2022.

Rania Briq, Michael Kamp, Ohad Fried, Sarel Cohen, and Stefan Kesselheim. The amazing stability of flow matching. In *EurIPS Workshop on Principles of Generative Modeling (PriGM)*, 2025.

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.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5453–5512. PMLR, 21–27 Jul 2024.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 2025.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/pdf?id=WE4qe9xlnQw.

Jing Chang and Jingkang Ye. Bidirectional generation of structure and properties through a single molecular foundation model. *Nat. Commun.*, 15:2323, 2024.

Junyoung Choi, Gunwook Nam, Jaesik Choi, and Yousung Jung. A perspective on foundation models in chemistry. *JACS Au*, 5(4):1499–1518, 2025. doi: 10.1021/jacsau.4c01160.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *International Conference on Machine Learning (ICML)*, 2016.

Ameya Daigavane, Song Eun Kim, Mario Geiger, and Tess Smidt. Symphony: Symmetry-equivariant point-centered spherical harmonics for 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Ian Dunn and David R. Koes. Flowmol3: Flow matching for 3d de novo small-molecule generation. *arXiv preprint arXiv:2508.12629*, 2025.

Alexandre Duval, Siddharth Betala, Samuel P Gleason, Andy Xu, Georgia Channing, Daniel Levy, Ali Ramlaoui, Clémentine Fourrier, Chaitanya K Joshi, Nikita Kazeev, et al. Lemat-genbench: Bridging the gap between crystal generation and materials discovery. In *NeurIPS AI for Accelerated Materials Design Workshop*, 2025.

Sathya Edamadaka, Soojung Yang, and Rafael Gomez-Bombarelli. Universally converging representations of matter across scientific foundation models. In *NeurIPS AI for Accelerated Materials Design Workshop*, 2025.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025.

R. Gao, E. Hoogeboom, J. Heek, V. de Bortoli, K. P. Murphy, and T. Salimans. Diffusion meets flow matching: Two sides of the same coin. https://diffusionflow.github.io/, 2024.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ralf W Grosse-Kunstleve, Nicholas K Sauter, and Paul D Adams. Numerically stable algorithms for the computation of reduced unit cells. *Foundations of Crystallography*, 60(1):1–6, 2004.

Charles Harris, Kieran Didi, Arian Jamasb, Chaitanya Joshi, Simon Mathis, Pietro Lio, and Tom Blundell. Posecheck: Generative models for 3d structure-based drug design produce unrealistic poses. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Alex Hernández-García, Alexandra Volokhova, Alexandre AGM Duval, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, Victor Schmidt, et al. Crystal-gfn: sampling materials with desirable properties and constraints. In *NeurIPS AI for Accelerated Materials Design Workshop*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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. PMLR, 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.

Mohammad Mohaiminul Islam, Rishabh Anand, David R Wessels, Friso de Kruiff, Thijs P Kuipers, Rex Ying, Clara I Sánchez, Sharvaree Vadgama, Georg Bökman, and Erik J Bekkers. Platonic transformers: A solid choice for equivariance. *arXiv preprint arXiv:2510.03511*, 2025.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

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.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Xin Jin, Kevin Maik Jablonka, Elias Moubarak, Yutao Li, and Berend Smit. Mofchecker: a package for validating and correcting metal–organic framework (mof) structures. *Digital Discovery*, 2025.

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 *Forty-second 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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *Forty-second International Conference on Machine Learning*, 2025.

Susumu Kitagawa et al. Metal–organic frameworks (mofs). *Chemical Society Reviews*, 43(16): 5415–5418, 2014.

Tobias Kreiman, Yutong Bai, Fadi Atieh, Elizabeth Weaver, Eric Qu, and Aditi S Krishnapriyan. Transformers discover molecular structure without graph priors. *arXiv preprint arXiv:2510.02259*, 2025.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in neural information processing systems*, 34:26548–26560, 2021.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Tuan Le, Frank Noe, and Djork-Arné Clevert. Representation learning on biomolecular structures using equivariant graph attention. In *The First Learning on Graphs Conference*, 2022.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *The Twelfth 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, et al. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. SymmCD: Symmetry-preserving crystal generation with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217, October 2023.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

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*, 2022a.

Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations*, 2022b.

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

Mason Minot and Gisbert Schneider. AIM: Adaptive intervention for deep multi-task learning of molecular properties. In *NeurIPS 2025 AI for Science Workshop*, 2025.

Vaibhav Mishra, Somaditya Singh, Dhruv Ahlawat, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, NM Krishnan, et al. Foundational large language models for materials research. *arXiv preprint arXiv:2412.09560*, 2024.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

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.

Edward O. Pyzer-Knapp, Matteo Manica, Peter Staar, Lucas Morin, Patrick Ruch, Teodoro Laino, John R. Smith, and Alessandro Curioni. Foundation models for materials discovery – current state and future directions. *npj Comput. Mater.*, 11:61, 2025. doi: 10.1038/s41524-025-01538-0.

Eric Qu and Aditi S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139030–139053. Curran Associates, Inc., 2024. doi: 10.52202/079017-4412.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Martin Siron, Inel Djafar, Ali Ramlaoui, Etienne du Fayette, Amandine Rossello, Edvin Fako, Matthew McDermott, Felix Therrien, Luis Barroso-Luque, Flaviu Cipcigan, et al. Lematbulk: aggregating, and de-duplicating quantum chemistry materials databases. *arXiv preprint arXiv:2511.05178*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for joint multi-ligand docking and binding site design. In *Forty-first International Conference on Machine Learning*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 71242–71262. Curran Associates, Inc., 2023.

Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Pedro Vélez, Luisa F. Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi S. M. Sajjadi. From image to video: An empirical study of diffusion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16948–16958, October 2025.

Venkat Venkatasubramanian. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*, 65(1), 2019.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 560–576. Springer, 2023.

Carlos Vonessen, Charles Harris, Miruna Cretu, and Pietro Lio. TABASCO: A fast, simplified model for molecular generation with improved physical quality. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Alexius Wadell, Anoushka Bhutani, Victor Azumah, Austin R Ellis-Mohr, Celia Kelly, Hancheng Zhao, Anuj K Nayak, Kareem Hegazy, Alexander Brace, Hongyi Lin, et al. Foundation models for discovery and exploration in chemical space. *arXiv preprint arXiv:2510.18900*, 2025.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1911.08251.

Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant CNNs. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1711.07289.

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. *Equivariant and Coordinate Independent Convolutional Networks*. WORLD SCIENTIFIC, 2025. doi: 10.1142/14143. URL https://maurice-weiler.gitlab.io/#cnn_book.

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, Sushree Jagriti Sahoo, Ammar Rizvi, Zachary Ward Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024.

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

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.

Andy Xu, Rohan Desai, Larry Wang, Gabriel Hope, and Ethan Ritz. Plaid++: A preference aligned language model for targeted inorganic materials design. *arXiv preprint arXiv:2509.07150*, 2025.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Luu Anh Tuan, and Chen Zhao. Diffusion vs. autoregressive language models: A text embedding perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4273–4303, 2025.

Maksim Zhdanov, David Ruhe, Maurice Weiler, Ana Lucic, Johannes Brandstetter, and Patrick Forré. Clifford-Steerable Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.14730.

APPENDICES

A Related Work

The concept of scientific foundation models, which are large, self-supervised networks pretrained on broad scientific data, has recently gained momentum in chemistry and materials science (Pyzer-Knapp et al., 2025; Choi et al., 2025; Edamadaka et al., 2025). Recent reviews emphasize the value of training on diverse chemical datasets and using generative or multi-task objectives to build broadly useful chemical representations (Pyzer-Knapp et al., 2025; Choi et al., 2025; Wadell et al., 2025). However, most existing works remain domain-specific: "foundation" efforts in molecules often focus on SMILES or 2D graph pretraining, while material science approaches largely target property prediction or synthesis planning rather than joint 3D modeling.

Generative modeling of 3D molecules. Several recent works propose equivariant diffusion or flow models for de novo 3D molecule design. Hoogeboom et al. (2022) introduce an E(3)-equivariant diffusion model (EDM) that denoises continuous atom coordinates and types together. Autoregressive graph models like Symphony (Daigavane et al., 2024) build molecules fragment-by-fragment with spherical harmonics features. Le et al. (2024) explore E(3)-equivariant diffusion losses (EQGAT-diff) to improve validity on QM9 and GEOM-Drugs. Latent diffusion approaches like GeoLDM (Xu et al., 2023) encode 3D structures into continuous latent spaces. More recent work explores the importance of equivariance: for instance, SemlaFlow uses flow matching with an equivariant graph neural network backbone (Irwin et al., 2025); FlowMol3 employs hybrid continuous-discrete flow matching with chirality-aware message passing (Dunn & Koes, 2025); and TABASCO shows that a non-equivariant transformer can match the validity of equivariant models while generating 3D drug-like molecules (Vonessen et al., 2025). In summary, these models each focus on molecule generation but do not generalize across materials.

Generative modeling of 3D materials. Likewise, generative models for periodic crystals have been developed separately. Score-based models (diffusion) and flow matching have been applied to crystal generation: e.g., Xie et al. (2022) propose CDVAE, a score-based model for generating periodic unit cells, and Jiao et al. (2023) introduce DiffCSP, an E(3)-equivariant diffusion model for predicting stable crystal structures. Recent work often uses variants of normalizing flows: Miller et al. (2024)'s FlowMM applies Riemannian flow matching over atom types and lattice geometry, and Yang et al. (2024) show that standard (non-equivariant) diffusion on a learnable crystal representation (UniMat) can scale to large collections of structures. Another parallel thread is generative models for inorganic materials like those of Zeni et al. (2025). FlowLLM combines a fine-tuned language model base with flow matching for crystals (Sriram et al., 2024), while the All-Atom Diffusion Transformer (ADiT) unifies molecule and material generation by learning a shared latent autoencoding of both domains and diffusing in that latent space (Joshi et al., 2025). Notably, none of these approaches simultaneously addresses both molecules and materials in one model without resorting to complex two-stage training approaches, such as latent diffusion or LLM-based data generation, nor do they leverage generative pretraining for downstream tasks.

Multimodal flow/diffusion architectures. Our work joins a growing class of multimodal generative models that handle discrete atom types and continuous 3D geometries jointly. For example, Miller et al. (2024) explicitly matches flows over categorical atom labels, continuous coordinates, and lattice parameters in a single framework. ADiT (Joshi et al., 2025) analogously embeds all-atom molecular and crystal representations into a shared latent space for latent diffusion. These approaches stand in contrast to earlier graph or point-based generative models (e.g., EDM (Hoogeboom et al., 2022), EQGAT-diff (Le et al., 2024)) that typically model only a small number of data modalities simultaneously. By jointly training on atom types, positions, and unit cell parameters, our approach builds on the multimodal flow concept of FlowMM while using a standard Transformer backbone for expressiveness, scalability, and speed.

Representation learning and pretraining in chemistry. Beyond full generative models, chemical representation learning has employed self-supervised pretraining on molecular graphs. For instance, GraphMVP (Liu et al., 2022a) uses contrastive learning between 2D molecular topology and 3D geometry. Masked autoencoder approaches and property prediction pretraining have also been explored. However, most existing pretraining does not integrate full sample generation. Even though image diffusion models have been shown to yield useful feature embeddings (Li et al., 2023; Vélez

et al., 2025), few works have investigated generative pretraining as a widely applicable paradigm in 3D chemistry or AI for Science more broadly. Accordingly, our work frames conditional multimodal flow matching (over atom types and geometry) as a self-supervised "pretraining" task for 3D chemical systems. By pretraining on joint molecule & material data, we obtain unified embeddings transferable to property, energy, and force prediction. This idea is similar to Chang & Ye (2024)'s structure-property molecular model, which learns a bidirectional generative model of molecular graphs and properties, but here we apply it broadly across both molecules and crystals.

Overview. Overall, prior work in molecular/material generation has focused either on discrete (graph/SMILES) models or on separate generative flows/diffusions for molecules or crystals (Hoogeboom et al., 2022; Xie et al., 2022). Our approach differs by unifying generative and predictive modeling for all-atom 3D chemistry. We draw on advances in flow matching and multimodal architecture design (Song et al., 2023; Miller et al., 2024; Vonessen et al., 2025), and we treat generative modeling itself as a pretraining strategy. This differs from previous scientific foundation models that often focus on language or graph representations (Pyzer-Knapp et al., 2025; Choi et al., 2025; Wadell et al., 2025). By explicitly modeling the ambient 3D space of atoms and elements together, we extend the flow-based generative paradigm to create a single foundation model across 3D molecular and material domains.

B PLATOM-1 – EXPLICITLY EQUIVARIANT TRANSFORMER ARCHITECTURE

A long-standing discussion in the molecular modeling community revolves around whether symmetries in the data should be learned or hard-coded into the model architecture. Given the considerable success of entirely non-equivariant models like AlphaFold3 (Abramson et al., 2024), the community's consensus has shifted toward the conclusion that equivariant models would be slower, harder to optimize, and generally inferior compared to architectures without such inductive biases. This appendix provides empirical evidence challenging this common belief, showing that – *when done right* – equivariance can unlock significant performance gains.

Section B.1 introduces the Trunk-based Flow Platoformer (TFP), a rotation equivariant counterpart of the Trunk-based Flow Transformer (TFT) architecture underlying ZATOM-1. Empirical results, showing faster convergence and higher validation scores of TFP, are presented in Section B.2 and Figure 4. For now, we focus exclusively on molecular tasks, since the current fractional coordinate-based invariant embedding of materials used in ZATOM-1 is at odds with the equivariant inputs expected by TFP.

B.1 TRUNK-BASED FLOW PLATOFORMER ARCHITECTURE (TFP)

The Trunk-based Flow Platoformer (TFP) is a rotation equivariant counterpart of our TFT architecture from Section 2.2, replacing its standard transformer layers with *Platonic Transformer* layers introduced by Islam et al. (2025). These layers are equivariant to discrete subgroups of roto-reflections $G < \mathrm{O}(3)$ which are symmetries of Platonic solids like the tetrahedron ($|G| = 12$) or cube ($|G| = 24$) (Cesa et al., 2022).

Rather than relying on $\mathrm{O}(3)$-irrep features (i.e., $\mathrm{O}(3)$-irreducible representations), as most equivariant point cloud message passing neural networks do, the features of TFP are associated with group elements, essentially encoding molecular geometry as viewed from $|G|$ different frames of reference. Mathematically, this is known as a *regular G-representation* (Weiler et al., 2025). In code, a point cloud of N features with C channels is represented by a tensor of shape $(N, |G|, C)$ rather than the usual shape (N, C). Such regular representation-valued features are more commonly used in vision models, most notably group convolutional networks (Cohen & Welling, 2016; Weiler et al., 2018; Bekkers et al., 2018).

TFP architecture. Before applying Platonic layers, inputs need to be lifted to regular G-representations. Scalars (e.g., atom types) are lifted to features that are invariant over the group axis. Vector-valued inputs, like Euclidean coordinates, lift to non-trivial features encoding directional information. Platonic linear layers satisfy a symmetry constraint which reduces their parameter count via $|G|$-fold weight sharing. The original linear embedding layers of TFT are replaced by such Platonic linear layers, applied after G-lifts.

Platonic attention layers utilize common scaled dot-product attention (Vaswani et al., 2017), and are thus compatible with IO-efficient Flash Attention kernels (Dao, 2024). The group axis of regular G-features is thereby reshaped into the attention heads axis rather than channels, i.e., $(H \cdot |G|,\ N,\ C)$ instead of $(H,\ N,\ |G| \cdot C)$ – this has the advantage that attention scores remain regular representation-valued rather than becoming G-invariant due to contracting over G (Islam et al., 2025). We use full softmax attention instead of the linear attention variants explored in the original Platonic Transformer paper, as these were required mainly to address training instabilities when encoding rotary positions in molecular modeling tasks.

A core component of the Platonic Transformer is a G-equivariant generalization of Rotary Position Embeddings (RoPE) (Su et al., 2024), which guarantees both translational and G-equivariance, i.e., approximate $\mathrm{E}(3)$-equivariance. As we aim to stick as close as possible to the original TFT design without *relative* RoPE attention, we omit this feature, relying merely on the *absolute* embedding of Euclidean coordinates via lifting and Platonic linear layers discussed above. Note that this design choice breaks translational symmetries but preserves G-equivariance. Since we zero-center non-periodic 3D atom coordinates, this is not an issue in practice.

We reimplement and extend many other operations from the original Platonic Transformer code, including cross-attention layers, new normalization modules, embedding layers, and optimized Platonic linear layers. Given these modules, we implement TFP in analogy to the TFT design summarized in Algorithm 1. Final predictions are projected back from G-valued features to scalar and vector outputs. Our implementation is available as part of the ZATOM-1 code repository.

Compute and parameter cost. As TFP operates on tensors of shape $(N, |G|, C)$ with an additional group axis, one may wonder how its computational, memory, and parameter complexities compare to those of the TFT baseline. Given that regular features have $|G|C$ effective dimensions, it is common practice to reduce their channel count C relative to the D channels of TFT's shape (N, D) features. The implied complexities for TFT's and TFP's linear and attention layers are summarized in Table 5. From these complexities, the following two heuristics to choose C to build "matching" architectures arise (Weiler & Cesa, 2019):

Compute-matching: To match the compute and memory cost of TFP and TFT, one sets $C = D/|G|$, essentially sticking with the same effective number of channels but splitting them into group and channel axes. However, this implies G times fewer trainable parameters in TFP.

Parameter-matching: Alternatively, one may match the number of parameters of TFP and TFT by setting $C = D/\sqrt{|G|}$. This leads to a $|G|$ fold increase in memory and Floating-point Operations Per Second (FLOPS).

As a middle ground between these two extremes, we choose $C = D/|G|^{2/3}$, resulting in $\sqrt{|G|}$ fewer parameters and $\sqrt{|G|}$ increased computational cost. For the tetrahedral group, $|G| = 12$ and $\sqrt{|G|} \approx 3.5$, such that a TFP analog to ZATOM-1 with 80M parameters has 23M parameters. Note that this model happens to be approximately compute-matched with ZATOM-1-XL, which has 300M parameters, i.e., 3.75 times the number of ZATOM-1's parameters.

B.2 EMPIRICAL RESULTS

Based on TFP, we implement PLATOM-1, the Platonic group equivariant counterpart of ZATOM-1. Due to limited computational resources, we specifically evaluate the (orientation-preserving) tetrahedral group with $|G| = 12$ rotations (no reflections) and 23M parameters as discussed above. Since PLATOM-1 is only equivariant w.r.t. discrete rotations, we use the same continuous $\mathrm{SO}(3)$ rotational data augmentation as used for ZATOM-1.

QM9 molecule generation. We first compare to the QM9-only pretrained ZATOM-1 baseline flow model in Figure 4. It is evident from these plots that PLATOM-1 converges much faster than ZATOM-1 in terms of training and validation losses, as well as validation metrics. Such improved convergence is common for equivariant models, as they do not need to implicitly learn symmetries from data (Weiler et al., 2025; Zhdanov et al., 2024). The best TFP checkpoint on the validation set is reached after 249 epochs, instead of 399 epochs for TFT, partially offsetting TFP's increased computational cost.

	Linear	G-linear	Attention	G-attention
Compute FLOPS	$\mathcal{O}(ND^2)$	$\mathcal{O}(N\|G\|^2C^2)$	$\mathcal{O}(N^2D)$	$\mathcal{O}(N^2\|G\|C)$
Memory	$\mathcal{O}(ND)$	$\mathcal{O}(N\|G\|C)$	$\mathcal{O}(ND)$	$\mathcal{O}(N\|G\|C)$
Parameters	D^2	$\|G\|\cdot C^2$	—	—

Table 5: **Compute, memory, and parameter complexity of TFT and TFP.** TFT and TFP operate on tensors of shapes $(N, |G|, C)$ and (N, D), respectively. Memory costs follow immediately from these shapes. Linear layers in TFT come with the usual quadratic number of FLOPS and parameters in their D channels. TFP is internally leveraging `nn.Linear` layers as well, however, after flattening features to shape $(N, |G|\cdot C)$, implying a computational cost quadratic in $|G|$ as well. Equivariance imposes a symmetry constraint on the matrices (weight sharing), resulting in a $|G|$ times-enhanced parameter efficiency. TFT's attention layers have the usual computational complexity linear in channels D, while those of TFP are linear in $|G|C$. Neither introduces additional parameters.

Table 6 confirms that these findings translate to all test metrics calculated for these checkpoints. In particular, PLATOM-1 achieves better results than ZATOM-1, prior equivariant QM9-only pretrained models Equivariant Diffusion (Hoogeboom et al., 2022) and Symphony (Daigavane et al., 2024), and the much larger and slower latent diffusion model ADiT (Joshi et al., 2025). It even outperforms our 300M parameter model ZATOM-1-XL, despite the latter being jointly trained on molecules and materials.

Joint molecule and material pretraining. One of the main hypotheses of ZATOM-1 is the benefit of positive transfer learning when pretraining jointly on molecular and material modalities. A similar mutual benefit should, in principle, be observable for PLATOM-1. However, our current encoding of materials is inherently incompatible with Platom-1's equivariant design, leading to training instabilities. Specifically, materials are not represented in terms of their (vector-valued) Euclidean coordinates, but rather in terms of fractional coordinates and lattice basis lengths and angles. All these quantities are scalars, and therefore lift to features that are invariant over the G-axis. Being blind to the original Euclidean geometry expected by PLATOM-1, the model fails to converge.

While ZATOM-1 can learn to leverage fractional coordinates inputs, we believe that this is not the ideal (equivariant) data representation for PLATOM-1, since it obfuscates the Euclidean geometry of the materials. As part of future work, we plan to experiment with TFP using a unified Euclidean coordinate representation of molecules and materials, for which we expect to see positive transfer learning results in a PLATOM-1-style model.

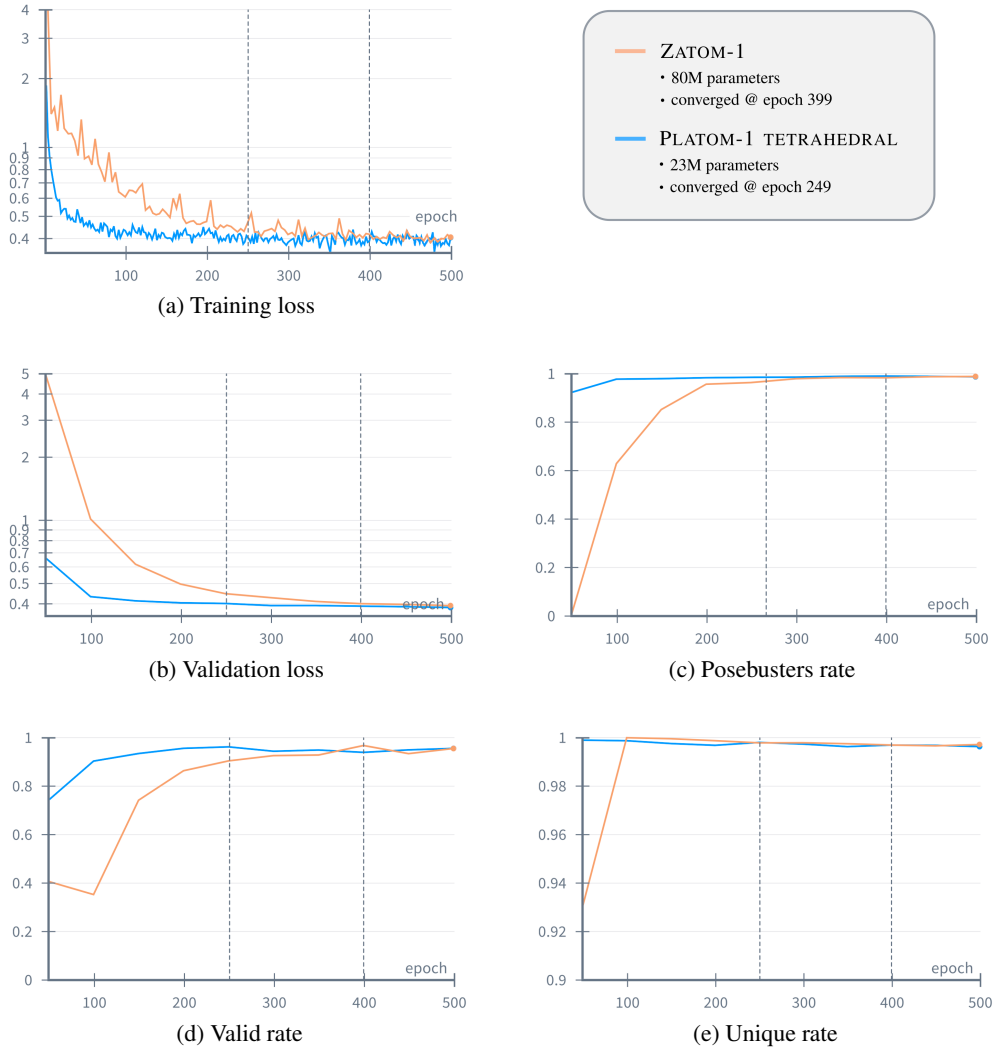
Table 6: **Molecule generation results on QM9 with explicit equivariance.** We report (a) validity and uniqueness rates as well as (b) % pass rates on 7 sanity checks from PoseBusters for 10,000 sampled molecules. All models explicitly generate hydrogen atoms. Notably, QM9-only PLATOM-1 achieves its results using 249 training epochs compared to QM9-only ZATOM-1 (80M), which requires 399 training epochs to converge. Despite its smaller size and not being jointly pretrained on molecules and materials, PLATOM-1 is competitive with ADiT and ZATOM-1-XL.

(a) **Validity results**

Model	**# Params**	**Validity (%) ↑**	**Unique (%) ↑**
One-stage training			
Equivariant Diffusion	20M	91.90	98.69
Symphony	2M	83.50	97.98
Two-stage training			
GeoLDM	20M	93.80	98.82
QM9-only ADiT	180M	92.19	97.90
Jointly trained ADiT	180M	94.45	97.82
Ours: One-stage training			
QM9-only ZATOM-1	80M	92.88	97.71
QM9-only PLATOM-1	23M	95.20	98.13
Jointly trained ZATOM-1	80M	94.94	97.16
Jointly trained ZATOM-1-L	160M	95.26	96.84
Jointly trained ZATOM-1-XL	300M	95.19	97.10

(b) **PoseBusters results**

Test (% pass) ↑	**Symphony** (QM9-only)	**Eq. Diff.** (QM9-only)	**ADiT** (jointly trained)	**PLATOM-1** (QM9-only)	**ZATOM-1** (QM9-only)	**ZATOM-1** (jontly trained)
Atoms connected	99.92	99.88	99.70	99.97	99.95	99.98
Bond angles	99.56	99.98	99.85	99.94	99.90	99.95
Bond lengths	98.72	100.00	99.41	99.96	99.89	99.97
Aromatic ring flat	100.00	100.00	100.00	100.00	100.00	100.00
Double bond flat	99.07	98.58	99.98	100.00	100.00	99.99
Internal energy	95.65	94.88	95.86	99.73	99.63	99.78
No steric clash	98.16	99.79	99.79	99.84	99.81	99.81



(a) Training loss

(b) Validation loss

(c) Posebusters rate

(d) Valid rate

(e) Unique rate

Figure 4: **QM9-only pretraining of ZATOM-1 vs. PLATOM-1.** These plots show the training dynamics of ZATOM-1 compared to the analogous tetrahedral group equivariant PLATOM-1 model described in Section B.1. Due to its equivariance, the Platonic model does not explicitly need to learn symmetries, resulting in significantly faster convergence while achieving improved evaluation metrics. These findings translate to the test results in Table 6.

C ADDITIONAL MODEL DETAILS

Hyperparameters. Table 7 reports the hyperparameters used to pretrain different sizes of ZATOM-1 on MP20 and QM9, and Table 8 shows which hyperparameters were used to pretrain ZATOM-1 on QMOF as well as on GEOM-Drugs jointly with other datasets. Likewise, Table 9 displays the hyperparameters used to finetune ZATOM-1 (80M) for the downstream tasks explored in this work.

For generative inference, Figure 5 shows the results of a hyperparameter sweep, where we find that using a limited number of integration steps and adding (varying levels of) white noise to each continuous modality during sampling generally improves sample validity and uniqueness. Notably, we find that for larger chemical systems, such as the molecules present in GEOM-Drugs, a smaller noise scale for atom positions (e.g., 0.01) is beneficial. Additionally, in all generative contexts, we unexpectedly find that classifier-free guidance worsens the quality of generated samples considerably (e.g., degrading 90% materials validity rates to 40%), motivating us to disable such inference guidance by default.

Table 7: Pretraining hyperparameters across different model sizes when training jointly on QM9 & MP20.

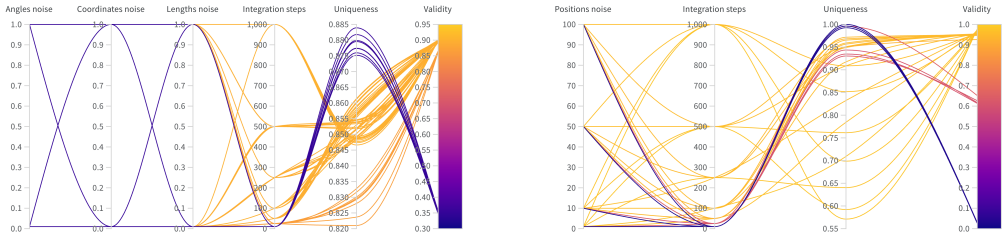
Hyperparam.	**ZATOM-1 (80M)**	**ZATOM-1-L (160M)**	**ZATOM-1-XL (300M)**
# Params.	77,350,264	162,615,544	293,887,096
Hidden size	512	768	1024
# Transformer blocks	16	16	16
# Attn. heads	8	8	8
Train t distribution	Beta(1.8,1)	Beta(1.8,1)	Beta(1.8,1)
$\lambda_{\text{discrete}}$	0.1	0.1	0.1
Learning rate	0.001	0.0005	0.00025
Learning rate schedule	N/A	N/A	N/A
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.0	0.0	0.0
EMA weight	0.999	0.999	0.999
Batch size	256	128	64
# Grad. accum. steps	1	2	4
# Rotation augs.	8	8	8
Effective batch size	32,768	32,768	32,768
# GPUs	16 (80GB)	16 (80GB)	16 (80GB)
Training duration	24h	48h	96h
# Sampling steps	100	100	100
$g(t)$	$\frac{1}{t+0.01}$	$\frac{1}{t+0.01}$	$\frac{1}{t+0.01}$
γ_g ($\rightarrow$ for small molecule coordinates)	0.01 ($\rightarrow$ 50.0)	0.01 ($\rightarrow$ 50.0)	0.01 ($\rightarrow$ 50.0)

Table 8: Pretraining hyperparameters when training ZATOM-1 (80M) on GEOM-Drugs or QMOF.

Hyperparam.	**GEOM-Drugs & MP20**	**QMOF**
# Params.	77,350,264	77,350,264
Hidden size	512	512
# Transformer blocks	16	16
# Attn. heads	8	8
Train t distribution	Beta(1.8,1)	Beta(1.8,1)
$\lambda_{\text{discrete}}$	0.1	0.1
Learning rate	0.001	0.001
Learning rate schedule	N/A	N/A
Optimizer	AdamW	AdamW
Weight decay	0.0	0.0
EMA weight	0.999	0.999
Batch size	32	32
# Grad. accum. steps	1	1
# Rotation augs.	8	8
Effective batch size	4,096	2,048
# GPUs	16 (80GB)	8 (80GB)
Training duration	168h	168h
# Sampling steps	100	100
$g(t)$	$\frac{1}{t+0.01}$	$\frac{1}{t+0.01}$
γ_g	0.01	0.01

Table 9: Finetuning hyperparameters across different downstream prediction tasks, optionally trained in sequential order.

Hyperparam.	**Mol. Properties (80M ❄ + 20M 🔥) →**	**Mol. / Mat. Properties (80M ❄ + 20M 🔥) →**	**Mol. / Mat. Energy & Forces (80M ❄ + 220M 🔥)**
# Dedicated params.	18,104,595	18,104,595	217,255,140
Aux. hidden size	512	512	1024
# Aux. transformer blocks	4	4	8
# Attn. heads	8	8	8
Train t distribution	max(Beta(1.8,1), 0.98)	max(Beta(1.8,1), 0.98)	max(Beta(1.8,1), 1.0)
λ_{forces}	N/A	N/A	5.0
Learning rate	0.001	0.001	0.0003
Learning rate schedule	N/A	N/A	N/A
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.0	0.0	0.0
EMA weight	N/A	N/A	N/A
Batch size	128	32	72
# Grad. accum. steps	2	4	1
# Rotation augs.	8	8	1
Effective batch size	4,096	2,048	1,152
# GPUs	2 (40GB)	2 (80GB)	16 (80GB)
Training duration	48h	192h	120h



(a) Crystals – MP20 (b) Molecules – QM9

Figure 5: **Tuning inference hyperparameters for best performance.** Best generative modeling results for crystals and molecules, balancing sample validity and uniqueness, are generally achieved with a limited number of integration steps (e.g., $T = 100$), a high white noise scale γ_g (e.g., 50) for continuous, non-periodic atom positions, and a low noise scale (e.g., 0.01) for continuous, periodic modalities such as fractional coordinates.

D EVALUATION METRICS

Model development metrics for assessing crystal generation quality. To evaluate the quality of generated crystals during (iterative) model *development*, we adopt the methodology proposed by Xie et al. (2022), Miller et al. (2024), and Joshi et al. (2025). This process involves generating a sample of 1,000 crystals and then assessing them based on their validity and uniqueness. These metrics are described below:

- **Structural Validity**: This metric represents the percentage of crystals where all pairwise atomic distances are at least 0.5 and the crystal volume is 0.1 or greater.
- **Compositional Validity**: In accordance with the SMACT framework (Davies et al., 2019), this is the percentage of crystal compositions that successfully maintain both charge neutrality and electronegativity balance.
- **Overall Validity**: This is the percentage of crystals that satisfy both the structural and the compositional validity criteria mentioned above. Note that this metric is adopted in Figures 3 and 5.
- **Unique**: This measures the percentage of crystals that are structurally unique. Uniqueness is determined via an all-to-all comparison using the Structure Matcher tool within PyMatGen (Ong et al., 2013).

Model evaluation metrics for assessing crystal generation quality. To evaluate the quality of generated crystals during (thorough) model *evaluation*, we adopt the methodology proposed by Duval et al. (2025). This process involves generating a sample of 2,500 crystals and then assessing them based on their validity, uniqueness, novelty, energy, and stability. These metrics are detailed below:

- **Charge Validity**: This metric ensures structures are charge-balanced using oxidation state analysis and bond valence calculations.
- **Distance Validity**: This validates that atomic distances exceed minimum thresholds based on atomic radii.
- **Coordination Validity**: This checks if coordination numbers match expected values for each element.
- **Physical Validity**: This validates density, lattice parameters, crystallographic format, and symmetry.
- **Overall Validity**: This represents the percentage of crystals meeting each of the evaluation validity criteria above.
- **Unique**: This measures the percentage of *valid* crystals that are also structurally unique. Uniqueness is determined via an all-to-all comparison using PyMatGen's Structure Matcher tool.
- **Novel**: This assesses the percentage of *valid*, *unique* crystals that are also structurally novel. Novelty is determined via an all-to-all comparison between the generated crystals and the LeMat-Bulk reference set (Siron et al., 2025) using the Structure Matcher tool of PyMatGen.
- **Formation Energy**: This denotes the average formation energy across multiple MLIPs, namely Orb-v3 (Rhodes et al., 2025), MACE-MP (Batatia et al., 2025), and UMA (Wood et al., 2025).
- **Mean Energy Above Hull**: This represents the average energy above hull across multiple MLIPs (i.e., Orb-v3, MACE-MP, UMA).
- **Relaxation Stability**: This signifies the root mean square deviation (RMSD) between original and relaxed atomic positions.
- **Stability Ratio**: This denotes the fraction of structures with energy above hull ≤ 0 eV/atom (thermodynamically stable).
- **Stable, Unique, & Novel (SUN)**: This represents the fraction of structures that are simultaneously valid, stable (energy above hull ≤ 0 eV/atom), unique, and novel, with the number of structures that meet each of these criteria shown in proceeding parentheses.

- **Metastability Ratio**: This signifies the fraction of structures with energy above hull ≤ 0.1 eV/atom (metastable).
- **Metastable, Unique, & Novel (MSUN)**: This denotes the fraction of structures that are simultaneously valid, metastable (energy above hull ≤ 0.1 eV/atom), unique, and novel, with the number of structures that meet each of these criteria shown in proceeding parentheses.

Model development/evaluation metrics for assessing molecule generation quality. For the evaluation of generated molecules, we utilize the protocol developed by Hoogeboom et al. (2022), Daigavane et al. (2024), and Joshi et al. (2025). From a sample of 10,000 molecules, we calculate rates for validity and uniqueness, and also determine success rates for seven distinct sanity checks provided by the PoseBusters toolkit (Buttenschoen et al., 2024):

- **Validity**: This measures the percentage of generated molecules for which a canonical SMILES representation can be successfully identified using RDKit (Landrum, 2013).
- **Uniqueness**: Among the molecules deemed valid, this metric calculates the percentage that possess a unique SMILES string.
- **All-atoms Connected**: This check quantifies the percentage of molecules in which every atom is part of a single connected component, meaning a continuous path of bonds links all atoms.
- **Reasonable Bond Angles/Lengths**: This is the fraction of molecules where all bond angles and lengths fall within a tolerance range, specifically between 0.75 and 1.25 times the standard values derived from distance geometry.
- **Aromatic Rings Flatness**: This metric counts the percentage of molecules where all atoms in 5- or 6-membered aromatic rings are positioned within 0.25Å of their nearest shared plane.
- **Double Bonds Flatness**: This is the percentage of molecules where the atoms of aliphatic carbon-carbon double bonds, along with their four neighboring atoms, lie within 0.25Å of the closest shared plane.
- **Reasonable Internal Energy**: This check determines the percentage of molecules whose calculated internal energy does not exceed 100 times the average energy observed in an ensemble of 50 conformations of the input molecule (Harris et al., 2023).
- **No Internal Steric Clash**: This metric represents the percentage of molecules where the distance between any pair of non-covalently bound atoms is greater than 0.8 times the lower bound established by distance geometry.

In summary, the validity and uniqueness metrics primarily confirm the chemical processability of the generated molecules by RDKit. In contrast, the PoseBusters sanity checks scrutinize the physical realism of the generated 3D structures, applying a range of criteria from geometric constraints to energetic plausibility.

E Additional Results

Exploring metal-organic framework generation. Table 10 investigates ZATOM-1's ability to generate metal-organic frameworks (MOFs), a novel class of materials with several promising applications (Kitagawa et al., 2014). With minimal hyperparameter tuning, we find that ZATOM-1 can generate valid MOFs despite their numerous chemical and geometric constraints, such as requiring a balance of metal, carbon, and hydrogen atoms without steric clashes. Nonetheless, the model (as currently evaluated) has not fully converged, indicating that its proportion of successfully generated MOFs could likely be improved with further training.

Balancing multi-task molecule and material property prediction. Table 11 presents the results of performing a second stage of finetuning for multi-task molecule and ($\Delta\varepsilon$) materials property prediction using QM9 and Matbench, starting from ZATOM-1's first stage of QM9 finetuning for multi-task molecule property prediction. In exchange for marginal molecule property prediction performance, these results demonstrate that ZATOM-1's generative *multi-task* property predictions can

Table 10: **Metal-organic framework generation results on QMOF.** We report sanity checks from MOFChecker (Jin et al., 2025) for 1,000 sampled MOFs, comparing methods trained only on QMOF for at least 10,000 epochs. (↑/↓ indicate higher/lower is better, respectively, and * denotes a model that has not fully converged)

Test%	QMOF ADiT	QMOF ZATOM-1*
Has carbon ↑	100.0	100.0
Has hydrogen ↑	99.6	100.0
Has atomic overlap ↓	8.3	12.8
Has overcoord. C ↓	23.6	3.2
Has overcoord. N ↓	1.5	0.2
Has overcoord. H ↓	1.0	3.1
Has undercoord. C ↓	60.0	75.3
Has undercoord. N ↓	39.1	31.5
Has undercoord. rare earth ↓	0.4	0.0
Has metal ↑	100.0	100.0
Has lone molecule ↓	72.9	88.5
Has high charge ↓	0.9	0.8
Has suspicious terminal oxo ↓	2.6	1.9
Has undercoord. alkali ↓	1.0	0.2
Has geom. exposed metal ↓	7.0	2.8
Validity rate (all passed) ↑	15.7	8.4

Table 11: **Property prediction for 3D molecules (QM9) / materials (Matbench) with joint molecule and materials finetuning.** Results are reported as *test* mean absolute errors (for ZATOM-1, with standard deviations over three runs with different random seeds). † denotes using different data partitions. / represents a task for which both (QM9) molecule and (Matbench) material property labels are available. The best (or second-best) results for each (optimized or unoptimized) category are in **bold** (or underlined). Notably, finetuning jointly for multi-task molecule and ($\Delta\varepsilon$) materials property prediction marginally trades off ZATOM-1's molecule property prediction performance for expanded applicability of its property predictions to both molecules and materials simultaneously.

Model / Metrics ↓	# Params	α (a_0^3)	$\Delta\varepsilon$ (meV)	$\varepsilon_{\text{HOMO}}$ (meV)	$\varepsilon_{\text{LUMO}}$ (meV)	μ (D)	C_ν (cal/mol K)	G^{ATOM} (meV)	H^{ATOM} (meV)	R^2 (a_0^2)	U^{ATOM} (meV)	U_0^{ATOM} (meV)	$ZPVE$ (meV)
Optimized single-task learning													
DimeNet++†	5M	.044	33.0 / **199**	25	20	.030	.023	8.00	7.00	.331	6.00	6.00	1.21
EGNN†	1M	.071	48.0 / —	29	25	.029	.031	12.0	12.0	.106	12.0	11.0	1.55
PaiNN	1M	.045	46.0 / —	28	20	.012	.024	**7.35**	**5.98**	.066	**5.83**	**5.85**	1.28
TorchMD-NET	7M	.059	36.0 / —	20	18	.011	.026	7.62	6.16	**.033**	6.38	6.15	1.84
SphereNet	2M	.046	32.0 / —	23	18	.026	**.021**	8.00	6.00	.292	7.00	6.00	**1.12**
SEGNN†	1M	.060	42.0 / —	24	21	.023	.031	15.0	16.0	.660	13.0	15.0	1.62
EQGAT	-	.053	32.0 / —	20	16	.011	.024	23.0	24.0	.382	25.0	25.0	2.00
Equiformer	4M	.046	30.0 / —	15	14	.011	.023	7.63	6.63	.251	6.74	6.59	1.26
EquiformerV2	11M	.050	**29.0** / —	**14**	**13**	.010	.023	7.57	6.22	.186	6.49	6.17	1.47
PϴNITA	-	**.038**	30.4 / —	16	15	.012	.024	8.63	8.04	.235	8.67	8.31	1.29
Platonic Transformer	-	.049	37.4 / —	22	17	**.010**	.024	12.0	12.0	.222	11.9	13.0	1.30
Unoptimized single-task learning													
AIM-STL†	-	.181	— / —	61	54	**.067**	.072	66.2	63.9	**.503**	64.2	58.8	**4.54**
Unoptimized multi-task learning													
AIM-MTL-Scalar†	-	.268	— / —	59	71	.089	.103	113	112	4.18	112	112	9.82
AIM-MTL-Matrix†	-	.251	— / —	61	72	.088	.103	112	118	4.07	116	116	12.2
QM9-pretrained ZATOM-1 (OURS)	80M ❄ + 20M 🔥	.105±.000	52.0±0.1 / 362±1	37±0.1	36±0.1	.102±0.001	.054±0.000	47.2±0.2	48.5±0.2	3.93±0.00	48.6±0.2	48.7±0.2	5.03±0.01
Jointly pretrained ZATOM-1 (OURS)	80M ❄ + 20M 🔥	**.095±.000**	**46.2±0.1 / 354±1**	**34±0.1**	**32±0.1**	.087±.000	**.047±.000**	**39.2±0.2**	**42.1±0.2**	3.33±0.01	**41.6±0.2**	**41.6±0.2**	4.62±0.02

effectively be applied to both 3D molecules and materials, to our knowledge a first-of-its-kind result. In contrast, for materials property prediction in Table 4 of the main text, we report zero-shot $\Delta\varepsilon$ results for the model weights performing best specifically for multi-task molecule property prediction, to illustrate the broad applicability of ZATOM-1's pretrained molecule and material embeddings.

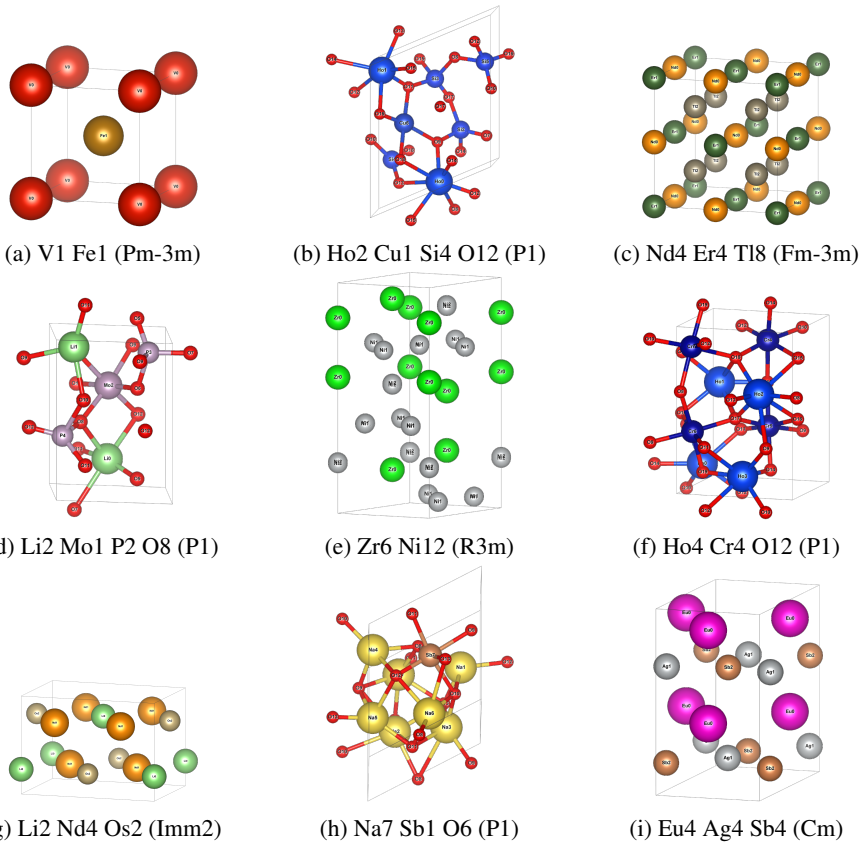
Extension to energies and forces. Table 12 explores ZATOM-1's ability to predict energies and forces to serve as an MLIP. Through this preliminary set of experiments, we find that the performance of ZATOM-1's material force predictions for MPtrj slightly surpasses that of Orb-v1, a state-of-the-art, pretrained model for this dataset. Results for OMol25 demonstrate that ZATOM-1 can effectively balance the performance of its force predictions for molecules and materials jointly, highlighting its potentially broad applicability in molecular and materials sciences. Despite these promising first steps, ZATOM-1's energy predictions could be notably improved, suggesting that its representations derived through generative pretraining on QM9 & MP20 are most informative of (equivariant) atomistic forces. As potential future directions, exploring bespoke additive noise schedules for each of the model's input modalities during finetuning, or pretraining using more structurally diverse datasets such as OMol25 and LeMat-Bulk, could potentially address this gap between energy and force finetuning performance for both molecules and materials.

Table 12: **Energy and force prediction for 3D materials (MPtrj) / molecules (OMol25-4M).** The results are reported as *validation* Energy (in meV or meV/Atom) and Force (in meV/Å) mean absolute error (MAE). All models directly predict forces, rather than using the gradient of their predicted energies with respect to input 3D coordinates for conservative force prediction, to considerably increase their computational efficiency.

Model	# Params	Energy MAE/Atom (meV) ↓	Energy MAE (meV) ↓	Force MAE (meV/Å) ↓
Material Interatomic Potential Prediction on MPtrj				
Non-pretrained Orb-v1	∼30M	-	350.00	122.00
Denoising-pretrained Orb-v1	∼30M	-	210.00	32.00
Jointly pretrained ZATOM-1	80M ❄ + 220M 🔥	112.81	2631.22	30.92
Molecule Interatomic Potential Prediction on OMol25-4M				
eSEN-md-d.	∼51M	1.32	77.13	6.78
Jointly pretrained ZATOM-1	80M ❄ + 220M 🔥	34.45	2512.78	62.99

F VISUALIZATION

Generated crystals, molecules, and MOFs. Figures 6 and 7 show representative crystals and small molecules, respectively, sampled from ZATOM-1 trained jointly on the QM9 and MP20 datasets. Additionally, Figure 8 visualizes MOFs generated by ZATOM-1 trained on the QMOF dataset. Overall, ZATOM-1 generates compositionally diverse crystals across multiple spacegroups and chemically valid, structurally diverse molecules, indicating that it learns a rich generative latent space rather than a redundant set of representations.



(a) V1 Fe1 (Pm-3m) (b) Ho2 Cu1 Si4 O12 (P1) (c) Nd4 Er4 Tl8 (Fm-3m)

(d) Li2 Mo1 P2 O8 (P1) (e) Zr6 Ni12 (R3m) (f) Ho4 Cr4 O12 (P1)

(g) Li2 Nd4 Os2 (Imm2) (h) Na7 Sb1 O6 (P1) (i) Eu4 Ag4 Sb4 (Cm)

Figure 6: Generated crystals from ZATOM-1 trained jointly on MP20 materials and QM9 molecules.

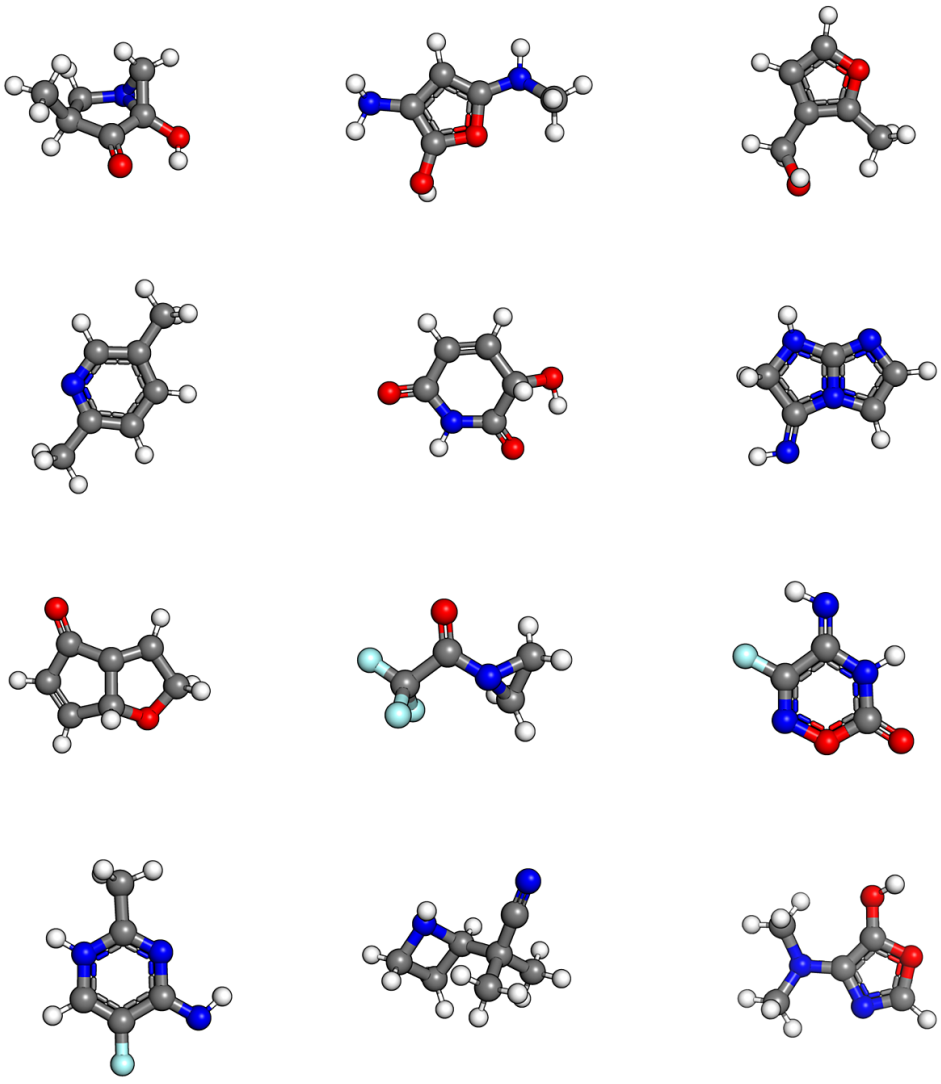
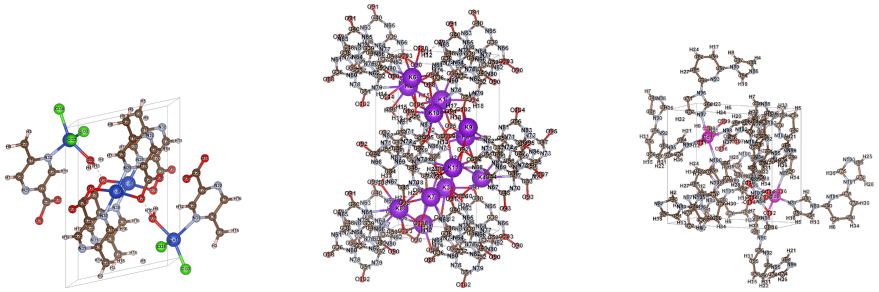
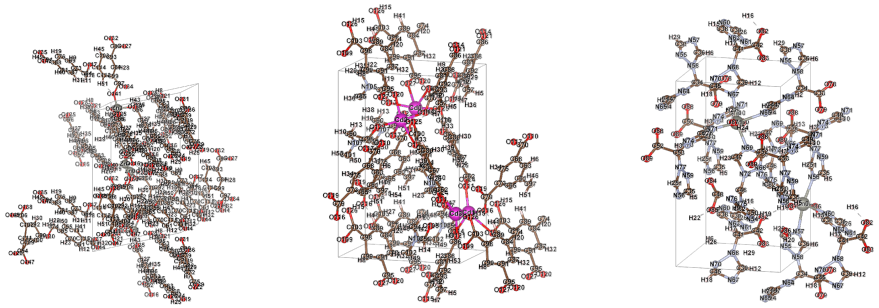


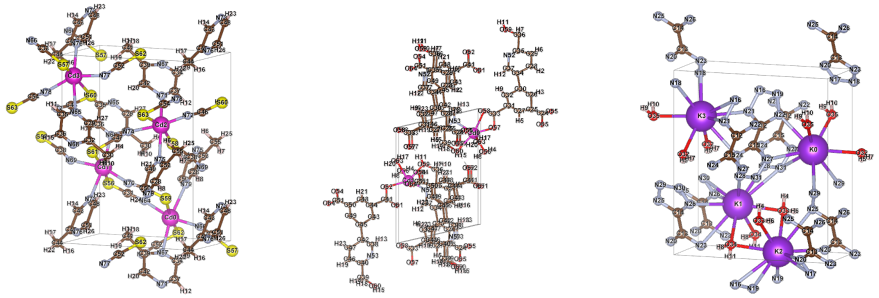
Figure 7: Generated molecules from ZATOM-1 trained jointly on MP20 materials and QM9 molecules.



(a) Cu3 H14 C12 N4 Cl2 O6 (P1) (b) K12 H24 C24 N28 O24 (P1) (c) Cd2 H36 C44 N24 O12 (P1)



(d) Zn4 H48 C64 O32 (P1) (e) Cd4 H52 C48 N4 O20 (P1) (f) Zn2 H28 C24 N24 O12 (P1)



(g) Cd4 H24 C28 S8 N16 (P1) (h) Cd2 H22 C28 N2 O12 (P1) (i) K4 H8 C4 N16 O4 (P1)

Figure 8: Generated metal-organic frameworks from ZATOM-1 trained on QMOF metal-organic frameworks.