COARSE-TO-FINE MOF GENERATION WITH PERIODIC U-DIT AND VERIFIER-GUIDED LINKER EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating novel metal–organic frameworks (MOFs) with application-relevant pore architectures is a vital challenge, requiring both diverse structural exploration and precise geometric control. Small pore apertures are particularly compelling targets given their prevalence in experimental MOFs and critical role in selective molecular separation. To pursue this objective, we introduce CoFi-MOF, a coarse-to-fine framework combining global MOF generation with target-directed linker editing for fine-grained control of small pore apertures, and evaluate its performance in the challenging ultra-microporous regime. At the global scale, a periodic diffusion Transformer generates pore geometry conditioned on the target pore-limiting diameter within a signed-distance-field latent space, which a learned topology-aware Constructor translates into discrete MOF structures. At the local scale, geometric measurements identify promising structures for local refinement by Multi-Edit-Former, which proposes chemically feasible linker modifications to adjust their apertures. Reassembly and verification establish whether these edits improve the realized pore match. Experiments convincingly show that CoFi-MOF generates structurally diverse MOFs with pore distributions concentrated around targets. We further expand PORMAKE's linker library with carboxylate-based building blocks and introduce MOF-Edit-DB, a large-scale dataset of edited MOFs to train and evaluate target-directed structural editing models. Together, CoFi-MOF enables precise small pore targeting in assembled MOFs and provides a reusable foundation for learning structural refinement.

1 INTRODUCTION

Metal–organic frameworks (MOFs) have emerged at the frontier of porous-material design because their modular construction offers extensive control over both chemical composition and periodic geometry (Yaghi et al., 2003; Furukawa et al., 2013). This tunability has made MOFs promising candidates for gas storage, molecular separation, and catalysis (Furukawa et al., 2013). However, the enormous number of possible combinations among nodes, linkers, and network topologies makes exhaustive exploration through experimental synthesis or atomistic simulation impractical. This challenge has catalyzed the development of generative models that learn directly from existing MOF structures and propose candidates at substantially greater scale (Yao et al., 2021). Recent advances include coarse-grained diffusion for generating novel and geometrically plausible frameworks in MOFDiff (Fu et al., 2024), torsion-aware flow matching for flexible linkers and novel building blocks in MOFFlow-2 (Kim et al., 2025a), pore-property-conditioned field generation and assembly in MOFFUSION (Park et al., 2025), and physical-reward optimization for targeted design in PRO-MOF (Liu et al., 2026). These methods have successfully captured complex structural distributions, expanded the accessible chemical and geometric design space, and increasingly supported property-directed MOF generation. These advances provide a strong foundation for exploring application-relevant pore-size regimes, where structural variety and precise aperture control are jointly valuable.

Small pore apertures offer a particularly compelling focus within this broader design space. Figure 1 compares selected experimental and hypothetical collections through the pore-limiting diameter (PLD), the diameter of the largest spherical probe that can traverse a periodic pore network (Willems et al., 2012). Among positive-PLD entries, 88.5% of the selected CoRE collection lie below 7 Å, compared with 21.7% of the processed MOFFUSION collection and 23.56% of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

analyzed BW-DB collection (Boyd et al., 2019). BW-DB is particularly relevant here as an upstream source of training structures for MOFDiff, MOFFlow and MOFFlow-2 (Fu et al., 2024; Kim et al., 2025b;a). This consistent contrast across widely used hypothetical collections further strengthens the experimental motivation for small-aperture generation. Fine control is also especially valuable for molecular separation, where modest aperture changes can enable molecular sieving or alter the dominant separation mechanism (Lin et al., 2018; Xue et al., 2018). Together, these distributional observations and application needs motivate dedicated generation within the small-aperture regime.

Generating MOFs within this regime is nevertheless demanding. A small PLD is determined by narrow bottlenecks along a connected periodic pore network, making aperture control particularly sensitive to the geometry surrounding these passages (Willems et al., 2012). Local changes can appreciably narrow an opening or obstruct a previously traversable channel, so fine control requires preserving both the requested clearance and pore connectivity. In a pore-field-based approach, small-aperture training data and PLD conditioning guide exploration, while the requested clearance must ultimately be realized through discrete building-block selection and assembly. Finite-resolution fields efficiently capture pore morphology, while closely related linker chemistries can remain difficult to distinguish (Park et al., 2025). The challenge is therefore to realize narrowly specified apertures while retaining sufficient chemical and topological freedom to explore distinct frameworks. This motivates periodic global generation coupled to direct geometric verification and chemically feasible local correction.

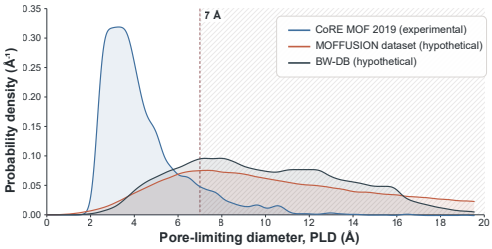


Figure 1: **PLD coverage motivates target-directed generation.** Positive-PLD entries from selected CoRE MOF 2019 ASR/FSR records (Chung et al., 2019), the processed MOFFUSION collection (Park et al., 2025), and BW-DB (Boyd et al., 2019). The dashed line marks 7 Å.

To address these challenges, we introduce **CoFi-MOF**, a verifier-guided coarse-to-fine framework for pore-targeted MOF generation. Periodic U-DiT explores target-conditioned pore geometries while modeling relationships across unit-cell boundaries. A learned topology-aware Constructor maps each field to candidate building-block assignments through node prediction and conditional linker identification. Geometric measurements rank the assembled candidates by target agreement, and Multi-Edit-Former (MEF) proposes topology-preserving linker modifications for local refinement. Reassembly and evaluation determine whether an edit improves the aperture match, with the original candidates retained during final selection. We further expand PORMAKE's carboxylate-based linker library and introduce **MOF-Edit-DB**, which pairs chemical transformations with measured framework-level pore responses.

Our main contributions are:

- ❶ We propose **CoFi-MOF**, a verifier-guided coarse-to-fine framework integrating periodic global generation, learned structural realization and measurement-driven local refinement for controllable MOF design within the experimentally relevant small-aperture regime.
- ❷ We develop a refinement procedure that selects assembled candidates by measured PLD and uses MEF to propose topology-preserving linker modifications. Reassembly and verification evaluate each modification while retaining the original candidates for final selection. This approach extends pore control beyond latent conditioning to direct measurement and correction of fully assembled frameworks.
- ❸ We expand PORMAKE's library with carboxylate-based linkers and build **MOF-Edit-DB**, a large-scale resource linking chemical edits to framework-level pore responses. The library and dataset support structural generation, property prediction and target-directed modification.
- ❹ We demonstrate accurate pore targeting over PLDs of 3–7 Å, achieving 88.15% Hit@0.5. Local editing further increases Hit@0.2 from 68.28% to 77.51% beyond candidate verification alone, showing its particular value for fine-grained aperture control.

2 Related Work

Property-directed MOF generation. CDVAE and DiffCSP encode crystal periodicity through boundary-aware representations and joint lattice–coordinate diffusion, respectively (Xie et al., 2022; Jiao et al., 2023). MOF generation has advanced from learned reticular representations to block-level diffusion and rigid-body or torsional flow matching (Yao et al., 2021; Fu et al., 2024; Kim et al., 2025b;a). Property control uses conditional pore-field diffusion (Park et al., 2025), predictor-guided reinforcement learning for direct air capture (Park et al., 2024b), and physical-reward optimization, including PLD targeting (Liu et al., 2026). Mofasa generates atoms, coordinates and lattices through all-atom latent diffusion (Simkus et al., 2025); EGMOF combines descriptor diffusion, Transformer decoding and predictor-based selection (Han et al., 2025). CoFi-MOF couples toroidal attention over signed-distance fields with direct pore measurement, linking periodic proposals to verified aperture control.

From pore geometry to verified structures. Precise aperture control links continuous pore geometry to discrete chemistry. MOFFUSION enables conditional pore-field generation, while identifying finite resolution and ambiguity between similar-looking functional groups as reconstruction challenges (Park et al., 2025). Assembly realizes these components as periodic structures (Lee et al., 2021; Park et al., 2024a), whose narrow-channel clearances strongly affect molecular access (Lin et al., 2018; Xue et al., 2018). Structural diversity is also an explicit optimization concern in PRO-MOF's reward design (Liu et al., 2026). CoFi-MOF connects periodic pore-field generation to learned, component-conditioned structural realization. Direct Zeo++ measurements guide aperture selection and assess local linker modifications (Willems et al., 2012).

Linker optimization and edit supervision. Linker modification provides a route to tune pore chemistry within isostructural families (Han et al., 2021). Library augmentation and predictor-guided optimization expand the available design choices (Yao et al., 2021; Cleeton & Sarkisov, 2025), while target-directed editing also benefits from supervision that links a specified chemical change to its realized pore response. MOF-Edit-DB provides these parent–edit pairs for training MEF.

3 Preliminaries

Pore-limiting and largest-cavity diameters. The internal geometry of a porous framework is commonly characterized by its pore-limiting diameter (PLD) and largest-cavity diameter (LCD). Let $\Omega(C)$ denote the void region of a framework C, and let $\rho_C(\mathbf{r})$ be the radius of the largest sphere centered at $\mathbf{r} \in \Omega(C)$ that does not intersect the framework surface. These descriptors can be expressed as

$$\mathrm{LCD}(C) = 2 \max_{\mathbf{r}\in\Omega(C)} \rho_C(\mathbf{r}), \qquad \mathrm{PLD}(C) = 2 \max_{\gamma\in\Gamma_{\mathrm{perc}}(C)} \min_{\mathbf{r}\in\gamma} \rho_C(\mathbf{r}), \tag{1}$$

where $\Gamma_{\mathrm{perc}}(C)$ denotes the set of continuous paths that traverse the periodic pore network. LCD therefore measures the widest cavity that can be accommodated within the framework, whereas PLD measures the largest spherical probe that can pass through a connected periodic channel (Willems et al., 2012). A framework may consequently contain spacious cavities while remaining accessible only through substantially narrower apertures. Because these bottlenecks directly govern molecular accessibility and sieving, we use PLD as the primary control variable and retain LCD as a complementary description of the internal pore volume. Both quantities are computed from realized structures using Zeo++.

Ultra-microporous MOFs have been described in terms of pore apertures below 5–7 Å (Adil et al., 2017). Here, we define the small-aperture regime by PLD below 7 Å, rather than requiring every cavity to be smaller than this threshold.

Periodic representation of MOF pore geometry. A crystalline MOF is an infinite periodic structure specified by a lattice matrix $\mathbf{A}$ and the atomic or building-block arrangement within one unit cell. For a fractional coordinate $\mathbf{u} \in [0, 1)^3$, its periodic images occupy

$$\mathbf{r}_{\mathbf{n}} = \mathbf{A}(\mathbf{u} + \mathbf{n}), \qquad \mathbf{n} \in \mathbb{Z}^3. \tag{2}$$

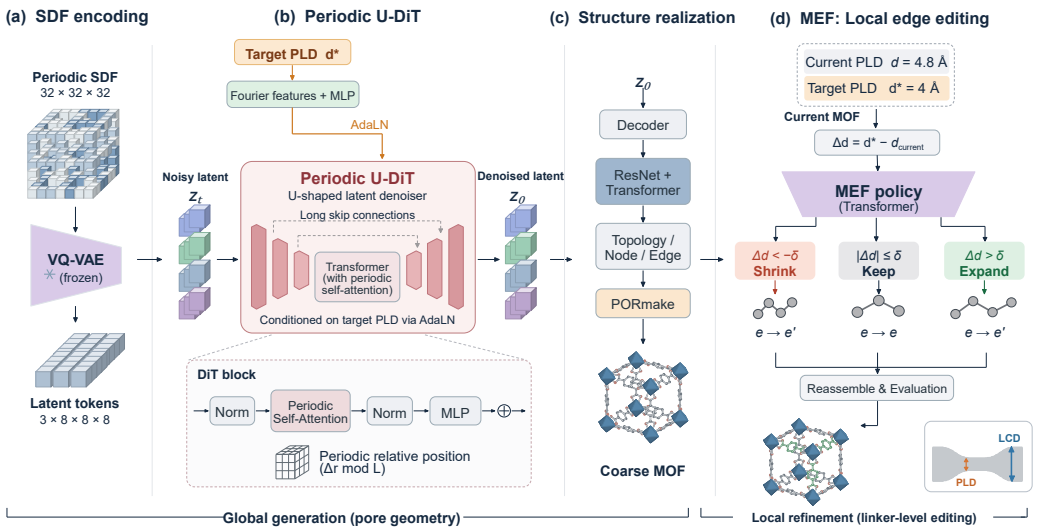


Figure 2: **Coarse-to-fine generation in CoFi-MOF.** (a) A frozen VQ-VAE encodes periodic signed-distance fields; (b) Periodic U-DiT generates target-conditioned pore geometry; (c) the learned Constructor realizes each field through node and conditional linker prediction; and (d) Zeo++ verification and MEF linker editing refine structures toward the requested PLD. Panel (d) illustrates pore contraction, retention and expansion; MEF predicts explicit chemical modifications. Algorithm 1 specifies the complete generation and single-batch refinement procedure.

The unit-cell domain is therefore topologically equivalent to a three-dimensional torus: opposing cell faces are identified, and sites that appear distant in an isolated voxel grid may be immediate neighbors across a periodic boundary. Following MOFFUSION (Park et al., 2025), we describe pore morphology using multichannel signed-distance fields (SDFs) sampled over this periodic domain, with distances evaluated against the periodically replicated framework surface. This representation compactly captures continuous pore geometry and cross-boundary connectivity, but it does not uniquely specify the discrete topology, nodes, or linkers needed to construct a chemically explicit MOF. Generating a periodic pore field and realizing it as a discrete framework therefore remain two closely connected but fundamentally distinct problems.

4 METHOD

CoFi-MOF targets precise apertures by translating a requested pore size into a discrete MOF while still exploring diverse chemistries and topologies. It does so in three stages: (1) PLD-conditioned Periodic U-DiT generates pore geometries with cross-boundary spatial modeling; (2) a topology-aware Constructor maps these geometries to candidate frameworks via node prediction and conditional linker identification; (3) geometric verification and MEF-based refinement apply geometric measurements to select promising assemblies and target-directed linker edits for finer aperture control. Together, these stages link global exploration of periodic pore geometry with local chemical refinement. Figure 2 illustrates the overall workflow.

4.1 PLD-CONDITIONED PERIODIC MOF GENERATION

Periodic SDF latent space. The global stage must increase access to small-aperture pore geometries without restricting generation to a narrowly filtered structural prior. Following MOFFUSION (Park et al., 2025), we represent each MOF by a four-channel periodic signed-distance field $x \in \mathbb{R}^{4\times32\times32\times32}$ and use a frozen VQ-VAE (van den Oord et al., 2017) to obtain an unquantized continuous latent $z_e \in \mathbb{R}^{3\times8\times8\times8}$. Diffusion (Ho et al., 2020; Rombach et al., 2022) is trained directly on z_e, without codebook quantization or latent normalization. At generation time, the sampled

latent is quantized once and decoded as $\widehat{x} = D(Q(\widehat{z}_e))$. This representation retains continuous pore morphology for global generation while deferring discrete chemical realization to the Constructor.

Periodic U-DiT. Because MOF pore channels extend across unit-cell boundaries, the latent denoiser must model long-range geometry without introducing artificial cell edges. We train an ϵ-prediction denoiser on $z_s = \sqrt{\bar{\alpha}_s} z_e + \sqrt{1-\bar{\alpha}_s}\epsilon$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\big[\|\epsilon - \epsilon_\theta(z_s, s, d^\star)\|_2^2\big]. \tag{3}$$

The 8^3 latent grid is treated as 512 tokens and processed by eight constant-resolution Transformer blocks with long feature skips (Bao et al., 2023). To make global attention consistent with the periodic grid, head h assigns sites r_i, r_j the attention logit

$$a_{ij}^{(h)} = \langle q_i^{(h)}, k_j^{(h)} \rangle / \sqrt{d_h} + b_h\big((r_i - r_j) \bmod 8\big). \tag{4}$$

The learned bias represents directed displacement on a discrete torus, including neighbors across opposite cell faces. No absolute spatial embedding is used. A Fourier encoding of $d^\star$ enters every block through adaptive layer normalization (Peebles & Xie, 2023). Spatially shared conditioning and token-wise operations preserve equivariance to integer cyclic shifts. Deterministic DDIM sampling (Song et al., 2021) produces $\widehat{z}_e$, which the frozen quantizer and decoder map to a target-conditioned periodic pore field. Detailed architecture and numerical probes are provided in Appendices A.2 and E.1.

4.2 Learned topology-aware structure realization

The Constructor completes the transition from periodic pore geometry to discrete chemistry. Given a generated field and its predicted topology, a Node Transformer identifies inorganic and optional organic nodes. An Edge Transformer then conditions on both the field and the proposed nodes to identify compatible linkers, explicitly accounting for component dependencies during structural realization (Vaswani et al., 2017). The models rank alternative building-block assignments, which PORMAKE assembles into periodic frameworks (Lee et al., 2021). These alternatives are chemical realizations of the same generated field. Zeo++ subsequently determines their agreement with the target aperture.

We expand PORMAKE's building-block library with carboxylate-based linkers informed by reviews of Zr-MOFs and titanium coordination polymers (Bai et al., 2016; Du et al., 2024). This extension supports a Constructor vocabulary of more than 600 linker classes and is packaged with the MOF-Edit-DB resources. Appendix A.3 describes the learned heads and their evaluation; Appendix C.1 details library construction.

4.3 Staged verification and MEF refinement

Measurement-based candidate selection. The global branch broadens access to small-aperture candidates, whereas precise target control is determined from their realized pore geometry. Learned scores prioritize which recipes and edits to evaluate; Zeo++ then verifies their actual pore apertures at both realization stages. We use this fixed geometric verifier (Willems et al., 2012) and define

$$V(C) = \text{PLD}_{\text{Zeo++}}(C), \qquad e(C; d^\star) = |V(C) - d^\star|. \tag{5}$$

Here, C is an assembled framework before force-field relaxation; unless stated otherwise, all measured apertures in this work refer to this unrelaxed state. Zeo++ verification takes approximately 0.3 s per structure in our CPU benchmark, making geometric measurement practical throughout structure realization and local refinement (Appendix B.4).

Let $\mathcal{C}$ contain the Constructor's successfully assembled structures with finite positive PLD. The closest initial match is

$$C_0 = \arg\min_{C \in \mathcal{C}} e(C; d^\star). \tag{6}$$

If no candidate has a valid measurement, generation is unsuccessful. Otherwise, refinement starts from the closest match whose linker supports the permitted chemical modifications. This source C_s can differ from C_0, which remains available for final selection. MEF receives the source's linker state, measured PLD, and target in a single call, including when the source already satisfies the reporting tolerance. This permits further improvement at finer pore-size resolution.

Target-directed linker correction. MEF broadens local chemical exploration beyond the Constructor's initial assignments by jointly ranking destination assignments over at most three selected linker sites. It conditions on the current chemical states, invariant linker geometry, measured PLD d_0, and requested target t. The difference $t - d_0$ provides the required correction direction and magnitude. For a legal assignment a and source state s, its joint score is

$$S_\phi(a) = \sum_i u_\phi(i, a_i, t) + \sum_{i<j} p_\phi(i, j, a_i, a_j) + g_\phi(a, t) - \lambda_{\text{edit}} d_H(a, s), \tag{7}$$

where unary, symmetric pairwise, and pooled global terms couple the site decisions, and d_H penalizes unnecessary changes. Deterministic chemistry and constructability masks remove unsupported assignments before ranking. MEF is trained on measured parent–edit relationships from MOF-Edit-DB using a multi-positive listwise objective; crucially, unmeasured assignments are not treated as negatives. MEF data are partitioned by linker lineage, with disjoint base-linker identities across training, development and held-out sets (Appendix A.4). During materialization and reassembly, the surrounding topology and node identities remain fixed. Architecture, supervision, and the frozen-model scope are detailed in Appendix A.4, with dataset construction described in Appendix C.

Verified pore refinement. MEF proposes up to M distinct linker modifications according to S_ϕ, excluding assignments that leave the source unchanged. Each proposal is reassembled with the original topology and nodes, and Zeo++ measures its resulting aperture. The learned score thus guides chemical exploration, while the measured pore response determines which modification, if any, should be retained. Let $\mathcal{F}$ contain all original and edited structures with finite positive measured PLD. We select

$$C_{\text{sel}} = \arg\min_{C \in \mathcal{F}} e(C; d^\star). \tag{8}$$

Retaining the original structures ensures that unsuccessful edits cannot displace a better initial match; ties favor the original candidate. If refinement produces no eligible alternative, C_0 is returned.

Each attempt begins with one generated pore field and returns at most one MOF after learned structural realization, geometric verification and local refinement. Both the generator and MEF accept continuous PLD conditions; the reported experiments use an integer target grid. Algorithm 1 in Appendix B provides the complete inference procedure.

5 Experiments

We evaluate small-aperture generation, the contributions of periodic modeling and geometric verification, and learned linker refinement. Component-level studies assess Periodic U-DiT and the Constructor separately, while the main ablations examine their integration with verification and MEF. Primary results use PLDs measured on assembled structures with Zeo++; Appendix E.4 assesses relaxation sensitivity.

5.1 PLD control benchmark

The main benchmark tests whether target conditioning and verified refinement concentrate the final pore distribution around each requested aperture. We consider target PLDs of $\{3, 4, 5, 6, 7\}$ Å, spanning the small-aperture range that motivates this work. Each method is allocated 1,000 attempts per target. For CoFi-MOF, 4,994 of the 5,000 attempts are complete at the reporting snapshot. Table 1 compares the resulting outputs with target-adapted GA, MOFDiff, MOFFlow-2, and MOFFUSION (Fu et al., 2024; Kim et al., 2025a; Park et al., 2025). Candidate selection is also used by the baselines: MOFDiff ranks a posterior latent pool by predicted target agreement, while MOFFUSION assembles top-5 Constructor proposals and retains the closest measured aperture. MOFFlow-2 uses learned PLD conditioning, and GA performs target-directed search. Model configurations, structural realization and data sources are specified in Appendices D.1 and D.3.

Target control and success. Let $N_{d^\star}$ be the number of completed attempts at target $d^\star$, and let $\mathcal{V}_{d^\star}$ contain attempts whose selected output has a finite positive measured PLD. Measurable yield is

Table 1: **Pore-targeted MOF generation across PLD targets of 3–7 Å.** Each method is allocated 1,000 attempts per target. Metrics use completed attempts and retain repeated outputs; completed failures count as misses. Yield and hits are percentages, while MAE and signed bias (Å) use measurable outputs. PLDs are measured on assembled, unrelaxed structures. Best yield, hit rates and MAE are bolded.

Method	Metric	3	4	5	6	7
GA	Yield	88.40	92.00	88.80	92.40	90.90
	Hit@0.2/0.5	36.10 / 59.10	59.90 / 71.10	64.00 / 67.20	71.40 / 72.40	59.00 / 72.00
	MAE/bias	1.270 / +1.135	1.193 / +1.155	0.920 / +0.754	1.396 / +1.164	0.954 / +0.625
MOFDiff + selection	Yield	89.90	91.90	93.20	91.50	92.60
	Hit@0.2/0.5	15.50 / 33.10	12.80 / 29.00	6.40 / 18.80	5.90 / 13.00	2.80 / 8.10
	MAE/bias	1.068 / +0.782	0.951 / –0.122	1.255 / –0.787	1.625 / –1.305	1.968 / –1.752
MOFFlow-2 (cond.)	Yield	91.10	91.90	93.70	93.80	93.80
	Hit@0.2/0.5	10.20 / 22.80	9.30 / 24.70	10.30 / 23.40	7.40 / 20.70	6.70 / 17.80
	MAE/bias	1.453 / +1.335	1.198 / +0.502	1.187 / –0.292	1.415 / –0.639	1.567 / –0.844
MOFFUSION (PLD)	Yield	83.50	85.70	87.30	88.80	89.40
	Hit@0.2/0.5	7.50 / 19.00	6.60 / 17.70	6.90 / 16.70	5.90 / 15.50	5.40 / 15.40
	MAE/bias	1.643 / +0.891	1.627 / +0.728	1.726 / +0.278	1.876 / +0.063	1.888 / –0.277
CoFi-MOF	**Yield**	**100.00**	**100.00**	**100.00**	**99.90**	**99.50**
	Hit@0.2/0.5	**85.50 / 94.50**	**85.79 / 93.99**	**79.90 / 90.70**	**73.02 / 83.65**	**63.33 / 77.86**
	MAE/bias (Å)	**0.126 / +0.013**	**0.134 / -0.013**	**0.178 / -0.068**	**0.269 / -0.192**	**0.393 / -0.296**

$|\mathcal{V}_{d^\star}|/N_{d^\star}$. The target hit rate is

$$H_\delta(d^\star) = \frac{1}{N_{d^\star}} \sum_{i \in \mathcal{V}_{d^\star}} \mathbf{1}[|d_i - d^\star| \le \delta]\,. \tag{9}$$

Hit@0.5 is the primary metric; Hit@0.2 and Hit@1.0 assess tighter and broader tolerances. Completed failures count as misses; pending attempts are excluded. MAE and signed bias, $d_i - d^\star$, use measurable outputs. Pooled results aggregate counts and errors across targets. Yield measures pore measurability, not synthetic feasibility.

Accurate pore realization across targets. Figure 3 shows strong target concentration. CoFi-MOF achieves 88.15% Hit@0.5 and 77.51% Hit@0.2, with 99.88% measurable yield and 0.220 Å MAE. Control is particularly strong at smaller apertures, with Hit@0.5 above 90% at 3–5 Å and 77.86% at 7 Å. The evaluated neural baselines reach at most 33.10% Hit@0.5 (MOFDiff at 3 Å) and 15.50% Hit@0.2 across targets, with measurable yields of 83.50–93.80%. GA reaches 59.10–72.40% attempt-level Hit@0.5, yet 74.7–78.7% of its proposals repeat a topology–node–linker combination within a target; CoFi-MOF keeps 98.8% of measured outputs in the initial batch structurally distinct (Appendix E.2). CoFi-MOF thus combines higher target agreement with higher measurable yield under the evaluated protocols.

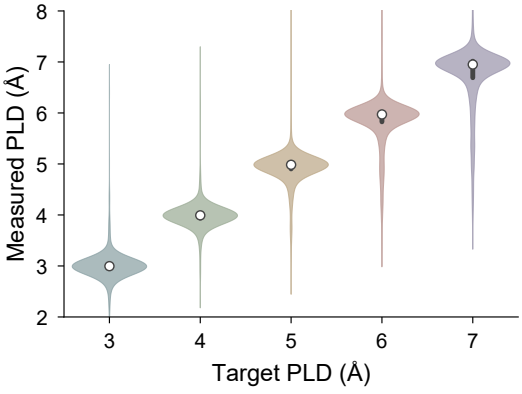


Figure 3: **Target-conditioned PLD distributions.** CoFi-MOF at 3–7 Å. Circles denote medians; bars, interquartile ranges. Widths show Gaussian KDEs (bandwidth 0.15 Å) using all observations, displayed over 2–8 Å.

In the available paired cohorts, relaxation moves CoFi-MOF Hit@0.5 from 93.6% unrelaxed to 20.9% relaxed, versus 12.3% for MOFFUSION, with relaxed MAE 1.239 versus 2.999 Å. These subset results distinguish precise control of assembled pore geometry from preservation of that match after relaxation (Appendix E.4).

Table 2: **Contributions of geometric verification and MEF.** All variants retain Periodic U-DiT and the learned Constructor and share 4,994 completed attempts at targets 3–7 Å. Without verification and MEF, the model directly returns the assignment with the highest joint node–linker score. Hits (%) include completed failures; MAE (Å) uses measurable outputs.

Variant	Hit@0.2	Hit@0.5	MAE
w/o verifier and MEF	7.63	17.56	1.734
w/o MEF	68.28	83.40	0.303
CoFi-MOF	77.51	88.15	0.220

5.2 Periodic generation and structural realization

To assess generation independently of candidate selection and editing, we compare unconditional Periodic U-DiT with a MOFFUSION-style 3D U-Net using matched data, a frozen VQ-VAE, 60,000 training updates and 500 shared initial noises. Periodic U-DiT increases assembly success from 64.0% to 71.2%, yielding 36 additional frameworks. Across 7,500 cyclic-shift probes, median relative equivariance error falls from 2.859×10^{-2} to 1.156×10^{-7}, showing substantially more consistent treatment of periodic geometry. Together, these results support improved structural realization and periodic consistency for the complete backbone. Appendix E.1 provides the protocols and detailed diagnostics. Complementary Constructor studies assess node and linker identification within the expanded building-block vocabulary and examine how component conditioning improves linker prediction (Appendix A.3).

5.3 Contributions of verification and linker refinement

Table 2 compares three variants on the same generated pore fields. CoFi-MOF w/o verifier and MEF directly returns the Constructor's globally highest-ranked assignment under the joint node–linker score, without measured candidate selection or local editing. CoFi-MOF w/o MEF adds Zeo++ selection among the Constructor's realized alternatives. The complete model further applies MEF and re-verifies its edits. Zeo++ characterizes all final outputs for evaluation, including the direct-output variant.

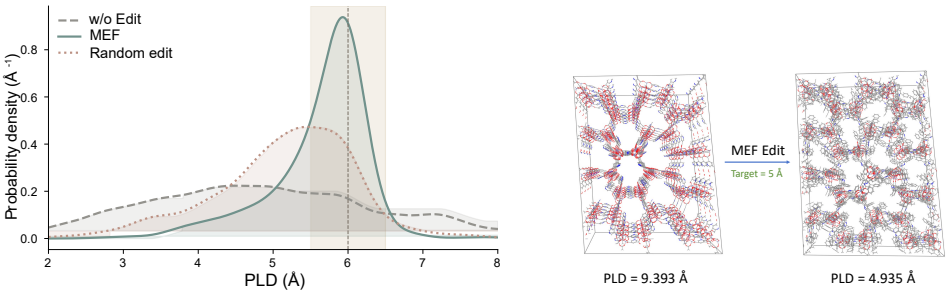
Verification completes pore-targeted realization. Combining the learned Constructor with geometric verification raises Hit@0.5 from 17.56% to 83.40% before local editing, while MAE decreases from 1.734 to 0.303 Å. The verified generation stage already exceeds the evaluated baselines at every target in Table 1. These results show the value of resolving alternative chemical realizations through their measured pore apertures, complementing the geometric information learned by Periodic U-DiT.

Local editing further sharpens pore control. Local refinement raises Hit@0.5 from 83.40% to 88.15%, adding 237 successful structures, while MAE falls from 0.303 to 0.220 Å, a 27.45% reduction. The gain is particularly pronounced at the tighter tolerance. Hit@0.2 increases from 68.28% to 77.51%, adding 461 precise matches. These results establish local chemical refinement as an effective complement to verified structural realization.

Hit@0.5 improves at all five targets and in both sampling batches, with original matches retained. The pooled signed error shifts from -0.036 to -0.111 Å, indicating a modest downward bias alongside improved absolute accuracy. Target-wise results and paired uncertainty estimates are reported in Appendix E.2.

5.4 Learned linker editing

To isolate the quality of learned edit proposals, we compare MEF with random editing on 937 paired tasks at a target PLD of 6 Å. Both editors act on the same selected source for each task, and their first proposals are evaluated without post-edit candidate selection. MEF achieves 68.20% Hit@0.5, substantially exceeding random editing at 33.62%, with a paired gain of 34.58 percentage points (95% bootstrap CI [31.38, 37.89]). MAE is also markedly lower, at 0.484 versus 1.004 Å. Figure 4



(a) First-proposal pore distributions

(b) Target-directed linker editing

Figure 4: **Learned edits concentrate pore apertures near the target.** (a) Measured PLDs at 6 Å for MEF and random first proposals from shared editable sources, with the direct output of the same pipeline (w/o verifier and MEF) as reference. Curves use 890 jointly measurable tasks; reported hit rates use all 937. (b) A representative edit brings PLD from 9.393 to 4.935 Å toward a 5 Å target. Further examples appear in Appendix E.3.

connects this sharper target concentration to a representative structural edit. All 937 tasks, including failed proposals, enter the hit-rate denominator; full results and the best-of-ten comparison are given in Appendix E.5.

5.5 MOF-Edit-DB and an expanded linker library

MOF-Edit-DB packages an expanded PORMAKE library of carboxylate-based linkers with a large-scale framework edit-response resource. The library directly extends the building blocks available to PORMAKE and supports both learned structural realization and chemically specified linker modification. It contains 138,771 force-field-relaxed framework records, of which 52.5% have PLDs below 7 Å, and 108,145 matched parent–edit comparisons. Construction links chemically specified linker modifications to complete periodic frameworks and their relaxed pore responses, with systematic MOFChecker-based screening and record-level quality annotations (Jin et al., 2025). These linked representations support pore-property prediction, target-conditioned edit ranking and small-aperture generative modeling, extending the resource's value beyond MEF training. Appendix C details library construction, quality annotations and data access; additional editing examples appear in Appendix E.3. The dataset review is included in the supplementary materials, and the link will be released upon completion of double-blind review.

6 Limitations

The reachable structures depend on PORMAKE's topologies, building blocks and assembly geometry. Our results establish control of assembled pore geometry, not stability or synthesizability. The primary evaluation concerns assembled unrelaxed structures; force-field relaxation changes pore apertures in the evaluated paired cohorts (Appendix E.4). Broader action coverage and task-specific physical validation remain important directions.

7 Conclusion

CoFi-MOF combines periodic pore-field generation, learned structural realization, and verified linker refinement for target-directed MOF design. Across PLD targets of 3–7 Å, verification-based realization achieves 83.40% Hit@0.5, and MEF further raises this to 88.15% while reducing MAE from 0.303 to 0.220 Å. At the tighter 0.2 Å tolerance, local editing adds 461 matches, increasing Hit@0.2 from 68.28% to 77.51%. These results show how global generation and measured local modifications jointly improve aperture targeting on assembled structures. The expanded PORMAKE linker library and MOF-Edit-DB also provide resources for learning and evaluating chemical edits and their framework-level pore responses.

Ethics Statement

This work is computational and involves no human subjects, personal data, or animal experimentation, so no IRB approval is required. The datasets are previously published or derived from public resources, and the provenance of our own contributions, including MOF-Edit-DB, is documented with the released materials. Generative models for porous materials are dual-use in principle. Our system produces hypothetical frameworks evaluated only in silico. We make no claims about synthetic feasibility, stability, or adsorption performance, and we assess the immediate misuse potential as low relative to the intended computational-discovery use. The authors acknowledge and adhere to the ICLR Code of Ethics.

Reproducibility Statement

The appendix records implementation details, evaluation definitions and data sources. MOF-Edit-DB links pore labels to edit recipes and construction procedures, with geometry availability and split details documented separately. An anonymized code snapshot containing model implementations, inference and evaluation procedures, experimental configurations, and scripts for recomputing the reported results from included run summaries is provided in the supplementary material. Its manifests specify the additional assets required for a full physical rerun.

AI-use Statement

We used LLMs to improve the readability and fluency of the manuscript. They did not contribute to the scientific ideas or algorithm design. All substantive content, reasoning, and conclusions were developed by the authors.

References

Matthew A. Addicoat, Nina Vankova, Ismot Farjana Akter, and Thomas Heine. Extension of the Universal Force Field to Metal–Organic Frameworks. *Journal of Chemical Theory and Computation*, 10(2):880–891, 2014. doi: 10.1021/ct400952t. URL https://doi.org/10.1021/ct400952t.

Karim Adil, Youssef Belmabkhout, Renjith S. Pillai, Amandine Cadiau, Prashant M. Bhatt, Ayalew H. Assen, Guillaume Maurin, and Mohamed Eddaoudi. Gas/vapour separation using ultra-microporous metal–organic frameworks: insights into the structure/separation relationship. *Chemical Society Reviews*, 46(11):3402–3430, 2017. doi: 10.1039/C7CS00153C. URL https://doi.org/10.1039/C7CS00153C.

Yan Bai, Yibo Dou, Lin-Hua Xie, William Rutledge, Jian-Rong Li, and Hong-Cai Zhou. Zr-based metal–organic frameworks: design, synthesis, structure, and applications. *Chemical Society Reviews*, 45(8):2327–2367, 2016. doi: 10.1039/C5CS00837A. URL https://doi.org/10.1039/C5CS00837A.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are Worth Words: A ViT Backbone for Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679, 2023. doi: 10.1109/cvpr52729.2023.02171. URL https://doi.org/10.1109/cvpr52729.2023.02171.

Peter G. Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P. Ireland, Thomas D. Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M. Mercedes Maroto-Valer, Jeffrey A. Reimer, Jorge A. R. Navarro, Tom K. Woo, Susana Garcia, Kyriakos C. Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature*, 576:253–256, 2019. doi: 10.1038/s41586-019-1798-7. URL https://www.nature.com/articles/s41586-019-1798-7.

Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S.

Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, 2019. doi: 10.1021/acs.jced.9b00835. URL https://doi.org/10.1021/acs.jced.9b00835.

Conor Cleeton and Lev Sarkisov. Inverse design of metal-organic frameworks using deep dreaming approaches. *Nature Communications*, 16(1):4806, 2025. doi: 10.1038/s41467-025-59952-3. URL https://doi.org/10.1038/s41467-025-59952-3.

Yafei Du, Qingqing Yan, and Sujing Wang. Progress and challenges of monometallic titanium coordination polymers. *Small*, 20(46):2403470, 2024. doi: 10.1002/smll.202403470. URL https://doi.org/10.1002/smll.202403470.

Xiang Fu, Tian Xie, Andrew Scott Rosen, Tommi S. Jaakkola, and Jake Allen Smith. MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0VBsoluxR2.

Hiroyasu Furukawa, Kyle E. Cordova, Michael O'Keeffe, and Omar M. Yaghi. The chemistry and applications of metal-organic frameworks. *Science*, 341:1230444, 2013. doi: 10.1126/science.1230444.

Seunghee Han, Yeonghun Kang, Taeun Bae, Junho Kim, Younghun Kim, Varinia Bernales, Alan Aspuru-Guzik, and Jihan Kim. EGMOF: Efficient Generation of Metal-Organic Frameworks Using a Hybrid Diffusion-Transformer Architecture. *arXiv preprint arXiv:2511.03122*, 2025. doi: 10.48550/arXiv.2511.03122. URL https://arxiv.org/abs/2511.03122v3.

Xiwei Han, Xiaoxian Yang, Chuan Yu, Shuyan Lu, Ehsan Sadeghi Pouya, Peng Bai, Jiafei Lyu, and Xianghai Guo. Fine-tuning the pore structure of metal–organic frameworks by linker substitution for enhanced hydrogen storage and gas separation. *CrystEngComm*, 23(16):3026–3032, 2021. doi: 10.1039/D1CE00087J. URL https://doi.org/10.1039/D1CE00087J.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal Structure Prediction by Joint Equivariant Diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0767. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/38b787fc530d0b31825827e2cc306656-Abstract.html.

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*, 4(6):1560–1569, 2025. doi: 10.1039/D5DD00109A. URL https://doi.org/10.1039/D5DD00109A.

Nayoung Kim, Seongsu Kim, and Sungsoo Ahn. Flexible MOF Generation with Torsion-Aware Flow Matching. In *Advances in Neural Information Processing Systems*, 2025a. doi: 10.52202/085713-0464. URL https://papers.neurips.cc/paper_files/paper/2025/hash/146b5b0feb14fe8e630669ad1faba25e-Abstract-Conference.html.

Nayoung Kim, Seongsu Kim, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. MOFFlow: Flow Matching for Structure Prediction of Metal-Organic Frameworks. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=dNT3abOsLo.

Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021. doi: 10.1021/acsami.1c02471. URL https://doi.org/10.1021/acsami.1c02471.

Rui-Biao Lin, Libo Li, Hao-Long Zhou, Hui Wu, Chaohui He, Shun Li, Rajamani Krishna, Jinping Li, Wei Zhou, and Banglin Chen. Molecular sieving of ethylene from ethane using a rigid metal–organic framework. *Nature Materials*, 17(12):1128–1133, 2018. doi: 10.1038/s41563-018-0206-2.

Zicheng Liu, Beng Fei, and Di Huang. PRO-MOF: Policy Optimization with Universal Atomistic Models for Controllable MOF Generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/1a1e46097c49a36ddc6fab8d14269902b8caba9f.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kyungmin Nam, Seunghee Han, and Jihan Kim. Interpretable inverse design of metal-organic frameworks with large language model agents. *arXiv preprint arXiv:2606.29459*, 2026. doi: 10.48550/arXiv.2606.29459. URL https://arxiv.org/abs/2606.29459.

Hyun Park, Xiaoli Yan, Ruijie Zhu, Eliu A. Huerta, Santanu Chaudhuri, Donny Cooper, Ian Foster, and Emad Tajkhorshid. A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture. *Communications Chemistry*, 7(1):21, 2024a. doi: 10.1038/s42004-023-01090-2. URL https://doi.org/10.1038/s42004-023-01090-2.

Hyunsoo Park, Sauradeep Majumdar, Xiaoqi Zhang, Jihan Kim, and Berend Smit. Inverse design of metal–organic frameworks for direct air capture of CO_2 via deep reinforcement learning. *Digital Discovery*, 3(4):728–741, 2024b. doi: 10.1039/D4DD00010B. URL https://doi.org/10.1039/D4DD00010B.

Junkil Park, Youhan Lee, and Jihan Kim. Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation. *Nature Communications*, 16 (1):34, 2025. doi: 10.1038/s41467-024-55390-9. URL https://doi.org/10.1038/s41467-024-55390-9.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/iccv51070.2023.00387. URL https://doi.org/10.1109/iccv51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/cvpr52688.2022.01042. URL https://doi.org/10.1109/cvpr52688.2022.01042.

Vaidotas Simkus, Anders Christensen, Steven Bennett, Ian Johnson, Mark Neumann, James Gin, Jonathan Godwin, and Benjamin Rhodes. Mofasa: A Step Change in Metal-Organic Framework Generation. *arXiv preprint arXiv:2512.01756*, 2025. doi: 10.48550/arXiv.2512.01756. URL https://arxiv.org/abs/2512.01756.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022. doi: 10.1016/j.cpc.2021.108171. URL https://doi.org/10.1016/j.cpc.2021.108171.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.nips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Thomas F. Willems, Chris H. Rycroft, Michaeel Kazi, Juan C. Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012. doi: 10.1016/j.micromeso.2011.08.020. URL https://doi.org/10.1016/j.micromeso.2011.08.020.

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

Dong-Xu Xue, Amandine Cadiau, Łukasz J. Weseliński, Hao Jiang, Prashant M. Bhatt, Aleksander Shkurenko, Lukasz Wojtas, Zhijie Chen, Youssef Belmabkhout, Karim Adil, and Mohamed Eddaoudi. Topology meets MOF chemistry for pore-aperture fine tuning: ftw-MOF platform for energy-efficient separations via adsorption kinetics or molecular sieving. *Chemical Communications*, 54(49):6404–6407, 2018. doi: 10.1039/C8CC03841D.

Omar M. Yaghi, Michael O'Keeffe, Nathan W. Ockwig, Hee K. Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423:705–714, 2003. doi: 10.1038/nature01650.

Zhenpeng Yao, Benjamín Sánchez-Lengeling, N. Scott Bobbitt, Benjamin J. Bucior, Sai Govind Hari Kumar, Sean P. Collins, Thomas Burns, Tom K. Woo, Omar K. Farha, Randall Q. Snurr, and Alán Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021. doi: 10.1038/s42256-020-00271-1. URL https://doi.org/10.1038/s42256-020-00271-1.

A MODEL ARCHITECTURE AND TRAINING

CoFi-MOF separates periodic pore-field generation from discrete framework realization and local linker refinement. This section details the representations and learned components; Appendix B specifies how they interact during inference.

A.1 PERIODIC PORE REPRESENTATION AND STRUCTURAL REALIZATION

Each MOF is represented by a four-channel 32^3 signed-distance field and encoded into a continuous latent of shape 3×8^3. The pretrained VQ-VAE remains fixed during diffusion training. The denoiser operates on unquantized encoder latents; sampled latents are quantized once before decoding into pore fields.

The learned Constructor converts each decoded field into discrete building-block assignments. PORMAKE realizes these assignments as periodic atomic structures, and Zeo++ measures their pore apertures. The component models and structural realization procedure are detailed below.

A.2 PERIODIC U-DIT ARCHITECTURE

Tokenization and network layout. The denoiser receives a noisy encoder latent $z_t \in \mathbb{R}^{3\times 8\times 8\times 8}$. Each spatial site becomes one three-channel token, giving 512 tokens; a shared linear projection maps each token to width 768. All eight transformer blocks retain this resolution and width. Each block has 12 attention heads of dimension 64 and a feed-forward expansion ratio of four. The U-shaped layout refers to long feature skips across depth, with no spatial downsampling or upsampling. A token-wise output head predicts three noise channels, which are reshaped to the input lattice. The unconditional backbone has 95,706,627 parameters; continuous PLD conditioning extends the same architecture.

Toroidal relative attention. For sites $r_i, r_j \in \{0,\dots,7\}^3$, define the directed wrapped displacement $\Delta_{ij} = (r_i - r_j) \bmod 8$. The lookup index is $64\Delta_{ij,x} + 8\Delta_{ij,y} + \Delta_{ij,z}$. Each block learns its own bias table $b \in \mathbb{R}^{12\times 512}$, with one scalar per head and wrapped displacement. A shared linear map produces queries, keys and values. For head h,

$$A_{ij}^{(h)} = \mathrm{softmax}_j\left(\frac{\langle q_i^{(h)}, k_j^{(h)}\rangle}{\sqrt{64}} + b_h[\Delta_{ij}]\right), \qquad o_i^{(h)} = \sum_j A_{ij}^{(h)} v_j^{(h)}. \tag{10}$$

Attention spans all 512 sites. Head outputs are concatenated and linearly projected to width 768. Directed offsets retain orientation while identifying displacements across opposite cell faces. The model has no absolute spatial embedding; the sinusoidal timestep embedding encodes diffusion time only. Bias tables are initialized to zero and learned independently across blocks.

Conditioning and AdaLN-Zero blocks. A 768-dimensional sinusoidal timestep embedding, with frequency base 10,000, passes through a $768\to 3072\to 768$ MLP with SiLU activation to obtain c_t. The unconditional model uses $c = c_t$. The PLD-conditioned model first normalizes its PLD target as $u = d^\star/15$ and constructs

$$\phi(d^\star) = \big[u, \sin(2\pi u), \cos(2\pi u), \sin(4\pi u), \cos(4\pi u), \sin(8\pi u), \cos(8\pi u)\big]. \tag{11}$$

A $7\to 128\to 128$ MLP uses SiLU after each layer; a linear $128\to 768$ projection gives c_d. The conditioned model uses $c = c_t + c_d$ in every block and the output head. The PLD projection is zero-initialized, so adding the conditioner initially preserves the unconditional computation. The implemented target domain is $[0, 15]$ Å. Conditioning is spatially broadcast through adaptive normalization, with no conditioning dropout or classifier-free guidance. The conditioner adds 116,608 parameters, giving 95,823,235 trainable parameters in total.

Each block maps c through SiLU and a linear $768\to 4608$ layer to obtain attention and MLP shift, scale and gate vectors $(\beta_a, \gamma_a, g_a, \beta_m, \gamma_m, g_m)$. For input H, the block computes

$$\widetilde{H} = H + g_a \odot \mathrm{Attn}\big((1+\gamma_a)\odot \mathrm{LN}_a(H) + \beta_a\big), \tag{12}$$

$$H' = \widetilde{H} + g_m \odot \mathrm{MLP}\big((1+\gamma_m)\odot \mathrm{LN}_m(\widetilde{H}) + \beta_m\big). \tag{13}$$

Condition vectors are broadcast over tokens. Both layer normalizations have no learned affine parameters and use $\epsilon = 10^{-6}$. The feed-forward MLP has dimensions $768 \to 3072 \to 768$ and tanh-approximated GELU activation. The modulation layer is zero-initialized, including its bias, giving zero initial residual gates (Peebles & Xie, 2023).

Long skips and output head. Let H_ℓ be the output of block ℓ. Outputs $H_1, \ldots, H_4$ are cached. Before block $\ell \in \{5, 6, 7, 8\}$, its current input $H_{\ell-1}$ is concatenated with $H_{9-\ell}$ and projected from 1536 to 768 dimensions using a separate learned linear layer. This implements skips $4 \to 5$, $3 \to 6$, $2 \to 7$ and $1 \to 8$ (Bao et al., 2023). In particular, block 5 receives the current and cached copies of H_4. The final token-wise layer normalization is modulated by a condition-dependent shift and scale, followed by a $768 \to 3$ output projection. Both the final modulation layer and output projection are zero-initialized. Training can use activation checkpointing within blocks.

Diffusion schedule and optimisation. Training uses a 1,000-step scaled-linear variance schedule with β_t interpolated between 10^{-4} and 2×10^{-2}, an ϵ-prediction objective, AdamW with learning rate and weight decay 10^{-4} and moment coefficients $(0.9, 0.999)$, linear warmup over 1,000 updates, BF16 mixed precision, a micro-batch of four with eight gradient-accumulation steps (effective batch 32) and gradient-norm clipping at 1.0. Neither exponential moving averages nor early stopping were used. Unconditional pretraining lasts 60,000 updates. Conditional training starts from these weights with a reset optimizer, initially runs for 60,000 updates, and continues to 300,000 conditional updates with the optimizer state, data order, objective, learning rate and noise schedule preserved. The evaluated inference configuration is specified in Appendix B.

Periodic interpretation. A cyclic grid shift permutes tokens while leaving their pairwise wrapped displacements unchanged. Shared token-wise operations, global conditioning and site-aligned skip fusion preserve this symmetry, so the complete network is equivariant to integer cyclic shifts along any combination of the three grid axes. This guarantee is specific to translations on the discrete 8^3 torus and does not imply rotation, continuous-translation or lattice-basis equivariance. This construction motivates the cyclic-shift diagnostics in Appendix E.1; physical realization is assessed separately.

Numerical verification. Numerical tests verified the cyclic-shift identity $f(T_{\mathbf{s}} z_t, t, y) = T_{\mathbf{s}} f(z_t, t, y)$ along each axis and under simultaneous multi-axis shifts within relative and absolute tolerances of 2×10^{-5}. Additional tests confirmed the absence of absolute position parameters, directional wrap-around indexing of the bias tables, exact equality between the unconditional model and the zero-initialized conditional model at initialization, gradient flow into the PLD conditioner, and rejection of non-finite or out-of-domain targets.

A.3 Learned Constructor architecture and evaluation

The Constructor is the learned structure-realization stage that follows a Periodic U-DiT pore-field sample. It maps a decoded four-channel 32^3 signed-distance field (SDF) to a topology and discrete building-block assignments, and expands the highest-scoring assignments into a compact candidate set for a single generated field.

Architecture. The Node Transformer reads the decoded field as 512 spatial tokens through a non-overlapping 4^3 projection. Two learned class tokens carry the inorganic and organic node heads, and a topology embedding conditions the shared representation. Four pre-norm Transformer layers of width 256, eight heads, feed-forward width 1024 and dropout 0.1 feed independent linear classifiers (approximately 3.98M parameters).

Training the node heads on an equal mixture of raw and reconstructed fields keeps the training input close to the reconstructed field that the model sees at inference, and improves accuracy at both heads (Table 3). Validation uses reconstructed fields, the predicted topology and a soft node-pair compatibility penalty.

The Edge Transformer conditions on the topology, both node assignments and the four-channel SDF. It trains on raw fields with ground-truth topology and node labels, which grounds each linker prediction in a specific node pair instead of the field alone. For the single-head comparison in Table 4,

Node head and training input	Top-1	Top-5
Inorganic node, raw SDF	46.27%	74.60%
Inorganic node, raw/reconstructed mixture	47.97%	75.52%
Organic node, raw SDF	96.96%	98.13%
Organic node, raw/reconstructed mixture	97.04%	98.17%

Table 3: **Node prediction and input-domain adaptation.** Both variants share one initialization, schedule and training budget, and are evaluated on reconstructed fields with the predicted topology and a soft node-pair compatibility penalty ($\lambda = 0.5$). Both heads are scored on the full 13,691-structure split.

Edge input condition	Top-1	Top-5
SDF-only control	36.08%	72.38%
Topology/node-conditioned	45.92%	83.32%

Table 4: **Component-conditioned linker identification.** Both models train under matched schedules and are evaluated at their best validation checkpoints on the same 13,691 reconstructed fields. The conditioned model receives ground-truth topology and node labels. Values are single-head classification accuracies.

both rows share the same 13,691-example validation split and the same reconstructed fields: the conditioned row receives ground-truth topology and node labels, and the SDF-only control receives neither, so the gap between the two rows isolates the contribution of component conditioning. The edge vocabulary contains 646 training classes, together with an `__unk__` output for out-of-vocabulary labels that is excluded from generation.

At inference, the topology is fixed at its Top-1 prediction. The Node stage produces the five highest-scoring joint node pairs, and the Edge stage predicts five conditional edge candidates for each pair. For a node pair n and linker ℓ, the combined score is

$$S(n, \ell) = \log P(n \mid \widehat{x}, \widehat{\tau}) + \log P(\ell \mid n, \widehat{\tau}, \widehat{x}). \tag{14}$$

The assignments are globally ranked by S, with duplicates removed, yielding at most 25 alternatives. The w/o verifier and MEF variant takes the global first-ranked assignment. Its node pair can therefore differ from the Node head's first-ranked pair, because the conditional linker score also contributes to selection.

MOFFUSION-style component reference. Table 5 reports a MOFFUSION-style ResNet Constructor that we retrained on our expanded framework corpus with our 646-class edge vocabulary. This reference uses raw ground-truth SDFs and ground-truth topology on the 13,691-structure test split. The Transformer studies above use reconstructed SDFs on the validation split and separately specify their conditioning inputs.

A.4 Multi-Edit-Former architecture and training

Task formulation. Given a source linker with $n \leq 3$ selected substitution sites, source states $s = (s_1, \ldots, s_n)$, source measured PLD d_0 and requested target t, Multi-Edit-Former (MEF) ranks *complete* destination assignments $a = (a_1, \ldots, a_n)$ jointly rather than selecting each site independently. The formal destination vocabulary is $\{\mathrm{H}, \mathrm{F}, \mathrm{Cl}, \mathrm{Br}, \mathrm{Me}, \mathrm{OH}, \mathrm{NH_2}, \mathrm{CN}, \mathrm{CF_3}, \mathrm{Ph}\}$; ethyl, methoxy and formyl groups are source-only states that can be retained or replaced but not newly installed. The complete Cartesian assignment space contains at most 10^3 formal assignments before deterministic single-site and pairwise constructability masks are applied, so MEF scores the enumerated assignments that satisfy these chemical constraints.

Invariant encoding and joint score. Each site is embedded by its source state together with eight invariant scalars relative to the two linker anchors; six pairwise features describe inter-site distance and relative axial/radial geometry, and a numeric context vector carries the normalized source measured PLD, anchor separation, atom count and the mean and standard deviation of the selected sites'

Table 5: **MOFFUSION-style Constructor retrained on our corpus.** Component classification on 13,691 test structures using raw SDFs. Building-block heads receive ground-truth topology. The edge head uses the reproduced 646-class vocabulary. Node 2 includes the empty-node class.

Model	Head	Top-1 (%)	Top-3 (%)	Top-5 (%)
Topology (10 epochs)	Topology	96.98	99.08	99.32
BB (10 epochs)	Node 1	28.12	50.98	60.95
	Node 2	96.33	96.74	96.96
	Edge	26.89	51.40	63.09
BB (40 epochs)	Node 1	37.38	59.04	67.66
	Node 2	96.69	97.14	97.52
	Edge	23.77	49.65	62.99

axial coordinates. A four-layer, eight-head Transformer ($d = 256$, feed-forward dimension 1024, dropout 0.1) encodes one context token and the site set, with pair geometry entering site-to-site attention through a learned symmetric bias that is shared across heads in the evaluated model. Because the encoder consumes distances and anchor-relative scalar projections, it is invariant to global translation and rotation; it is not an SE(3)-equivariant coordinate network. The target encoder represents the requested PLD and its difference from the source through two normalized scalars and eight Gaussian radial basis functions centred at the integer targets from 3 to 10 Å. Every legal assignment receives the score

$$S_\theta(a \mid s, x, d_0, t) = \sum_i u_\theta(h_i, E'_s(s_i), E_a(a_i), q) + \sum_{i<j} p_\theta(h_i, h_j, E_a(a_i), E_a(a_j), \bar{g}_{ij}) + g_\theta\left(\sum_i [h_i + E_a(a_i)], q\right) - \lambda_{\text{edit}} d_H(a, s), \tag{15}$$

where E_a is the destination-state embedding, $\bar{g}_{ij} = (g_{ij} + g_{ji})/2$ symmetrizes each pair term, d_H is the Hamming edit count and $\lambda_{\text{edit}} = 0.05$. The unary and pooled global terms are target-conditioned; the pair term is target-independent in the evaluated model. Exact score ties are resolved by fewer edits and then by a stable assignment key. An auxiliary PLD residual head has zero weight and does not contribute to action selection.

Training objective. Training tasks are constructed from measured PLD endpoints, and only measured legal destinations enter the denominator; unmeasured assignments are not treated as negatives. A measured destination is positive when it lies within 0.5 Å of the target, or, when no destination reaches that window, when it lies within 0.15 Å of the best measured target error; for source–target gaps above 3 Å, direction-improving positives are retained when available. MEF is trained with a multi-positive listwise ranking loss

$$\mathcal{L}_{\text{rank}} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\sum_m C_{bm} P_{bm} \exp S_{bm}}{\sum_m C_{bm} \exp S_{bm}}, \tag{16}$$

where C marks measured legal candidates and P marks positives. It was warm-started from a geometry-aware first-stage checkpoint and calibrated for two epochs over $82{,}334$ target-augmented tasks ($1{,}287$ steps per epoch), with $4{,}366{,}260$ parameters, AdamW (Loshchilov & Hutter, 2019) (learning rate 3×10^{-4}, weight decay 10^{-4}) and gradient clipping at 1.0.

Lineage-disjoint data partitions. The historical MEF benchmark groups related linker assets into lineages using base-asset identity and graph matching, then assigns complete lineages to training, development or held-out partitions. The three partitions contain 92, 12 and 11 lineages, respectively, and 475, 58 and 47 framework contexts. All pairwise intersections are zero for linker lineages, base-linker keys and base-XYZ identities. Training tasks draw exclusively from training contexts, with candidate menus restricted to measured, chemically legal destination states in the corresponding context.

Table 6: **MEF architecture and training configuration.** The same trained editor is used throughout inference.

Component	Setting
Selected sites	1–3
Formal destination states	H, F, Cl, Br, Me, OH, NH_2, CN, CF_3, Ph
Source-only states	Et, OMe, CHO
Formal-only menu	at most 10^3, before constructability masking
Site / pair features	8 (or 6) scalars + 16-centre RBF of anchor or inter-site distances
Encoder	4 layers, 8 heads, $d = 256$, FFN 1024, dropout 0.1
Pair attention bias	symmetric, shared across heads in the evaluated model
Assignment score	unary + symmetric pair + pooled global $-0.05\times$ edit count
Calibration	$82{,}334$ tasks, 2 epochs, $1{,}287$ steps per epoch
Parameters	$4{,}366{,}260$
Optimizer	AdamW, lr 3×10^{-4}, weight decay 10^{-4}, clip 1.0
Benchmark target grid	integer PLDs 3–10 Å
Training endpoint	measured PLD
Preservation gate	10^{-3} Å for anchors and retained backbone

Feature	Train	Dev	Held-out	Train∩Dev	Train∩Held	Dev∩Held
Framework contexts	475	58	47	0	0	0
Linker lineages	92	12	11	0	0	0
Base-linker keys	92	12	11	0	0	0
Base-XYZ identities	92	12	11	0	0	0
Topologies	147	35	32	32	26	12
Topology–node scaffolds	336	52	43	35	19	10

This partitioning evaluates transfer to new linker lineages under partially shared framework environments. Of the 32 held-out topologies and 43 topology–node scaffolds, 26 and 19 also occur in training. Here, scaffold denotes the MOF topology–node combination. The historical held-out partition has also supported subsequent diagnostics. This training-data audit is distinct from the paired generated-source editing experiment in Appendix E.5.

Continuous target conditioning. The target encoder receives numerical PLD values through normalized scalars and fixed Gaussian bases centred at 3–10 Å. These fixed centres provide a continuous expansion of the target and its difference from the source PLD. The chemical assignment menu is discrete, while conditioning varies continuously with the requested pore diameter. The historical evaluation uses integer targets on the 3–10 Å grid, obtained as $t = \mathrm{round}(d_{\mathrm{raw}})$ in target construction. This grid standardizes candidate evaluation and hit-rate reporting; it does not restrict the encoder to categorical target labels.

Execution and scope. The selected assignment is materialized by a deterministic source-preserving editor that supports H→FG, FG→H and FG_1 →FG_2 transitions; anchor and retained-backbone displacements must stay below 10^{-3} Å, and failed chemistry, clash or structural checks count as failed proposals and do not trigger substitution by a lower-ranked action. The edited linker is reassembled and geometrically evaluated as described in Appendix B. Assembly, relaxation and Zeo++ are outside the differentiable training objective. The model receives source-state tokens, linker and site geometry, source measured PLD and the requested target, but no categorical topology/node inputs or additional base-edge descriptors. Its geometric information is therefore local to the edited linker. It is trained on measured PLD, with continuous numerical target conditioning and a standardized integer evaluation grid from 3 to 10 Å. Joint scoring represents non-additive assignment preferences; the resulting physical pore response is assessed through reassembly and Zeo++ rather than inferred from the score decomposition.

B VERIFIER-GUIDED GENERATION AND REFINEMENT

B.1 DIRECT GEOMETRIC VERIFICATION

Learned scores prioritize candidate recipes and linker edits, while Zeo++ evaluates their realized pore apertures (Willems et al., 2012). For an assembled framework C, the verifier returns $V(C) = \mathrm{PLD}_{\mathrm{Zeo++}}(C)$, and candidate selection minimizes $e(C; d^\star) = |V(C) - d^\star|$. All structure-selection and edit-refinement decisions use measured apertures.

This role differs from pre-assembly property prediction. LLM4MOF uses MOF2Zeo to rank topology–node–linker combinations before realization and relaxation, then uses Zeo++ to characterize the resulting structures (Nam et al., 2026). CoFi-MOF applies its verifier after atomic frameworks are available, so direct geometric measurement can guide selection and assess each linker modification without training an additional surrogate.

B.2 CANDIDATE SELECTION AND LOCAL REFINEMENT

Algorithm 1 describes one generation attempt. The PLD-conditioned model, trained for 300,000 updates, samples one pore field with 200-step deterministic DDIM. The learned Constructor predicts one topology, retains five joint inorganic/organic node-pair assignments and predicts five conditional linkers for each pair. The resulting assignments are deduplicated and ranked by their combined model scores before PORMAKE assembly. This procedure realizes alternative chemical structures from the same generated field. Successfully assembled structures with finite positive measured PLD enter verification-based selection; assembly failures remain recorded.

Refinement starts from the measured candidate closest to the target whose linker supports the permitted edits. This editable source can differ from the best initial pore match, which remains available for final selection. Every successfully materialized and assembled edit is measured by Zeo++; failed proposals remain recorded. Editing is performed even when the source already satisfies the reporting tolerance, allowing improvement at finer resolution.

The final output is the closest measured pore match among the original and edited structures, with ties favoring an original candidate. If no valid edit is available, the best original structure is retained; if no original structure has a valid measurement, the attempt fails. Retaining the initial candidates makes the best measured aperture error non-increasing within the attempt. The selected output is not fed back into a second editing round.

B.3 COMPONENT ABLATIONS

The three variants in Table 2 share the same generated fields and learned Constructor. CoFi-MOF w/o verifier and MEF directly outputs the globally highest-scoring assignment under the combined node and conditional linker log-probabilities, without aperture-based selection or editing. Zeo++ is applied afterward to evaluate the output, without changing the selected structure. CoFi-MOF w/o MEF uses direct geometric verification to select among the Constructor's assembled alternatives. Full CoFi-MOF additionally proposes at most ten distinct, non-KEEP linker edits and measures all successfully assembled edits before final selection. Each attempt produces at most one output.

The comparison isolates the successive roles of measured structural selection and local correction within the complete pipeline. Appendix E.5 separately evaluates learned edit ranking against random legal modifications from shared sources, using both first proposals and verified best-of-ten edited products. Supplementary relaxation characterizes selected structures without retrospective reselection.

B.4 DIRECT VERIFICATION RUNTIME

In a serial CPU benchmark over 40 structures from the MOFDiff and MOFFlow-2 pre-relaxation CIF evaluation package, Zeo++ 0.3 required an average of 316 ms per structure. The test used a shared server with 256 CPU cores and 503 GB RAM, without GPU resources. The approximately 0.3 s figure describes this benchmark configuration, not end-to-end CoFi-MOF inference time.

Algorithm 1 CoFi-MOF: verifier-guided generation and linker refinement

Require: Target PLD $d^\star$; learned Constructor; at most 10 linker edits

Draw attempt noise $\xi \sim \mathcal{N}(0, I)$; $\widehat{z}_e \leftarrow \mathrm{DDIM}_\theta(\xi, d^\star)$
Decode one SDF $\widehat{x} = D(Q(\widehat{z}_e))$
$\widehat{\tau} \leftarrow \mathrm{PredictTopology}(\widehat{x})$
$\mathcal{N} \leftarrow \mathrm{NodeTransformer}(\widehat{x}, \widehat{\tau})$ ▷ five ranked node pairs
$\mathcal{P} \leftarrow \varnothing$
for each node pair $n \in \mathcal{N}$ **do**
    $\mathcal{L}_n \leftarrow \mathrm{EdgeTransformer}(\widehat{x}, \widehat{\tau}, n)$ ▷ five conditional linkers
    Add $(\widehat{\tau}, n, \ell)$ for $\ell \in \mathcal{L}_n$ to $\mathcal{P}$
end for
Rank and deduplicate $\mathcal{P}$ using node and conditional linker scores
$\mathcal{C} \leftarrow \mathrm{AssembleAndVerify}(\mathcal{P})$
if $\mathcal{C} = \varnothing$ **then**
    return failure
end if
$\mathcal{E} \leftarrow \varnothing$; $\mathcal{S} \leftarrow \mathrm{EditableSources}(\mathcal{C})$
if $\mathcal{S} \neq \varnothing$ **then**
    $C_s \leftarrow \arg\min_{C \in \mathcal{S}} e(C; d^\star)$
    $\mathcal{A} \leftarrow \mathrm{MEF}(C_s, V(C_s), d^\star)$ ▷ ranked edits; unchanged assignments excluded
    $\mathcal{E} \leftarrow \mathrm{AssembleAndVerify}(\mathrm{Apply}(C_s, \mathcal{A}))$
end if
$\mathcal{F} \leftarrow \mathcal{C} \cup \mathcal{E}$
$C_{\mathrm{sel}} \leftarrow \arg\min_{C \in \mathcal{F}} e(C; d^\star)$ ▷ ties favor the original structure
return $(C_{\mathrm{sel}}, V(C_{\mathrm{sel}}))$ ▷ assembled structure and measured PLD

C Datasets and edit-response supervision

C.1 MOF-Edit-DB construction

MOF-Edit-DB connects chemically specified linker modifications to the pore responses of their realized frameworks. We expand the PORMAKE edge library (Lee et al., 2021) with carboxylate-based linkers informed by reviews of Zr-MOFs (Bai et al., 2016) and titanium coordination polymers (Du et al., 2024). This carboxylate-based extension contributes additional building blocks to PORMAKE's structural assembly workflow and is distributed with the MOF-Edit-DB resources. Functional-group attachment and removal further diversify these linkers, following the broader linker-design strategies explored in inverse design and local optimization (Yao et al., 2021; Cleeton & Sarkisov, 2025). Each edit records its substitution sites, fragment identity and parent-linker mapping. The full base-edge catalog contains 844 entries, comprising 695 geometry-asset entries and 149 crosswalk aliases. The Constructor's supervised prediction vocabulary uses the 646 linker classes observed in its training split. The catalog, training vocabulary and chemically distinct linker count describe different populations. Table 7 separately summarizes base-edge and edited-edge resources.

A paired example associates a framework context c and linker edit a with

$$(c, a) \longmapsto (d_{\mathrm{parent}}, d_{\mathrm{edited}}, \Delta d), \qquad \Delta d = d_{\mathrm{edited}} - d_{\mathrm{parent}}. \tag{17}$$

The context identifies the topology, inorganic and optional organic nodes, and parent linker. Parent and edited frameworks share this chemical and topological context, while their realized geometries may differ. The resulting action–response representation supports learning how local chemical changes affect framework-level pore apertures.

C.2 Pore labels, paired records and quality screening

Frameworks are assembled with PORMAKE, relaxed with UFF4MOF in LAMMPS (Addicoat et al., 2014; Thompson et al., 2022), and characterized with Zeo++ (Willems et al., 2012). The released pore labels therefore describe force-field-relaxed structures. PLD and LCD correspond to the Zeo++ D_f and D_i fields; PLD does not exceed LCD in any released record.

Table 7: MOF-Edit-DB v1.2. Framework counts are records; linker counts are catalog entries. Canonical paired responses are the subset used in Figure 5.

Resource	Count	Role
Framework records	138,771	Pore labels with framework context
Parent records	26,058	Reference structures
Edited records	112,713	Products of annotated linker modifications
Canonical paired responses	108,145	Before/after PLD supervision
Edited-edge catalog	78,961	Edit identities and asset mappings
Base-edge catalog	844	695 asset entries and 149 aliases

The master table contains 138,771 records: 26,058 parents and 112,713 edited frameworks. Among the edited records, 108,145 have a canonical parent PLD and a paired response; the remaining 4,568 retain their own pore labels but are excluded from paired-response statistics. Counts refer to records rather than independently deduplicated crystal structures. The collection spans 584 topology labels and 37,428 canonical scaffold keys. Table 7 summarizes its framework and linker resources.

Response statistics use the master table's `delta_pld`, which agrees with `modified_pld − base_pld` within 10^{-4} Å wherever populated. Canonical identifiers and scaffold mappings define the parent–edit associations; auxiliary comparisons with alternative parent references are documented separately in the data README.

MOFChecker-based screening addresses atomic overlaps and detached molecular fragments (Jin et al., 2025). The released quality annotations contain no positive overlap or lone-molecule flags, although these individual flags are unrecorded for 201 and 1,731 records, respectively. The retention policy includes 9,729 overcoordination-only records and 342 records flagged as not three-dimensionally connected. These annotations support stricter task-specific subsets without treating every retained record as satisfying an identical set of structural checks.

C.3 Pore coverage and edit responses

Using each record's canonical `modified_pld`, 72,855 of 138,771 records (52.5%) have $0 < \mathrm{PLD} < 7$ Å. All records have finite positive values, so this proportion also describes the full positive-PLD population. The collection therefore provides substantial coverage of the small-aperture regime together with chemically specified edit responses. Across all records, PLD has mean 7.933 Å and median 6.765 Å, with 5th and 95th percentiles of 2.548 and 17.104 Å. Figure 5 shows the full empirical distributions. The canonical pairs comprise 65,198 functional-group attachment examples and 42,947 removal examples. Attachment has median ΔPLD -2.709 Å, with 9.1% positive responses; removal has median $+0.423$ Å, with 77.8% positive responses.

These contrasting shifts provide examples of both pore contraction and expansion, while the within-direction variation makes the framework context informative: the edit category alone does not determine the sign or magnitude of the response. The two edit groups contain different parent populations, so their distributions describe the sampled actions rather than a matched comparison of attachment and removal.

C.4 Learning tasks and data access

The paired records support edit-response prediction, target-conditioned action ranking, and transfer evaluation across scaffold or linker families. Family-disjoint evaluation should keep each parent and its related edits together; a standardized public benchmark split is not supplied in the current release. The review package provides data tables, edit catalogs, quality annotations and a structure manifest, with geometry availability documented in the accompanying README. The compact package does not contain every referenced geometry.

MOF-Edit-DB is distinct from the global generator's mixed training corpus, which combines selected small-pore MOFFUSION structures (Park et al., 2025) with self-generated frameworks. Its release size is not the training-set size of each CoFi-MOF component. Likewise, the relaxed labels summarized here are distinct from the unrelaxed-structure labels used to train MEF.

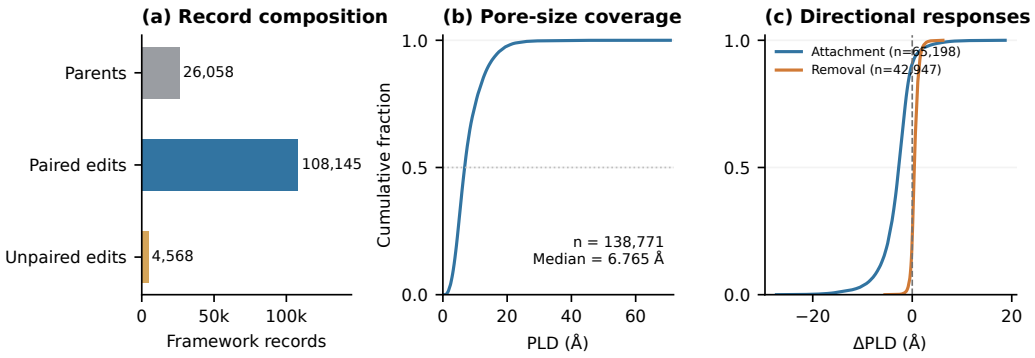


Figure 5: **Coverage and edit responses in MOF-Edit-DB.** (a) Parent records, paired edits and edits without a canonical parent comparison. (b) PLD coverage over all 138,771 records. (c) Canonical parent-to-edited ΔPLD for functional-group attachment and removal, normalized separately. Curves show empirical cumulative distributions over the full observed ranges; the vertical line denotes no change. All response statistics use the master table.

C.5 Data underlying the PLD motivation figure

Figure 1 compares the selected experimental CoRE collection with the processed MOFFUSION collection and a BW-DB release derived from the hypothetical MOFs of Boyd et al. (2019). For CoRE MOF 2019 v2.0.1, we use the source tables `ASR_20241119.csv` and `FSR_20241119.csv`, retaining entries marked `CR=Y`. The 6,603 ASR and 3,764 FSR entries have finite positive PLDs. Pooling them gives 10,367 entries, of which 9,173 (88.5%) lie strictly below 7 Å. ASR and FSR denote removal of all solvents and free solvents, respectively (Chung et al., 2019); the pool counts preparation variants rather than unique experimental frameworks. The processed MOFFUSION table contains 247,744 rows, with 247,732 finite positive PLDs and 12 missing or invalid values; 53,788 positive values (21.7%) are below 7 Å.

For BW-DB, we analyze `MetalOxo.lmdb` from the `bw_db.tar.gz` release distributed with MOFDiff[1]. This release contains 308,962 records, of which 308,961 have a PLD value. Excluding the 39 nonpositive values leaves 308,922 positive-PLD entries. Among these, 72,797 satisfy $0 < \mathrm{PLD} < 7$ Å, giving 23.56%. BW-DB is an upstream structural resource for MOFDiff, MOFFlow and MOFFlow-2 (Fu et al., 2024; Kim et al., 2025b;a). These methods apply their own preprocessing and data selection, so the figure characterizes the analyzed BW-DB collection rather than each model's final training split.

All percentages use the corresponding positive-PLD population and are calculated from counts below the strict 7 Å threshold, independently of curve smoothing. The curves use a reflected Gaussian kernel with bandwidth 0.25 Å and a 0.02 Å evaluation grid. Normalization retains the full positive-PLD population, including the tail beyond the displayed 20 Å range. Source-reported descriptors and preparation settings are retained.

The contrast is consistent across both hypothetical collections, strengthening the motivation for dedicated generation in the experimentally prevalent small-aperture regime. These statistics describe aperture coverage through PLD, rather than the size of every cavity within a framework.

D Experimental setup and evaluation

D.1 Structural realization and pore measurement

Recipes are assembled with PORMAKE (Lee et al., 2021) using topology, inorganic and optional organic node, and linker identifiers. Augmented-edge geometries supply building blocks absent from

[1] `https://zenodo.org/records/10806179`

the original library. Frameworks are written as P1 CIFs with assembly-defined bonds. A minimum-image check excludes structures with any atom pair closer than 0.1 Å before pore characterization.

Zeo++ provides the largest cavity diameter, pore-limiting diameter and largest free-path diameter. Finite positive PLD determines eligibility for target-error selection; missing and non-finite measurements are failures rather than zero-valued PLDs.

Supplementary relaxation uses UFF4MOF generated by `lammps-interface`, with `units real`, `atom_style full` and periodic boundaries. Three conjugate-gradient stages perform fixed-cell minimization, isotropic cell relaxation to zero pressure, and final fixed-cell minimization. Each uses energy and force tolerances of 10^{-6} and 10^{-8}, with a 180-second wall-clock limit. Relaxation failures affect this supplementary analysis, not the primary endpoint.

D.2 Evaluation populations and statistics

Targets are $3, 4, 5, 6, 7$ Å, with 1,000 allocated generation attempts per target. For CoFi-MOF and its matched variants, the evaluated population contains 4,994 completed attempts; six remain pending. Completed assembly and measurement failures remain in the hit-rate denominator, while pending outcomes are identified separately. Hit@ τ denotes an absolute PLD error no greater than τ; MAE and signed bias are calculated over selected structures with finite positive PLD.

The initial batch contains 100 attempts per target, followed by a continuation allocating 900 per target. Pooled rates aggregate counts across these batches. Paired uncertainty for the component ablations and editing comparisons resamples shared noise identifiers using a block bootstrap with 2,000 resamples and a fixed seed; batch-specific intervals are reported separately. The periodic-backbone comparison uses its own 500-pair design and uncertainty calculation, described in Appendix E.1.

Within-batch structural uniqueness measures distinct frameworks among measured outputs and is reported only for the 500-attempt batch evaluated by structure matching. It is distinct from novelty relative to a training collection. Baseline-specific sampling, selection and deduplication are detailed below; the table populations retain those distinctions.

D.3 Baseline configurations

The external-model comparison evaluates targets of 3–7 Å with 1,000 allocated attempts per method and target. For MOFFlow-2 and MOFDiff, pore characterization uses retained pre-relaxation CIFs from the formal runs: 911–938 and 899–932 assembled structures per target, respectively, with finite positive PLDs. MOFFUSION uses the retained raw assemblies from its target-conditioned validation runs; six target-3 structures have zero PLD and are excluded from the finite-positive population.

GA: target-conditioned search. The genetic algorithm searches combinations of 30 topologies, 208 metal-cluster nodes and 3,334 linkers, using reference data curated for the 0–7 Å pore range. Its target-dependent fitness is

$$\mathrm{fitness}(g, d^\star) = \exp\big(-((\mathrm{PLD}(g) - d^\star)/0.5)^2\big), \tag{18}$$

where PLD is obtained from a measured library or an on-demand Zeo++ evaluation. The population size is 24, with four elites, tournament size three, single-field crossover probability 0.8 and single-field mutation probability 0.25. Each target is allocated 1,000 scored proposals.

Table 1 evaluates the full combinatorial search space using 1,000 scored proposals per target. Repeated proposals are retained in all attempt-level metrics, with cached geometric measurements reused for repeated combinations. Failed proposals count as misses, and MAE and bias are averaged over proposals with finite positive measured PLD, retaining their multiplicities. Across targets of 3–7 Å, the 1,000 proposals contain 253, 222, 236, 213 and 232 distinct topology–node–linker combinations, respectively, corresponding to repetition rates of 74.7–78.7%. Of these distinct combinations, 144, 154, 146, 152 and 149 have finite positive PLD. The run required 883 new geometric measurements, and 99.9% of its unique combinations were absent from the preloaded measured library. These combination counts characterize repetition separately from the attempt-level targeting results.

MOFDiff: target-directed pool selection. We adapt MOFDiff (Fu et al., 2024) using its building-block diffusion model and a separately trained PLD prediction head. For each requested target, the head scores 15,351 posterior latents from the official held-out validation split, and the 1,000 latents with smallest predicted target error are decoded and assembled. Zeo++ characterizes these assemblies before force-field relaxation.

This adaptation is labeled *MOFDiff + selection*: its target-directed retrieval-and-realization procedure uses a PLD predictor to choose promising latent candidates before decoding and assembly. Thus, the comparison includes predictor-guided candidate selection rather than unselected unconditional samples. Target subsets may overlap; targets 3 and 4 share 780 selected latents. Each selected latent counts as one final-stage attempt, with shared-pool scoring constituting an additional inference cost.

MOFFlow-2: learned PLD conditioning. The MOFFlow-2 adaptation (Kim et al., 2025a) uses the released data and checkpoints, initializing the sequence generator from its unconditional checkpoint and fine-tuning continuous PLD conditioning through property embeddings and cross-attention. Labels use the database's `largest_free_sphere_diameter [A]` field; the filtered training and validation sets contain 187,028 and 10,242 structures. Sequence-model fine-tuning lasts two epochs, using checkpoint `epoch_1-step_17540-loss_0.1640.ckpt`, while the structural generator retains its released weights.

Each target produces 1,000 attempts in five batches of 200. generation and measurement failures remain in the allocated denominator. This adaptation uses a trained condition rather than posterior-pool retrieval or target-based output selection.

MOFFUSION: conditional pore-field generation. The MOFFUSION adaptation (Park et al., 2025) is trained on the mixed corpus for 300,000 cumulative updates. Its target condition is scaled by $1/100$. Sampling uses 200-step deterministic DDIM with seed $20260915 +$ target. Following the original Constructor's top-5 prediction scheme, candidate recipes are assembled with PORMAKE and evaluated with Zeo++; the realized structure with PLD closest to the requested target is retained. The baseline therefore includes measured candidate selection rather than rank-1 realization alone. The experiment generates 1,000 attempts at each integer target from 3 to 11 Å; Table 1 evaluates targets 3–7 Å using the retained 841–894 raw assemblies per target.

E Additional results

The following analyses examine periodic consistency and structural realization, the roles of geometric verification and local editing, and the resulting pore-targeting accuracy. Structural-state sensitivity is reported separately from the primary evaluation.

E.1 Periodic consistency and structural realization

We examine whether explicitly periodic latent generation improves downstream MOF realization. We compare a MOFFUSION-style 3D U-Net with zero-padded convolutions against unconditional Periodic U-DiT (Appendix A.2). Figures denote these models by B0 and B1, respectively, and the PLD-conditioned variant by D2. The comparison controls the data, latent representation, objective, optimizer and training budget: both backbones were trained for 60,000 updates on the same 109,529 training structures encoded by the same frozen VQ-VAE encoder, using the ϵ-prediction objective, AdamW (10^{-4} learning rate and weight decay, 1,000 warmup updates, gradient-norm clip 1.0) and an effective batch size of 32 in BF16 on one RTX 4090, under one common seed. The test split was not accessed.

Matched sampling and paired statistics. Evaluation draws on a bank of 500 shared Gaussian noise tensors: attempt i from each backbone start from the same initial noise and are decoded with deterministic DDIM (200 steps, $\eta = 0$), then pass through an identical assembly–relaxation–Zeo++ chain with no stage-specific changes. Paired design removes variation from the diffusion noise draw, so stage-wise differences are assessed with the exact McNemar test on discordant pairs, and the final-validity difference is summarized with a paired-bootstrap 95% confidence interval over 10,000 replicates.

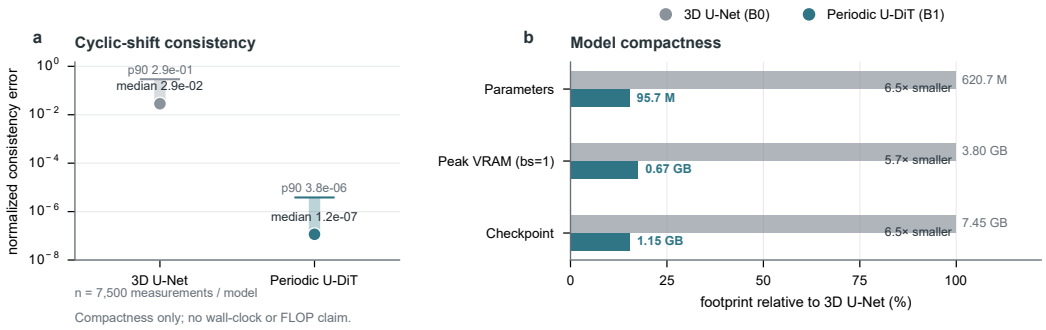


Figure 6: **Denoiser periodicity and model footprint.** (a) Normalized cyclic-shift consistency error over 7,500 probes per model (medians and 90th percentiles). (b) Parameter count, peak forward memory (batch size 1) and checkpoint size relative to the 3D U-Net. The footprint panel reports compactness only and makes no wall-clock or operation-count claim.

Table 8: **Spatial consistency of denoiser outputs.** Relative cyclic-shift error is aggregated over 7,500 probes. Uniform-input CV is measured at $t = 50$.

Model	Mean E_{eq}	Median E_{eq}	Uniform-input CV
3D U-Net	1.034×10^{-1}	2.859×10^{-2}	4.524×10^{-1}
Periodic U-DiT	1.227×10^{-6}	1.156×10^{-7}	6.105×10^{-6}
PLD-conditioned U-DiT	5.024×10^{-7}	1.067×10^{-7}	2.811×10^{-6}

Cyclic-shift and uniform-input probes. For shift T_Δ, we measure

$$E_{\mathrm{eq}} = \frac{\|\epsilon_\theta(T_\Delta z_t, t, c) - T_\Delta \epsilon_\theta(z_t, t, c)\|_2}{\|\epsilon_\theta(z_t, t, c)\|_2 + 10^{-12}}. \tag{19}$$

The protocol uses 500 held-out validation latents, timesteps $\{50, 500, 950\}$ and five shifts, yielding 7,500 measurements per model. The unconditional and PLD-conditioned periodic models have median relative errors near 10^{-7} (Table 8 and Figure 6), consistent with cyclic-shift equivariance up to numerical precision. A complementary uniform-input probe broadcasts the per-channel validation-latent mean to all sites; at $t = 50$ the coefficient of variation of token-wise output norms is 0.4524 for the 3D U-Net and below 10^{-5} for both periodic models.

Generated latent fields. Across the 500 paired generations, per-site variance over samples reveals whether boundary artefacts propagate into the generated distribution. The 3D U-Net suppresses variance near the unit-cell faces, with face-to-interior ratios $(0.8870, 0.9518, 0.9014)$ across the three latent channels, whereas Periodic U-DiT closely matches faces and interior at $(1.0008, 1.0043, 1.0053)$ and reduces the mean variance-map coefficient of variation from 0.0893 to 0.0561 (Figure 7). The sampled latent statistics thus complement the equivariance probes, showing substantially more uniform spatial variation across the unit cell.

Paired structural realization. Both models decode all 500 latents successfully, so the comparison emerges only after realization. Successful attempts increase from 320 for the 3D U-Net to 356 for Periodic U-DiT at assembly, from 309 to 349 after LAMMPS relaxation, and from 290 to 337 after Zeo++ characterization (Figure 8). End-to-end validity rises from 0.580 to 0.674 ($+9.4$ percentage points; paired-bootstrap 95% CI $[3.6, 15.0]$; one-sided exact McNemar $p = 8.2 \times 10^{-4}$). The difference is already evident at assembly and persists through the subsequent stages. Discordant final outcomes include 84 successes unique to the 3D U-Net and 131 unique to Periodic U-DiT, yielding 47 additional valid structures.

Resource footprint. Periodic U-DiT is also substantially more compact: 95.7 M parameters versus 620.7 M for the 3D U-Net ($6.5\times$ smaller), 0.67 GB peak forward memory versus 3.80 GB

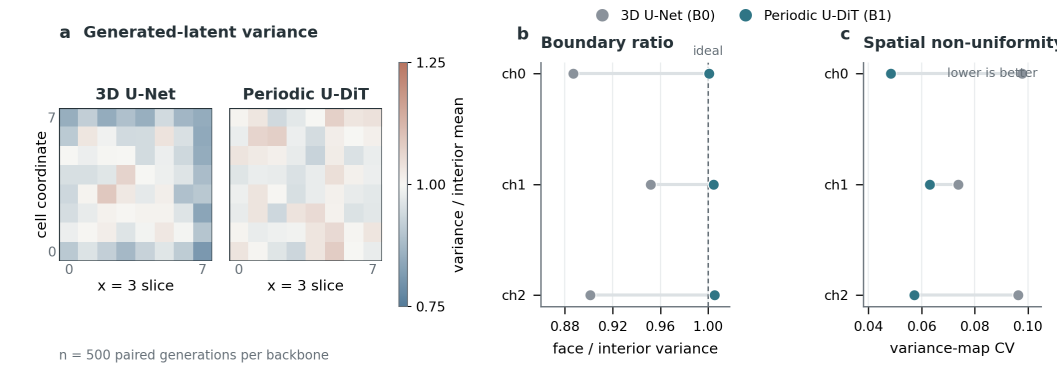


Figure 7: **Generated-latent spatial statistics over 500 paired generations.** (a) Per-site variance maps (x=3 slice). (b) Face-to-interior variance ratio per channel; cyclic-shift-invariant sampling gives a population ratio of 1, with finite-sample deviations. (c) Variance-map coefficient of variation; lower is more spatially uniform.

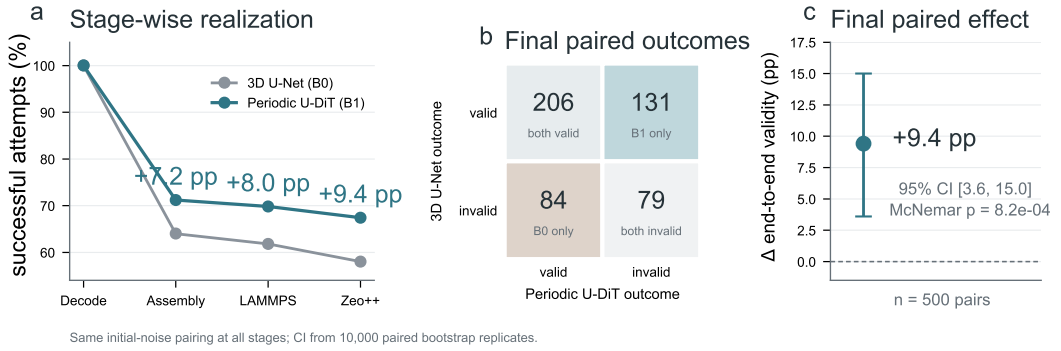


Figure 8: **Paired structural realization on 500 shared initial noises.** (a) Stage-wise success. (b) Discordant final outcomes; discordant pairs determine the McNemar test. (c) End-to-end validity difference with the paired-bootstrap 95% CI.

(5.7×), and a 1.15 GB checkpoint versus 7.45 GB (Figure 6b). A batch-one forward pass of Periodic U-DiT uses 668.5 MB peak memory and 69.40 GFLOPs. The smaller parameter and memory footprint coexists with a higher forward-operation count from attention over 512 tokens; throughput is not measured here.

Interpretation. Together, the probes establish periodic spatial consistency, and the paired comparison shows improved structural realization for the complete architecture. Backbone family, width and skip structure change alongside the attention bias; without an ordinary U-DiT control, the gain cannot be attributed to the toroidal bias alone. Position-free attention can also satisfy the two consistency probes. The results cover one training seed and one matched-noise panel. At 60,000 updates, denoising MSE is 0.07796 for the 3D U-Net and 0.08062 for Periodic U-DiT, so the observed structural advantage is not a lower training loss or demonstrated training-speed gain.

E.2 Geometric verification and local pore refinement

CoFi-MOF w/o verifier and MEF, recorded as the direct output in the archived run summaries, retains the learned Constructor and returns the global top-ranked assignment under the joint node–linker score. CoFi-MOF w/o MEF retains geometric verification but omits local linker refinement.

Evaluation coverage. The initial 500-attempt batch and the subsequent 4,500-attempt continuation cover targets of 3–7 Å. At the reporting snapshot, 4,994 attempts are complete; the six pending attempts are distributed as 0, 1, 0, 3, and 2 across the five targets. Table 2 uses completed attempts as hit-rate denominators, including completed failures. Metric definitions and pooling are given in Appendix D.2. With all allocated attempts as denominator, the confirmed Hit@0.5 yields are 17.54%, 83.30%, and 88.04% for CoFi-MOF w/o verifier and MEF, CoFi-MOF w/o MEF, and full CoFi-MOF, respectively; these are lower bounds while six outcomes remain unresolved.

Target-wise improvement. MEF increases Hit@0.5 at every target, adding 71, 50, 38, 33, and 45 hits at 3–7 Å, respectively. The final rates range from 94.50% at 3 Å to 77.86% at 7 Å. The pooled gain is 237 hits (4.75 percentage points); at 0.2 Å, it is 461 hits (9.23 points). No baseline hits are lost because the coarse pool is retained. Paired 95% intervals for the Hit@0.5 gain are [4.20, 8.80] percentage points in the initial batch and [3.92, 5.28] in the continuation, using shared-noise blocks. These are separate batch estimates, not independent training-seed replications or a pooled confidence interval.

Editing coverage. Across the 5,000 attempt manifests, 4,713 identify an editable source, 281 report no editable source, and six have no measured coarse candidate. These are source-availability categories, not selected-output success counts. An editable source is selected from the Constructor's measured structures. When no eligible linker modification is available, the model retains the best original pore match.

Supplementary outcomes for selected structures are reported separately in Appendix E.4; they do not replace or redefine the primary endpoint here.

Structural uniqueness. In the initial 500-attempt batch, structural uniqueness is 436/476 (91.6%) for CoFi-MOF w/o verifier and MEF, 490/499 (98.2%) for CoFi-MOF w/o MEF, and 493/499 (98.8%) for the complete model. These measurements describe distinct structures among measured outputs in that batch, not novelty relative to a training collection. The uniqueness analysis is restricted to this initial batch.

E.3 Representative linker edits

Figure 9 illustrates the structural basis of local refinement: linker modifications alter the accessible aperture while preserving the surrounding topology and node identities. These examples complement the aggregate improvements above by showing how local chemical changes reshape the realized pore geometry.

E.4 Sensitivity to structural state and measurement settings

We compare 1,280 selected MEF-edited endpoints from CoFi-MOF with 3,506 MOFFUSION outputs for target pore sizes of 3–7 Å, keeping only structures with finite positive PLDs at both endpoints. MOFFUSION records are paired by target and attempt identity. Table 9 reports unrelaxed and relaxed measurements for identical structures under both methods. Hit@0.5 uses the paired population as its denominator, while retention denotes the fraction of unrelaxed hits that remain hits after relaxation.

Both cohorts exhibit substantial pore shifts after relaxation. CoFi-MOF's paired Hit@0.5 decreases from 93.6% to 20.9%, whereas MOFFUSION drops from 19.7% to 12.3%. Despite this degradation, CoFi-MOF achieves higher relaxed target agreement, over twice the hit retention (21.6% versus 9.7%), and lower relaxed MAE (1.239 versus 2.999 Å). Its relaxed MAE remains lower across all evaluated target sizes. This comparison demonstrates that stronger initial pore control translates to better post-relaxation aperture agreement in the tested cohorts.

CoFi-MOF operates on static framework geometries: generation, verification and editing all act on assembled, unrelaxed structures. This design keeps measurement-driven search practical, as a Zeo++ check takes approximately 0.3 s per structure (Appendix B.4) and can be repeated across dozens of realizations and edits within one attempt. In contrast, force-field relaxation introduces an extra simulation stage that would dominate screening cost. Because MEF conditions on the measured PLD of the input framework rather than an assumed structural state, the same refinement

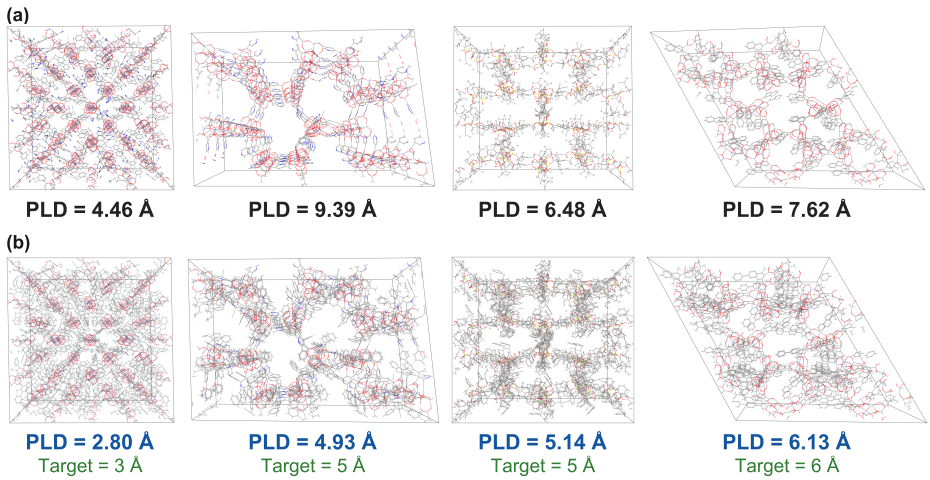


Figure 9: **Editing-controlled pore-limiting diameter on representative MOFs.** (a) Source structures before editing; (b) the corresponding structures after CoFi-MOF editing, aligned column-wise. Under each structure, the measured PLD is reported in Å.

Table 9: **Paired pore measurements before and after relaxation.** Each method uses its own available paired cohort. *Both* denotes a hit at both endpoints; MAE is in Å. CoFi-MOF includes selected edited structures, and rates correspond to these paired cohorts rather than all generation attempts.

Target	Paired n	Unrelaxed hits	Relaxed hits	Both	Retention (%)	Unrelaxed MAE	Relaxed MAE
CoFi-MOF							
3	297	279	50	48	17.2	0.127	1.167
4	283	270	47	44	16.3	0.131	1.321
5	247	237	63	62	26.2	0.091	1.201
6	216	203	47	47	23.2	0.135	1.241
7	237	209	60	58	27.8	0.223	1.270
All	1,280	1,198	267	259	21.6	0.140	1.239
MOFFUSION							
3	671	161	94	20	12.4	1.582	2.810
4	692	143	88	19	13.3	1.609	2.965
5	712	139	91	12	8.6	1.725	2.870
6	708	123	79	9	7.3	1.809	3.213
7	723	126	79	7	5.6	1.872	3.123
All	3,506	692	431	67	9.7	1.722	2.999

pipeline can accept relaxed measurements as descriptors. Pore shifts induced by relaxation remain non-negligible. For workflows editing pre-relaxed MOF frameworks, it is advisable to select structures that exhibit small geometric changes upon relaxation, rather than enforcing pre-screening for energetic conformational stability. Under such conditions, the deployed geometry stays close to the regime where the edit response was learned, enabling MEF to function as intended.

E.5 Paired evaluation of learned linker editing

We evaluate MEF's proposal quality on 937 source-matched tasks at a target PLD of 6 Å. Each task supplies the same editable source to MEF and random editing. Sources are selected from the verified Constructor pool, and all edited products are assembled and measured with Zeo++. The w/o verifier and MEF entry is the direct output of the same attempt: the globally highest-ranked assignment under the joint node–linker score, without measured candidate selection or local editing. It is a component-removed variant of our pipeline rather than an external baseline; the actual editable parent is reported separately in Table 10.

Table 10: **Learned and random editing on 937 paired tasks at 6 Å.** All tasks enter the hit-rate denominator, including failures. MAE and bias use measurable outputs and are reported in Å. Parent structures are shared by both editors; w/o verifier and MEF shows the direct output of the same attempts.

Output	Measured	Failed	Hits	Hit@0.5 (%)	MAE	Bias
w/o verifier and MEF	906	31	140	14.94	1.8622	−0.6729
Editable parent	937	0	682	72.79	0.4541	−0.1348
MEF first proposal	932	5	639	**68.20**	**0.4837**	−0.3710
Random first proposal	923	14	315	33.62	1.0040	−0.7922
MEF best of ten	937	0	732	**78.12**	**0.3523**	−0.2653
Random best of ten	937	0	611	65.21	0.5029	−0.3942

First-proposal and verified-edit comparisons. The first-proposal comparison evaluates the model's highest-ranked edit and the first randomly proposed edit without selecting by measured outcome. Assembly or measurement failures count as misses. MEF reaches 639/937 hits (68.20%), compared with 315/937 (33.62%) for random editing. The paired difference is 34.58 percentage points, with a 95% bootstrap interval of [31.38, 37.89] from 5,000 paired resamples grouped by batch and source identity. This establishes a clear advantage for learned proposal ranking at the primary tolerance.

The best-of-ten comparison applies the same measured-error selection to each editor's set of up to ten proposals. It considers edited products only, without falling back to the parent or the initial Constructor pool. MEF retains a clear advantage after verification, reaching 78.12% Hit@0.5 versus 65.21% for random editing, with MAE of 0.3523 versus 0.5029 Å. The main CoFi-MOF pipeline additionally retains its original candidates during final selection, as specified in Appendix B.

Distribution visualization. Figure 4(a) uses the common subset of 890 tasks with measurable direct outputs, both first proposals and both best-of-ten outputs. Gaussian kernel densities use a shared 0.22 Å bandwidth and reflection at zero. The 2–8 Å display window retains all observations in density estimation without window-specific renormalization. In this subset, first-proposal hits are 613/890 for MEF and 309/890 for random editing, while the direct output reaches 136/890. The direct output has 20 observations below 2 Å and 98 above 8 Å; MEF and random first proposals have none below 2 Å and one and eleven above 8 Å, respectively. The plotted densities therefore describe a common measurable subset, whereas Table 10 reports all-task performance.