Supplementary Information

MolCrystalFlow: Molecular Crystal Structure Prediction via Flow Matching

Cheng Zeng*, Harry W. Sullivan, Thomas Egg, Maya M. Martirossyan, Philipp Höllmer, Jirui Jin, Richard G. Hennig, Adrian Roitberg, Stefano Martiniani, Ellad B. Tadmor, Mingjie Liu*

March 2, 2026

Table of contents

*c.zeng@ufl.edu, mingjieliu@ufl.edu

S-1 Mathematical notations

We list the mathematical notations used throughout this work below.

- h is used to denote the node embedding for both atoms and molecular building blocks.
- F, R, and L denote fractional coordinates, rotational orientation in the format of rotation matrices, and lattice matrices, respectively. $\mathcal{C}$ denotes a molecular crystal.
- X denotes the atomic positions of molecular building blocks, and Y denotes centered positions.
- m_ξ, where $\xi \in \{F, R\}$, is used to denote the message passing related to the corresponding modalities, including fractional coordinates F and rotational orientations R. m_i denotes atomic mass for atom i.
- t denotes the time steps in a flow matching setting, in the range between 0 and 1.
- $u_t(\xi_t)$, where $\xi \in \{F, R\}$, is used to denote the velocity fields at the time step t for different modalities.
- ω denotes the angle of rotation in their corresponding axis-angle representation, and $\vec{e}$ denotes the unit vector for the corresponding rotational axis. ρ denotes the azimuthal (longitude) angle and κ denotes the inclination (colatitude) angle for a projection of $\vec{e}$ in a polar coordinate frames.
- χ denotes the binary axis-flip states for internal molecular building block frames.
- s_{u_ξ}, where $\xi \in \{F, R\}$, denotes the velocity annealing scaling factors for velocity fields u_ξ.
- Q denotes a general rotational operation.
- $\tau(\cdot)$ represents a wrapping function to map any values back to a torus in the range of $[0, 1)$.
- A denotes the atomic species in molecular building blocks.
- D denotes a displacement vector connecting geometric centroid to atomic-mass weighted centroid (or to the nearest neighboring atom when the molecule is symmetric).
- w_i denotes the learnable weight for atom i used to construct the building block embedding.
- ϕ_ξ, where $\xi \in \{m, h, w\}$, denotes a multilayer perceptron (MLP) that operates on the corresponding components.
- $\psi_{\text{FT}}(\cdot)$ represents a Fourier embedding based on trigonometric functions.
- $\Psi(\cdot)$ denotes a concatenated embedding for multiple different embeddings.
- $\Delta(\cdot)$ denotes a change of state.
- Z denotes the number of molecules in a crystal.

- α, β, and γ are used to denote the molecular crystal polymorphs ordered by their PBE-MBD energies. $\beta_{1,2}$ are used to denote the momentum parameters for an AdamW optimizer [1].
- d_{ij} denotes the Euclidean distance between atom i and j.
- λ_ξ denotes the loss function weight for the corresponding modality ξ.
- θ denotes the neural network parameters.
- u_k denote the eigenvectors obtained from PCA, and ν_k denote the projection of displacement vector D on to the eigenvectors, where $k \in \{1, 2, 3\}$.

S-2 Model hyperparameters

Figure S1 shows a list of hyperparameters used in the building block embedder network, in the MolCrystalNet model for flow matching and for training.

Table S1: Hyperparameters used for MolCrystalFlow and MOFFlow models.

Component	**Hyperparameters**
EGNN BB embedder	Node/edge dimension: 64; unique elements: 7 (Thurlemann), 12 (OMC25-MCF); cutoff radius: 6 Å; 6 graph convolution layers
Attention pooling	3-layer MLP with sigmoid output; BB embedding dimension: 128
MolCrystalNet for flow matching	Time embedding dimension: 1024; input BB embedding dimension: 128; Update blocks: 6; node/edge dimensions: 256 / 128
Centroid/Angular embeddings	Fourier embedding: 32 basis functions; inclination-angle embedding dimension: 32
Batch size	Training: 128 (MolCrystalFlow), 160 (MOFFlow); Inference: 512
Training epochs	300 (Thurlemann); 500 (OMC25-MCF)
Optimizer	AdamW; learning rate: 1×10^{-4}; weight decay: 1×10^{-2}; Momentum parameters: $\beta_1 = 0.9$, $\beta_2 = 0.98$

S-3 Exploratory data analysis for the two open-source molecular datasets

S-3.1 The structure filtered out in the Thurlemann Molecular Crystal Data.

Figure S1 shows the crystal structure filtered out from the Thurlemann dataset.

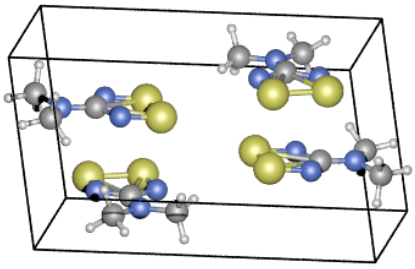

Figure S1: Outlier structure in the Thurlemann data filtered out. This structure is identified to exhibit four molecules in the crystal but the original data indicates a $Z = 2$. To exercise caution, we do not include this structure for training MolCrystalFlow and MOFFlow models. C, H, N, and S are in gray, white, blue and yellow, respectively.

S-3.2 Exploratory data analysis for the Thurlemann molecular crystal data

Figure S2 shows the exploratory data analysis on the Thurlemann molecular crystal data. It comprises seven elements with elemental compositions of 41.6, 43.4, 8.1, 1.1, 0.7, 5.0, 0.01% for respective C, H, O, S, Cl, N, and F elements (Figure S2a). The number of atoms per molecular building block ranges from 5 to 89 atoms, with the most frequent number of atoms located at 25 atoms (Figure S2b). The majority of crystals consist of 2 moleucles in one crystal, followed by 4 and 1 crystals and with a handful cases comprising 3 and 8 molecules (Figure S2c). The more molecules in the crystal, the smaller numbers of molecular building block will be (Figure S2d).

S-3.3 Exploratory data analysis for the OMC25-MCF data

Figure S2 shows the exploratory data analysis on the selected subset of OMC25 dataset (OMC25-MCF). It comprises 12 elements, including, B, Br, C, Cl, F, H, I, N, O, P, S, Si where C, H, O and N are the top-4 most frequent elements (Figure S2a). The number of atoms per molecular building block ranges from 4 to 164 atoms, with the most frequent number of atoms located at 36 atoms (Figure S2b). The majority of crystals consist of 2 and 4 moleucles in one crystal, followed by 8 and 1 crystals and with a small number of cases comprising 6 and 16 molecules (Figure S2c). Likewise, more molecules in a crystal correspond to smaller numbers of atoms per molecular building blocks (Figure S2d).

S-4 Supporting results

S-4.1 Building block embedding from trained EGNN

Figure S4 shows the UMAP of building block embeddings [2]. One can see that there is weak correlation between atom counts of each BB and the first two dimensions of UMAP (Figure S4a). Pearson correlation coefficients with number of atoms per building blocks are 0.05 and 0.43 for respective UMAP-1 and UMAP-2 components. Moreover, the test data set are within the training data distribution in these two dimensions (Figure S4b), suggesting that a generalized representation across training and test data points has been learned.

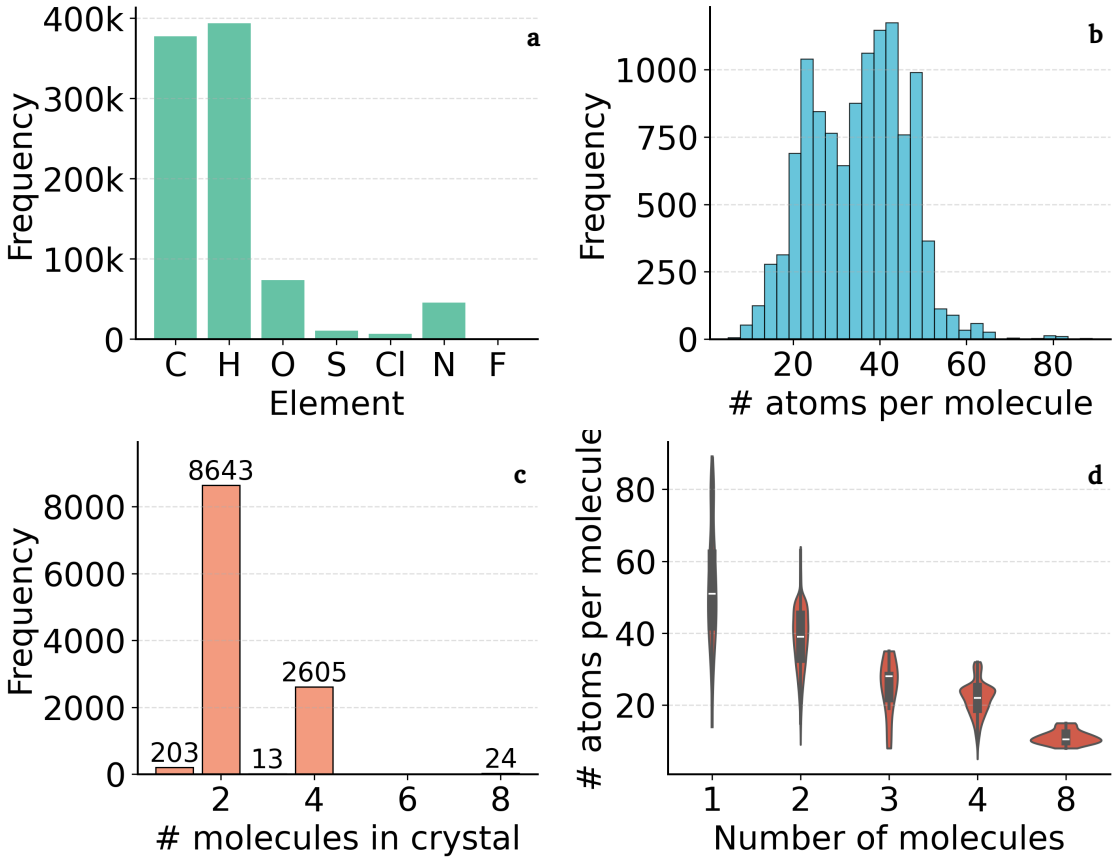

Figure S2: EDA on the Thurlemann data. a, Frequency of the seven elements C, H, O, S, Cl, N and F. b, Number of atoms per molecular building blocks. c, Number of molecules in one molecular crystal. d, Distribution of numbers of atoms per building block against numbers of molecules in the crystal.

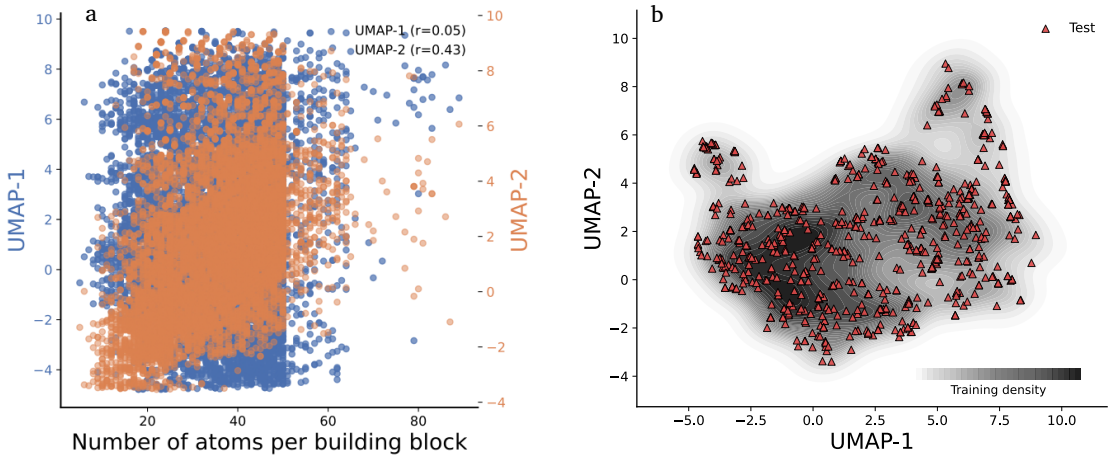

Figure S4: UMAP of EGNN-derived building block (BB) embedding. a, scatter plots of UMAP-1 (blue) and UMAP-2 (orange) against number of atoms per building block. b, UMAP-2 versus UMAP-1 for both training data (kernel density fit density) and test data (red triangles).

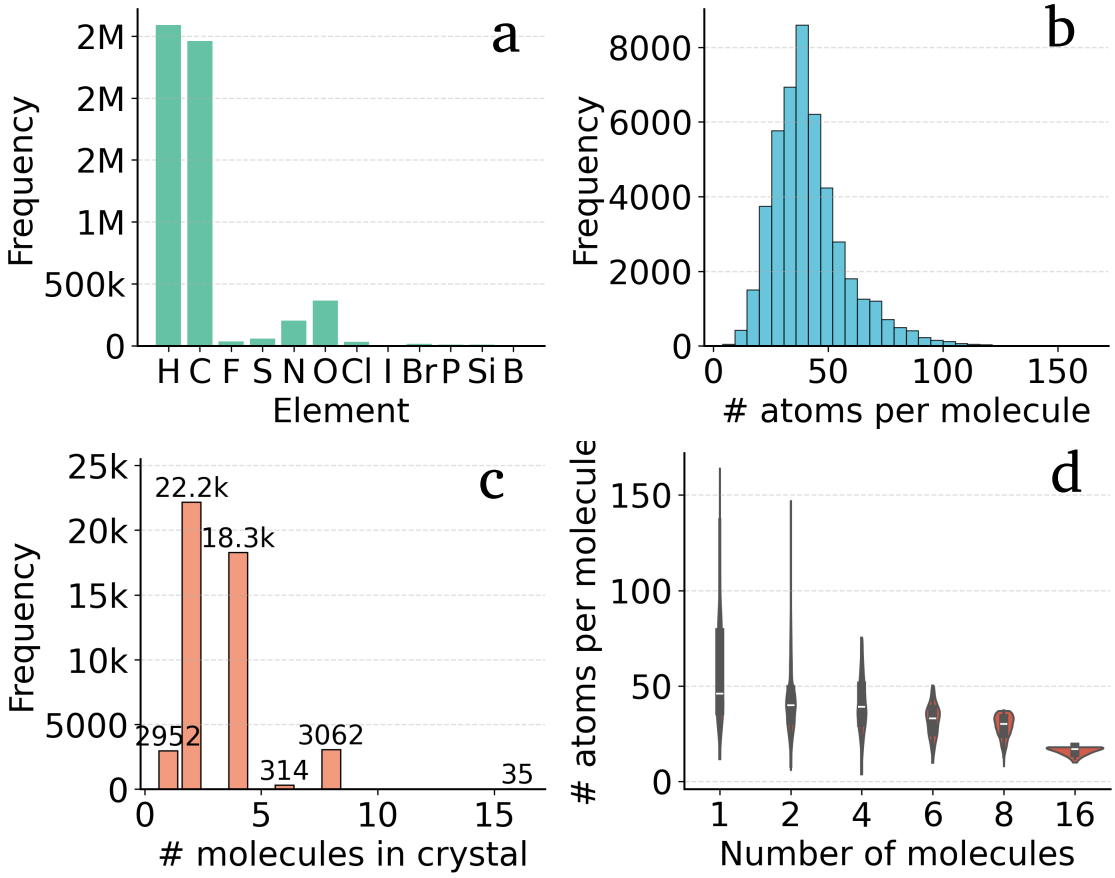

Figure S3: EDA on the OMC25-MCF data. a, Frequency of the 12 elements b, Number of atoms per molecular building blocks. c, Number of molecules in one molecular crystal. d, Distribution of numbers of atoms per building block against numbers of molecules in the crystal.

S-4.2 Effect of integration timesteps for MolCrystalFlow models

Figure S5 shows the variation of structure matching rates as a function of the number of integration time steps. Below 40 timesteps, the structure matching rates increase gradually, and then reach a small plateau around 50 timesteps. After that, one can see fluctuations of matching rates between 50 and 100 time steps and a slight increase after 100 steps and it drops again after 500 steps. To strike a good balance between efficiency and accuracy, we chose 50 integration timesteps throughout the main text.

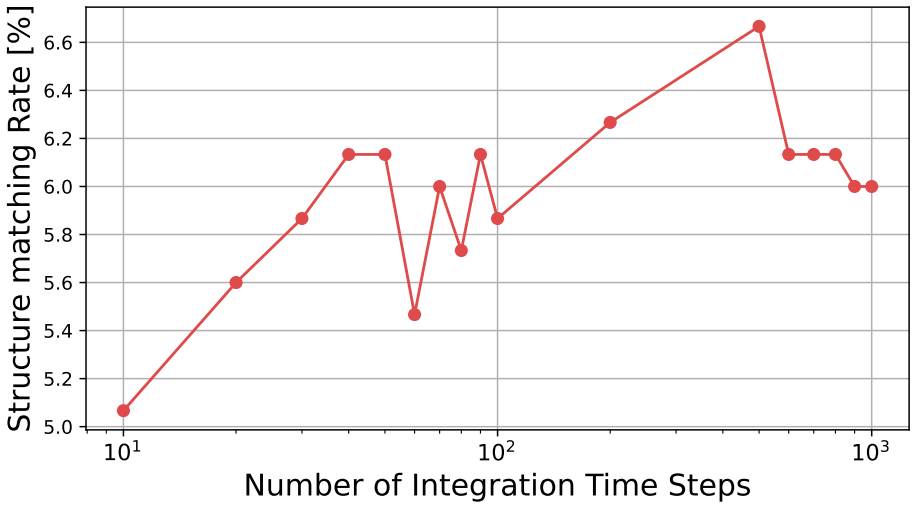

Figure S5: Integration timestep effect. Structure matching rates on the Thurlmann test data are evaluated over 10 samples, and the matching criteria use stol=0.8, ltol=0.3 and atol=10.

S-4.3 Effect of number of samples for MolCrystalFlow and baseline models

Figure S6 shows the matching rate variation as a function of numbers of samples. Figure S6a demonstrates that the substantial performance gap between MolCrystalFlow and MOFFlow (exceeding 10% at saturation) cannot be closed simply by increasing the number of samples for MOFFlow. Figures S6b and c further show that Genarris-3 with rigid-press achieves higher matching rates under a tighter site tolerance ($\mathrm{stol} = 0.8$) as the number of samples increases, whereas MolCrystalFlow performs comparatively better under a looser tolerance ($\mathrm{stol} = 1.0$). These results suggest that MolCrystalFlow is more effective at broadly exploring the configurational space, while Genarris-3 tends to yield more geometrically precise structures among the initially generated and optimized candidates.

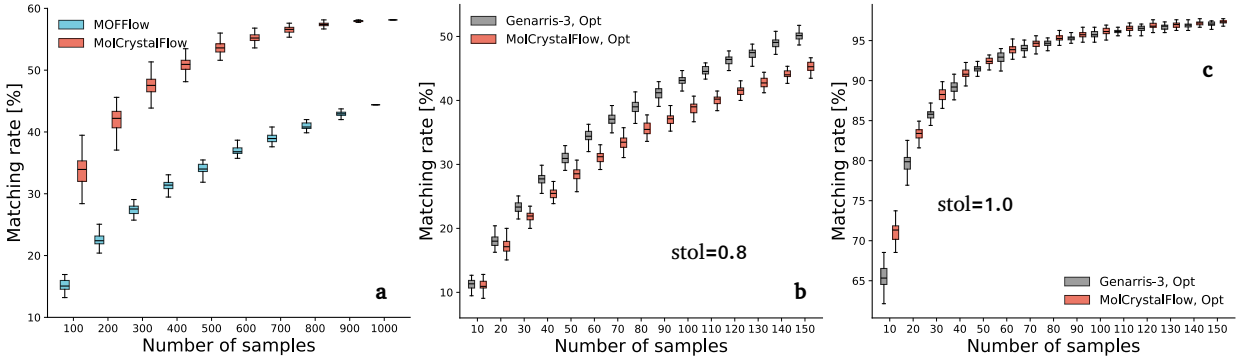

Figure S6: Effect of sample counts. a, MolCrystalFlow versus MOFFlow at different numbers of samples over directly generated structures using stol=0.8. b, MolCrystalFlow against Genarris-3 at different numbers of samples over generated, rigid-body optimized structures using stol=0.8. c, MolCrystalFlow against Genarris-3 at different numbers of samples over generated, rigid-body optimized structures using stol=1.0. The box plot statistics are obtained by bootstrapping without replacement for 50 times.

S-4.4 Lattice volume deviation of example test structures

The relative lattice volume deviation for sixty structures evenly indexed among the 750 test structures for the Thurlemann molecular crystal data. Overall, most structures show a mean lattice volume deviation below 5%, and the three example structures with the highest prediction deviations either consist of 4 molecules or exhibit complex building blocks.

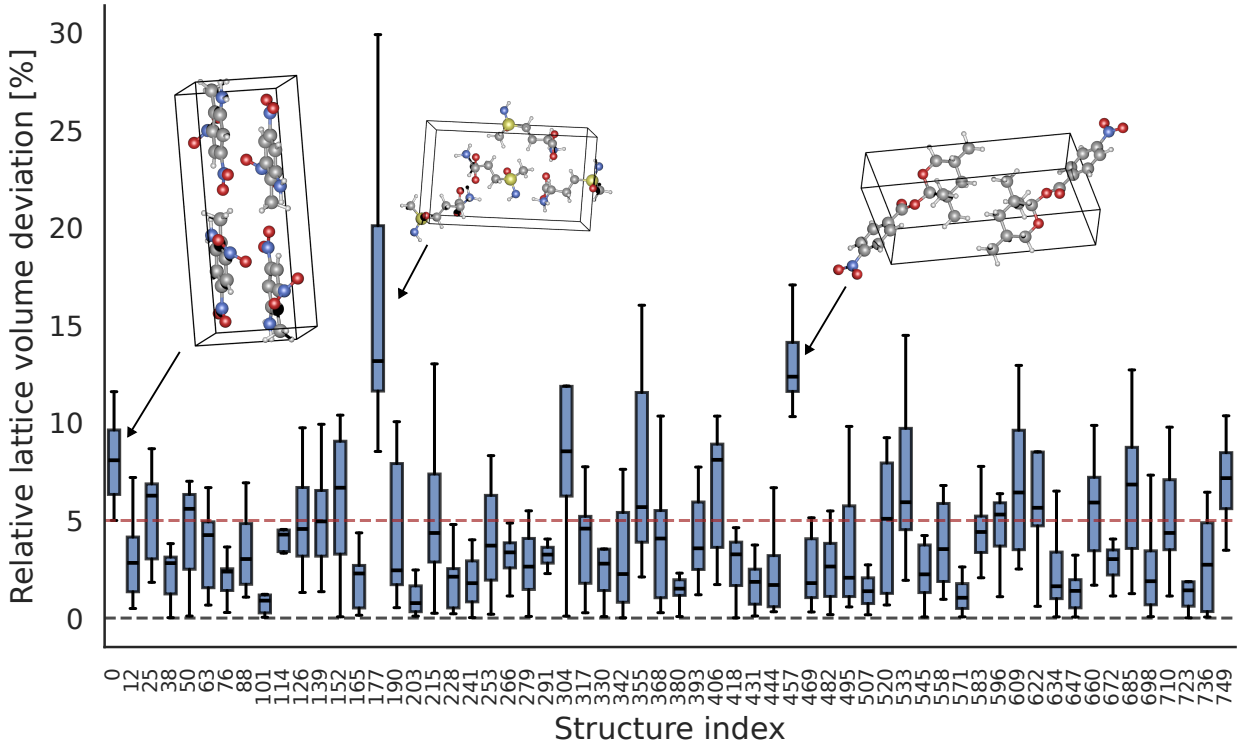

Figure S7: Relative lattice volume prediction per structure. Sixty structures were chosen and the box plot summarizes the statistics of the same structure across ten samples. Three ground-truth structures whose prediction deviations are among the highest are shown in the Figure inset.

S-4.5 Results for models trained on the OMC-MCF data

Figure S8 shows the structure matching rates and lattice volume deviation results. The matching rates are 5.4% and 50.8% for respective stol=0.8 and stol=1.0. Average of lattice volume deviations is 6.35%, much larger than that of the Thurlemann data (3.86%) since the lattice volume spans in a much wider range up to 4000 $Å^3$.

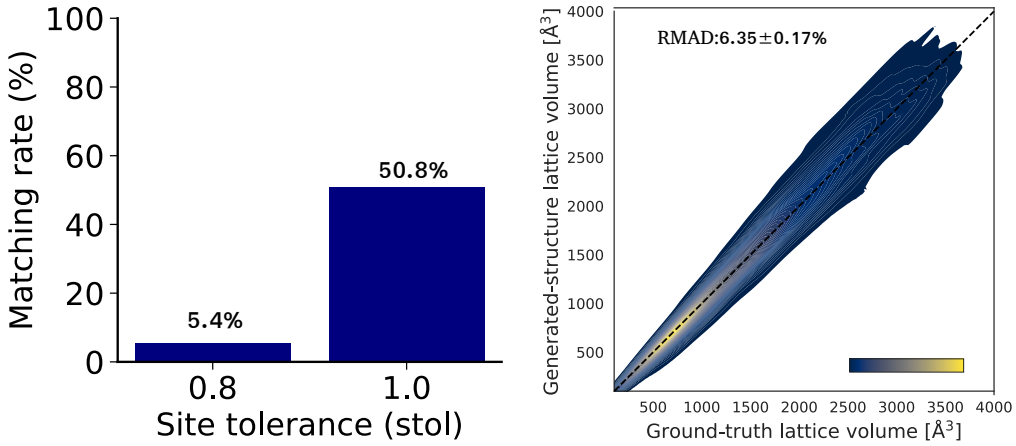

Figure S8: Structure matching and lattice volume comparison results evaluated on the 1000 test structures. Structure matching uses ltol=0.3, atol=10 and stol=0.8 or stol=1.0 with ten samples. Relative mean absolute deviation is defined by the absolute difference divided by the ground-truth volume.

As a comparison, Figure S9 shows the structure matching rates and lattice volume deviation results using MOFFlow trained on the same OMC25 data. The matching rates with 10 samples are 1.4% and 20.7% using stol=0.8 and stol=1.0, respectively. The mean of lattice deviation is 32%, much larger than that using MolCrystalFlow.

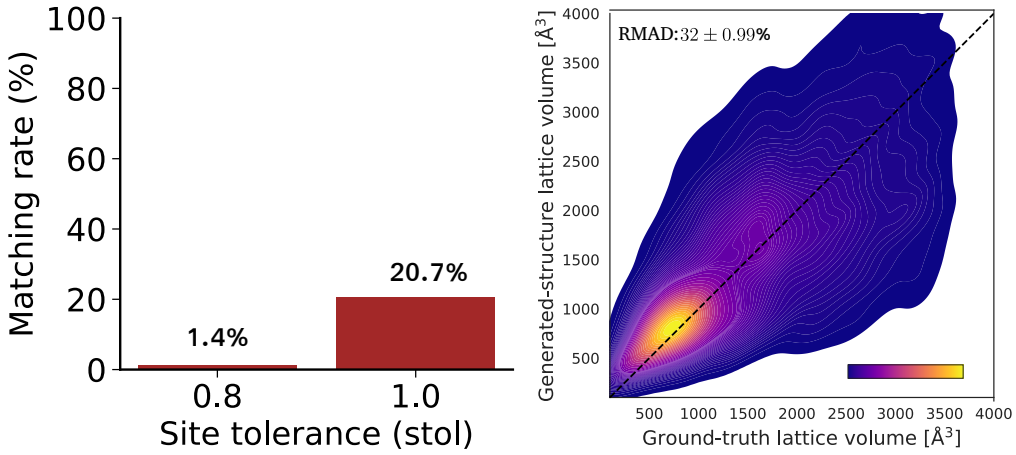

Figure S9: Structure matching and lattice volume deviation results using MOFFlow trained on the same OMC25 data. Structure matching uses ltol=0.3, atol=10 and stol=0.8 or stol=1.0 with ten samples. Relative mean absolute deviation is defined by the absolute difference divided by the ground-truth volume.

S-5 Crystal structure prediction using the 3rd CCDC CSP competition compounds

S-5.1 UMA-OMC Energy landscapes for target VIII, X and XI over MolCrystalFlow generated crystals

Figure S10 shows the energy landscape obtained with the UMA-OMC model. The majority of generated structures remain within the 20 kJ/mol to the lowest-energy structures for small molecules such as C3H4N2O2 (>70%) and C3H7N ($\sim$100%), while around 25% to 45% data points for the much larger molecule C9H9N3O5, suggesting the high quality of sampled structures. The relative lattice energy shows that most sampled structures are concentrated at relatively high lattice energies and a small fraction of structures have energies close to the experimental structures. In addition, the stochasticity plays a role in the generation as Round 1 identifies a structure whose energy is almost identical to the experimental one for the molecule C3H4N2O2, whereas Round 2 identifies a structure for the large molecule C9H9N3O5. One thing to note is that although UMA-OMC does not show structures close to the experimental energy for target XI, but it is due to the inaccuracy of UMA-OMC on these structures since the DFT calculatations actually confirm that many of these structures are with low energies.

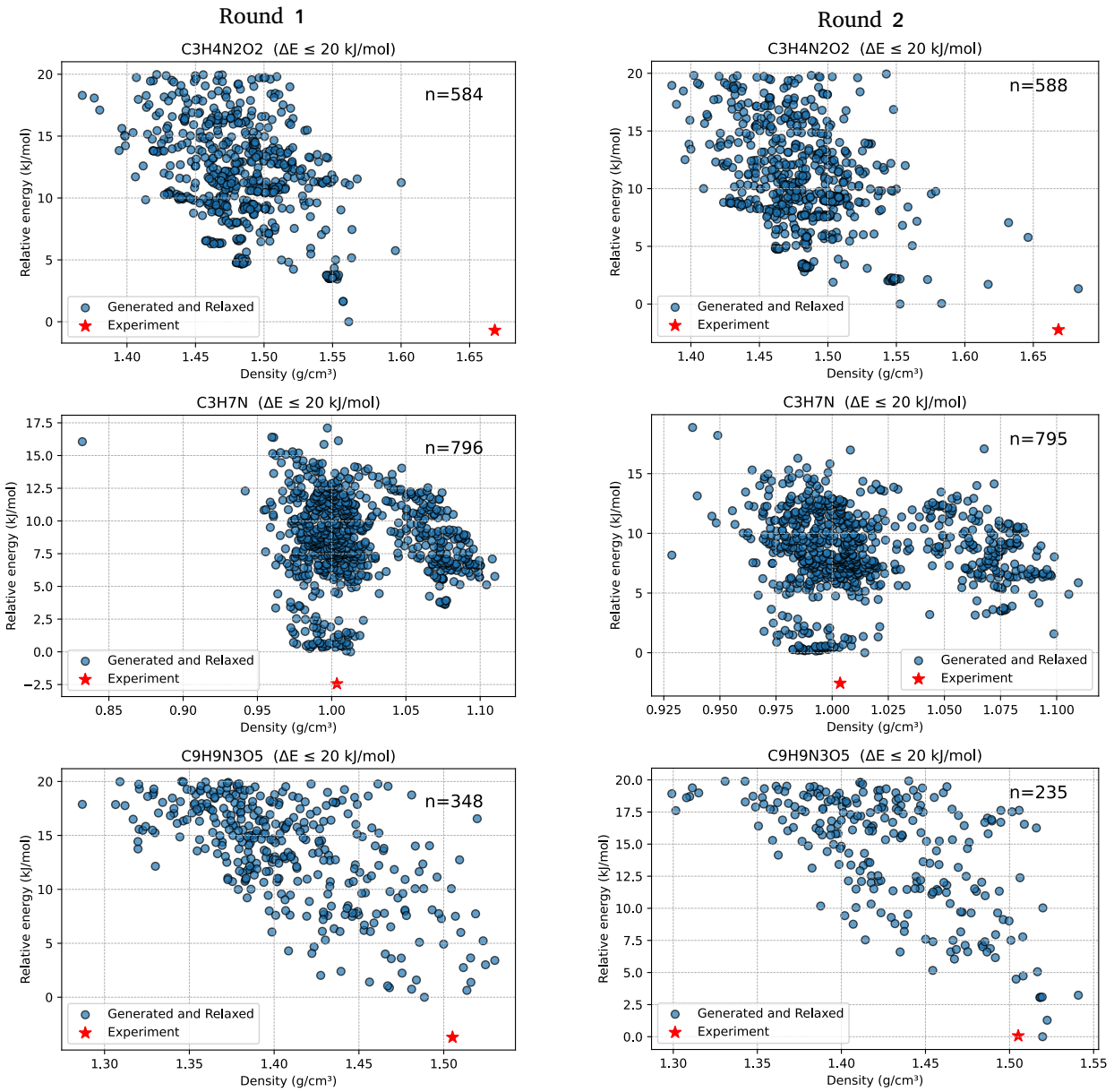

Figure S10: UMA-OMC Energy landscapes at 0K for target VIII, X and XI over MolCrystalFlow generated crystals Structures are generated using MolCrystalFlow and optimized via the UMA-OMC model. Energy of experimental structure is evaluated on the UMA-OMC relaxed structure, while the density is calculated directly on the experimental structure. Number of remaining data points whose lattice energy is within 20 kJ/mol to that of the minimum are shown by the number 'n' as an inset. Run 1 and 2 used different random seeds for structure generation. Total number of initially generated structures is 800.

S-5.2 UMA-OMC Energy landscapes for target VIII, X and XI over Genarris generated crystals

Figure S11 shows the energy landscape obtained with the UMA-OMC model for Genarris-3 generated structures. Overall the success rates of relaxed structures with energies to the minimum below

20 kJ/mol are smaller than the corresponding ones generated by the MolCrystalFlow model. However, Generris is able to locate more data points close to experimental structures for the smaller target XI while the MolCrystalFlow model samples more structures close to experiments for the more challenging target X, even though Genarris-3 samples 800 more structures for $Z = 8$.

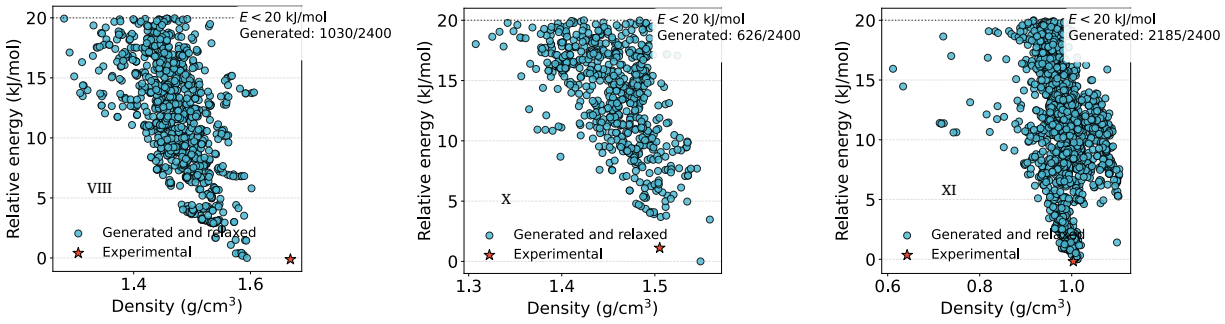

Figure S11: UMA-OMC Energy landscapes at 0K for target VIII, X and XI over Genarris generated crystals Structures are generated using Genarris-3 with $Z \in \{2, 4, 8\}$ with subsampled space groups. Each Z has 800 structures and there are 2400 structures in total. Energy of experimental structure is evaluated on the UMA-OMC relaxed structure, while the density is calculated directly on the experimental structure. Number of remaining data points whose lattice energy is within 20 kJ/mol to that of the minimum are shown as inset.

S-5.3 Genarris generated PBE-MBD lowest-energy polymorphs compared to experimental structures for target VIII, X, and XI

Figure S12 shows the packing similarity analysis for the PBE-MBD identified lowest-energy polymorphs compared to the experimental structures. Genarris-3 was able to find the experimental polymorph for target VIII while failing to do so for target XI and X. The RMSEs for target XI and X are similar to the values of polymorphs identified by MolCrystalFlow.

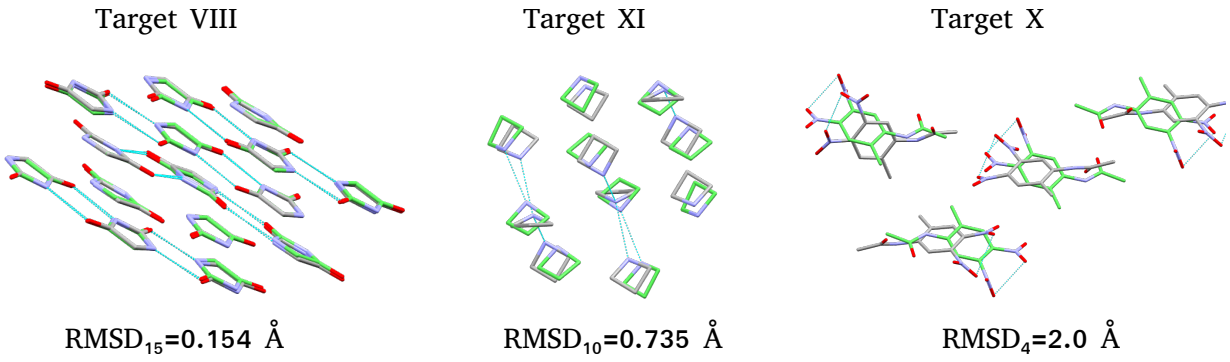

Figure S12: RMSE between PBE-MBD identified lowest-energy structures generated by Genarris-3 and the corresponding experimental structures. The RMSE is evaluated using COMPACK program implemented in the CSD Python API with 15 molecules using distance and angular tolerances of 50%.

S-6 Proof: Symmetric-rotation prior is SE(3) invariant

It is natural for the prior to be translationally invariant, since the space of rotation matrices and the space of Euclidean centroids form a direct product for molecular crystals; that is, the rotational and translational degrees of freedom are independent.

Moreover, the geodesic distance between any two rotation matrices on SO(3) is invariant under left multiplication by a fixed rotation. Therefore, if R is a rotation matrix distributed uniformly on SO(3), then applying any rotation operation $Q \in SO(3)$ preserves the distribution. It implies $p(QR) = p(R)$. This result readily extends to multiple rotations. If several rotation matrices are sampled independently from the uniform distribution on SO(3), their joint distribution remains SE(3)-invariant by independence.

For the symmetric-rotation prior, it suffices to consider the case of two building blocks, as configurations with additional building blocks follow directly by induction. By Bayes' rule, the joint distribution of the two rotation matrices can read as :

$$p(R_1, R_2) = p(R_1) \times p(R_2|R_1) \tag{1}$$

Since R_2 is deterministically defined once R_1 is specified under the symmetric-rotation construction, we have $p(R_2|R_1) = 1$. Therefore, the joint distribution reduces to $p(R_1)$, which is SE(3)-invariant. This completes the proof.

S-7 Gaussian smearing of interatomic distances

To embed continuous interatomic distances into a fixed-dimensional representation in the EGNN building block embedder, we employ a Gaussian smearing (radial basis function) expansion. For a pair of atoms i and j, the interatomic distance r_{ij} is mapped to a vector

$$\boldsymbol{\phi}_d(r_{ij}) = [\phi_1(d_{ij}), \phi_2(d_{ij}), \ldots, \phi_K(d_{ij})]\,, \tag{2}$$

where each component is defined as

$$\phi_k(d_{ij}) = \exp\left(-\frac{(d_{ij} - \mu_k)^2}{2\sigma^2}\right), \qquad k = 1, \ldots, K. \tag{3}$$

The Gaussian centers $\{\mu_k\}_{k=1}^{K}$ are uniformly spaced between a minimum and maximum cutoff distance, and $K = 64$ in this work.

$$\mu_k = d_{\min} + (k-1)\Delta, \qquad \Delta = \frac{d_{\max} - d_{\min}}{K-1}, \tag{4}$$

and the bandwidth is chosen as Δ.

With this choice, the smearing corresponds to a set of overlapping radial basis functions that provide a smooth and differentiable embedding of interatomic distances. The resulting representation is invariant to global translations and rotations and is well suited for message-passing graph neural networks operating on molecular building blocks.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[2] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018.