Check for updates

Advanced Intelligent Systems

ADVANCED INTELLIGENT SYSTEMS Open Access

www.advintellsyst.com

RESEARCH ARTICLE OPEN ACCESS

GANg: A Target-Conditioned Generative Adversarial Network for Node Synthesis in Metal–Organic Framework Similarity Graphs

Mehrdad Jalali[1] | Andre Orth[2] | Markus Reischl[2]

[1]Department of Applied Data Science and Artificial Intelligence, School of Arts, Information and Media, SRH University of Applied Sciences Heidelberg, Heidelberg, Baden-Württemberg, Germany | [2]Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen, Baden-Württemberg, Germany

Correspondence: Mehrdad Jalali (mehrdad.jalali@srh.de) | Markus Reischl (markus.reischl@kit.edu)

Received: 21 July 2026 | **Revised:** 21 October 2026 | **Accepted:** 31 October 2026

Keywords: data augmentation | generative intelligent systems | graph-node synthesismaterials informatics | metal–organic frameworks | mode collapse | similarity graphs | synthetic scientific data | trustworthy generative AI

ABSTRACT

GANg is a target-conditioned generative adversarial network for node synthesis in metal–organic framework (MOF) similarity graphs: a Wasserstein GAN with gradient penalty (WGAN-GP) generates 106-dimensional MOF descriptors that are inserted into QMOF similarity graphs by cosine k-nearest-neighbor attachment, so graph topology is assigned rather than learned. Under matched controls and a strict inductive protocol, a conditional WGAN-GP baseline emitting the full descriptor passes every range screen while concentrating its samples and showing no conditional response (window hit rate 0.201, chance 0.200). GANg avoids these failures by emitting only an independent descriptor core—a sparsemax composition and six cell scalars—from which all derived descriptors are recomputed; it also uses a full-space critic with coverage-based checkpoint selection. It raises coverage of held-out MOFs from 0.14 to 0.81, makes every sample descriptor-schema consistent, and doubles the conditional hit rate (0.42). Pseudo-label augmentation yields modest gains only when labels are scarce (GraphSAGE −0.032 eV at 500 MOFs), is matched by simpler generators, and is driven mainly by pseudo-label distillation. GANg provides an auditable framework for evaluating synthetic graph-node hypotheses, not physically validated materials.

1 | Introduction

Generative artificial intelligence (AI) is becoming an important tool for materials exploration, but the evidentiary standard for materials generation is substantially higher than the standard for producing plausible vectors or visually compelling candidates. For metal–organic frameworks (MOFs), a useful candidate must ultimately satisfy composition, coordination chemistry, topology, periodicity, geometry, porosity, stability, and target-property constraints simultaneously [1–8]. A descriptor vector can be valuable for hypothesis generation while still requiring structural decoding or retrieval, physical sanity checks, relaxation, electronic-structure assessment, and eventually experimental evidence. This creates a need for generative workflows that are explicit about the material representation they generate and the validation that representation can support.

Recent generative AI work for MOFs and porous materials has increasingly moved from enumeration and screening toward models that propose latent, building-block, molecular, or periodic-structure candidates. Examples include variational autoencoder (VAE)-based inverse design of reticular materials, diffusion-based porous-material generation, MOF-focused molecular diffusion frameworks for carbon-capture candidates, and

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Advanced Intelligent Systems* published by Wiley-VCH GmbH.

closed-loop design pipelines that connect generative proposal with computational or automated validation stages [3, 9–11]. These approaches represent structure-oriented generation and motivate complementary methods that operate at earlier, more easily audited representation levels. The main contribution here is a target-conditioned, auditable graph-node generation and evaluation framework, not evidence of broader descriptor-space coverage than simpler statistical generators.

Alongside model development, MOF discovery increasingly depends on curated data infrastructure. Databases such as MOFX-DB, the Cambridge Structural Database MOF subset, CoRE MOF 2019, and MOFSimplify demonstrate how standardized structures, adsorption data, stability annotations, metadata, and reproducibility practices support high-throughput screening and machine-learning workflows [12–15]. These resources show that generative MOF research is not only a modeling problem but also a data-infrastructure problem: generated candidates or hypotheses must be traceable to well-defined descriptors, structures, metadata, and validation records. GANg follows this view by constructing target-column-excluded QMOF-derived similarity networks, preserving explicit feature policies, and interpreting generated descriptor-space nodes through nearest real QMOF analogs.

MOF property prediction has also advanced through specialized graph-learning, atomistic, and transformer-based representations. MOFormer uses self-supervised learning on MOF string representations for efficient property prediction, ALIGNN-based MOF models use atomistic line graphs for adsorption and structural-property prediction, and multiscale crystal graph networks model adsorption-relevant interactions at multiple structural scales [16–18]. These models show that high-performing MOF prediction depends strongly on representation choice, pretraining, periodic structure handling, and task-specific graph construction. GANg addresses a different question: whether descriptor-space graph-node hypotheses can be generated, inserted into a MOF similarity network, interpreted through nearest real analogs, and tested under conservative graph-learning diagnostics.

MOF similarity networks provide a natural intermediate representation between tabular descriptor modeling and atomistic structure generation. In MOFGalaxyNet- and MOFSocialNet-style analyses, material collections are treated as networks whose neighborhoods, bridges, communities, and homophily encode relationships among known MOFs [19, 20]. However, existing work does not provide a claim-bounded workflow for generating synthetic descriptor-space nodes, inserting them into such networks, and testing both their local plausibility and their downstream utility.

GANg addresses this gap using similarity networks derived from the quantum-chemical MOF database (QMOF) [21]. GANg is a graph-node generative adversarial network framework: a target-conditioned WGAN-GP generates a synthetic MOF descriptor vector, and cosine k-nearest-neighbor (kNN) insertion subsequently places that vector as a node in the existing similarity network. Each QMOF entry is one graph node, node features are target-column-excluded MOF-level descriptors, and edges encode cosine similarity in descriptor space. The generator does not learn or generate a complete graph; graph context is assigned after descriptor generation through proposed similarity edges. The resulting object is a descriptor-space graph-node hypothesis, not a chemical-bond graph, decoded crystal structure, atomic-coordinate set, or crystallographic information file (CIF). Evaluation covers descriptor validity, edge plausibility, novelty, topology, nearest-real-analog interpretation, and validation-selected graph neural network (GNN) augmentation diagnostics, allowing validity and downstream utility to be tested as separate claims.

This work makes five principal contributions. First, it presents a target-conditioned graph-node synthesis framework for MOF descriptor-space hypotheses. Second, it establishes a layered evaluation protocol that separates descriptor validity, novelty, graph compatibility, topology, interpretation, and downstream utility, and extends it with matched adversarial and statistical controls, strict inductive evaluation, and independent descriptor-consistency and conditional-fidelity diagnostics. Third, it adds a nearest-real-QMOF analog layer that connects synthetic descriptors to auditable structural context without treating retrieved structures as generated structures. Fourth, it diagnoses sample concentration, descriptor inconsistency, and a weak conditional response in a full-vector conditional WGAN-GP baseline, and introduces GANg, a constraint-aware generator that substantially mitigates these limitations and makes every generated descriptor internally consistent with the enforced descriptor relations. Fifth, learning-curve and label-strategy experiments locate where descriptor augmentation helps and why: modest gains appear only when real labels are scarce, are matched by unconditional generation and local resampling, and arise mainly from distillation of the pseudo-label model, which clarifies the validity–utility gap rather than assuming it away. The central contribution is therefore an auditable descriptor-space generation and diagnostic framework, not a claim that adversarial augmentation generally improves GNN prediction.

GANg denotes the constraint-aware generator together with its graph-insertion and evaluation framework. As a reference, we evaluate a conditional WGAN-GP baseline with the same conditioning input and graph insertion that generates the full 106-dimensional descriptor and relies on coordinate-wise clipping (Algorithm 1). In code and file names, GANg carries the label `v2` (for example `gang_v2.py`).

1.1 | Intended Use and Validation Logic

GANg provides a controlled and auditable descriptor-space layer for studying generated MOF-like graph nodes before structure-level claims are made. It supports synthetic descriptor-space data analysis, MOF similarity-network augmentation diagnostics, nearest-real-analog triage, validity-aware generative evaluation, and preparation of conditioning records for future structure-aware decoders. Validation proceeds from descriptor and edge checks through novelty and filtering to downstream GNN evaluation on held-out real nodes; nearest-real-QMOF analog retrieval and the retrieved structures are used for interpretation only. The generated objects remain descriptor-space hypotheses; structure-level and experimental assessment would require a subsequent decoding workflow.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Retrieved analogs are existing QMOF entries; their structures describe the neighborhood of a hypothesis, not the hypothesis itself.

The contribution is therefore a transparent descriptor-space framework rather than a physical MOF-discovery claim. No force-field relaxation, density functional theory (DFT) calculation, synthesis-ready assessment, molecular-dynamics validation, grand-canonical Monte Carlo validation, or experimental validation is performed.

2 | Results

2.1 | A QMOF Similarity Network With Target-Excluded MOF Descriptors

We constructed QMOF-derived similarity networks using 106 MOF-level descriptor features that exclude target and output columns from graph construction, generator training, and downstream GNN inputs [21]. The descriptors include formula-derived composition fractions, chemistry-role descriptors, metal-family indicators, atomic statistics, and scalar QMOF metadata available without using target/output columns; the full feature audit is provided in Section S2 and Table S1 and the source CSV. The 1000-node bandgap study provides the primary connected case, while target-column-excluded 1000-node and 5000-node extensions evaluate four Perdew–Burke–Ernzerhof (PBE) targets: bandgap, conduction-band minimum (CBM), valence-band maximum (VBM), and total energy. In the 1000-node study, the initial cosine kNN graph contains 9998 undirected weighted edges and one connected component. The corresponding 5000-node bandgap graph contains 51 906 edges before sparsification, enabling a larger-scale test of the same descriptor policy. Figure 1 summarizes the complete path from real QMOF descriptors to generated descriptor-space nodes, layered validation, and held-out-real-node evaluation.

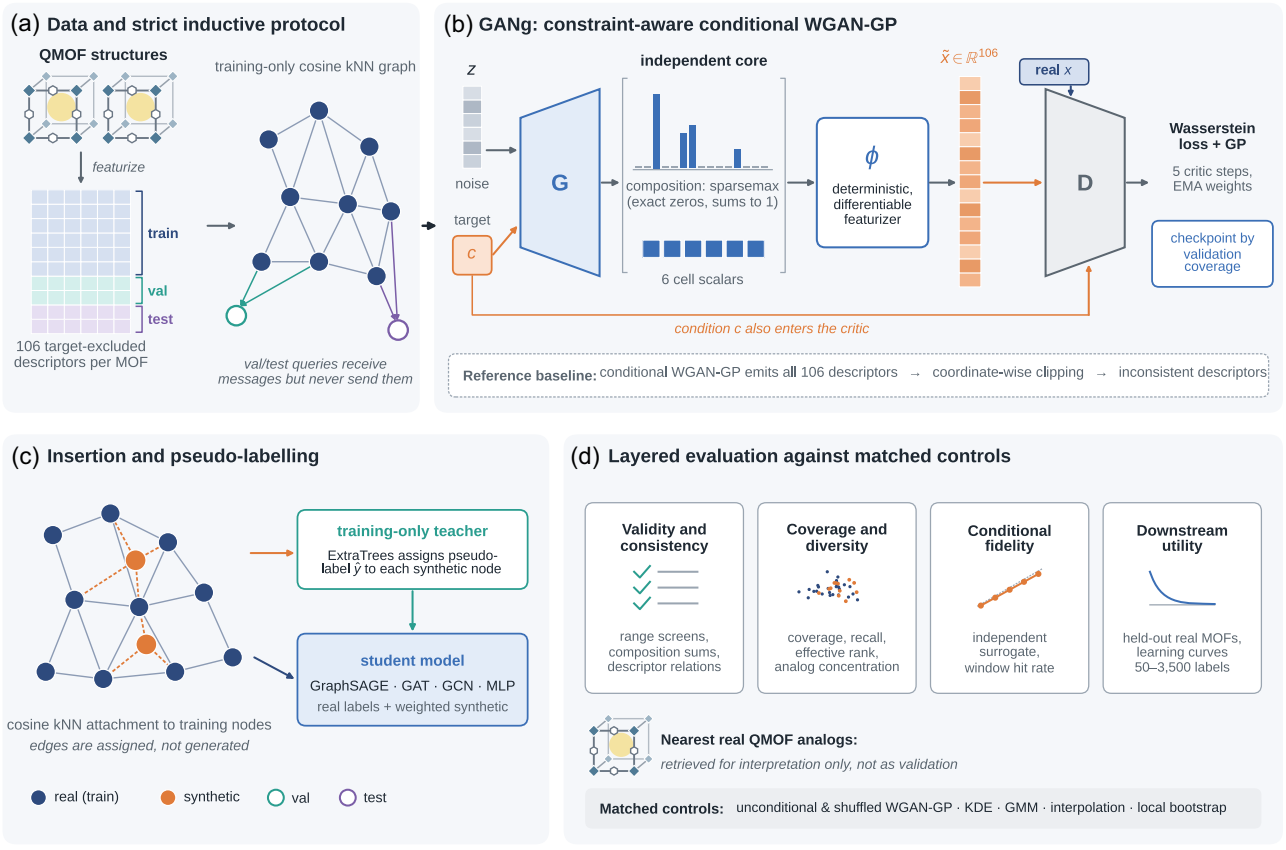


FIGURE 1 | GANg workflow. (a) QMOF descriptors (106 target-column-excluded features) are split into training, validation, and test sets; the cosine kNN reference graph is built from training nodes only, and validation and test queries receive messages but never send them. (b) GANg is a constraint-aware conditional WGAN-GP. From noise z and target c, the generator emits only an independent descriptor core, a sparsemax composition and six cell scalars, from which a deterministic, differentiable featurizer ϕ rebuilds the full descriptor $\tilde{x} \in \mathbb{R}^{106}$, consistent with the descriptor schema by construction. The critic scores $\tilde{x}$ against real training descriptors, conditioned on c, and the checkpoint is selected by coverage of validation descriptors. The reference baseline emits all 106 descriptors directly and relies on clipping. (c) Synthetic nodes are attached to training nodes by cosine kNN (edges are assigned, not generated) and receive pseudo-labels from a training-only ExtraTrees model before student training. (d) Generated nodes are evaluated for validity and consistency, coverage and diversity, conditional fidelity, and downstream utility on held-out real MOFs, always against matched controls that share the same splits, seeds, pseudo-label model, and graph insertion. Nearest real QMOF analogs are used for interpretation only. "Generated" means computationally generated in descriptor space, not synthesized, structurally decoded, or physically validated.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.2 | Sparsification Retains Local Graph Structure While Reducing Density

Sparsification is graph-preparation evidence rather than the main novelty of GANg: it controls computational density before generation and message passing. In the 1000-node analysis, the selected graph contains 6651 edges, reducing edge count by 33.48% while preserving one connected component and all real nodes in the largest component. The selected top-$k = 10$ setting is justified by unsupervised graph criteria—connectivity, retained edge count, largest-component preservation, and average degree (Table S2). Bandgap-bin homophily, the fraction of edges joining nodes assigned to the same bandgap bin, changes from 0.435 before sparsification to 0.450 after sparsification and is reported as a post hoc descriptive diagnostic. Figures 2 and 3 document the density-control step used by later diagnostics [22].

2.3 | Baseline-Generated Nodes Satisfy the Implemented Descriptor, Edge, and Novelty Screens

The conditional WGAN-GP baseline uses a WGAN-GP generator and critic to synthesize MOF-level descriptor vectors [23, 24]. Its generator maps latent noise and a conditioning scalar to descriptor space; generated descriptors are clipped to training feature ranges and attached to the real graph by cosine nearest neighbors as specified in Algorithm 1. In the connected 1000-node study, implemented screens report descriptor-valid rate 1.000, edge-valid rate 1.000, duplicate rate 0.000, training-set novelty-pass rate 1.000, and filtered-pass rate 0.500. These rates lie between 0 and 1 and are defined in Methods. Although all 1000 generated nodes passed the implemented descriptor, edge, nonduplication, and training-set novelty screens, the downstream filtered set retained 500 candidates according to the configured combined filtering and uncertainty-selection policy. The filtered-pass rate therefore measures downstream retention, not basic descriptor invalidity.

The training-set novelty-pass rate of the WGAN-GP baseline and the harmonized generator-comparison novelty rate use different reference sets and should not be compared directly. The core screen evaluates similarity to training nodes only, whereas the baseline comparison evaluates separation from the broader real-QMOF reference pool. Consequently, the WGAN-GP baseline passes the training-set novelty screen while showing a full-reference-pool novelty-pass rate of 0.013 (novelty-threshold sensitivity in Table S3).

Perfect range compliance is partly guaranteed by construction, so we tested it against independent consistency checks that the screens do not enforce. In the matched 1000-node bandgap experiments (Methods), every raw sample of the WGAN-GP baseline required clipping, affecting 32.5% of all descriptor coordinates. After clipping, elemental fractions summed to one within 0.001 for only 1.0% of generated descriptors, compared with 98.3% of real training descriptors, and the two independently emitted carbon-fraction coordinates disagreed by 0.038 on average (0.000 for real data). Range validity therefore does not imply that the generator has learned coherent descriptor combinations; Section 2.8 describes GANg, for which these relations hold exactly.

2.4 | Baseline-Generated Nodes Occupy a Concentrated Region of Descriptor Space

Figure 4 separates computational screening from distributional diagnostics. All baseline-generated nodes pass the implemented descriptor, edge, nonduplication, and training-set novelty screens, while 50% are retained by the combined filtering policy (Figure 4A). For the post hoc distribution audit, feature means, standard deviations, and the two-component principal component analysis (PCA) basis were fitted using real training nodes only. The baseline-generated descriptors overlap part of the real-training PCA cloud but occupy a visibly narrower region (Figure 4B). Ranked feature-wise standardized mean

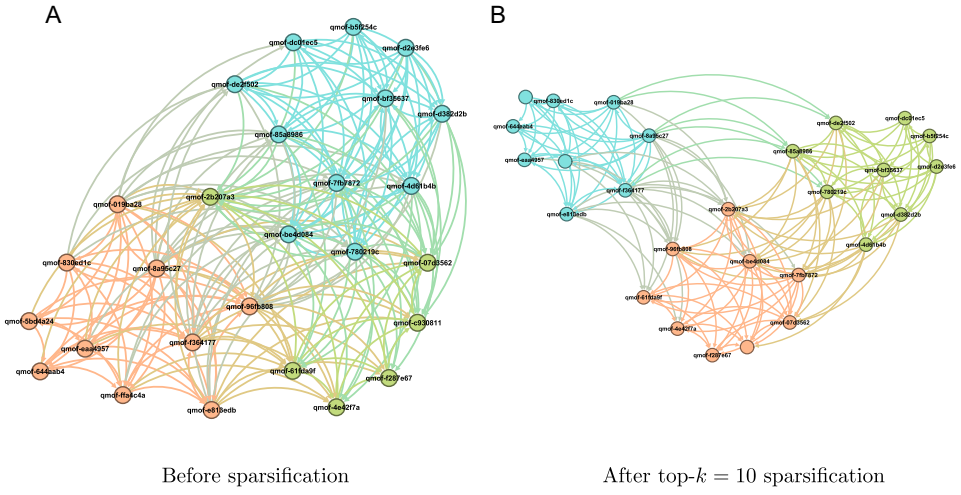


FIGURE 2 | Representative labeled QMOF-derived similarity-network views before and after sparsification. The panels visualize graph connectivity among QMOF nodes in descriptor-similarity space; node positions are network-layout coordinates, not atomic or crystal-structure coordinates. The before-sparsification view shows the denser cosine kNN similarity graph, while the after-sparsification view shows the selected top-$k = 10$ graph used for later GANg diagnostics.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

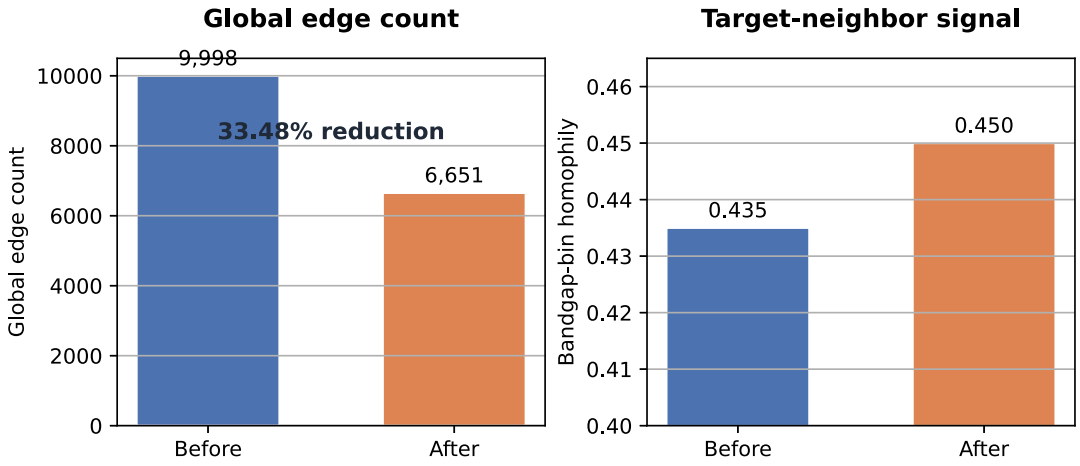


FIGURE 3 | Quantitative sparsification diagnostics. The global graph decreases from 9998 to 6651 edges (33.48% reduction), while bandgap-bin homophily changes from 0.435 to 0.450 and one connected component is preserved. This density-control step supports later GANg diagnostics but is not the main methodological novelty.

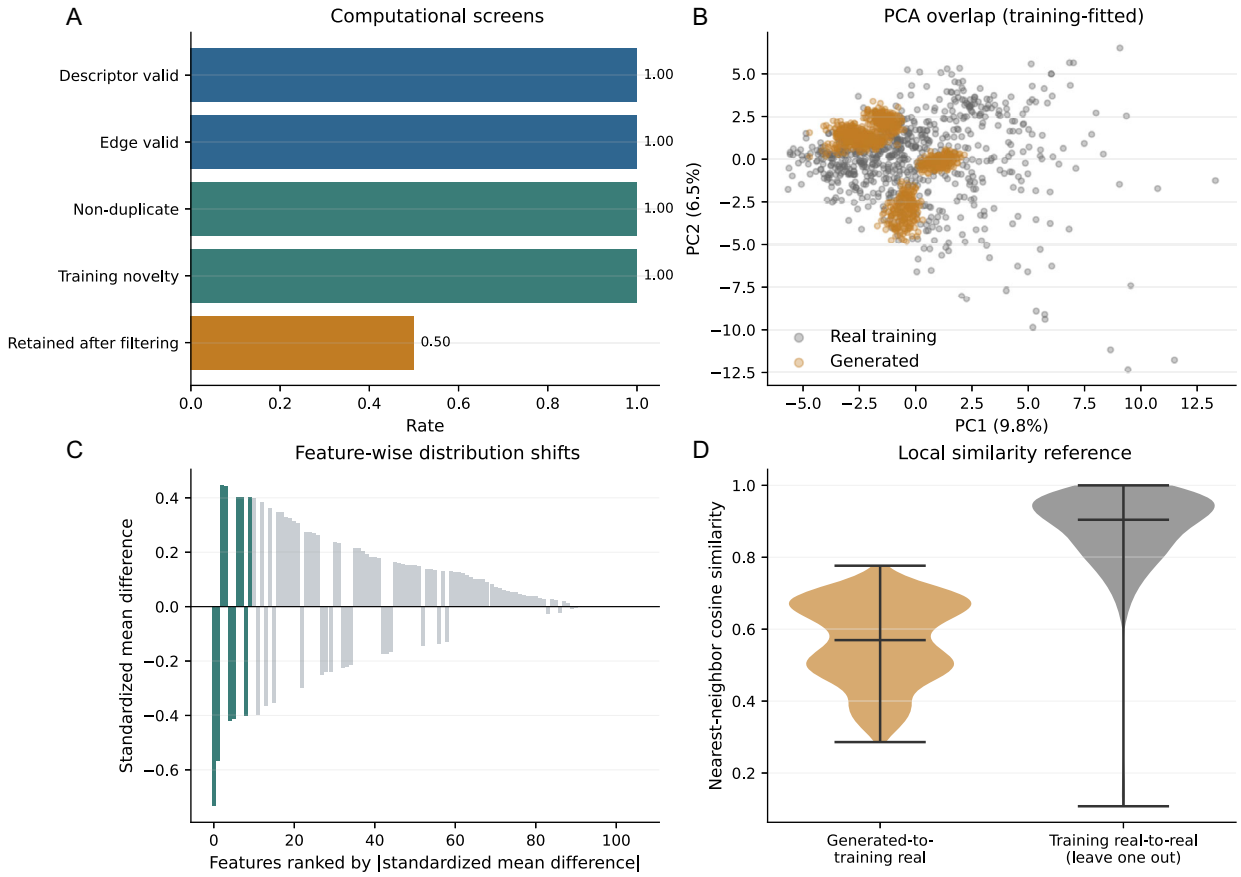


FIGURE 4 | Computational screening and descriptor-space distribution diagnostics for nodes generated by the conditional WGAN-GP baseline. (A) Descriptor-valid, edge-valid, nonduplicate, training-set-novelty, and retained-after-filtering rates. (B) PCA projection of real training and generated descriptors; centering, scaling, and PCA fitting use the real training partition only. (C) Feature-wise standardized mean differences, calculated as the generated-minus-training mean divided by the real-training standard deviation and ranked by absolute magnitude; the ten largest absolute shifts are highlighted in teal, with their full descriptor names and values reported in the Figure 4 source-data table and Supporting Information. (D) Generated-to-training-real nearest-neighbor cosine similarity compared with leave-one-out training-real-to-training-real similarity in the same training-standardized representation. These are computational descriptor-space diagnostics, not evidence of structural, energetic, synthetic, or experimental validity.

differences identify which of the 106 descriptors contribute most strongly to this shift (Figure 4C). Generated-to-training-real nearest-neighbor similarities are also lower than leave-one-out similarities among real training nodes (Figure 4D), consistent

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with local separation rather than broad distributional coverage. These diagnostics characterize descriptor-space behavior only and do not establish structural plausibility or physical validity.

We quantified this concentration with matched diagnostics. Across six split–seed replicates of the 1000-node bandgap experiment, the covariance effective rank of baseline-generated descriptors is 19.4, compared with 39.0 for real training descriptors, and mean pairwise distance is 1.6 versus 12.1 standardized units. Within each of five training-derived bandgap windows, the pairwise diversity of baseline samples is only 13.8% of the diversity of real MOFs in the same window. The parameter-matched unconditional (14.3%) and shuffled-condition (14.5%) WGAN-GP controls contract to the same degree, whereas KDE/bootstrap, GMM, interpolation, and target-local bootstrap retain 93%–103% of real within-window diversity (Section S34, Figure S3 and Table S14). The concentration is therefore a property of the baseline's adversarial training configuration, consistent with partial mode collapse, and not a deliberate or beneficial consequence of conditioning.

2.5 | Nearest-Neighbor Interpretation Turns Generated Nodes Into Auditable Hypotheses

Each generated node is interpreted through nearest real QMOF neighbors. QMOF identifiers are database IDs for known entries and their associated reference structures. For example, baseline-generated node gang1000_bandgap_seed0_gen052 passes descriptor validity, edge validity, nonduplication, and training-set novelty filtering; its nearest real QMOF neighbors are qmof-21760ea, qmof-83de5c7, and qmof-4a1e9e7, and its top-5 neighbor bandgap mean is 1.887 eV with standard deviation 0.595 eV. This means that the generated node is not a structure; it is a synthetic descriptor vector whose closest known QMOF examples provide context for interpretation. Figure 5 visualizes this bridge from generated features to known materials using a descriptor-vector icon for the generated node and QMOF labels for retrieved real analogs.

Analog retrieval also exposes how concentrated the WGAN-GP baseline is. For 256 samples per replicate, the single most frequent nearest QMOF analog accounts for 58.7% of baseline samples and the five most frequent analogs for 95.0%; only 3.5% of retrieved analog IDs are distinct, compared with 82%–84% for KDE/bootstrap, interpolation, and target-local bootstrap. We find no evidence of exact training-row copying under the reported nearest-reference comparison: the near-exact training-match rate is zero, and generated samples have comparable distances to equally sized training and held-out reference pools (4.72 versus 4.71 standardized units; Table S17). This does not rule out all forms of local concentration. Repeated analogs indicate concentration around a few training neighborhoods, which we treat as a limitation of the WGAN-GP baseline rather than as evidence of local plausibility. Retrieved analogs are interpretability aids only; they are not generated CIFs and are not evidence of novelty or of a generated physical structure.

2.6 | Generator Baselines and Target-Conditioned Descriptor-Space Generation

We compared the conditional WGAN-GP baseline with simpler descriptor-space generators: Gaussian/noise, nearest-neighbor interpolation, KDE/bootstrap, GMM, and a lightweight VAE (Figure S1). All five simpler generators have descriptor-valid rate 1.000 and edge-valid rate 1.000 after feature clipping and cosine-kNN attachment. Full-reference-pool novelty-pass rates vary substantially under the harmonized baseline-comparison metric: KDE/bootstrap reaches 0.128, GMM reaches 0.111, the WGAN-GP baseline reaches 0.013, and Gaussian/noise, nearest-neighbor interpolation, and the lightweight VAE remain near the real-data manifold by this broader-pool diagnostic. These results indicate

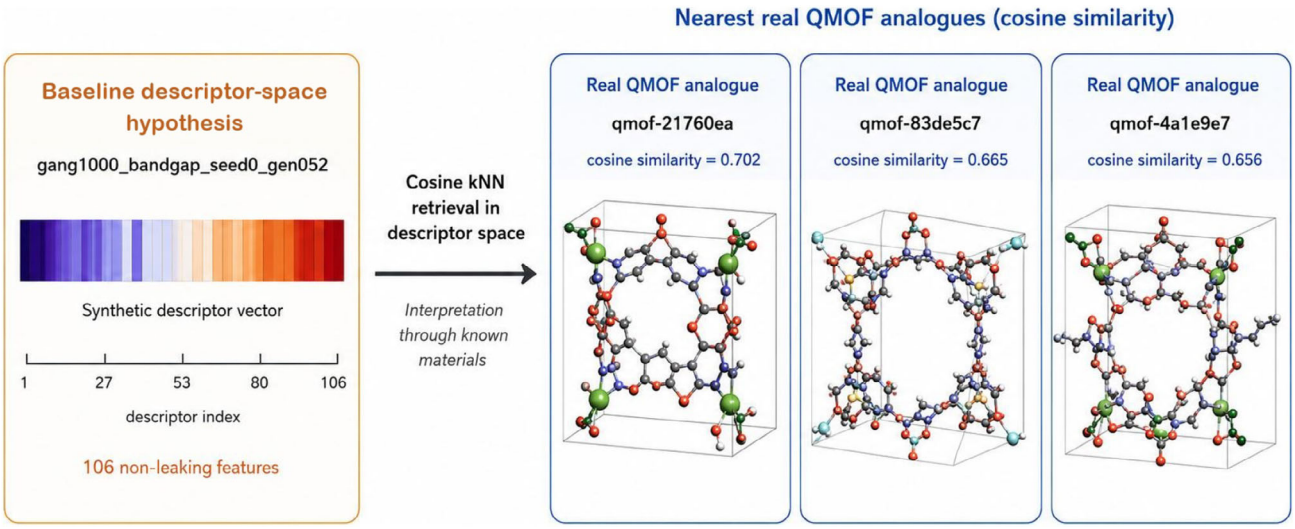


FIGURE 5 | Descriptor-space interpretation of generated node `gang1000_bandgap_seed0_gen052`. This node was generated by the conditional WGAN-GP baseline; the identifier is a file name. The generator output is represented as a synthetic descriptor vector rather than a crystal structure. Cosine *k*-nearest-neighbor retrieval links this hypothesis to three real QMOF reference analogs: `qmof-21760ea` (0.702), `qmof-83de5c7` (0.665), and `qmof-4a1e9e7` (0.656). The displayed analog structures provide structural context for the retrieved real QMOF entries only. The top-five analog bandgap mean is 1.887 ± 0.595 eV. No generated CIF, atomic-structure prediction, or physical validation is implied.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that descriptor validity is easy to satisfy in this bounded feature space, while full-reference-pool novelty and the implemented descriptor-diversity score are more discriminating generator-quality diagnostics. Under this harmonized comparison, the WGAN-GP baseline shows lower full-reference-pool novelty and descriptor diversity than KDE/bootstrap and GMM, indicating concentrated sampling in a relatively narrow region of descriptor space and limiting any claim that it provides broad coverage of the real-data distribution. Generator-quality diagnostics and downstream utility are evaluated separately; no downstream superiority is inferred from baseline validity or diversity alone.

Because KDE/bootstrap and GMM are unconditional, a fair test of conditioning requires adversarial controls with the same architecture and budget. We therefore trained a parameter-matched unconditional WGAN-GP, whose condition input is fixed at zero in both generator and critic, and a shuffled-condition WGAN-GP, which keeps the conditional architecture but destroys the pairing between descriptors and targets; a target-local bootstrap that resamples training MOFs near the requested value provides a nonadversarial conditional reference. All methods share training rows, clipping, sample counts, pseudo-label model, graph insertion, downstream models, splits, and seeds (Methods). Conditional fidelity was scored by an independent random-forest evaluator [25] trained on real training rows (test MAE 0.474 eV, $R^2 = 0.647$) in five training-quantile windows. The conditional WGAN-GP reaches a window hit rate of 0.201, indistinguishable from the unconditional (0.200) and shuffled (0.198) controls and from the chance level of 0.200, whereas target-local bootstrap reaches 0.621. The conditional WGAN-GP baseline therefore shows no measurable conditional response (Section S34 and Table S15).

We also generated descriptor-space target-window candidates with the WGAN-GP baseline for ten predefined numerical windows spanning bandgap, CBM, VBM, and total energy (Table 1). Window boundaries were estimated from training nodes only, and nearest-analog target summaries used training nodes only. Hit rate is the fraction of generated hypotheses whose nearest-analog target summary falls inside the requested window. The strongest window-level result is 0.845. The matched controls above show that these nearest-analog hit rates of the WGAN-GP baseline do not isolate a conditioning effect: concentrated samples whose fixed analog neighborhood happens to fall inside a window score well regardless of the requested value, and the high total-energy rates largely reflect size correlations.

2.7 | Retrieved Existing QMOF Reference Analogs

Because raw QMOF metadata and CIFs are available, each generated descriptor-space hypothesis can be linked to nearest real QMOF analog CIFs. The top-5 analog CIF matching rate is 1.000 for the processed baseline-generated nodes. To provide a structure-grounded interpretation check, we selected 20 baseline-generated descriptor vectors from the rank-1 bandgap analog retrieval workflow. This rank-1 structure-grounded workflow uses a broader real-QMOF reference pool and a different normalization/retrieval procedure than the illustrative Figure 5 descriptor-space card, so its similarity values are workflow-specific and should not be compared directly with the Figure 5 similarities. For each generated descriptor vector, the nearest real QMOF analog CIF was retrieved and checked with pymatgen, a Python materials-analysis library, and the Atomic Simulation Environment (ASE), a Python toolkit for atomistic structure handling [26, 27]. All 20 retrieved analog CIFs parsed successfully and passed the implemented basic sanity screen. These checks therefore evaluate retrieved real analog structures, not generated CIFs.

Figure 6 shows nearest-analog bandgap context, minimum interatomic distance, density, and number of sites. It provides structure-grounded interpretation for the descriptor hypotheses; it does not physically validate generated structures. The corresponding parsing and sanity results and generated-node interpretation card are provided in the Supporting Information (Figure S2). Because the retrieved CIFs are existing QMOF reference structures that already passed curation, their parsing and sanity results carry no information about the realizability of an

TABLE 1 | Leakage-corrected target-window descriptor-space generation diagnostics for the conditional WGAN-GP baseline. Numerical intervals are training-data quantile windows and include units of eV. Hit rate is determined from the nearest-training-real-analog target summary, not from a generated or physically validated structure. Total energy is extensive and its high hit rates may reflect system-size and composition correlations.

Target	Requested window, eV	Hit rate	Generated nodes	Nearest-analog target mean, eV
Bandgap	[0.036, 1.733]	0.440	200	1.898
Bandgap	[1.733, 2.820]	0.620	200	2.393
Bandgap	[2.820, 5.917]	0.300	200	2.483
CBM	[−1.456, 1.854]	0.600	200	1.610
CBM	[2.736, 5.502]	0.535	200	2.786
VBM	[−3.228, −0.405]	0.495	200	− 0.490
VBM	[0.379, 3.583]	0.395	200	0.227
Total energy	[−3329.363, −739.412]	0.845	200	− 1057.262
Total energy	[−739.412, −497.468]	0.730	200	− 660.074
Total energy	[−497.468, −110.014]	0.810	200	− 421.808

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

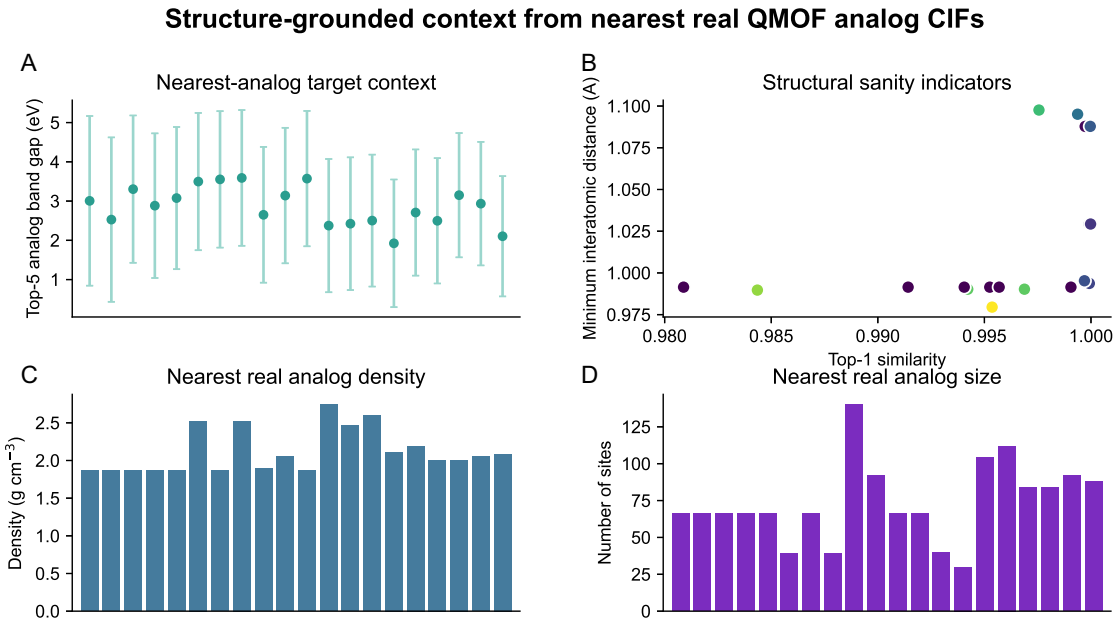


FIGURE 6 | Structure-grounded interpretation for 20 baseline-generated descriptor vectors (conditional WGAN-GP baseline) selected from the rank-1 bandgap analog retrieval workflow. (A) Top-5 bandgap context for the selected descriptor vectors. (B) Nearest-real-QMOF cosine similarity versus minimum interatomic distance, colored by density. (C) Density of the retrieved real analog CIFs. (D) Number of sites in the retrieved real analog CIFs. The checked CIFs are retrieved real QMOF analog structures, not generated CIFs. This figure provides structure-grounded interpretation, not physical validation.

independently generated descriptor vector, and they are not a validation stage for generated candidates. They are retrieval-based interpretability aids, not proof of novelty or of a generated physical structure.

2.8 | GANg's Constraint-Aware Design Mitigates Concentration and Enforces Descriptor-Schema Consistency

The diagnostics above identify three limitations of the conditional WGAN-GP baseline: concentrated sampling, internally inconsistent descriptors, and a weak conditional response. They trace to its training configuration and to an unconstrained output layer. GANg addresses them through a constraint-aware descriptor core, full-space critic evaluation, and coverage-based checkpoint selection (Methods), while sharing the baseline's conditioning input, graph insertion, and evaluation protocol. First, in the matched experiments the baseline's budget of one critic update per generator update for 100 epochs at batch size 128 amounts to roughly 600 generator updates on 700 training MOFs (about 1100 at the batch size of 64 used in the single-split package); GANg uses five critic updates per generator update, gradient-penalty weight 10, an exponential moving average of generator weights, and up to 10^4 generator updates, with the checkpoint chosen by coverage of validation descriptors [28] rather than by a fixed epoch count. Second, 100 of the 106 descriptors are deterministic functions of the 75 elemental fractions and the formula atom count, and the remaining scalars obey three exact relations in QMOF (cell atom count is an integer multiple Z of the formula atom count; density equals Z times formula mass over cell volume; pore-limiting diameter does not exceed the largest cavity diameter). GANg therefore emits only an independent core—elemental fractions through a sparsemax layer [29], which yields an exact simplex with exact zeros, plus formula atom count, cell atom count, volume, largest cavity diameter, the pore-diameter ratio, and space group—and recomputes every derived descriptor with the same featurizer used for real MOFs. These relations are descriptor and metadata constraints: snapping the space-group descriptor to a training value, for example, does not establish real crystallographic symmetry. Third, the critic judges the full 106-dimensional training-standardized descriptor, rebuilt differentiably from the core, so that the generator is penalized in exactly the representation used downstream.

Figure 7 and Table 2 report a cumulative ablation in which components are added progressively; because several changes enter together at each step, the ablation does not isolate the causal contribution of individual components. The first step, which combines a longer training budget, more critic updates, a weight average, and coverage-based checkpoint selection, reduces the dominant-analog share from 0.48 to 0.02 but leaves descriptors inconsistent and coverage at 0.50. Adding the constrained core and descriptor-space critic restores real-level spread (mean pairwise distance 11.6 versus 11.6 for real data;

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

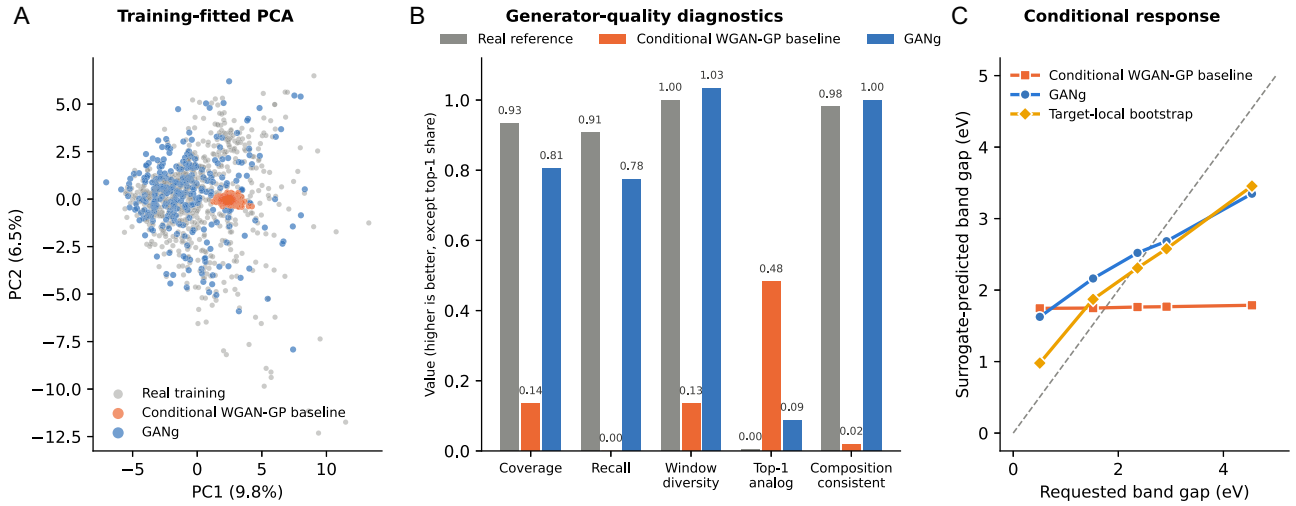


FIGURE 7 | Conditional WGAN-GP baseline versus GANg, 1000-node bandgap cohort, split 42. (A) PCA of real training descriptors and 256 generated descriptors per generator; standardization and the PCA basis are fitted on real training nodes only, as in Figure 4B. (B) Coverage and recall of held-out real MOFs [28], within-window diversity relative to real MOFs, share of samples whose nearest training analog is the single most frequent one, and fraction of descriptors whose elemental fractions sum to one; mean over two model seeds. (C) Mean independent-surrogate prediction against requested bandgap in five training-quantile windows; the dashed line is identity. Surrogate scores are descriptor-space diagnostics, not property measurements of generated MOFs.

TABLE 2 | Cumulative ablation from the conditional WGAN-GP baseline to GANg; components are added progressively, so rows do not isolate individual causal contributions (1000-node bandgap cohort, split 42; mean of two model seeds; 256 samples). Pair distance and effective rank use training-standardized descriptors. Coverage and recall are measured against held-out real MOFs. "Consistent" gives the fraction of descriptors whose elemental fractions sum to one within 10^{-3}, whose cell atom count is an integer multiple of the formula atom count, and whose density matches Z-scaled formula mass over volume within 1%. Real data contain a few elements outside the 75-element basis, hence 0.98. Hit rate (chance 0.20) and within-window diversity ratio as in Section S34. Constrained outputs are not clipped, because clipping individual coordinates would break the enforced relations. Consistency refers to the enforced descriptor relations, not to chemical or structural validity. Values refer to split 42 with two model seeds; the real reference is a bootstrap of 256 training rows. They therefore differ from the six-replicate matched audit in Section 2.4 and Table S14; per-seed values are given in Table S20. All entries are means rounded half up.

Generator	Pair dist.	Eff. rank	Top-1 analog	Coverage	Recall	Consistent (frac./ Z/density)	Hit rate	Window div.
Real reference	11.64	35.8	0.004	0.93	0.91	0.98/1.00/0.98	—	1.00
Conditional WGAN-GP baseline	1.61	20.1	0.482	0.14	0.00	0.02/0.00/0.01	0.20	0.13
+ Training budget, EMA, coverage selection	6.66	20.9	0.020	0.50	0.25	0.01/0.00/0.02	0.24	0.56
+ Constrained core, descriptor-space critic (GANg)	11.63	21.4	0.088	0.81	0.78	1.00/1.00/1.00	0.42	1.03
+ Mode-seeking loss, minibatch std.	12.29	22.3	0.090	0.74	0.81	1.00/1.00/1.00	0.40	1.02

within-window diversity ratio 1.03), raises coverage of held-out MOFs from 0.14 to 0.81 and recall from 0.00 to 0.78, and makes composition, Z-multiplicity, and density relations hold for every sample. We find no evidence of exact training-row copying under the reported nearest-reference comparison: generated samples have comparable distances to training and held-out reference pools (4.89 versus 4.85); this does not rule out all forms of local concentration. Although GANg uses the same conditioning mechanism as the baseline, the surrogate conditional hit rate doubles from 0.20 to 0.42 and the surrogate response becomes monotonic in the requested bandgap (Figure 7C), approaching the target-local bootstrap. This condition-response metric is a surrogate descriptor-space diagnostic, limited by the accuracy of the independent predictor ($R^2 = 0.647$); a generated descriptor is not evidence that a realizable MOF would possess the requested property. An additional mode-seeking regularizer [30] with a minibatch-standard-deviation critic feature [31] gave no further improvement. GANg therefore substantially mitigates, but does not remove, concentration: effective rank remains below the real reference (21.4 versus 35.8 for the split-42 bootstrap reference; the six-replicate audit in Section 2.4 gives 39.0), and coverage (0.81) remains below that of a real bootstrap sample (0.93), so some descriptor directions are still under-represented.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.9 | Downstream GNN Diagnostics Quantify Augmentation Behavior

Downstream tests in the single-split, transductive protocol augment training with nodes generated by the conditional WGAN-GP baseline and use the same train/validation/test split across all modes and evaluate performance only on held-out real QMOF nodes. "Real-only" denotes the baseline in which no generated nodes are added during training. Checkpoint selection within a single training mode uses validation MAE. Cross-mode selection is then performed separately by choosing, for each architecture, the baseline augmentation mode with the lowest validation MAE and evaluating that selected mode on the held-out test set. In the leakage-corrected connected 1000-node bandgap study, GraphSAGE obtains a real-only MAE of 0.606 ± 0.039, whereas the validation-selected conditional-target mode obtains 0.655 ± 0.037. GAT obtains a real-only MAE of 0.690 ± 0.036, whereas the filtered-baseline mode obtains 0.716 ± 0.008. GCN obtains a real-only MAE of 0.690 ± 0.014, whereas the filtered-baseline mode obtains 0.730 ± 0.012. Thus, real-only training remains the lower-error reference for all three architectures in this bandgap case.

The architectures are standard GNN variants: GraphSAGE learns inductive neighborhood aggregations, GAT weights neighbors by learned attention, and GCN applies normalized graph convolutions [32–36]. The paired five-seed analyses quantify initialization sensitivity under one fixed split and are therefore treated as exploratory descriptive diagnostics rather than definitive performance estimates. We therefore emphasize effect sizes, confidence intervals, and seedwise consistency rather than threshold-based significance or equivalence claims. Figure 8 visualizes a validity–utility gap: nodes satisfying the implemented descriptor-space screens are not automatically beneficial for supervised augmentation. Full root mean squared error (RMSE), coefficient of determination (R^2), Spearman, Kendall, confidence-interval, and seedwise diagnostics are reported in the Supporting Information.

2.10 | Matched Downstream Comparison Under a Strict Inductive Protocol

The analysis above compares modes of the WGAN-GP baseline only with real-only training, uses a single split, and exposes test descriptors during message passing. We therefore repeated the downstream comparison under a matched and stricter protocol (Methods): three random 70/15/15 splits with two paired model seeds each, at 1000 and 5000 nodes, for all five target formulations, comparing the conditional WGAN-GP baseline and its unconditional and shuffled-condition controls with KDE/bootstrap, GMM, interpolation, and target-local bootstrap, using GraphSAGE, GAT, GCN, and a graph-free multilayer perceptron (MLP). Every generator draws 25% synthetic nodes, labeled by one shared training-only ExtraTrees model [37]. In the primary strict inductive setting, the graph used for fitting contains only training and synthetic nodes, and no test descriptors, labels, or nodes enter fitting, graph construction, checkpoint selection, or tuning; a held-out query receives messages from its 15 nearest training neighbors but sends none, so reference predictions are invariant to query insertion, which unit tests verify for every architecture. The matrix comprises 60 size–target–split–seed units and 2616 downstream fits. The single-split results of Sections 2.9, 2.13, and 2.14 come from the single-split, transductive protocol, which differs from this matched protocol in training budget (30 epochs, learning rate 0.01, and early-stopping patience 12, versus 80 epochs, learning rate 0.003, and patience 15), synthetic-label source (mode-specific labels from nearest-neighbor or property-head pseudo-labeling or the conditioning target, versus one shared ExtraTrees model), and mode selection (validation-selected baseline mode, versus fixed 25% augmentation; ratio sensitivity in Table S13). The difference in sign between Section 2.9 (+0.049 eV for GraphSAGE) and this section (−0.008 eV inductive, −0.024 eV transductive) is therefore attributed to these protocol differences, and the matched results in Sections 2.10–2.12 are the primary evidence.

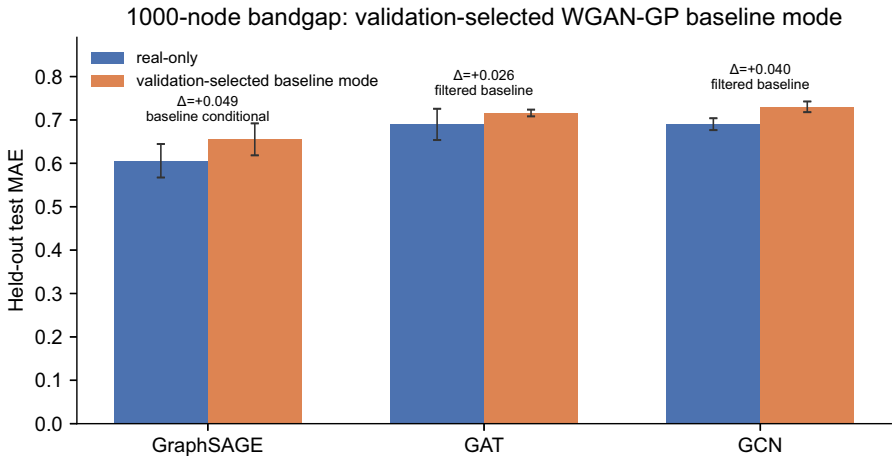


FIGURE 8 | Leakage-corrected downstream GNN diagnostics for the 1000-node bandgap analysis (single-split, transductive protocol). Augmentation modes of the conditional WGAN-GP baseline are selected using validation MAE within each architecture and then evaluated on held-out real QMOF test nodes. Generated nodes are used only for training-time augmentation, and pseudo-labels use training-node labels only. The comparison quantifies model-dependent augmentation behavior under a conservative real-only reference. The paired five-seed analyses quantify initialization sensitivity under one fixed split and are therefore treated as exploratory descriptive diagnostics rather than definitive performance estimates.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

No generator yields a consistent gain (Figure 9A,B; Table 3). For 1000-node bandgap, the paired effect of the conditional WGAN-GP baseline is −0.008 eV [−0.051, +0.036] for GraphSAGE, against −0.014 eV for the unconditional and −0.001 eV for the shuffled control; for GraphSAGE all statistical generators lie within ±0.015 eV of real-only. Across the 40 size–target–architecture comparisons, the WGAN-GP baseline has lower mean MAE in 14 and higher in 26. At 5000 nodes every bandgap effect for GraphSAGE and the MLP lies within 0.01 eV of zero. The training-only ExtraTrees model itself reaches 0.460 eV at 1000 nodes, lower than every graph or graph-free student, a point we return to below.

Three ablations address the role of the graph. Exposing all real descriptors during message passing (transductive access; Table S11) changes the effects of the WGAN-GP baseline by at most 0.016 eV (−0.008 versus −0.024 eV for GraphSAGE at 1000 nodes), so these conclusions do not depend on test-descriptor access. Replacing cosine attachment of baseline-generated nodes by random attachment of equal degree, or leaving synthetic nodes isolated so that they contribute only to the loss, gives GraphSAGE MAEs of 0.519, 0.532, and 0.525 eV, respectively (Table S12), and the graph-free MLP with the same synthetic data reaches 0.512 eV: message passing through inserted nodes adds no measurable value beyond their supervised loss. Finally, splitting by exact reduced formula, so that no composition crosses partitions, gives a GraphSAGE effect of the WGAN-GP baseline of −0.006 eV [−0.036, +0.024], consistent with the random splits (Section S34 and Table S18).

2.11 | Augmentation Helps Modestly Only When Real Labels Are Scarce

The matched comparison uses 700 or 3500 training MOFs, a regime in which the controls suggest that additional descriptors carry little new information. To test whether augmentation helps when labels are scarce, we drew nested training subsets of 50, 100, 250, 500, 1000, and 3500 MOFs from the 5000-node training partition, kept a fixed 750-MOF test set per split, and scaled the labeled validation set to 20% of the training size (minimum 30). All generators, scalers, and the pseudo-label model are refitted on each subset, and the augmentation ratio (25% or 100%) is chosen on validation MAE. Besides overall MAE we report tail MAE on test MOFs whose bandgap lies outside the central 80% of the training targets (Methods).

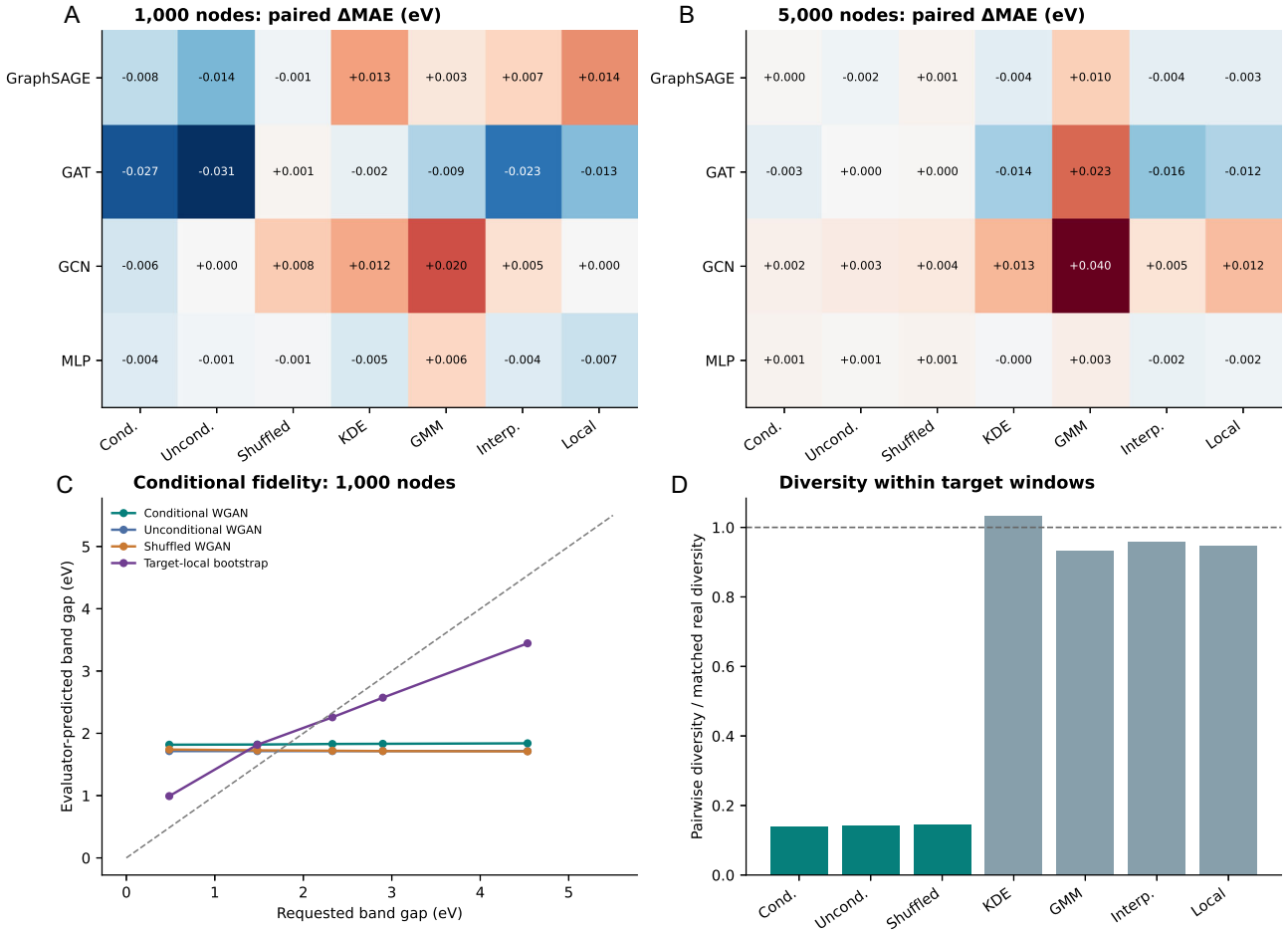


FIGURE 9 | Matched strict-inductive bandgap comparison at 25% augmentation. (A,B) Paired mean ΔMAE relative to real-only at 1000 and 5000 nodes for the conditional WGAN-GP baseline ("Conditional WGAN"), its unconditional and shuffled-condition controls, and four statistical generators; three splits, two seeds per split; negative favors augmentation. (C) Independent-surrogate prediction against requested bandgap for the WGAN-GP baseline, its adversarial controls, and target-local bootstrap; the dashed line is identity. (D) Within-window pairwise diversity relative to real MOFs in the same window. Intervals for all targets and models are given in Section S34.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | Mean strict-inductive bandgap test MAE (eV) over three splits and two initializations per split, 25% augmentation, shared pseudo-label model and attachment policy. "Conditional WGAN" is the conditional WGAN-GP baseline; "Unconditional WGAN" and "Shuffled WGAN" are its controls. Paired intervals are in Section S34 (Table S10).

Nodes	Method	GraphSAGE	GAT	GCN	MLP
1000	Real-only	0.527	0.616	0.614	0.516
1000	Conditional WGAN	0.519	0.590	0.608	0.512
1000	Unconditional WGAN	0.513	0.585	0.614	0.515
1000	Shuffled WGAN	0.526	0.617	0.622	0.516
1000	KDE/bootstrap	0.540	0.614	0.626	0.512
1000	GMM	0.531	0.607	0.634	0.523
1000	Interpolation	0.534	0.593	0.619	0.512
1000	Target-local bootstrap	0.542	0.603	0.614	0.509
5000	Real-only	0.473	0.525	0.539	0.464
5000	Conditional WGAN	0.474	0.522	0.541	0.466
5000	Unconditional WGAN	0.471	0.526	0.542	0.465
5000	Shuffled WGAN	0.474	0.526	0.543	0.465
5000	KDE/bootstrap	0.469	0.512	0.552	0.464
5000	GMM	0.483	0.548	0.579	0.467
5000	Interpolation	0.469	0.509	0.545	0.462
5000	Target-local bootstrap	0.470	0.513	0.551	0.463

Pseudo-label augmentation with GANg lowers mean overall MAE at every training size, although only the 500-MOF GraphSAGE interval excludes zero (Figure 10): for GraphSAGE by 0.037 eV [−0.109, +0.035] at 100 MOFs and 0.032 eV [−0.064, −0.000] at 500 MOFs (about 6% of real-only MAE), for the MLP by about 0.015 eV between 100 and 500 MOFs, and by at most 0.004 eV at 3500 MOFs. The WGAN-GP baseline gives smaller and less consistent effects (MLP within 0.003 eV of zero beyond 50 MOFs; GraphSAGE between −0.020 and +0.013 eV). These effects are, however, matched by simple resampling: target-local bootstrap gives −0.031 eV [−0.035, −0.026] for GraphSAGE at 500 MOFs, KDE/bootstrap −0.016 eV [−0.029, −0.003] at 1000 MOFs, and the unconditional GANg control −0.037 eV [−0.060, −0.013] at 500 MOFs. Conditioning on the empirical or on a flat target distribution does not change the picture, and tail errors rise slightly for almost all generators (for example +0.019 eV [+0.002, +0.036] for GANg with the MLP at 3500 MOFs). In the low-label regime, pseudo-label augmentation with GANg therefore yields modest gains; however, comparable effects from local resampling and the unconditional control indicate that this benefit is not uniquely GAN-specific, and it weakens or vanishes at larger label budgets.

2.12 | Synthetic Labels, Not Generated Descriptors, Determine Augmentation Utility

In both experiments, every synthetic node is labeled by the training-only ExtraTrees model, which is more accurate than every student at every training size (Table S16) (for example 0.688 versus 0.797 eV for GraphSAGE at 50 MOFs and 0.417 versus 0.473 eV at 3500 MOFs). This suggests that augmentation mainly distills the pseudo-label model [38] into the student. Tree ensembles cannot extrapolate beyond their training targets, so pseudo-labels of GANg samples vary less than real bandgaps (standard deviation 0.71–0.76 versus 1.05–1.16 eV), which would explain the tail degradation. We tested this directly by changing only the labels of synthetic nodes (Methods): the requested condition, the mean of condition and pseudo-label, the same mean restricted to samples on which the two agree within the pseudo-label model's validation MAE, and, for bootstrap generators, the real label of the source MOF. The augmentation ratio and a synthetic loss weight (0.25 or 1) were chosen on validation for every strategy.

GANg does respond to its condition in this surrogate sense: the pseudo-label correlates with the requested bandgap at 0.56–0.76 across training sizes. Nevertheless, condition labels remove the gain (Figure 11; GraphSAGE −0.002 to +0.015 eV versus −0.012 to −0.032 eV with pseudo-labels), blended and agreement-filtered labels perform like pseudo-labels, and source labels for jittered bootstrap copies give effects near zero. Agreement filtering was the only strategy that lowered tail error with an interval excluding zero, and only for GraphSAGE at 500 MOFs (−0.025 eV [−0.038, −0.012] with empirical and −0.018 eV [−0.029, −0.007] with flat conditions), an isolated result. Pseudo-labeled GANg remains the most consistent configuration, but pseudo-labeled KDE/bootstrap performs comparably. Across concentrated and diverse generators, conditional and unconditional controls, and adversarial and statistical generators, the downstream effect therefore tracks the labels rather than the generator.

2.13 | 5000-Node Scaling Across Four PBE Targets

We executed a 5000-node QMOF-derived package with the conditional WGAN-GP baseline (single-split, transductive protocol)

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

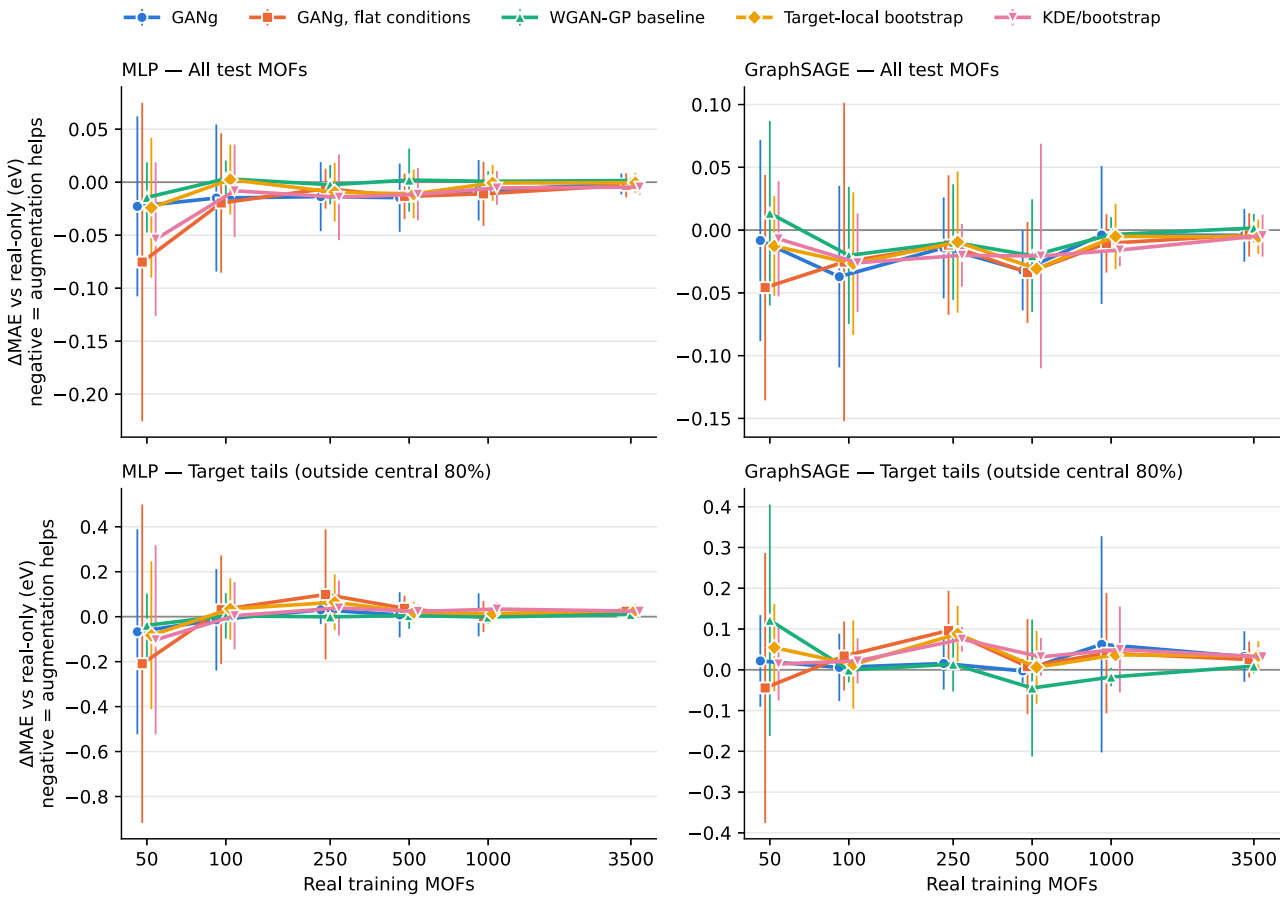


FIGURE 10 | Bandgap learning curves, strict inductive protocol, 5000-node cohort. Paired ΔMAE relative to real-only at each real training size, for GANg, the WGAN-GP baseline, and two resampling baselines, for all test MOFs (top) and for target tails outside the central 80% of training bandgaps (bottom); augmentation ratio selected on validation; mean over three splits (two seeds averaged within split) with exploratory 95% *t*-intervals; negative favors augmentation. "Flat conditions" draws requested bandgaps uniformly over the training range. Tables for all generators are given in Section S35 (Table S21).

for four PBE regression targets using the same target-column-excluded 106-feature policy and fixed reproducible split. The 5000-node analyses used the same training-only preprocessing and partition-restricted label policy as the leakage-corrected 1000-node workflow. The split assigns 70% of nodes to training, 15% to validation, and 15% to testing. The graph-scaling quantities in Table 4 are generated directly from the recorded graph-construction and sparsification output tables. The larger bandgap graph has a larger retained edge set but a lower post-sparsification graph density than the connected 1000-node reference. Message passing cannot cross component boundaries, so node representations in smaller components receive no information from the largest component; the downstream diagnostics are interpreted accordingly and the connected 1000-node case remains the cleanest connectivity-controlled reference.

Across all four 5000-node targets, the WGAN-GP baseline generated descriptor-valid and edge-valid nodes under implemented screens, with duplicate rate 0.000, training-set novelty-pass rate 1.000, and mean filtered-pass rate approximately 0.501 (range 0.500–0.503). Nearest real QMOF analog retrieval was available for generated nodes; for 20 selected bandgap descriptor-space hypotheses, the nearest real QMOF analog CIFs parsed with pymatgen and ASE and passed the implemented basic sanity checks (20/20).

Total energy is an extensive target and is partly linked to system size and composition descriptors such as atom count and volume. We therefore retrained all generators and downstream models independently for energy per atom (cell total energy divided by the cell atom count), and fitted validation-tuned ridge baselines on cell atom count, element counts reconstructed from fractions and atom count, or elemental fractions alone. Raw total energy is almost fully explained by size and composition: in the training data it correlates with cell atom count at −0.994, and an element-count ridge model reaches 4.73 eV test MAE at 5000 nodes, compared with 34.4 eV for atom count alone (Table S19). After normalization, augmentation effects are small and of mixed sign (Table 5); at 5000 nodes the paired change for the WGAN-GP baseline with GraphSAGE is +0.001 eV/atom (0.0464 and 0.0477 eV/atom), whereas the same model changes raw-energy MAE by +2.3 eV. Raw-energy differences are therefore treated only as a size-dominated stress test, and neither target is interpreted as formation energy or relative stability. The large absolute raw-energy changes (Figure 12B) reflect the extensive scale of the target and its size and composition correlations;

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

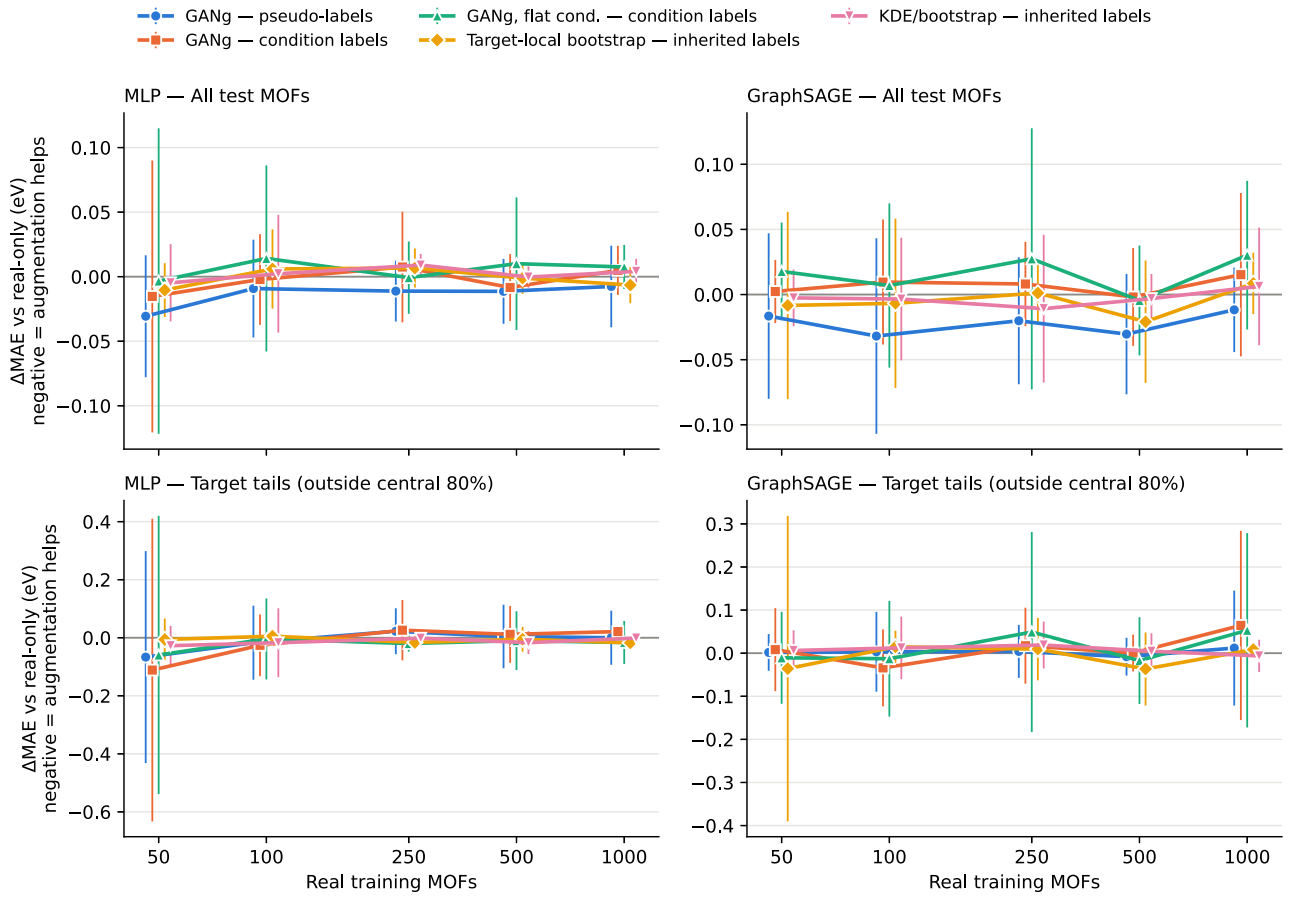


FIGURE 11 | Effect of the synthetic-label strategy on bandgap augmentation, strict inductive protocol, 5000-node cohort. Paired ΔMAE relative to real-only for GANg with pseudo-labels or condition labels (empirical and flat conditions) and for bootstrap generators with inherited source labels; augmentation ratio and synthetic loss weight selected on validation; mean over three splits with exploratory 95% *t*-intervals. All fifteen generator–label combinations are reported in Section S36 (Table S22).

TABLE 4 | Implementation-grounded graph-scaling summary for the 1000-node and 5000-node bandgap studies. Values and derived interpretations are calculated from the graph-construction and sparsification output tables.

Graph characteristic	Smaller study	Larger study	Scaling interpretation
Node count	1000	5000	5.00× nodes
Edges before sparsification	9998	51 906	5.19× edges
Edges after sparsification	6651	41 909	6.30× retained edges
Edge reduction	33.48%	19.26%	1.74× removed-fraction ratio
Avg. degree after sparsification	13.302	16.764	1.26× average degree
Density after sparsification	0.01332	0.00335	0.25× density
Components after sparsification	1	5	More components in larger study
Largest-component size	1000	4923	98.5% of larger-study nodes in largest component
Sparsification configuration	top-k $k = 10$	top-k $k = 12$	Study-specific settings differ

they should not be read as exceptional energetic-property prediction, and the electronic targets remain the primary evidence.

The 5000-node downstream results are target- and model-dependent after validation-selected cross-mode evaluation (Table S7). Figure 12 reports $\Delta\text{MAE} = \text{MAE}_{\text{validation-selected baseline}} - \text{MAE}_{\text{real-only}}$: negative values favor the selected baseline mode, whereas positive values favor real-only. Bandgap and CBM effects are mixed by architecture, VBM favors real-only for all three architectures, and raw total energy has lower mean MAE for validation-selected baseline modes across all three architectures; this reflects the extensive scale and size correlation of the target and disappears after normalization (Table 5). The raw total-energy analysis is included primarily as an architectural stress test of scaling behavior for an extensive, unnormalized target rather than as a physical energetic model. These raw total-energy differences may substantially reflect system-size and

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 | Independently retrained energy-per-atom models (eV/atom), strict inductive protocol, 25% augmentation, three splits with two seeds each. A ridge model on elemental fractions alone reaches 0.044 eV/atom. This is a normalization control, not a stability or formation-energy benchmark. "Conditional WGAN" denotes the conditional WGAN-GP baseline. Δ is the paired mean over splits and seeds and may differ from the difference of rounded means.

Nodes	Model	Real-only	Conditional WGAN	Δ [95% interval]
1000	GraphSAGE	0.064	0.062	−0.002 [−0.006, +0.002]
1000	GAT	0.160	0.157	−0.003 [−0.005, −0.001]
1000	GCN	0.189	0.189	+0.000 [−0.002, +0.003]
1000	MLP	0.056	0.056	−0.000 [−0.001, +0.000]
5000	GraphSAGE	0.046	0.048	+0.001 [−0.000, +0.003]
5000	GAT	0.125	0.126	+0.001 [−0.005, +0.006]
5000	GCN	0.162	0.165	+0.003 [−0.002, +0.008]
5000	MLP	0.044	0.045	+0.001 [−0.000, +0.002]

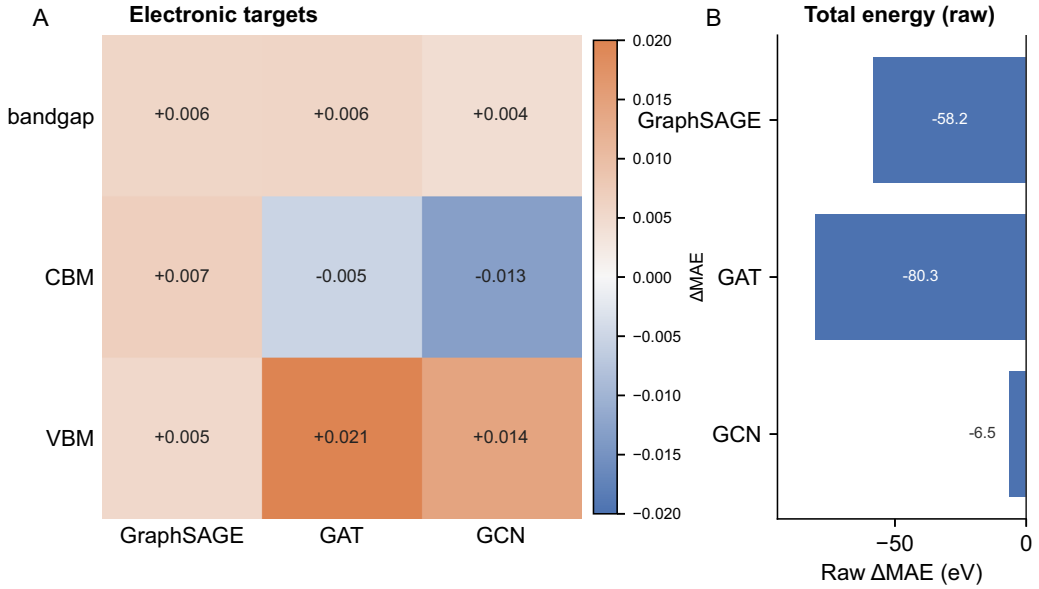


FIGURE 12 | 5000-node multitarget downstream diagnostic after validation-selected mode selection for the conditional WGAN-GP baseline (single-split, transductive protocol). (A) The electronic-target panel reports ΔMAE, defined as validation-selected baseline-mode MAE minus real-only MAE, for bandgap, CBM, and VBM on a symmetric zero-centered diverging linear scale; blue indicates lower error for the baseline mode and amber indicates real-only lower error. (B) Total energy is shown separately on the right with its own raw zero-referenced linear scale because its ΔMAE values are numerically much larger. Panel (B) annotations give raw ΔMAE in eV only. Raw total energy is a size-dominated stress test; after normalization to energy per atom, effects are small and of mixed sign (Table 5). Negative ΔMAE = baseline lower error; positive ΔMAE = real-only lower error. A logarithmic scale is not used because ΔMAE includes positive, negative, and near-zero values. The raw total-energy analysis is included primarily as an architectural stress test of scaling behavior for an extensive, unnormalized target rather than as a physical energetic model. Raw total-energy differences may substantially reflect system-size and composition effects. The paired five-seed analyses quantify initialization sensitivity under one fixed split and are therefore treated as exploratory descriptive diagnostics rather than definitive performance estimates.

composition effects and should not be interpreted as evidence of improved transferable energetic modeling. The energy-per-atom control above confirms this interpretation. The paired five-seed analyses quantify initialization sensitivity under one fixed split and are therefore treated as exploratory descriptive diagnostics rather than definitive performance estimates.

Figure 13 visualizes only the edge counts before and after sparsification for the 1000-node and 5000-node bandgap studies. Other graph-scaling quantities, including node count, average degree, graph density, connected components, and largest-component size, are reported in Table 4. A linear scale is retained because the figure compares directly interpretable edge counts rather than values spanning many orders of magnitude.

2.14 | Multitarget PBE Extension in the 1000-Node Setting

Using the same 106-feature target-column-excluded descriptor policy and fixed split seed 42, we leakage-corrected and reran the 1000-node package for four PBE targets (single-split,

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

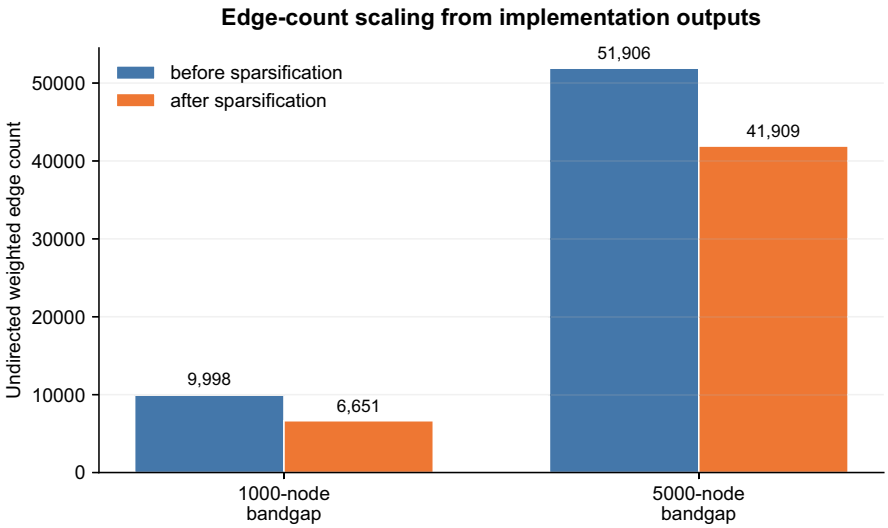


FIGURE 13 | Edge-count scaling comparison between the connected 1000-node bandgap study and the 5000-node bandgap study. The bars show only implementation-reported edge counts before and after sparsification; graph density, average degree, connected-component count, and largest-component size are reported separately in Table 4.

transductive protocol). For this predefined target-wise control, the filtered mode of the WGAN-GP baseline was applied consistently across all four targets and was not selected using test-set performance. For bandgap, GraphSAGE real-only gives MAE 0.606 ± 0.039, while the filtered baseline gives MAE 0.645 ± 0.024. For CBM, the values are 0.543 ± 0.014 and 0.569 ± 0.020; for VBM, 0.417 ± 0.021 and 0.419 ± 0.011; and for total energy, 284.1 ± 6.6 and 273.5 ± 19.5, respectively. The raw total-energy result remains an extensive-target stress test rather than evidence of transferable energetic learning; its absolute MAE scale is not comparable with the electronic targets. These target-wise controls show that generated-node use must be benchmarked separately for each property rather than inferred from descriptor validity alone.

2.15 | Claim Boundary

GANg generates descriptor-space graph nodes and similarity edges in QMOF-derived MOF similarity networks; the implemented descriptor, edge, and nonduplication screens were evaluated on nodes generated by the conditional WGAN-GP baseline, all of which passed them. The generated objects are descriptor-space graph hypotheses unless decoded and physically validated. The structure–decoder extension is provided as future-facing software infrastructure and does not change the descriptor-space scope of the present study. No generated CIF, DFT validation, force-field relaxation, synthesis-ready assessment, or experimental validation is claimed in this work. Graph edges are assigned by deterministic cosine attachment after descriptor generation and are not generated. GANg descriptors satisfy composition, Z-multiplicity, density, and pore-diameter relations by construction; this is consistency with the enforced descriptor relations, not chemical, crystallographic, or structural validity. GANg does not generate CIFs, atomistic structures, crystal topology, or physically validated materials. Augmentation gains are reported only where measured—modest, in the low-data regime, comparable to unconditional generation and simple resampling—and no GAN-specific or universal predictive superiority is claimed.

3 | Discussion

GANg establishes a descriptor-space generative layer for MOF similarity networks and exposes a validity–utility gap. Rather than treating a generated descriptor vector as a material, the framework asks whether the vector is plausible as a graph node, whether its proposed similarity edges are consistent with the real network, whether it is distinct from training examples, whether it can be interpreted through neighboring real QMOFs, and whether validation-selected augmentation actually helps held-out real-node prediction. This framing is important for materials AI because it creates an auditable intermediate representation between tabular generation and atomistic structure generation.

The principal intelligent-system contribution is the integration of target-conditioned generation, descriptor-space graph-node synthesis, similarity-edge insertion, validity and novelty screening, topology diagnostics, nearest-real-material analog retrieval, and held-out-real-node evaluation in one reproducible workflow. Auditability arises from preserving the representation and feature policy, tracing numerical claims to source artifacts, and keeping computational, structural, and physical claims separate.

The matched experiments show why this separation matters. Every sample of the conditional WGAN-GP baseline passed the descriptor, edge, and novelty screens, yet the matched diagnostics revealed concentration consistent with partial mode collapse, inconsistent descriptor combinations, and no measurable conditional response. None of these limitations was visible to range-based validity checks, and the repeated nearest analogs were a symptom of concentration rather than evidence of plausibility. The same diagnostics motivated GANg's design: a longer, coverage-selected training schedule and a constrained output

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that reuses the real-data featurizer. We regard this as a general lesson for descriptor generation in materials science: wherever descriptors are deterministic functions of a smaller core of independent descriptors, generating the core and recomputing the rest is simpler and more reliable than penalizing violations after the fact.

The nearest-analog retrieval results strengthen this interpretation. Generated nodes can be linked to real QMOF neighborhoods, target-property summaries, and available reference analog CIFs. The 20/20 structural sanity result for selected nearest real analogs does not validate the generated descriptors as physical structures, but it does provide a structure-grounded context for hypothesis triage. In practical terms, a generated node can be inspected as a region of known materials space: which real MOFs surround it, how similar they are, what property values they carry, and which reference structures might guide subsequent decoding or retrieval workflows.

Nearest-real-analog retrieval makes GANg hypotheses auditable, but it should be interpreted with the same care as other similarity-based analog methods. Prior work on MOF diversity has shown that MOF databases and chemical-design spaces can contain coverage biases and that similarity measures may emphasize some chemical or structural aspects while under-representing others [39]. Related read-across studies likewise show that graph-similarity methods can help identify analogs, but analog suitability depends on the similarity definition, data coverage, and target context [40]. In GANg, nearest QMOF analogs therefore provide a structure-grounded interpretation layer and triage context, not independent validation of generated descriptors as physical MOFs.

3.1 | Relationship to Structure-Generative Materials Models

GANg is complementary to VAE, diffusion, and structure-aware materials generators rather than a replacement for them (Table S8). Representative examples include supramolecular VAE-based inverse design of reticular materials [9], diffusion-based porous-material inverse design [10], MOF-focused molecular diffusion for carbon-capture design [11], CDVAE-style periodic material generation [41], broader MOF generative-AI workflows [3], analyses of diffusion-model limitations in unexplored materials space [42], symmetry-aware Wyckoff-position generation [43], and all-atom diffusion transformers [44]. These approaches generally target latent structure spaces, linker or building-block design, periodic material generation, symmetry-aware periodic candidate modeling, or atomistic candidate generation, whereas GANg operates at the descriptor-space graph-node level in a MOF similarity network.

This difference is both a limitation and a strength. GANg does not produce CIFs or periodic structures, but its graph-node representation is transparent, scalable, auditable, and directly connected to nearest real QMOF analogs. A generated node can therefore be inspected through its descriptor position, proposed similarity edges, nearest-neighbor property context, and retrieved reference analogs before any structure-level claim is considered.

A future structure-aware pipeline could use descriptor-space hypotheses, target windows, and analog context as conditioning signals for a latent VAE, diffusion, or symmetry-aware decoder. Such a pipeline would require structural decoding, geometry relaxation, DFT validation for selected decoded candidates, stability and synthesizability screening, and eventual experimental feedback. These stages are not implemented in the present work. Establishing a materials-design claim would additionally require checking that decoded structures reproduce the requested descriptors and targets; retrieved-analog inspection cannot substitute for any of these steps.

Consistent with this direction, the repository includes a modular structure–decoder extension interface. The implementation exports generated hypotheses into descriptor-to-structure conditioning records, links them to nearest real QMOF analog CIFs, provides safe adapter interfaces for external CDVAE-, MOFFUSION-, MOFDiff-, WyckoffDiff-, and ADiT-style decoders, and includes CIF sanity-validation utilities. This addition makes the route from descriptor-space hypothesis generation to future structure-aware decoding explicit while keeping the present study focused on auditable descriptor-space generation. The constrained core of GANg—a sparse composition, formula and cell atom counts, volume, pore diameters, and space group—is a natural conditioning record for such decoders, because each record already satisfies the stoichiometric and density descriptor relations that a decoded structure must reproduce; it is not itself a structure.

The downstream experiments add a second level of evidence and permit a more specific explanation of the validity–utility gap. They show that descriptor-space plausibility and graph insertion should be evaluated separately from supervised augmentation utility; the GraphSAGE, GAT, GCN, and MLP experiments are conservative augmentation diagnostics, not claims to outperform specialized MOF property predictors. Three hypotheses can be assessed. Message-passing perturbation is not the explanation: random or absent attachment and a graph-free model give the same results as cosine attachment. Concentrated sampling is not the explanation either: substantially mitigating the concentration makes the low-label gains somewhat more consistent but no larger than those of unconditional generation and local resampling. The explanation best supported by the data is the information content of the synthetic labels. Synthetic descriptors carry no label information of their own; with a shared pseudo-label model, augmentation transfers that model's predictions to the student. This helps a little when the student is data-starved and the pseudo-label model is more accurate, disappears once real labels suffice, and harms target tails, where tree-based pseudo-labels shrink toward the mean. Labels derived from the condition are noisier still, because a requested property is not a measured one. Under these conditions, no generator, however diverse or well-conditioned, can add predictive information that its labels do not contain, which explains why conditional, unconditional, and statistical generators converge.

The raw total-energy analysis was included as an architectural stress test of an extensive, unnormalized target. The energy-per-atom rerun and the size and composition ridge controls

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

confirm that raw-energy improvements reflect system size and composition and must not be read as transferable energetic learning or physical validation. Several limitations remain. The cohorts are nested database-order subsets of QMOF, and three partitions with two initializations each describe partition sensitivity rather than population uncertainty. Generators use fixed rather than extensively tuned hyperparameters. Conditional fidelity is scored by a surrogate model of limited accuracy ($R^2 \approx 0.65$), and the learning-curve and label-strategy experiments are restricted to bandgap. Composition-disjoint splitting tests one form of transfer but not generalization to unseen structural families.

These results also clarify where target conditioning is useful. In GANg the condition shapes generation measurably in surrogate diagnostics, but synthetic training data are not the right outlet for that capability. Targeted generation is: proposing consistent descriptor hypotheses in under-represented property ranges for retrieval, decoding, or screening, with any predictive use resting on measured rather than requested labels, and with the caveat that surrogate condition-response scores do not show that a realizable MOF would possess the requested property.

3.2 | From Descriptor Hypotheses to Validated Materials

GANg deliberately occupies an intermediate layer between generative descriptors and validated materials. This position reflects a broader validation gap in data-driven materials generation: statistical plausibility, descriptor-range validity, or nearest-neighbor similarity do not by themselves establish physical correctness. Recent materials-generative and inverse-design discussions emphasize that predictive and generative workflows require domain constraints, robustness checks, interpretability, transferability assessment, transparent pipelines, and ultimately physical or experimental confirmation before deployment-level claims are justified [1, 6]. In GANg, this gap is handled by separating descriptor validity, graph-edge plausibility, novelty, nearest-real-analog interpretation, and downstream GNN diagnostics from structure-level validation.

This balanced outcome is scientifically useful. It demonstrates scalable descriptor-space graph-node synthesis, transparent filtering, nearest-neighbor interpretability, structure-grounded analog retrieval, and conservative downstream benchmarking in a single reproducible workflow. Future work should integrate atomistic or topology-aware descriptors, calibrated property heads, uncertainty-aware pseudo-labeling, structure-aware decoding, force-field relaxation, DFT validation of selected decoded candidates, external-dataset validation, and experimental feedback. Stronger tabular generators with mode-specific normalization [45] and labels from physics-based calculations for selected synthetic descriptors are the most direct routes to augmentation that adds information rather than redistributing it. These additions could connect descriptor hypotheses to later structure-aware materials-design stages without changing the scope of the present evidence.

4 | Methods

4.1 | Node Table and Feature Policy

The connected 1000-node study uses a fixed random split for reproducibility, with the same split policy applied to the multi-target and 5000-node extensions. Descriptor features include formula-derived composition fractions, chemistry-role descriptors, metal-family indicators, atomic statistics, and scalar structural metadata available without using target/output columns. PBE target quantities are excluded from node features and used only as supervised labels, conditioning controls, or post hoc interpretation values. This target-column-excluded policy is essential because the graph itself is constructed from descriptor similarity; including target-derived outputs as node features would contaminate both similarity edges and downstream evaluation. Section S2 summarizes the feature groups, and the distributed feature-audit CSV lists all 106 feature names and their use in graph construction, generator training, and downstream GNNs.

4.2 | Similarity Graph Construction and Sparsification

Descriptor features are normalized before graph construction. Each MOF is represented as one node, and weighted undirected edges are created from cosine nearest-neighbor similarity in descriptor space. Cosine similarity is appropriate for this representation because the descriptors combine heterogeneous normalized quantities and the graph is intended to capture relative feature orientation rather than direct Euclidean distance alone. Sparsification compares top-k, mutual-kNN, threshold, and connectivity-preserving variants. For the primary 1000-node study, the selected connected graph balances reduced density, component preservation, retained average degree, and computational tractability. The 5000-node extension retains five connected components, with 4923 of 5000 nodes in the largest component, and is therefore interpreted as scale-up evidence rather than a connectivity-controlled replacement for the primary study. Message passing cannot cross component boundaries, and no largest-component-only rerun was performed. Bandgap-bin homophily is reported after graph construction as a descriptive target-neighborhood diagnostic, not as the primary graph-selection criterion.

4.3 | Conditional WGAN-GP Baseline

The conditional WGAN-GP baseline uses a WGAN-GP objective because Wasserstein training with gradient penalty is well suited to continuous descriptor generation and provides a stable critic signal for bounded feature spaces [24]. The generator maps latent noise and a conditioning scalar to MOF-level descriptor vectors. The critic scores real and generated descriptors under the Wasserstein objective, encouraging generated nodes to occupy plausible regions of the training descriptor distribution. Generated descriptors are clipped to training feature bounds and inserted by cosine nearest-neighbor attachment to real training nodes following Algorithm 1. The inserted edges are descriptor-space similarity proposals, not chemical bonds or structural connectivity. Generator fitting, descriptor preprocessing, feature-

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ALGORITHM 1 | Conditional WGAN-GP baseline: full-descriptor generation and cosine-kNN insertion (see Section 4.4 for GANg).

Require: Real MOF descriptor matrix X, real similarity graph $G = (V, E)$, target-conditioning values c, trained WGAN-GP generator G_θ, attachment degree k

Ensure: Generated descriptor-space nodes $\hat{V}$ and proposed similarity edges $\hat{E}$

1: Normalize real descriptors and construct/sparsify the MOF similarity graph.

2: **for** each requested generated node **do**

3: Sample latent vector z and conditioning value c.

4: Generate descriptor vector $\hat{x} = G_\theta(z, c)$.

5: Clip $\hat{x}$ to training descriptor bounds and reject nonfinite descriptors.

6: Compute cosine similarity between $\hat{x}$ and the real training-node descriptors.

7: Attach $\hat{x}$ to the top-k nearest real training nodes as proposed similarity edges.

8: Store $\hat{x}$ as a descriptor-space graph node, not as a decoded crystal structure.

9: **end for**

10: Apply descriptor-validity, edge-plausibility, duplication, novelty, distance, degree, and uncertainty filters.

range estimation, and pseudo-label model fitting were performed exclusively using the training partition. Validation labels were used for checkpoint and augmentation-mode selection, while held-out test labels were accessed only for final evaluation. Under the transductive setting, the full real-node graph and its target-column-excluded features were available during message passing. In the baseline, generator and critic each have two hidden layers of width 128; a 64-dimensional Gaussian latent and the standardized target enter the generator; training uses 100 epochs, batch size 64 in the single-split package and 128 in the matched experiments, Adam (learning rate 2×10^{-4}, $\beta = (0.5, 0.9)$), gradient-penalty weight 5, one critic update per generator update, and a soft range penalty $0.05\,\mathrm{mean}[\max(|x| - 3, 0)]$ in standardized coordinates.

Algorithm 1 operationalizes descriptor-space generation and cosine-kNN insertion into the real similarity graph for the conditional WGAN-GP baseline, which generates the full descriptor vector and relies on coordinate-wise clipping. GANg (Section 4.4) instead generates a constrained descriptor core and recomputes all derived descriptors, so coordinate-wise clipping is not used; graph insertion (steps 6–7) is identical in both.

4.4 | GANg Generator

GANg uses a conditional WGAN-GP objective, the scalar conditioning input, and cosine-kNN graph insertion, as does the baseline; it generates descriptor vectors only, and graph edges are attached deterministically afterwards. In code and file names, GANg carries the label `v2` (for example `gang_v2.py`). The generator (three hidden layers of width 256, LeakyReLU) has two heads. The composition head produces 75 logits, one per element in the descriptor basis; logits of elements absent from the training rows are masked, and a sparsemax transformation [29] maps the remainder onto the probability simplex with exact zeros. The scalar head produces six z-scored values: the logarithms of formula atom count, cell atom count, cell volume, and largest cavity diameter (LCD), the logit of the ratio of pore-limiting diameter (PLD) to LCD, and the space-group number. For the critic, a differentiable implementation of the descriptor featurizer reconstructs all 106 descriptors from this core—group fractions, counts, ratios, and composition-weighted means and ranges of atomic number, mass, and electronegativity, plus density as cell atom count times mean atomic mass over volume—and standardizes them with the training statistics; the six core scalars are appended. A unit test confirms that the differentiable featurizer and the standard featurizer agree on real QMOF entries and that real entries are recovered exactly from their core. The critic (three hidden layers of width 256) receives this vector and the condition. At sampling time, element counts are obtained by rounding fractions times the formula atom count, a formula string is written, and all composition descriptors are recomputed with the standard featurizer; Z is the rounded ratio of cell to formula atom count, the cell atom count is set to Z times the formula count, density is recomputed from Z, formula mass, and volume, PLD is the predicted ratio times LCD, and the space group is snapped to the nearest training value; this is a metadata constraint and does not establish real crystallographic symmetry. Volume and LCD are bounded to the training range; no other clipping is applied. Training uses five critic updates per generator update, gradient-penalty weight 10, Adam (learning rate 10^{-4}, $\beta = (0.5, 0.9)$), batch size 128, an exponential moving average of generator weights (decay 0.999), and 6000–10 000 generator updates. Generator and critic parameters are fitted on training rows only. Every 250 updates the averaged generator is evaluated by precision, recall, density, and coverage [28] ($k = 5$) of the validation descriptors, using validation targets only as conditions; the checkpoint with the highest coverage plus $0.1\,\min(\mathrm{density}, 1)$ is retained. No test descriptors or test labels enter fitting or checkpoint selection; test descriptors are used only in post hoc evaluation diagnostics. The unconditional GANg control sets the condition to zero throughout. The ablation in Table 2 adds a

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

mode-seeking term [30] with weight 1 and a minibatch-standard-deviation critic feature [31].

4.5 | Matched Controls and Strict Inductive Evaluation

The matched experiments use the first 1000 and 5000 distinct QMOF entries with finite bandgap, CBM, VBM, and total energy and a positive cell atom count; the smaller cohort is nested in the larger. Three random split seeds assign 70%/15%/15% of entries to training, validation, and testing, each with two paired model seeds. Imputation, scaling, feature bounds, target windows, generators, the pseudo-label model, and downstream model parameters are fitted on training rows and nodes only. Validation rows are used only for early stopping, checkpoint selection, and the choice of augmentation ratio and synthetic loss weight. Test descriptors, test labels, and test nodes are used for none of fitting, graph construction, checkpoint selection, or tuning, and enter only final scoring. The conditional WGAN-GP in these experiments is the baseline of Section 4.3. The unconditional WGAN-GP has the same parameter count and budget as the conditional model with the condition fixed to zero; the shuffled-condition model uses one fixed permutation of training targets. Statistical generators are a Gaussian-kernel bootstrap (bandwidth 0.15), an eight-component diagonal GMM, convex interpolation between a training MOF and one of its five nearest cosine neighbors, and target-local bootstrap, which samples among the training MOFs nearest the requested target (10%, at least 20) before the same jitter. Synthetic nodes, 25% of the training count unless stated otherwise, receive pseudo-labels from one training-only ExtraTrees regressor (80 trees, minimum leaf size 3) [37]. The graph is the union of cosine 15-nearest-neighbor links among training nodes; synthetic nodes attach only to training nodes. In strict inductive inference, each validation or test query receives messages from its 15 nearest training neighbors and itself and sends none, and reference degrees and representations stay fixed. The transductive control exposes all real descriptors during message passing. GraphSAGE, GCN (two layers), GAT (one attention layer and a hidden linear layer), and an MLP use hidden width 64, Adam (learning rate 0.003, weight decay 10^{-4}), at most 80 epochs, and early stopping on validation MAE with patience 15. Conditional fidelity uses an independent random forest (80 trees, minimum leaf size 3) [25] fitted on training rows and scored in five training-quantile windows with 128 samples each. Energy per atom divides cell total energy by the cell atom count; ridge baselines select regularization from $\{0.01, 0.1, 1, 10, 100\}$ on validation MAE. The composition-disjoint control splits groups of identical reduced formula.

4.6 | Learning-Curve and Label-Strategy Experiments

For each split seed of the 5000-node cohort, training subsets of 50–3500 MOFs are the first entries of a split-specific permutation of the training partition, so smaller subsets are nested in larger ones; the validation set is a fixed permutation of the validation partition truncated to 20% of the training size (minimum 30), and the 750-MOF test partition is unchanged. Tails are test MOFs whose target lies below the 10th or above the 90th percentile of the training subset. Flat conditions are drawn uniformly between the minimum and maximum training target. In the label-strategy experiment, synthetic loss weights λ multiply the squared error of synthetic nodes, and the loss is normalized by the total weight. Agreement filtering keeps synthetic samples whose pseudo-label lies within the pseudo-label model's validation MAE of the requested condition, and inherited labels assign each jittered bootstrap sample the real label of its source MOF. With $\lambda = 1$, the weighted trainer reproduces the unweighted trainer exactly.

4.7 | Statistical Summaries

The primary paired effect is $\Delta\mathrm{MAE} = \mathrm{MAE}_{\mathrm{augmented}} - \mathrm{MAE}_{\mathrm{real-only}}$ within the same size, target, architecture, protocol, split, and model seed. Model seeds are averaged within each split before computing the mean over three splits and an exploratory 95% Student's t interval with two degrees of freedom. Because the splits share a fixed cohort, these intervals describe partition sensitivity rather than independent population uncertainty, and we make no multiple-comparison significance claims.

4.8 | Generated-Node Evaluation

Generated-node evaluation is separated into descriptor validity, edge plausibility, novelty, filtering, and interpretation. Descriptor validity checks finite values, training-range constraints, composition sanity, scalar-descriptor sanity, and general constraint satisfaction. Edge plausibility checks generated degree, edge-weight range, cosine consistency, and connectivity after insertion. Novelty checks duplicate rate, nearest-training distance, and training-set novelty-pass rate. In the implemented evaluation code, real and generated descriptor matrices are normalized using the real-training descriptor mean and standard deviation before cosine attachment; nearest-training similarity is the maximum cosine similarity to training descriptors, novelty score is 1−nearest-training similarity, duplicate status is assigned when nearest-training similarity exceeds 0.999, and training-set novelty-pass status uses novelty score >0.01. Analog retrieval is a separate interpretation step and may use a broader real QMOF reference pool after train-set normalization; a generated node can therefore pass the training-set novelty screen while still having a close retrieved real analog outside the training-only novelty comparison set. The final filtering stage combines these signals with a similarity window, edge checks, and median uncertainty thresholding to identify generated nodes suitable for downstream augmentation or interpretation. Nearest-neighbor interpretation reports top real QMOF analogs, similarity scores, top-neighbor target summaries, and feature-difference comments, thereby converting a generated vector into an auditable graph-space hypothesis.

4.9 | Metric Definitions

Let N_g be the number of generated nodes. Descriptor-valid rate is $N_{\mathrm{descriptor\ pass}}/N_g$; edge-valid rate is $N_{\mathrm{edge\ pass}}/N_g$; duplicate rate is $N_{\mathrm{duplicates}}/N_g$; training-set novelty-pass rate is N_{novel}/N_g under the

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

training-node comparison set; and filtered-pass rate is N_{retained}/N_g. Hit rate is the fraction of generated hypotheses whose nearest-real-analog target summary falls inside the requested target window. These rates lie in [0, 1].

For the Figure 4 audit, each feature is standardized using the real-training mean and standard deviation. PCA is fitted to these standardized real-training descriptors and then applied without refitting to generated descriptors. For feature j, the standardized mean difference is $\left(\bar{x}_{g,j} - \bar{x}_{\text{tr},j}\right)/s_{\text{tr},j}$. Generated-to-real similarity is the maximum cosine similarity between each generated descriptor and the real-training descriptors in this representation; the real-data reference is the leave-one-out maximum cosine similarity among real-training descriptors.

In the matched and GANg diagnostics, effective rank is $\exp\left(-\sum_i p_i \log p_i\right)$ with p_i the normalized eigenvalues of the covariance of 256 generated descriptors in training-standardized coordinates. The within-window diversity ratio divides the mean pairwise Euclidean distance of 128 samples requested at a window midpoint by that of real training MOFs in the same window, averaged over five windows. The surrogate window hit rate is the fraction of those samples whose independent-surrogate prediction falls inside the window. Analog concentration counts the nearest training MOF of each of 256 samples under cosine distance: the top-1 share is the largest count divided by 256, and the distinct-analog fraction is the number of different nearest MOFs divided by 256. Precision, recall, density, and coverage follow Naeem et al. [28] with $k = 5$, comparing generated descriptors with the held-out real descriptors.

Bandgap-bin homophily is the fraction of graph edges whose endpoints share the same discretized bandgap bin and also lies in [0, 1]. For n held-out real nodes with targets y_i and predictions $\hat{y}_i$,

$$\text{MAE} = \frac{1}{n}\sum_i |y_i - \hat{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

Both are non-negative and lower is better. $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, with $R^2 \le 1$ and values below zero possible. Spearman and Kendall rank correlations lie in [−1, 1], where larger positive values indicate stronger rank agreement. Finally,

$$\Delta\text{MAE} = \text{MAE}_{\text{augmentation}} - \text{MAE}_{\text{real-only}}$$

Negative ΔMAE means the augmentation mode has lower error; positive ΔMAE means real-only remains better.

4.10 | Structure-Grounded Analog Retrieval

For generated nodes with available nearest real QMOF identifiers, the retrieval step links each descriptor-space hypothesis to nearby real QMOF analog structures. Retrieval is an interpretability aid only and provides no evidence of novelty or of a generated physical structure. Retrieved CIFs are treated strictly as reference analogs from the QMOF source, not as generated structures. The workflow selected the first 20 rank-1 bandgap analog records for the reported structural inspection. Their CIFs were parsed with pymatgen and ASE and screened for positive site count, volume, and density and for no interatomic distance below 0.5 Å when a distance matrix was available [26, 27]. This step grounds descriptor-space hypotheses in known structural context while preserving the distinction between analog retrieval and structure generation. Because analog retrieval depends on the selected descriptor representation, normalization, similarity metric, and database coverage, nearest-neighbor outputs are used as interpretation aids rather than validation labels.

4.11 | Downstream Graph Learning

GraphSAGE, GAT, and GCN are standard GNN architectures for neighborhood aggregation, attention-weighted aggregation, and graph convolution, respectively [32–34]. They are trained across five seeds under real-only, random synthetic, feature-noise, nearest-neighbor interpolation, baseline-generated, filtered-baseline, property-head pseudo-label, and conditional-target label modes; the last four use nodes generated by the conditional WGAN-GP baseline. Within each mode, the checkpoint is selected using validation MAE. Across these baseline modes, the principal manuscript comparison selects the mode with the lowest validation MAE for each target–architecture pair and then reports the selected mode's held-out test metrics. Generated nodes are used only for training-time augmentation; test metrics are computed only on held-out real QMOF nodes. Training uses the fixed training settings summarized in Section S12 and Tables S4 and S5. This transductive evaluation design exposes the full target-column-excluded real graph during message passing while restricting all fitted preprocessing and label use to their designated partitions. It treats downstream learning as a conservative diagnostic rather than as a claim to outperform specialized MOF predictors [16–18]. The matched, learning-curve, and label-strategy experiments use the strict inductive protocol described above as their primary setting.

Algorithm 2 operationalizes the validity-aware downstream utility diagnostic and keeps predictive evidence separate from physical-material claims.

4.12 | Structure–Decoder Extension Interface

The repository includes a modular structure–decoder extension interface for future structure-aware workflows (Table S9). It exports generated descriptor-space hypotheses into JSONL conditioning records, links them to nearest real QMOF analog CIFs, provides adapter classes for external structure-generative models, and validates available CIFs using basic pymatgen/ASE sanity checks. The current implementation produced 1000 conditioning records from baseline-generated nodes and exported 20 nearest real QMOF analog CIFs for basic sanity checks. Detailed scope and claim-boundary information are provided in the Supporting Information.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ALGORITHM 2 | Validity-aware utility diagnostic for descriptor-space graph generation.

Require: Real-only scores, generated-node validity tables, augmentation scores, paired random seeds

Ensure: Claim category for descriptor-space graph-node augmentation

1: Compute descriptor-valid, edge-valid, nonduplicate, novelty-pass, and filtered-pass rates.

2: Interpret generated nodes through nearest real QMOF neighbors and target summaries.

3: **for** each model in GraphSAGE, GAT, and GCN **do**

4: Select checkpoints within each mode using validation MAE.

5: Select the augmentation mode across candidate modes using validation MAE only.

6: Evaluate real-only and the validation-selected augmentation mode on the same held-out real nodes.

7: Compute paired ΔMAE, confidence intervals, effect sizes, and seedwise consistency across seeds.

8: **end for**

9: Report descriptor-space plausibility separately from augmentation utility.

10: State that generated nodes are descriptor-space hypotheses, not validated MOF structures.

4.13 | Reproducibility

The author-controlled evidence package contains the workflow code, processed artifacts, configurations, figure-generation resources, evidence tables, and manuscript sources used for this study. The main workflow scripts build the QMOF node table and similarity graph, sparsify the graph, train the generator, generate and evaluate synthetic nodes, train downstream GNNs, retrieve real analogs, and regenerate figures. Inputs are QMOF metadata/structures plus the supplied configuration files; outputs include descriptor tables, graph edge lists, generated-node and generated-edge tables, analog records, GNN result tables, and figures. The observed 5000-node main pipeline runtime was approximately 8 min per target, or 32 min for four targets, on the recorded CPU-only environment; auxiliary aggregation and figure steps required a few additional minutes. The recorded environment used a MacBook Pro Mac14,9 with Apple M2 Pro, 10 CPU cores, 32 GB memory, macOS-26.3 on Apple arm64, Python 3.9.6, torch 2.4.0, no CUDA, and no torch MPS availability; additional package details are listed in Section S20 and Table S6. Runtime depends on hardware and configuration. Numerical claims are traceable to source tables in the evidence package. The matched-control, ablation, learning-curve, and label-strategy experiments are run by separate scripts with per-unit completion manifests (matched controls, the cumulative ablation from the baseline to GANg, learning curves, and label strategies; Section S37); completed units are reused only when configuration and code hashes match, and the single-split results are not pooled with them.

5 | Conclusion

GANg synthesizes target-conditioned descriptor-space nodes, inserts them into MOF similarity networks, and evaluates them through layered validity, novelty, topology, analog-retrieval, and real-node diagnostics. It deliberately does not generate CIFs, atomic coordinates, periodic structures, or validated materials. Matched controls showed that a conditional WGAN-GP baseline that emits the full descriptor passed every range screen while concentrating its samples around a few training neighborhoods, emitting internally inconsistent descriptors, and showing no measurable conditional response, exposing a validity–utility gap. GANg's constraint-aware design substantially mitigates these limitations: it increases sampling diversity and coverage, satisfies the enforced composition and cell-metadata relations by construction, and improves the surrogate conditional response, although effective rank remains below that of real data. Better descriptor-space diagnostics did not, however, translate into markedly better prediction. Pseudo-label augmentation with GANg helps modestly when real labels are scarce, is matched by unconditional generation and local resampling, weakens or vanishes at larger label budgets, and is attributable mainly to pseudo-label distillation rather than to adversarial conditional generation. GANg does not establish GAN-specific or universal predictive superiority, and it generates descriptor-space hypotheses, not CIFs, atomistic structures, or physically validated materials. GANg therefore demonstrates that responsible generative intelligent systems for scientific networks should be evaluated not only by whether they produce valid synthetic entities but also by how those entities integrate, remain interpretable, preserve network behavior, and affect learning on held-out real data—and that the information carried by synthetic labels must be assessed as carefully as the synthetic descriptors themselves.

Author Contributions

M.J. conceived the study, designed and implemented the computational workflow, performed the analyses, prepared the figures and manuscript, and curated the repository artifacts. A.O. reviewed the manuscript and contributed feedback on synthetic-data validation, robustness, statistical interpretation, and presentation. M.R. contributed expert feedback on synthetic-data validation, robustness, and interpretation and reviewed the manuscript.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2154 – Project ID 390874152 (POLiS Cluster of Excellence). Open access funding enabled and organized by SRH University of Applied Sciences Heidelberg.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The code and derived artifacts supporting this study are publicly available in the GANg GitHub repository: https://github.com/MehrdadJalali-AI/GANg. Raw QMOF structures and metadata should be obtained from the original QMOF source [21] and are not redistributed.

References

1. L. Chen, W. Zhang, Z. Nie, S. Li, and F. Pan, "Generative Models for Inverse Design of Inorganic Solid Materials," *Journal of Materials Informatics* 1 (2021): 4, https://doi.org/10.20517/jmi.

2. E. Xie, X. Wang, J. I. Siepmann, H. Chen, and R. Q. Snurr, "Generative AI for Design of Nanoporous Materials: Review and Future Prospects," *Digital Discovery* 4 (2025): 2336–2363, https://doi.org/10.1039/D5DD00221D.

3. C. Duan, A. Nandy, S. C. Pal, et al., "The Rise of Generative AI for Metal-Organic Framework Design and Synthesis," *Matter* 9, no. 5 (2026): 102748, https://doi.org/10.1016/j.matt.2026.102748.

4. Y. Song, J. Li, D. Chi, et al., "Ai-Driven Advances in Metal–Organic Frameworks: From Data to Design and Applications," *Chemical Communications* 61, no. 82 (2025): 15972–16001, https://doi.org/10.1039/D5CC04220H.

5. R. Jabbar and S. Kamoun, "Recent Progress in Generative Adversarial Networks Applied to Inversely Designing Inorganic Materials: A Brief Review," *Computational Materials Science* 213 (2022): 111612, https://doi.org/10.1016/j.commatsci.2022.111612.

6. Y. Zhang, T. Long, and H. Zhang, "Generative Deep Learning for the Inverse Design of Materials," in *Artificial Intelligence and Intelligent Matter: Nanoscience, Soft Matter, Philosophy* (Springer, 2026), 127–166.

7. S. Badrinarayanan, R. Magar, A. Antony, R. S. Meda, and A. B. Farimani, "MOFGPT: Generative Design of Metal–Organic Frameworks Using Language Models," *Journal of Chemical Information and Modeling* 65, no. 17 (2025): 9049–9060, https://doi.org/10.1021/acs.jcim.5c01625.

8. H. Kneiding and L. Morán-González, "Nishamol Kuriakose, Ainara Nova, and David Balcells. Inverse Design of Inorganic Compounds With Generative AI," arXiv preprint arXiv:2604.11827 (2026), https://doi.org/10.48550/arXiv.2604.11827.

9. Z. Yao, N. S. B. Benjamin Sanchez-Lengeling, B. J. Bucior, et al., "Inverse Design of Nanoporous Crystalline Reticular Materials With Deep Generative Models," *Nature Machine Intelligence* 3, no. 1 (2021): 76–86, https://doi.org/10.1038/s42256-020-00271-1.

10. J. Park, A. P. S. Gill, S. M. Moosavi, and J. Kim, "Inverse Design of Porous Materials: A Diffusion Model Approach," *Journal of Materials Chemistry A* 12 (2024): 6507–6514, https://doi.org/10.1039/D3TA06274K.

11. H. Park, X. Yan, R. Zhu, et al., "A Generative Artificial Intelligence Framework Based on a Molecular Diffusion Model for the Design of Metal-Organic Frameworks for Carbon Capture," *Communications Chemistry* 7 (2024): 21, https://doi.org/10.1038/s42004-023-01090-2.

12. N. S. Bobbitt, K. Shi, B. J. Bucior, et al., "MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials," *Journal of Chemical and Engineering Data* 68, no. 2 (2023): 483–498, https://doi.org/10.1021/acs.jced.2c00583.

13. P. Z. Moghadam, A. Li, S. Wiggin, et al., "Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future," *Chemistry of Materials* 29, no. 7 (2017): 2618–2625, https://doi.org/10.1021/acs.chemmater.7b00441.

14. Y. G. Chung, E. Haldoupis, B. J. Bucior, et al., "Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019," *Journal of Chemical and Engineering Data* 64, no. 12 (2019): 5985–5998, https://doi.org/10.1021/acs.jced.9b00835.

15. A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, and H. J. Kulik, "MOFSimplify, Machine Learning Models With Extracted Stability Data of Three Thousand Metal–Organic Frameworks," *Scientific Data* 9 (2022): 74, https://doi.org/10.1038/s41597-022-01181-0.

16. Z. Cao, R. Magar, Y. Wang, and A. B. Farimani, "MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction," *Journal of the American Chemical Society* 145, no. 5 (2023): 2958–2967, https://doi.org/10.1021/jacs.2c11420.

17. K. Choudhary, T. N. Yildirim, A. G. K. Daniel Siderius, A. McDannald, and D. Ortiz-Montalvo, "Graph Neural Network Predictions of Metal Organic Framework CO_2 Adsorption Properties," *Computational Materials Science* 210 (2022): 111388, https://doi.org/10.1016/j.commatsci.2022.111388.

18. L. Li and H. Yu, "A Multi-Scale Graph Neural Network for the Prediction of Multi-Component Gas Adsorption," *Engineering* 52 (2025): 102–111, https://doi.org/10.1016/j.eng.2025.08.012.

19. M. Jalali, A. D. D. Wonanke, and C. Wöll, "MOFGalaxyNet: A Social Network Analysis for Predicting Guest Accessibility in Metal–Organic Frameworks Utilizing Graph Convolutional Networks," *Journal of Cheminformatics* 15 (2023): 94, https://doi.org/10.1186/s13321-023-00764-2.

20. M. Jalali, M. Tsotsalas, and C. Wöll, "MOFSocialNet: Exploiting Metal–Organic Framework Relationships via Social Network Analysis," *Nanomaterials* 12, no. 4 (2022): 704, https://doi.org/10.3390/nano12040704.

21. A. S. Rosen, S. M. Iyer, D. Ray, et al., "A Quantum-Chemical Database of Metal-Organic Frameworks," *Matter* 4, no. 3 (2021): 915–924, https://doi.org/10.1016/j.matt.2021.02.015.

22. D. A. Spielman and N. Srivastava, "Spectral Sparsification of Graphs," *SIAM Journal on Computing* 40, no. 6 (2011): 1913–1926, https://doi.org/10.1137/080734029.

23. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems* 27 (2014): 2672–2680.

24. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in Advances in Neural Information Processing Systems, Vol. 30 (2017).

25. L. Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32, https://doi.org/10.1023/A:1010933404324.

26. S. P. Ong, W. D. Richards, A. Jain, et al., "Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis," *Computational Materials Science* 68 (2013): 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028.

27. A. H. Larsen, J. J. Mortensen, J. Blomqvist, et al., "The Atomic Simulation Environment—A Python Library for Working With Atoms," *Journal of Physics: Condensed Matter* 29, no. 27 (2017): 273002.

28. M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable Fidelity and Diversity Metrics for Generative Models," in Proceedings of the 37th International Conference on Machine Learning, Vol. 119 (PMLR (Proceedings of Machine Learning Research), 2020), 7176–7185.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29. A. F. T. Martins and R. F. Astudillo, "From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification," in Proceedings of the 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research (PMLR, 2016), 1614–1623.

30. Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE, 2019), 1429–1437.

31. T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in International Conference on Learning Representations (OpenReview.net, 2018).

32. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in Advances in Neural Information Processing Systems, Vol. 30 (Curran Associates, 2017): pp. 1024–1034.

33. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in International Conference on Learning Representations (OpenReview.net, 2018).

34. T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in International Conference on Learning Representations (OpenReview.net, 2017).

35. K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful Are Graph Neural Networks?," in International Conference on Learning Representations (OpenReview.net, 2019).

36. M. Fey and J. E. Lenssen, "Fast Graph Representation Learning With Pytorch Geometric," in ICLR Workshop on Representation Learning on Graphs and Manifolds (2019), arXiv:1903.02428.

37. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Machine Learning* 63, no. 1 (2006): 3–42, https://doi.org/10.1007/s10994-006-6226-1.

38. G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531 (2015).

39. S. M. Moosavi, A. Nandy, K. M. Jablonka, et al., "Understanding the Diversity of the Metal-Organic Framework Ecosystem," *Nature Communications* 11 (2020): 4068, https://doi.org/10.1038/s41467-020-17755-8.

40. B. Hagan, I. Shah, and G. Patlewicz, "Can Graph Similarity Metrics Be Helpful for Analogue Identification as Part of a Read-Across Approach?," *Computational Toxicology* 34 (2025): 100353, https://doi.org/10.1016/j.comtox.2025.100353.

41. T. Xie, X. Fu, O. E. Ganea, R. Barzilay, and T. S. J. Crystal, "Crystal Diffusion Variational Autoencoder for Periodic Material Generation," in International Conference on Learning Representations (OpenReview.net, 2022).

42. S. Kim, G. Jeon, S. Hwang, et al., "Are Diffusion Models Ready for Materials Discovery in Unexplored Chemical Space?," *Patterns* 7, no. 6 (2026): 101537, https://doi.org/10.1016/j.patter.2026.101537.

43. F. E. Kelvinius, O. B. Andersson, A. S. Parackal, D. Qian, R. Armiento, and F. Lindsten, "Wyckoffdiff – a Generative Diffusion Model for Crystal Symmetry," in Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research (PMLR, 2025), 15130–15147.

44. C. K. Joshi, X. Fu, Y.-L. Liao, et al., "All-Atom Diffusion Transformers: Unified Generative Modelling of Molecules and Materials," in Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research (PMLR, 2025), 28393–28417.

45. L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data Using Conditional GAN," in Advances in Neural Information Processing Systems, Vol. 32 (Curran Associates, 2019): 7335–7345.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70639 by Karlsruher Institut Fur Technologie, Wiley Online Library on [18/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License