REPRESENTATIONS FROM PRETRAINED MACHINE-LEARNING INTERATOMIC POTENTIALS AS COARSE COORDINATES FOR MATERIAL GENERATION AND EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative machine learning is increasingly tested for inorganic materials generation. Most models and the corresponding evaluation approaches thereby rely on simple forms of crystal structure representations. In this paper, we showcase the power of integrating atom-averaged features from pretrained Machine-Learning Interatomic Potentials (MLIPs), such as MACE, into these tasks. We first introduce the Coarse-Fine Transport Distance (CFTD), a distance measure that captures both quality and novelty of materials created with generative models in a single distribution-based evaluation framework. It uses two different featurizers, directly deriving the quality component from coarse MACE representations. We showcase CFTD's versatility in capturing crystal structure quality while also detecting memorization and compare it with the recently introduced continuous SUN metrics. We further illustrate the power of MLIP representations by directly adding them to the generation process itself and show that coarse MACE features can be used as helpful guidance for a material generative model.

1 INTRODUCTION

Materials are a foundation of technological progress, and finding new ones with desired properties is crucial for future developments across many fields. Inorganic crystals form a specific kind of material and underpin many devices and applications, but because the space of possible crystal structures is immense, identifying suitable novel candidates is difficult, time-consuming, and expensive. With recent advancements in generative machine learning, various models for creating inorganic crystal structures from scratch have emerged (De Breuck et al., 2025). Multiple publications (e.g., Zeni et al. (2025); Xie et al. (2022); Wu et al. (2026); Ekström Kelvinius et al. (2025)) have proposed architectures with the potential to generate realistic materials not present in the training data, and new models are published frequently.

Conversely, actual synthesis of crystals originating from generative models remains rare, and the novelty of such discoveries has been heavily debated (Juelsholt, 2026). Additionally, and perhaps related, metric development for evaluating materials generative models seems harder than in other domains: in particular, *novelty* and *stability* are crucial for model design and evaluation. Determining these two characteristics is not straightforward, and, in the absence of alternatives, most material generators are evaluated using SUN-metrics (Zeni et al., 2025). This now-established evaluation criterion measures stability, uniqueness, and novelty at the instance level (structure-by-structure), which would allow, e.g., a highly collapsed generative model to still produce novel and stable materials. Since generative models are usually trained to match the training data distribution, distributional metrics offer a complementary approach to evaluation.

Other evaluation criteria target specific properties such as atom or space-group diversity or the volume of the generated crystals (e.g., in Xie et al. (2022); Betala et al. (2026)). These are more distributional but yield a reliable picture of performance only when combined. The recently introduced Transport Novelty Distance (TNovD) (Hagemann et al., 2025) also has a distributional flavor and uses optimal transport (OT) (Peyré & Cuturi, 2020) and a contrastive GNN featurizer to evaluate novelty and quality in a single framework. We adapted their approach and extended this metric

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

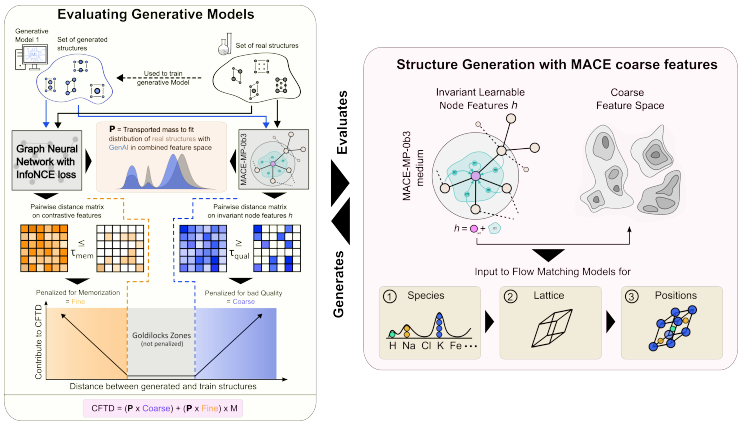


Figure 1: Overview of the CFTD metric and the generative pipeline. We propose using random projections of mean-pooled, invariant h features from the message-passing step of a pretrained MACE model for both evaluation and generation, a practice vital for materials design.

because, with a single featurizer, every improvement in quality necessarily reduces novelty, and vice versa. To overcome this disadvantage, we propose an addition inspired by the Fréchet Inception Distance (FID) (Heusel et al., 2017). In their paper, the authors argued that an image classifier also provides a good feature representation of images. A possible analog for materials are the hidden features of pre-trained Machine-Learning Interatomic Potentials (MLIPs), such as MACE (Batatia et al., 2025), which learn energetic representations of materials from atomic environments.

We used these representations to better evaluate the quality of generated structures and overcome the described drawback of TNovD. We did this by establishing a Goldilocks zone between a memorization and a quality regime that includes structures not considered for evaluation at all. The two separated regimes are established through two different featurizers: a MACE featurizer for quality and a contrastive GNN featurizer for novelty evaluation. Since each featurizer captures either coarse or fine information about the investigated structure, we call the new metric Coarse-Fine Transport distance (CFTD), borrowing terminology from the vision community, where coarse and fine features capture different levels of visual detail. We discuss how the classical coverage-novelty tradeoff relates to our new metric and, by replicating toy experiments from Hagemann et al. (2025), show that including an MLIP featurizer makes CFTD chemically more sensible and captures stability much better than TNovD. To investigate the differences to more instance-based metrics, e.g., continuous SUN (Negishi et al., 2026), we added experiments that resemble mode-collapsing generative models. In such cases, our newly introduced CFTD behaves complementarily to instance-based metrics, which do not capture mode collapse.

We then use the CFTD to benchmark sets of 10,000 generated crystal structures, each created with a different materials generative model (Zeni et al., 2025; Jiao et al., 2023; 2024; Xie et al., 2022; Joshi et al., 2025; Park et al., 2025; Park & Walsh, 2025). We benchmark before and after relaxing the resulting crystal structures, and we discuss qualitative behavior in comparison to Negishi et al. (2026). In particular, we observe that relaxing the generated structures with a pre-trained MLIP generally improves model quality but often also increases memorization. This suggests that, to some extent, generative models push existing materials slightly away from their energy minima, rather than generating truly novel and stable structures.

In the Appendix, we show additional experiments on the geometries of the fine contrastive featurizer and the coarse MACE featurizer (Appendix D.2) and investigate which informational content the MACE features possess (Appendix D). The results indicate that the MLIP representations carry structural, chemical, and physical information, even in compressed form.

These results led us to further investigate a possible application of hidden MACE features in the generation process itself. Thereto, we created an MLIP-conditional material generative model and trained it by including empirically sampled hidden MACE representations from the training set as an additional conditioning signal. The model overfits when trained on only a small subset of the training set, a behavior that is detected by the CFTD. We then trained our generative model with

and without MACE feature conditioning on the full training set and evaluated both models with the CFTD and the recently introduced LeMat-GenBench benchmark (Betala et al., 2026). The results show that the architecture benefits from MACE conditioning, as the generated materials obey the relation to their condition. The MACE-conditional model exhibits a very different stability-novelty profile, further confirming our thesis. Fig. 1 provides a detailed overview of our idea on how to include MACE representations into the generation and the evaluation process.

1.1 Related Work

Evaluation of generative models In imaging, early work used precision-recall (Simon et al., 2019; Cheema & Urner, 2023) or nearest-neighbor checks. Then, feature-based metrics like LPIPS (Zhang et al., 2018) for single images or FID (Heusel et al., 2017) became more popular because they better align with human visual judgments.

For materials generative model evaluation, the most common methods are the SUN metrics, which, to the best of our knowledge, were first introduced in Zeni et al. (2025). These have been refined recently in Negishi et al. (2025), which treats stability, uniqueness, and novelty not as binary objectives but as continuous ones. Further, Ekström Kelvinius et al. (2025) proposes an FID-style evaluation in the activation space of a GNN, which is close in spirit but does not capture novelty or stability. In this paper, we follow up on Hagemann et al. (2025), which introduced the first distribution-based quality-novelty metric. We refine their metric with a dual-featurizer framework to avoid the otherwise opposing behavior of novelty and quality. We can show that these changes enable the new metric to capture stability much better than Hagemann et al. (2025) and mode collapse much better than Negishi et al. (2025).

MLIP-conditional generation Li & Walsh (2025) make a strong case for using the hidden feature space of pretrained foundation models for conditioning generative models, instead of generating in the raw materials space: they show that different MLIPs seem to converge to the same representation, once anchored. This is consistent with a commentary in Didi (2025), which advocates using feature-based representations for generative modeling, as is common in the vision community (Oquab et al., 2024; Rombach et al., 2022).

Le et al. (2025) trained an encoder-decoder framework to conditionally generate ligands in structure-based drug design, and their encoder is partly learned from training representations. In Li et al. (2025), they learn a conditional/hierarchical model for generating organic molecules based on the UniMol representation. This is similar to what we do in practice, but their conditioning focuses on semantic meaning, whereas ours uses hidden features from the MACE foundation model. Li et al. (2026) use MLIP force fields in a reinforcement learning-style guidance procedure, highlighting their effectiveness for generating molecular structures. Pinede et al. (2025) also use MLIP representations by regularizing the Boltzmann generators' hidden states to match a pretrained MACE model. Recently, Hessmann et al. (2026) draw an analogy between the denoising behavior of diffusion (Song et al., 2021) and MLIP relaxation by learning a pseudo force field whose pseudo forces can be used in the score function of the diffusion process.

1.2 Contributions

Our main contribution is integrating MLIP representations (specifically, hidden features of the MACE foundation model) to guide and evaluate inorganic material generators. Our evaluation metric is inspired by hierarchical image generation, where a coarse-to-fine approach is often pursued (e.g., Kadkhodaie et al. (2023); Tian et al. (2024)). We show that a single metric that captures both novelty and quality requires a dual-featurizer geometry, in which each component measures a distinct property. For the coarse component (capturing material quality), we introduce specific MACE representations and show that they also help to guide a material generator. In detail, we contribute as follows:

- We introduce the Coarse-Fine Transport Distance (CFTD), which extends the recently published TNovD (Hagemann et al., 2025). It resolves the quality-novelty trade-off intrinsic to metrics that capture both with a single featurizer, and we relate this to classical coverage and novelty.

- For the coarse (or quality) component, we take random projections of pooled hidden MACE features and investigate what information these features capture. For the fine component (capturing the materials novelty), we build upon the contrastive GNN from Hagemann et al. (2025) and refine it using substitutions with chemically similar elements (Glawe et al., 2016).
- We show that, in contrast to cSUN (Negishi et al., 2025), the CFTD reliably tracks stability and mode collapse, respectively. Using toy experiments, we verify that it detects quality errors and memorization, and show that it provides a better representation of stability and chemical composition than TNovD (Hagemann et al., 2025).
- We benchmark different material generators using CFTD, and show that it captures memorization in an overfitted material generator.
- We show that a material generative model benefits from integrating MACE features and that these features can serve as a guidance signal.

2 The Coarse-Fine Transport Distance

In this section, we review established methods for measuring generative-model performance and motivate our proposed Coarse-Fine Transport Distance (CFTD). We first discuss more classical coverage and novelty metrics and how they relate to the quality and novelty of structures. We then extend the TNovD (Hagemann et al., 2025) by three things: we argue that novelty and coverage should be judged in two different geometries, suggest the use of MACE features for capturing stability, and introduce a "Goldilocks" zone of no penalization. Note that when we talk about stability in this work, we mean *implicit* stability: stability-related information encoded in the hidden MACE features introduced below, rather than an explicit thermodynamic quantity such as energy above hull. Our perturbation experiments and correlations with polymorph formation energies in Appendix D.1 indeed suggest that they are a useful proxy. Direct stability measures, such as energy above hull, are complementary but often prohibitively costly.

The classical Coverage-Novelty tradeoff One of the main issues in generative modeling is finding a model that captures the statistics of the data distribution without overfitting to specific training examples. In essence, we want to generate *novel* samples that still resemble the features from the data distribution. Classically, this is often summed up in the terms *coverage* and *novelty*, which we define in Appendix B. In the general case, coverage describes the amount of training data for which a close generated sample can be found, while novelty describes the amount of generated samples for which no close training sample exists. These two points conflict when measured under the same notion of closeness. In materials generation, we are especially interested in coverage with respect to material quality: whether a training structure has a close generated sample in a feature space that reflects whether a material is approximately valid. We refer to this as quality coverage. Novelty refers to structural and compositional resemblance of a generated structure to the training data. These two things do not necessarily need to be measured in the same reference space (i.e., under the same notion of closeness), and we therefore resolve the classical tradeoff by using two featurizers that decouple quality coverage from novelty, allowing us to evaluate both at the same time.

To measure novelty, we rely on the identity featurizer F_{id} in Hagemann et al. (2025) and use their GNN trained with an InfoNCE loss for contrastive learning. This model uses positive pairs of augmented samples to learn a feature space where similar structures are close, while the distance to dissimilar structures is maximized. Appendix C provides the specific augmentation procedure and justifies the loss function. To resolve the novelty-quality tradeoff of TNovD, we introduce a second featurizer F_{mace} to explicitly capture a structure's quality parameters. It is directly derived from MACE-MP-0b3 "medium", an MLIP foundation model (Batatia et al., 2022b;a; 2025) that captures stability information. Instead of using MACE's final output, we use layer statistics that work similarly to FID (Heusel et al., 2017) and are independent of the number of atoms in the structure. More specifically, let $h^k = (h^k_1, ..., h^k_{n_{atoms}})$ be the hidden states of its k-th interaction block, which are of size (n_{atoms}, h_{dim}). We apply mean pooling to obtain features of the same size, regardless of the number of atoms in a structure, and then concatenate the resulting layers. This yields $H = (\mathrm{mean}(h^1), ..., \mathrm{mean}(h^L))$, of which we take random projections to reduce dimensionality. F_{mace} can be seen as an indirect measure of chemical and physical quality, which we show in Appendix D, together with a detailed description of the featurizer.

Algorithm 1 The Coarse-Fine Transport Distance

Input: Empirical measures $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\nu = \frac{1}{m}\sum_{j=1}^{m}\delta_{g_j}$, validation/calibration set, featurizers F_{id}, F_{mace}, parameter β.

Step 1: Calibration.

1. Compute the optimal transport plan between the arithmetic mean of the identity and the MACE representation distances of train and validation sets.
2. Choose τ_{mem} and τ_{qual} so that $\frac{1-\beta}{2}$ of the coupled mass lies in the memorization and low-quality regimes, respectively. Choose M so that both error terms are balanced; see Appendix F.

Step 2: Evaluation.

$$C_{i,j}^{id} = \|F_{id}(x_i) - F_{id}(g_j)\|, \quad C_{i,j}^{mace} = \|F_{mace}(x_i) - F_{mace}(g_j)\|, \quad C_{i,j} = \tfrac{1}{2}C_{i,j}^{id} + \tfrac{1}{2}C_{i,j}^{mace}.$$

Calculate the optimal transport plan $\pi(\mu,\nu)$ for train and generated samples with respect to the pairwise distances $C_{i,j}$. Define

$$\mathrm{CFTD}(\mu,\nu) = \sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{i,j}(\mu,\nu)\big[M\,\mathrm{ReLU}(\tau_{mem} - C_{i,j}^{id}) + \mathrm{ReLU}(C_{i,j}^{mace} - \tau_{qual})\big].$$

The CFTD algorithm Given training samples $(x_i)_{i=1}^n$ and generated samples $(g_i)_{i=1}^m$, we compute the optimal transport coupling using the arithmetic mean of the identity and the MACE representation distances as the transport cost. Together with the defined thresholds τ_{mem} and τ_{qual}, this coupling is used to measure whether a pair of "matched" generated and training data is counted towards the memorization error, the quality error, or if it is not penalized at all. For typical samples, we implemented a "Goldilocks" zone separating novelty and quality coverage into two distinct regimes. This zone defines the percentage of structures that are not penalized in a reasonable train split. We call this proportion β. We then proceed with the steps described in 1. We first determine τ_{mem} and τ_{qual}, the two thresholds for penalizing memorized or low-quality data, and choose a scale M that acts as a tuning knob balancing the two parts. We then calculate the optimal transport plan for a weighted average of both featurizers and use it to penalize coupled generated and training set pairs by assigning a cost to generated samples that are either too close in the identity feature space (below τ_{mem}) or too far away in the MACE feature space (above τ_{qual}). This procedure allows us to use different notions of closeness to detect memorization (fine) and low quality (coarse) in the generated samples, resolving the novelty-quality coverage tradeoff. In the final metric, lower CFTD values indicate better generative models.

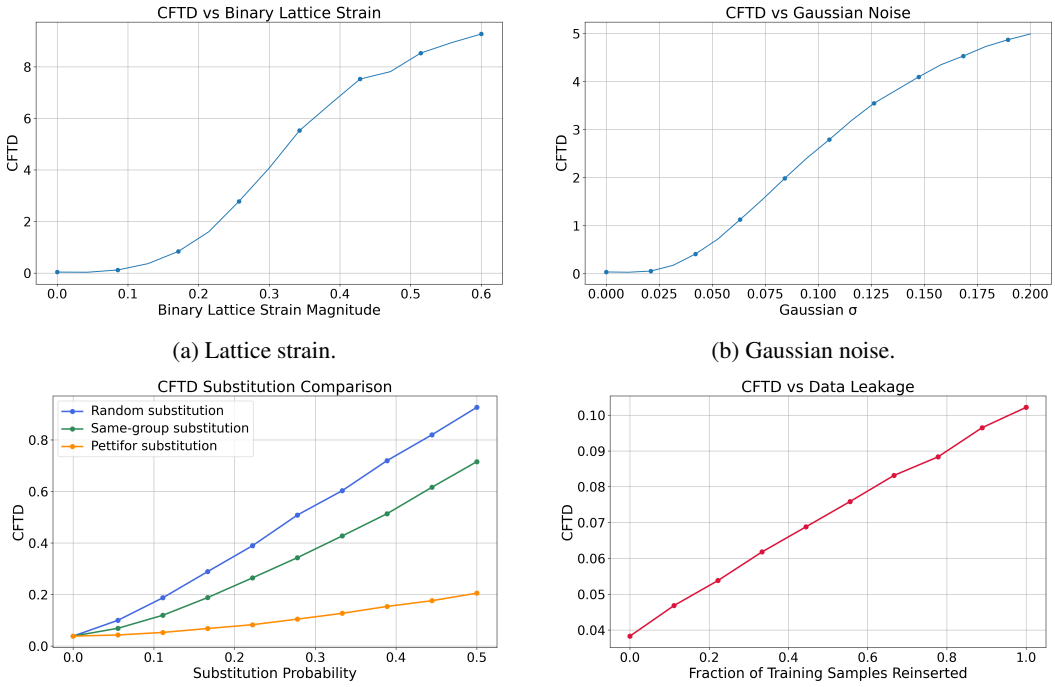
At heart, the metric remains distributional, but the optimal transport coupling now accounts for two distinct featurizers: novelty and quality coverage. Informally, quality coverage (Cov) and novelty (Nov) approach 1 as the CFTD vanishes. We state this informally here, with the full statement in Appendix B.2. This means that CFTD penalizes mass outside the Goldilocks zone.

Proposition 2.1. For suitable radii $r_{\mathrm{qual}} > \tau_{\mathrm{qual}}$ and $r_{\mathrm{mem}} < \tau_{\mathrm{mem}}$, we have in the feature space that $\mathrm{Nov}_{r_{mem}} \to 1$ and $\mathrm{Cov}_{r_{qual}} \to 1$ as the $\mathrm{CFTD} \to 0$ under mild conditions on the coupling.

3 CFTD EXPERIMENTS

Quality Coverage-Novelty Checks To show that the CFTD can indeed detect issues of generative models with regard to novelty and quality coverage, we repeat some experiments from Hagemann et al. (2025). For TNovD, the contrastive GNN did not heavily penalize unstable materials, as even strong structural distortions resulted in only small increases (Hagemann et al., 2025). To test this for the CFTD, we strain the lattice, add Gaussian noise to the fractional coordinates, substitute atomic sites with neighbors on the modified Pettifor scale (Glawe et al., 2016), elements of the same group, and random elements, and iteratively replace an increasing number of test-set data points with training data. Fig. 2 shows the behavior of these tests.

We observe that Gaussian noise and the diagonal lattice strain both (mainly) increase the CFTD for large values. This follows the expected behavior, since both experiments create structures with increasingly poorer quality. Fig. 2 (d) shows the results of iteratively replacing more and more test-set structures with structures from the train set. This simulates the memorization of structures,



(a) Lattice strain.

(b) Gaussian noise.

(c) Substitutions.

(d) Train leakage.

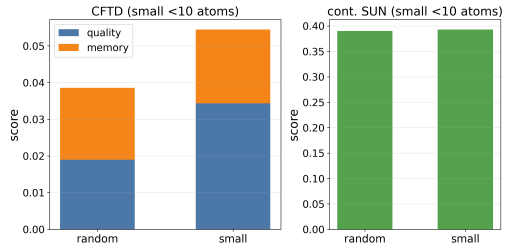
Figure 2: CFTD behavior of the MP20 train and test set across different perturbation settings of the test set.

and the results show the expected increase in the CFTD. Interestingly, the CFTD is very sensitive to atom substitutions (Fig. 2 (c)). Random substitutions are penalized more strongly than in-group substitutions, which, in turn, are more strongly penalized than modified Pettifor substitutions. This is reasonable, as structures created with Pettifor substitutions are expected to be of higher quality than structures created with in-group or random substitutions. This quality gain outweighs the increased memorization in structures with Pettifor substitutions, which comes from including modified Pettifor neighbors in the contrastive GNN's augment function (see Appendix C). The Pettifor substitutions are penalized more strongly in the memorization component, which is not seen in the final CFTD value of Fig. 2 (c), due to the higher quality of the structures. This shows that, with this change, the CFTD can detect chemical similarity very well. Further, our approach seems to penalize Gaussian noise and lattice strain more heavily than TNovD, indicating a better treatment of stability.

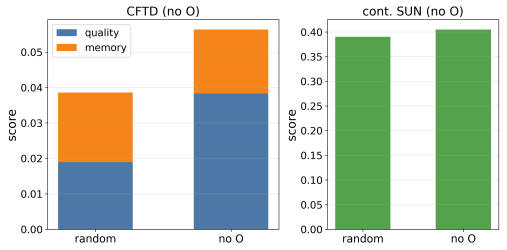
Although the experiments yield highly divergent CFTD values that prevent straightforward interpretation, the results in Fig. 2 show that our metric indeed penalizes problematic structures and data leakage. We also repeat the same experiment with the Perov-5 dataset (Castelli et al., 2012a;b; Xie et al., 2022) in the Appendix I and reach similar conclusions.

Comparison with continuous SUN Because our CFTD bears similarities to the continuous SUN metrics (Negishi et al., 2026), we propose a small toy experiment to highlight their differences. We use their implementation in xtalmet (Negishi et al., 2025) and create three subsets of 1,000 structures each. Two subsets were created by randomly drawing 1000 structures from the MP20 train and validation sets, respectively. For the third subset, only structures from the validation set with fewer than ten atoms were considered. In the toy experiment, the first subset was treated as the training set, while the other two simulated the generated structures for which the CFTD and cSUN metrics were calculated. This experiment artificially introduces a strong distributional shift between generated and training data, since the third subset covers only a specific part of the full distribution of structures (only those with at most ten atoms). As shown in Fig. 3 (a), the resulting cSUN values barely differ between subsets two and three. This is not surprising, since, at an instance level, there is no reason to assume that materials with fewer than ten atoms are less unique or novel than those with

more atoms. Nevertheless, if a model trained on the complete dataset generates only structures with fewer than ten atoms, it would be considered mode-collapsed. The CFTD penalizes memorization and quality by optimizing the transport plan between the training and generated distributions. It is therefore primed to detect the artificially induced distributional shift, and, as indicated by the higher CFTD values in Fig. 3 (a), it indeed does so. This shows that CFTD complements existing metrics by clearly detecting distributional shifts between training and generated samples. Since the training objective of generative models is to create samples that fit the training distribution without exactly mirroring it, this helps to identify potentially problematic models that score well in other metrics. To simulate a different kind of distributional shift, we perform a second toy experiment by creating a subset of 1,000 randomly drawn structures from the validation set that do not contain oxygen. As shown by the results in Fig. 3 (b), we observe the same behavior.

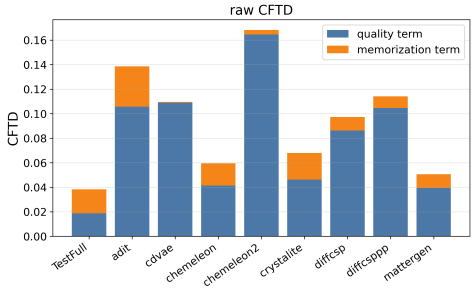


(a) Comparison of a reference set versus a subset with fewer than 10 atoms.

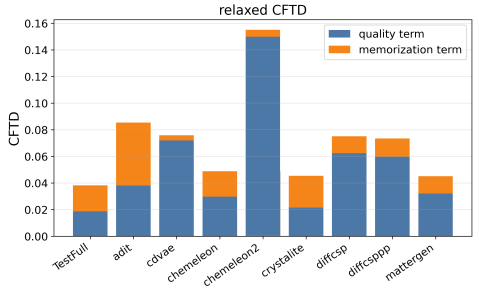


(b) Comparison of a reference set versus a subset without oxygen.

Figure 3: Comparison of continuous SUN versus CFTD.



(a) CFTD for unrelaxed structures from different generative models.



(b) CFTD for relaxed structures from different generative models.

Figure 4: Comparison of common material generative models: relaxed and unrelaxed.

Evaluation of current material generators We also benchmark current material generators with the CFTD and samples provided in Negishi et al. (2026). The same MP20 dataset was used to train the generators and our contrastive GNN, enabling us to use the generated structures directly without retraining the identity featurizer.

We follow our proposed hyperparameter fitting procedure (see Appendix F) using the MP20 train and validation sets and provide evaluations of the materials generative models MatterGen (Zeni et al., 2025), ADiT (Joshi et al., 2025), Chemeleon (Park et al., 2025), Chemeleon2 (Park & Walsh, 2025), CdVAE (Xie et al., 2022), DiffCSP (Jiao et al., 2023), DiffCSP++ (Jiao et al., 2024), Crystalite (Veljković et al., 2026), as well as the MP20 test set.

To investigate the influence of relaxation (i.e., a structural optimization of the atomic positions and unit cell parameters), we relaxed all generated structures with the MACE-MP-0b3 foundation model "medium" for 100 steps and 0.02 Å. As shown in Fig. 4, relaxed structures from the generative models achieve lower CFTD values (i.e., can be considered better) but display a higher memorization component. Our metric consistently ranks the MP20 test-set structures as best, with MatterGen second. The Chemeleon and Crystalite models also perform well, albeit with a higher memorization

term than MatterGen. In particular, Crystalite and MatterGen are very close in the relaxed case and mainly differ in their novelty-quality trade-off.

Compared with the evaluation in Negishi et al. (2026), the biggest difference seems to be the performance of Chemeleon2. It is ranked highest in Negishi et al. (2026) but scores the lowest on our metric. As a reinforcement-learning-based model, it deliberately overweights rare events, thereby shifting the generated distribution away from the training distribution. This is consistent with analysis in Negishi et al. (2026), where they write that Chemeleon2 employs "reward hacking" under their cSUN. We hypothesize that this is caused by overconcentration, which is penalized by a distributional metric like CFTD, but mostly ignored by an *instance-level* one.

4 GUIDING A MATERIAL GENERATOR USING MLIP FEATURES

In addition to evaluating generative models, we investigated whether the coarse MACE features carry a usable signal for generation itself and therefore designed a MACE-conditional generator. Our modeling ideas build on recent work on material generators, and we do not aim to propose a SOTA generator. Instead, we want to verify two things: that a generator can use MACE representations, and that it faithfully obeys them as a condition, i.e., generates materials that respect the conditional input. Both of these are necessary conditions for a *hierarchical* generator, which we briefly describe in Appendix M.

In the following, we describe a material as a tuple $(n, (a_i)_{i=1}^n, L, (x_i)_{i=1}^n, h)$, where n describes the number of atoms, a_i describes the (categorical) atom types, L the lattice, and (x_i) the fractional coordinates of atom positions. Compared with existing generative models, we include an additional parameter h, which corresponds to our coarse mace features. We choose the factorization

$$p(n, h, L, a, x) = p(n, h)p(a|n, h)p(L|n, h, a)p(x|a, n, h, L),$$

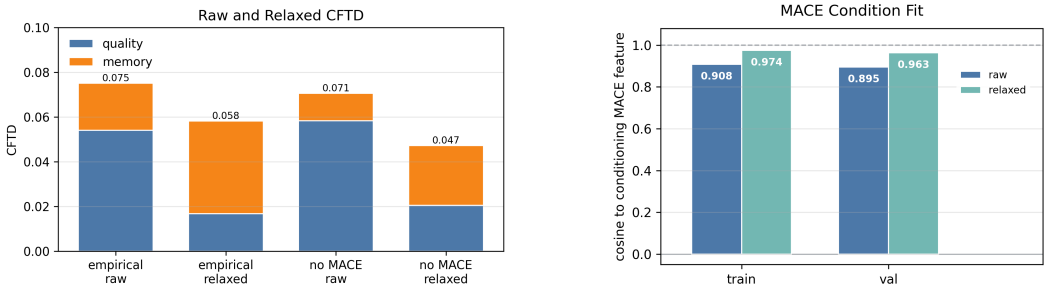
This means that, based on the MACE representations and the number of atoms, we generate the atom types, then the lattice, and then the positions. In particular, we learn all the conditional probabilities $p(a|n, h)$, $p(L|a, n, h)$ and $p(x|L, a, n, h)$ using the means of flow matching (Albergo & Vanden-Eijnden, 2023; Lipman et al., 2023; Liu et al., 2023; Gat et al., 2024). Given the variable types, we use continuous flow matching for position and lattice, and discrete flow matching for atom types. Details are provided in Appendix K. We verify in Appendix L that this model indeed can overfit on small data sizes (2000 samples) and that this is detected by the CFTD.

To investigate whether coarse MACE features can guide a material generative model, we tested if adding empirical MACE features improves performance on LeMat-GenBench (Betala et al., 2026), i.e., whether the guidance provides a useful downstream signal. More precisely, we compare the MACE-conditional model from above against an otherwise identical model that does not use MACE features. We trained both models for 1,000 epochs and saved the model with the lowest validation loss. In the MACE-conditional model, generation is initialized by sampling the pair (n,h), where (n) is the number of atoms in the unit cell and (h) is the corresponding pooled MACE feature vector projection. In the empirical variant considered here, we use the empirical distribution of the MP20 train set as the prior over these conditions, i.e., $p(n, h) = \frac{1}{N}\sum_{i=1}^{N} \delta_{(n_i,h_i)}$. The remaining components of the generator then sample atom types, lattice, and fractional coordinates conditional on this sampled pair.

Since the conditions are taken directly from the empirical distribution of the training set, one may expect increased memorization. This can potentially be improved by generating the MACE feature conditions as well, and we discuss possible updates of the generator in Appendix M. At the same time, because MACE feature representations encode structural, chemical, and physical information, they may guide the generator toward improved (meta)stability. To investigate such potential upsides, we compare the conditional and unconditional models using both CFTD and LeMat-GenBench metrics. Furthermore, we check whether the generated materials indeed comply with the MACE feature conditioning. This means that when we apply the MACE featurizer to the generated samples, we check whether the resulting features match the condition we fed to the model.

In Fig. 5 (a), we see that the empirical MACE-conditional model indeed exhibits higher memorization but better quality coverage (i.e., a lower quality value). The cosine distances in Fig. 5 (b) indicate a very high degree of agreement between the features of the generated materials and the

conditions, as they barely change after relaxing the structures to their energetic ground state with MACE-MP-0b3.



(a) CFTD for the final models (empirical and none), raw and relaxed.

(b) Cosine similarity for the empirical distribution between generated materials and condition.

Figure 5: CFTD of our generative models. Cosine similarities are for materials generated conditioned on empirical features, and they are tested to verify that they satisfy the condition.

On LeMat-GenBench, the empirical MACE-conditioned model improved validity (V), stability (S), and metastability (M) over the No-MACE baseline ($V = 98.5$ vs. 97.2, $S = 11.7$ vs. 7.7, $M = 51.6$ vs. 44.4), while the No-MACE model attains slightly higher novelty ($N = 61.0$ vs. 51.2). Both models have similar SUN and MSUN profiles; the empirical MACE-conditioned model has SUN 0.8, No-MACE has 0.5. For MSUN, the empirical MACE-conditioned model reaches 16.7, No-MACE reaches 18.0. This implies that the empirical MACE-conditioned model does precisely what we expected: guide the models towards more stable but less novel samples (since we fed in MACE features from the training set). It also shows high potential for learning a novel MACE feature space that combines stability improvements with greater novelty.

Note that while the CFTD evaluation of the MACE-conditioned model shares the same underlying MACE representation, LeMat-GenBench uses a more thorough pre-relaxation and evaluation procedure with three MLIPs.

5 Conclusions

In this paper, we introduced a dual featurizer framework for evaluating material generative models, extending existing distribution-based metrics. We related our featurizers to the more classical terms of coverage and novelty and showed that the resulting Coarse-Fine Transport Distance behaves consistently across different toy experiments. We used it to benchmark existing material generators and showed that it can detect mode collapse, indicating a more distributional flavor than the cSUN metrics. We also showed that our metric detects overfitting, making it a desirable model checkpointing signal. However, this needs further investigation, especially compared with other checkpointing strategies, such as validation loss or MSUN, as in Veljković et al. (2026).

We further investigated the utility of using MACE representations beyond generative model evaluation. We show that they inform both composition and structure and can be used to guide a conditional material generative model. This opens an interesting line of research for a *hierarchical* model that first generates coarse MACE features and then builds a generative model for structure creation on top of them. Another benefit of conditioning a generative model on MLIP features is that it opens a route to property-conditional materials generation: such a feature generator could be steered toward regions in the MLIP feature space corresponding to structures with the desired property, yielding a property-conditional generator without retraining a separate conditional model for each target. For the *hierarchical* model, one needs to investigate the learnability-identifiability tradeoff, i.e., how well our choices can inform the generator and how easy it is to learn the MACE feature space. Other summary statistics might be viable, such as taking atom-level features or the max or standard deviation instead of pooling. This offers a new way to design material generators, which we believe is a fruitful area for future research. We discuss potential extensions to conditional generation in Appendix M.

AI USE STATEMENT

We used generative AI tools in this work for coding assistance, proofreading and discussion of (experimental) results and ideation, generation of mathematical arguments and code prototyping. All LLM-generated code and mathematical arguments were reviewed, verified, and tested for correctness. We take responsibility for the final content of this work, including code produced with the aid of generative AI.

ETHICS STATEMENT

Our paper submission does not raise concerns regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the source code for the Coarse-Fine Novelty Distance and the MACE-conditioned generative models, as well as the code for the most important toy experiments is submitted as supplemental material. Multiple experiments that back our findings are provided in the Appendix. All data necessary for reproduction is open source and the links for downloading are included in the code. The Appendix also contains the mathematical proofs for claims made in the main text.

REFERENCES

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

Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e(3)-equivariant atom-centered interatomic potentials. *arXiv preprint arXiv:2205.06643*, 2022a.

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

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas Van Der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, November 2025. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0297006.

Siddharth Betala, Samuel P. Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K. Joshi, Sékou-Oumar Kaba, Félix Therrien,

Alex Hernandez-Garcia, Rocío Mercado, N. M. Anoop Krishnan, and Alexandre Duval. Lematgenbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2026.

Ivano E. Castelli, David D. Landis, Kristian S. Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F. Jaramillo, and Karsten W. Jacobsen. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy Environ. Sci.*, 5:9034–9043, 2012a. doi: 10.1039/C2EE22341D. URL http://dx.doi.org/10.1039/C2EE22341D.

Ivano E. Castelli, Thomas Olsen, Soumendu Datta, David D. Landis, Søren Dahl, Kristian S. Thygesen, and Karsten W. Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.*, 5:5814–5819, 2012b. doi: 10.1039/C1EE02717D. URL http://dx.doi.org/10.1039/C1EE02717D.

Fasil Cheema and Ruth Urner. Precision recall cover: A method for assessing generative models. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6571–6594. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/cheema23a.html.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching, 2025. URL http://jmlr.org/papers/v26/24-0586.html.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Pierre-Paul De Breuck, Hai-Chen Wang, Gian-Marco Rignanese, Silvana Botti, and Miguel A. L. Marques. Generative AI for crystal structures: A review. *npj Computational Materials*, 11(1): 370, December 2025. ISSN 2057-3960. doi: 10.1038/s41524-025-01881-2.

Kieran Didi. The unification of representation learning and generative modelling, 2025. URL https://kdidi.netlify.app/blog/ml/2025-12-31-r4g/.

Filip Ekström Kelvinius, Oskar B. Andersson, Abhijith S Parackal, Dong Qian, Rickard Armiento, and Fredrik Lindsten. WyckoffDiff – a generative diffusion model for crystal symmetry. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15130–15147. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ekstrom-kelvinius25a.html.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=GTDKo3Sv9p.

Samuel Gerber and Mauro Maggioni. Multiscale strategies for computing optimal transport. *Journal of Machine Learning Research*, 18(72):1–32, 2017. URL http://jmlr.org/papers/v18/16-108.html.

Henning Glawe, Antonio Sanna, E K U Gross, and Miguel A L Marques. The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining. *New Journal of Physics*, 18(9):093011, sep 2016. doi: 10.1088/1367-2630/18/9/093011. URL https://doi.org/10.1088/1367-2630/18/9/093011.

Paul Hagemann, Simon Müller, Janine George, and Philipp Benner. Transport novelty distance: A distributional metric for evaluating material generative models. *arXiv preprint arXiv:2512.09514*, 2025.

Stefaan Simon Pierre Hessmann, Khaled Kahouli, Stefan Gugler, Michael Plainer, Frank Noé, Klaus-Robert Müller, and Niklas Wolf Andreas Gebauer. Generative pseudo-force fields for molecular generation. *arXiv preprint arXiv:2605.19050*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

Konstantin S. Jakob, Aron Walsh, Karsten Reuter, and Johannes T. Margraf. Learning Crystallographic Disorder: Bridging Prediction and Experiment in Materials Discovery. *Advanced Materials*, 38(5):e14226, January 2026. doi: 10.1002/adma.202514226.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DNdN26m2Jk.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jkvZ7v4OmP.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28393–28417. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/joshi25a.html.

Mikkel Juelsholt. Continued challenges in high-throughput materials predictions: MatterGen predicts compounds from the training dataset. *Mater. Horiz.*, 13(11):5672–5679, June 2026.

Zahra Kadkhodaie, Florentin Guth, Stéphane Mallat, and Eero P Simoncelli. Learning multi-scale local conditional probability models of images. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VZX2I_VVJKH.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In H. Wallach,

H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0234c510bc6d908b28c70ff313743079-Paper.pdf.

Tuan Le, Julian Cremer, Djork-Arne Clevert, and Kristof T Schütt. Equivariant diffusion for structure-based de novo ligand generation with latent-conditioning. *Journal of Cheminformatics*, 17(1):90, 2025.

Yunyang Li, Lin Huang, Luojia Xia, Wenhe Zhang, and Mark Gerstein. Elign: Equivariant diffusion model alignment from foundational machine learning force fields. *arXiv preprint arXiv:2601.21985*, 2026.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *arXiv preprint arXiv:2512.05349*, 2025.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=79O2XccGXZ.

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

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P7O1Vt1BdU.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Masahiro Negishi, Hyunsoo Park, Kinga O. Mastej, and Aron Walsh. Continuous uniqueness and novelty metrics for generative modeling of inorganic crystals. In *AI for Accelerated Materials Design - NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=PiKMmLHbEH.

Masahiro Negishi, Hyunsoo Park, Kinga Oliwia Mastej, and Aron Walsh. Continuous SUN (stable, unique, and novel) metric for generative modeling of inorganic crystals. *Machine Learning: Science and Technology*, 7(3):035064, June 2026. ISSN 2632-2153. doi: 10.1088/2632-2153/ae7d85.

Ivan S Novikov, Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. The mlip package: moment tensor potentials with mpi and active learning. *Machine Learning: Science and Technology*, 2(2):025002, dec 2020. doi: 10.1088/2632-2153/abc9fe. URL https://doi.org/10.1088/2632-2153/abc9fe.

Kento Nozawa, Pascal Germain, and Benjamin Guedj. Pac-bayesian contrastive unsupervised representation learning. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 21–30. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/nozawa20a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick

Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *arXiv preprint arXiv:2511.07158*, 2025.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nat Commun*, 16, 2025. doi: https://doi.org/10.1038/s41467-025-59636-y.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2020.

Chris J. Pickard. Airss data for carbon at 10gpa and the c+n+h+o system at 1gpa, 2020. URL https://doi.org/10.24435/materialscloud:2020.0026/v1.

Lucas Pinede, Soojung Yang, Juno Nam, and Rafael Gomez-Bombarelli. Unifying force prediction and molecular conformation generation through representation alignment. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. URL https://openreview.net/forum?id=yzkHGHvC74.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5799–5808. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simon19a.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0bFXbEMz8e.

Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gojL67CfS8.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling. *arXiv preprint arXiv:2604.02270*, 2026.

Cédric Villani. *Optimal transport – Old and new*, volume 338, pp. xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Liming Wu, Qi Li, Mingze Li, Songyou Li, Shifeng Jin, and Wenbing Huang. DMFlow: Disordered materials generation by flow matching. 2026. URL https://openreview.net/forum?id=hECemWeuhT.

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

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

A Primer on Optimal Transport

We will briefly recall the basics of Optimal Transport (OT), following Peyré & Cuturi (2020); for a more theoretical, abstract overview, see Villani (2008).

Optimal Transport is concerned with the task of how to efficiently transport mass from one probability measure μ to another one ν. Specifically, it tries to find a coupling $\pi \in \Pi(\mu, \nu)$ such that $\pi = \text{argmin}_{\pi \in \Pi(\mu,\nu)} \int \|x - y\| d\pi(x, y)$, where $\Pi(\mu, \nu)$ is the set of probability measures with marginals μ and ν.

For empirical, disjoint measures of the same size, this reduces to finding the optimal permutation π, i.e., which point x_i should be transported to which other point y_j, which can be solved using the Hungarian algorithm (Kuhn, 1955). In practice, we use the POT package (Flamary et al., 2021) with the standard ot.emd function.

Note that OT is also available when the two probability measures have different numbers of samples. However, we then need to split the mass. For instance, consider $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{-2}$, then "half" of the mass of δ_0 gets mapped to 0, whereas the other half gets mapped to -2. This also explains how OT penalizes mode collapse. If a point is overrepresented in one (like the 0) but not in the other measure, it manifests as a transportation error. This is also the reason why TNovD and CFTD penalize distributional shifts; see Hagemann et al. (2025) for an illustration.

Regarding runtime, Hungarian matching (or the LP solver we use) scales as n^3 in the number of particles (Peyré & Cuturi, 2020) and is therefore more expensive than the Sinkhorn algorithm (Cuturi, 2013). However, for a larger number of particles, our method could use similar heuristical couplings as done in OT flow matching (Pooladian et al., 2023). To scale to more particles, we outline two potential ways.

Expected Slice Transport To overcome poor scaling, we propose using expected slice transport, introduced in Liu et al. (2025). The main idea is to use slicing to reduce dimensionality when calculating the optimal transport plan. Assume we want to find a transport between the measure $\mu = \frac{1}{n}\sum_i \delta_{x_i}$ and $\nu = \frac{1}{m}\sum_i \delta_{y_j}$, then we define the sliced probability measure with respect to a vector $\|\theta\| = 1$ as $\theta_{\#}\mu = \frac{1}{n}\sum_i \delta_{\theta^T x_i}$ and similarly for ν. Now for each θ, find an optimal transport plan $\tilde{\gamma}_\theta \in \Pi(\theta_{\#}\mu, \theta_{\#}\nu)$. Since this trick makes it a one-dimensional problem, it can now easily be calculated via sorting, before being lifted to a genuine transport between the original points x_i and y_j, which we call γ_θ (note that this is easiest to understand in the case $n = m$ but also works in $n \neq m$). All transport plans are averaged to find the so-called *Expected Sliced Transport* (EST)

$$\gamma = \mathbb{E}_{\theta \sim S^{d-1}}[\gamma_\theta].$$

Note that Liu et al. (2025) give a more general construction, which we can disregard when working with empirical measures and assuming no projection clashes.

Multiscale Optimal Transport In Gerber & Maggioni (2017), another strategy is outlined that reduces scaling to $n \log n$. There we find a multiscale partition of training and generated features, i.e., we denote the data by X_j at scales $j = 0, ..., J$, with 0 being the coarsest scale and J the full data. Each scale is indexed further by a number of clusters, i.e., $X_j = \{(X_j)_k\}_{k=1}^{K_j}$. Each cluster has a representative point (the center of mass), and we solve an OT problem between each scale represented by this point. We now solve OT between training points $(X_j)_k$ at scale j and generated points $(G_j)_k$. Since each finer scale cluster $(X_{j+1})_{k'}$ is contained in one coarse scale cluster, we can propagate that we only solve at scale $j + 1$ with the clusters that were supported at scale j. Depending on the internal structure, this reduces run time to $n \log n$ in the case of a genuine tree partitioning.

These two directions could support scaling our transport distance to much larger sample sizes. However, neither yields OT couplings and the implications for novelty detection remain to be investigated, as they could match suboptimally or smear mass.

B Transport Novelty Distance and Coverage-Novelty Bounds

Here, we discuss the recently introduced Transport Novelty Distance (Hagemann et al., 2025) and firstly introduce the mathematical notions of coverage and novelty.

Coverage and Novelty In early generative modeling, people proposed precision and recall, as well as continuous versions thereof, to evaluate generative models (Kynkäänniemi et al., 2019; Simon et al., 2019; Cheema & Urner, 2023). Following Cheema & Urner (2023), precision is defined as the fraction of the generated distribution Q that is in the support of the true data distribution P. More specifically, these are generated samples that lie within the support of the training samples. Conversely, all samples from the true data distribution P that fall within the generated distribution Q yield the recall. This precision–recall viewpoint motivates the two quantities used in our analysis. It provides the conceptual starting point for our analysis, because it separates two complementary failure modes of generative models: generating samples outside the true data distribution and failing to cover parts of the true data distribution. To connect this viewpoint to our transport-based metric, we use the closely related radius-based notions of coverage and novelty. *Coverage* measures whether each data sample is approximated by at least one generated sample, and is therefore aligned with recall. *Novelty* measures whether generated samples remain outside a prescribed neighborhood of the training data, allowing us to distinguish genuinely new samples from memorized ones. Thus, coverage and novelty provide the concrete quantities that CFTD is designed to control. More specifically, we want high coverage regarding the quality of the generated structures (the quality parameters of the generated samples are supposed to cover the training data distribution), while we also want high novelty regarding the structural and compositional aspects (the generated structures are supposed to not exactly memorize structures from the training set).

We define the *coverage* (Naeem et al., 2020) of the measure $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m}\sum_{i=1}^m \delta_{g_i}$ of radius r as

$$\mathrm{Cov}_r(\mu, \nu) = \frac{1}{n}\sum_{i=1}^{n} 1_{\exists j : \|x_i - g_j\| \leq r}.$$

If the distance of at least one generated sample g_j to the investigated sample x_i is smaller than or equal to r, the respective sample x_i is counted as covered by the generated distribution. *Coverage* therefore indicates, how much of the true data distribution μ is covered by the generated distribution ν.

Similarly, we define the *novelty* (Negishi et al., 2026) with radius r as

$$\mathrm{Nov}_r(\mu, \nu) = \frac{1}{m} \sum_{j=1}^{m} 1_{\forall i: \|x_i - g_j\| > r}.$$

If all samples x_i from the true data distribution are farther than r from the generated sample g_j, then g_j is counted as novel. For the final *novelty*, we sum over all samples g_j, for which we cannot find any x_i from the true data distribution within a distance of r. As such, *novelty* controls how much of the generated distribution ν is not included in the training distribution μ.

Note that Naeem et al. (2020) defines the error using distances between k-nearest neighbors, precisely because the threshold must be chosen data-adaptively.

Transport Novelty Distance We quickly recall the main definition of TNovD (Hagemann et al., 2025), where the authors aimed to find a quality-novelty metric for unconditional generative models that jointly penalized generated distributions that are either unrealistic or copies of the training set.

With available training samples $(x_i)_{i=1}^n$ and generated samples $(g_i)_{i=1}^m$, the TNovD follows two steps:

1. Step 1: Calculate the pairwise distance matrix $C_{i,j} = \|x_i - g_j\|$ and the OT plan $\pi(\mu, \nu) \in \mathbb{R}^{n,m}$ between $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m} \sum_{i=1}^m \delta_{g_i}$. This is realized as a doubly stochastic matrix.
2. Step 2: Calculate the actual TNovD metric as

$$\mathrm{TNovD}(\mu, \nu) = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{i,j}(\mu, \nu)(\mathrm{ReLU}(C_{i,j} - \tau) + M\ \mathrm{ReLU}(\tau - C_{i,j})).$$

During step 2, τ serves as a threshold that splits $C_{i,j}$ into a memorization and a quality regime, which are supposed to reflect the above-introduced novelty and quality coverage, respectively. Since the memorization values are below the threshold τ, they are naturally lower than those of the quality regime. To adapt to this, a hyperparameter M is tuned to scale the memorization values. The two hyperparameter τ and M are therefore crucial for the success of this algorithm.

We will now show that we can bound the quality coverage and novelty using the TNovD, i.e., a low TNovD translates to a high quality coverage *and* a high novelty. While the coverage bound is natural and trivial to formulate, novelty requires a non-degeneracy condition that ensures that whenever a generated sample is close to a training sample, the OT plan also assigns mass to it. In the following simple lemma, we highlight the connections between TNovD and the established metrics coverage and novelty.

Lemma B.1. *Let $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m} \sum_{j=1}^m \delta_{g_j}$ be as above and assume that $r > \tau$. Then,*

$$1 - \mathrm{Cov}_r(\mu, \nu) \leq \frac{\mathrm{TNovD}(\mu, \nu)}{r - \tau}.$$

Now assume $r < \tau$, and assume that there exists $\alpha > 0$, $\min_i C_{i,j} \leq r \implies \sum_{i: C_{i,j} \leq r} \pi_{i,j} \geq \frac{\alpha}{m}$. Then we obtain that

$$1 - \mathrm{Nov}_r(\mu, \nu) \leq \frac{\mathrm{TNovD}(\mu, \nu)}{\alpha\ M\ (\tau - r)}.$$

This is a positive statement for TNovD: It shows that it can indeed control both novelty and coverage. On the other hand, to increase both novelty and quality coverage, we would need to choose $r > \tau$ and $r < \tau$ simultaneously.

Proof. Fix the coupling $\pi(\mu, \nu)$. We now rewrite

$$\mathrm{TNovD}(\mu,\nu) = \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{i,j}(C_{i,j}-\tau)1_{C_{i,j}>\tau} + M(\tau - C_{i,j})1_{C_{i,j}\leq\tau}.$$

We can drop the second part, then this implies

$$\mathrm{TNovD}(\mu,\nu) \geq \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{i,j}(C_{i,j}-\tau)1_{C_{i,j}>\tau}.$$

Hence, for $r > \tau$ that

$$\mathrm{TNovD}(\mu,\nu) \geq \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{i,j}(r-\tau)1_{C_{i,j}>r}.$$

Using this, we directly get

$$\sum_{i,j} \pi_{i,j}1_{C_{i,j}>r} \leq \frac{\mathrm{TNovD}}{r-\tau}.$$

By properties of the OT plan, namely that $\sum_{j=1}^{m}\pi_{i,j} = \frac{1}{n}$ we get

$$\sum_{i,j} \pi_{i,j}1_{C_{i,j}>r} \geq \sum_{i=1}^{n} 1_{\forall j:\|x_i-g_j\|>r}\frac{1}{n} = 1 - \mathrm{Cov}_r(\mu,\nu).$$

For the second part, we define the set of non-novel points, i.e., $S = \{j : \min_i C_{i,j} \leq r\}$. These are the generated points that are r-close to a training point.

In the same manner, we bound

$$\mathrm{TNovD}(\mu,\nu) \geq \sum_{j\in S}\sum_{i:C_{i,j}\leq r} M\,\pi_{i,j}(\tau - r) \geq \frac{|S|\alpha}{m}M(\tau-r).$$

Since S is precisely the set of generated points that are r-close to training samples, we can redefine novelty via

$$\mathrm{Nov}(\mu,\nu) = \frac{1}{m}\sum_{j=1}^{m} 1_{j\in S^c} = \frac{1}{m}|S^c|.$$

Using $|S| + |S^c| = m$, we have $1 - \frac{|S|}{m} = \mathrm{Nov}_r$.

This overall implies the claim by rearranging that

$$1 - \mathrm{Nov}(\mu,\nu) \leq \frac{\mathrm{TNovD}(\mu,\nu)}{\alpha M(\tau - r)}.$$

□

Note that for the CFTD, we now redefine

$$\mathrm{Cov}_r(\mu,\nu) = \frac{1}{n}\sum_{i=1}^{n} 1_{\exists j:\|F^{mace}(x_i)-F^{mace}(g_j)\|\leq r},\ \mathrm{Nov}_r(\mu,\nu) = \frac{1}{m}\sum_{j=1}^{m} 1_{\forall i:\|F^{id}(x_i)-F^{id}(g_j)\|>r}.$$

Theorem B.2. *Define the cost functions $C^{id}_{i,j} = \|F_{id}(x_i) - F_{id}(g_j)\|$ and $C^{mace}_{i,j} = \|F_{mace}(x_i) - F_{mace}(g_j)\|$ for the empirical measures $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\nu = \frac{1}{m}\sum_{j=1}^{m}\delta_{g_j}$.*

Further, denote π an optimal coupling associated to the cost $C = \frac{1}{2}C^{id} + \frac{1}{2}C^{mace}$.

Denote $r_{mem} < \tau_{mem}$ and similarly $r_{qual} > \tau_{qual}$ (Goldilocks Zone).

We then have the coverage (quality) as

$$1 - \mathrm{Cov}_{r_{qual}}(\mu, \nu) \leq \frac{\mathrm{CFTD}(\mu, \nu)}{r_{qual} - \tau_{qual}}$$

Assume that there exists $\alpha > 0$ *such that for every* j $\min_i C^{id}_{i,j} \leq r_{mem} \implies \sum_{i: C^{id}_{i,j} \leq r_{mem}} \pi_{i,j} \geq \frac{\alpha}{m}$*, then we have novelty as*

$$1 - \mathrm{Nov}_{r_{mem}}(\mu, \nu) \leq \frac{\mathrm{CFTD}(\mu, \nu)}{\alpha\, M(\tau_{mem} - r_{mem})}.$$

This shows that we can indeed control both novelty and quality coverage independently, which allows us to explicitly look at the coverage of structural quality and structural memorization detached from one another. It also shows that both parameters are bound by the CFTD.

Proof. Denote the CFTD in its probabilistic form via

$$\mathrm{CFTD}(\mu, \nu) = \mathbb{E}_\pi[M\mathrm{ReLU}((\tau_{mem} - C^{id})) + \mathrm{ReLU}(C^{mace} - \tau_{qual})].$$

Note that here C^{id} and C^{mace} denote the random variables which are obtained from sampling $(i, j) \sim \pi$ and evaluating the costs. $\mathbb{E}_\pi$ is the expectation value with respect to π

On the event $\{C^{id} \leq r_{mem}\}$, we have

$$\mathrm{CFTD}(\mu, \nu) \geq M(\tau_{mem} - r_{mem})\pi(C^{id} \leq r_{mem}),$$

and similarly,

$$\mathrm{CFTD}(\mu, \nu) \geq (r_{qual} - \tau_{qual})\pi(C^{mace} \geq r_{qual}).$$

We can proceed as in the case of the proof of Lemma B.1 and show that

$$\pi(C^{mace} \geq r_{qual}) = \sum_{i,j} \pi_{i,j} 1_{C^{mace}_{i,j} \geq r_{qual}} \geq \frac{1}{n} \sum_{i=1}^{n} 1_{\forall j: C^{mace}_{i,j} \geq r_{qual}} = 1 - \mathrm{Cov}_{r_{qual}}(\mu, \nu).$$

For the novelty we proceed similarly by defining $S = \{j : \min_i C^{id}_{i,j} \leq r_{mem}\}$.

By assumption we have that $\sum_{i: C^{id}_{i,j} \leq r_{mem}} \pi_{i,j} \geq \frac{\alpha}{m}$, hence

$$\pi(C^{id} \leq r_{mem}) = \sum_{j \in S} \sum_{i: C^{id}_{i,j} \leq r_{mem}} \pi_{i,j} \geq \frac{|S|\alpha}{m},$$

from which we can infer the claim in the same way as before. □

C On the fineness of the Identity features

In Hagemann et al. (2025), the featurizer F_{id} is a GNN trained using contrastive learning. We keep this featurizer for novelty evaluation.

More precisely, we used the InfoNCE loss (van den Oord et al., 2018), which maximizes the similarity between pairs $(z^1_i, z^2_i) = (F_{id}(x^1_i), F_{id}(x^2_i))$ by minimizing the loss

$$L(\theta) = -\sum_{i=1}^{n} \log\left(\frac{\exp(z^1_i \cdot z^2_i)}{\sum_{j \in N_i} \exp(z^1_i \cdot z^2_j)}\right),$$

where N_i is the set of indices without z^1_i.

For the similarity calculation, we augment a structure and treat the original structure and its augmentation as a pair. We thereby need to specify the $\mathrm{augment}$-function, which returns these *positive* pairs for the contrastive procedure. In Hagemann et al. (2025), the authors used rotation and translation.

However, to make the identity featurizer F_{id} more chemically aware and more robust towards noise, we added a small amount of random zero-mean Gaussian noise with a standard deviation of 0.02 to the fractional coordinates with probability 0.5, so that small differences in atomic positions between generated samples and training materials are not penalized as heavily. Additionally, we expanded the augmentation by substitution with similar elements, replacing each site of an element with either its nearest or its next-nearest neighbors on the modified Pettifor scale (Glawe et al., 2016) with a certain probability (0.3 and 0.01, respectively). Pettifor neighbors are often chemically similar atom types that have a high probability of substituting for one another in a crystal structure. We therefore expect similar bonding properties and consider the adapted structure to be a comparable material, or at least a material for which we do not require a generative artificial intelligence model to discover it. This is also relevant to substitutional disorder in crystal structures (Jakob et al., 2026), which generative models for crystals often neglect. The modification will therefore penalize generative models that mostly create new crystal structures by substituting supercells of well-known crystal structures with low concentrations of chemically similar elements. We note that the augmentation can be flexibly adapted to the specific definition of identical materials or to the materials discovery use case.

We provide a brief mathematical argument explaining why a contrastive loss (van den Oord et al., 2018) is a good choice for the fine featurizer (i.e., to detect novelty). We want to show that if the InfoNCE loss is small, the probability that materials are mapped to the same features vanishes. This is by no means a new technical argument, but helpful for our story. A similar result, albeit in a much different language, has been shown in Nozawa et al. (2020). It basically says that a small InfoNCE loss yields good separation between distinct materials, thereby preventing false positives.

Proposition C.1. Let x be a material, with positive augment x^+ (should be close) and negative augments $x_1, ..., x_K$ (should not be close) with feature versions $z = F_{id}(x), z^+ = F_{id}(x^+), z_1, ..., z_K = F_{id}(x_1), ..., F_{id}(x_K)$ Further denote the InfoNCE loss for x by $l_{nce} = -\log\left(\frac{\exp(z^T z^+)}{\exp(z^T z^+)+\sum_{k=1}^K \exp(z^T z_k)}\right)$. Then the probability of the event $A_\gamma = \{\exists j : z^T z_j \geq z^T z^+ - \gamma\}$ can be bounded by

$$\mathbb{P}(A_\gamma) \leq \frac{\mathbb{E}(l_{nce})}{\log(1+\exp(-\gamma))}.$$

Proof. On the event A_γ, we have that by definition for some fixed j that

$$\frac{\exp(z^T z^+)}{\exp(z^T z^+) + \sum_{k=1}^K \exp(z^T z_k)} \leq \frac{\exp(z^T z^+)}{\exp(z^T z^+) + \exp(z^T z_j)} \leq \frac{1}{1+\exp(-\gamma)}.$$

Taking logs yields that

$$l_{nce} \geq \log(1+\exp(-\gamma)).$$

Hence, by Markov, we have

$$\mathbb{P}(A_\gamma) \leq \mathbb{P}(l_{nce} \geq \log(1+\exp(-\gamma))) \leq \frac{\mathbb{E}(l_{nce})}{\log(1+\exp(-\gamma))}.$$

□

This exemplifies the role of the fine featurizer: it defines an augmentation-dependent identity geometry in which structures related by the chosen augmentations are kept close, while other materials are encouraged to be separated. One could also take more classical approaches, such as the novelty "featurizers" employed in Negishi et al. (2026); however, they need to be adapted to have a fixed length to fit the OT framework.

Fine Featurizer Parameters Summary The fine featurizer was trained for 10 epochs using an equivariant GNN with batch size 128, a contrastive InfoNCE loss, and an augmentation function with modified Pettifor nearest-neighbor and next-nearest-neighbor substitution probabilities of 0.3 and 0.01, respectively, and Gaussian coordinate noise applied with probability 0.5 and scale 0.02.

D ON THE COARSENESS OF MACE FEATURES

The featurizer in TNovD (Hagemann et al., 2025) is trained with a contrastive loss function determining whether two structures are identical. However, there is no reason to expect it to directly capture a structure's quality parameters.

For this, we added F_{mace}, which considers pooled "node features" of MACE-MP-0b3 "medium": Let $h^k = (h^k_1, ..., h^k_{n_{atoms}})$ be the hidden states of its k-th interaction block, which are of size (n_{atoms}, h_{dim}). We apply mean pooling to obtain features of the same size, regardless of the number of atoms in a structure, and then concatenate the resulting layers. This yields $H = (\mathrm{mean}(h^1), ..., \mathrm{mean}(h^L))$.

To create the final feature space, we use a random projection by considering a fixed set of projections $(\xi_1, ..., \xi_{n_{proj}})$ with $\|\xi_i\| = 1$ and Gaussian initialization. From this, we then obtain the map $F_{mace}(x) = \xi H(x)$, where x is the material and $\xi \in \mathbb{R}^{n_{proj}, dim(H)}$. This operation allows us to reduce the dimensionality of the pooled MACE features while retaining a specified number of random dimensions. The result is an n_{proj}-dimensional representation of the coarse feature space that encodes quality information of the structure.

We verify its coarse character and discuss what information the MACE features actually contain by reconstructing the space group, lattice, and composition of the test set, as well as the formation energies of polymorphs from it.

Then, we will contrast the fine (for novelty) and coarse (for quality) features by probing how they treat augmentations, whether they separate them, and by comparing their cosine similarities.

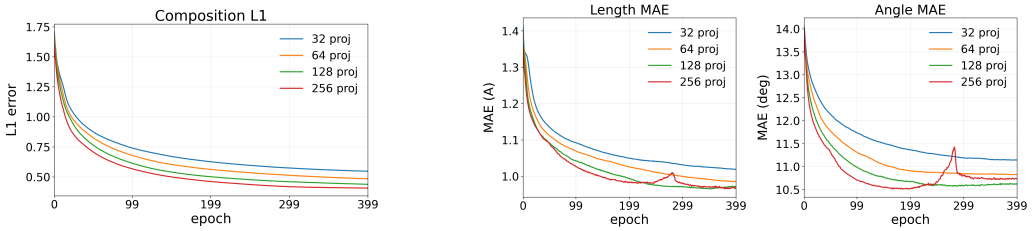
D.1 WHAT INFORMATION IS CONTAINED IN THE MACE FEATURES?

To further elucidate the role of the MACE features, we examine whether we can reconstruct space groups, composition, and lattice parameters by simply learning a standard feedforward neural network from the MACE feature space to the respective structural parameters. We perform these analyses on the MP20-dataset, the results are shown in Fig. 6 and Fig. 7 (a).

For the composition, we measure the $L1$-Distance between the predicted fraction of elements of a structure and the true fraction of elements. For the lattice, we compute the MAE between predicted and true lattice vectors and angles. The results for the validation set can be seen in Fig. 6 and show that the composition is quite predictable, whereas an accurate prediction of the lattice is harder. Generally, both improve as the number of projections increases. Regarding the space group reconstruction, we report the accuracy of the predicted top space group for MP20 in Fig. 7 (a). The resulting accuracy is around 0.35-0.40.

While space group and lattice parameters show good reconstruction values, composition is the best-controlled part of a material, i.e., the easiest to reconstruct from the F_{mace} feature space. To check whether the reconstruction of space groups is inferred by the good prediction of the composition, we repeat the space group prediction on the Carbon24 data set (Pickard, 2020; Xie et al., 2022). This is interesting, since each material in this dataset consists solely of carbon, and the results tackles the hypothesis that MACE features encode only composition and no structural information. The results for Carbon24 are shown in Fig. 7 (b) and compared to MP20 (Fig. 7 (a)), the accuracy values of around 0.55 are slightly better. Given the architecture and success of pretrained MACE models, the results for both datasets were somewhat expected. However, since we apply mean pooling to the features and restrict ourselves to a certain number of random projections to create the F_{mace} feature space, we needed to assess the influence of these changes on the final representations. Further, we test four different projection lengths of the mean-pooled invariant MACE features (32, 64, 128, and 256), with larger projections generally conveying more information.

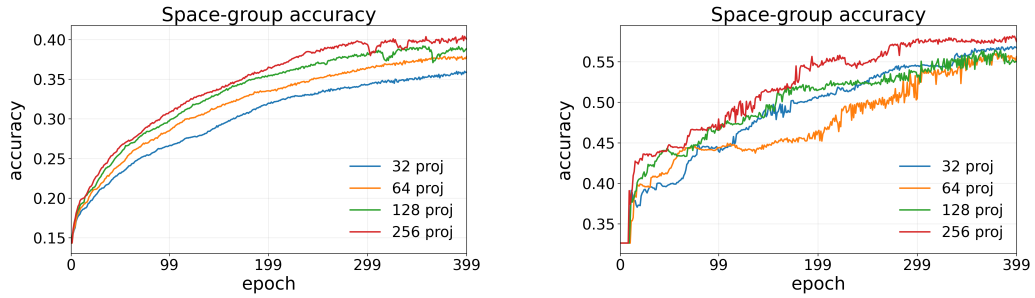
These experiments show that the projections of the mean-pooled MACE features capture both chemical and geometric information. However, they do not suffice to perfectly reconstruct any parameter. The closest one is chemical composition. This underlines the coarse character of the MACE features. Generally, information improves as we add more projections.



(a) Composition curve.

(b) Lattice curve.

Figure 6: Composition and lattice reconstruction curves on MP20.



(a) Space group accuracy for MP20.

(b) Space group curve for Carbon24.

Figure 7: Space group reconstruction curves.

D.2 Comparison to the Identity featurizer

The role of the Identity featurizer is to provide a different kind of "geometry" differentiating between positive (those structures that should be close) and negative pairs (those structures that should not be close) as indicated by Proposition C.1. Therefore, we investigate whether the contrastive GNN indeed separates positive and negative pairs differently than the MACE featurizer. To test this, we randomly select 5000 samples from the MP20 train set and sweep over γ as in the Proposition. We measure the number of negative pairs that exceed the threshold. More specifically, we generate for each material a positive augment and check whether there exists a different material that exceeds the augment cosine similarity with a subtracted γ. Hence, as γ increases, the probability of hitting a negative pair (a different material) instead of the augment increases. We do this for both the MACE and the Identity featurizer, for which the results can be seen in Fig. 8. In general, the contrastive GNN provides much better separation, as indicated by the much lower probability of margin violation. This explains the role of the fine Identity featurizer, which induces a completely different geometry than the coarse MACE featurizer.

D.3 A cosine investigation

To further understand the different featurizers, we randomly drew two times 5000 samples from the MP20 train set and formed random pairs between the two created subsets. These samples were then transformed into the two feature spaces. The features were not only calculated on "clean" structures directly drawn from the MP20 train set, but also on "noisy" structures. We induced this noise by altering the fractional coordinates with Gaussian noise and applied a relatively high standard deviation of 0.05. Fig. 9 compares the cosine similarity for each featurizer and randomly selected feature pairs of, in total, four different sets. The sets cover feature pairs for clean vs. clean, clean vs. noisy, noisy vs. clean, and noisy vs. noisy structures. For the Identity featuzier, cosine similarity values are mainly focused around 0 and high similarities are the absolute exception for all four sets. The low similarity values indicate that the contrastive GNN is indeed trained to achieve separation of the different samples, even when noise is present. For the MACE featurizer, we see a different behavior: for the clean-clean pairs, we observe that relatively few have a cosine similarity below 0.5.

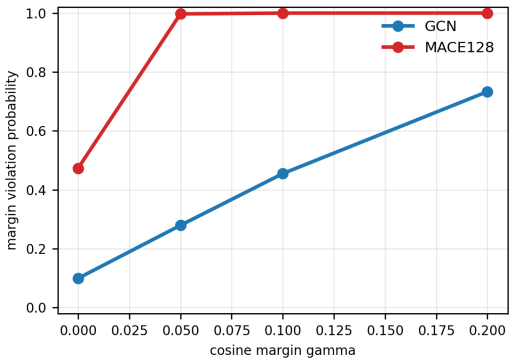


Figure 8: Separation between the MACE featurizer and the Identity featurizer.

The MACE tail, however, changes for clean vs. noisy or noisy vs clean feature pairs. This indicates that the MACE feature space also accounts for perturbations, as it is trained not only on ”clean” and stable data. The plots in Fig. 9 thereby further underline the different geometries learned by both featurizers.

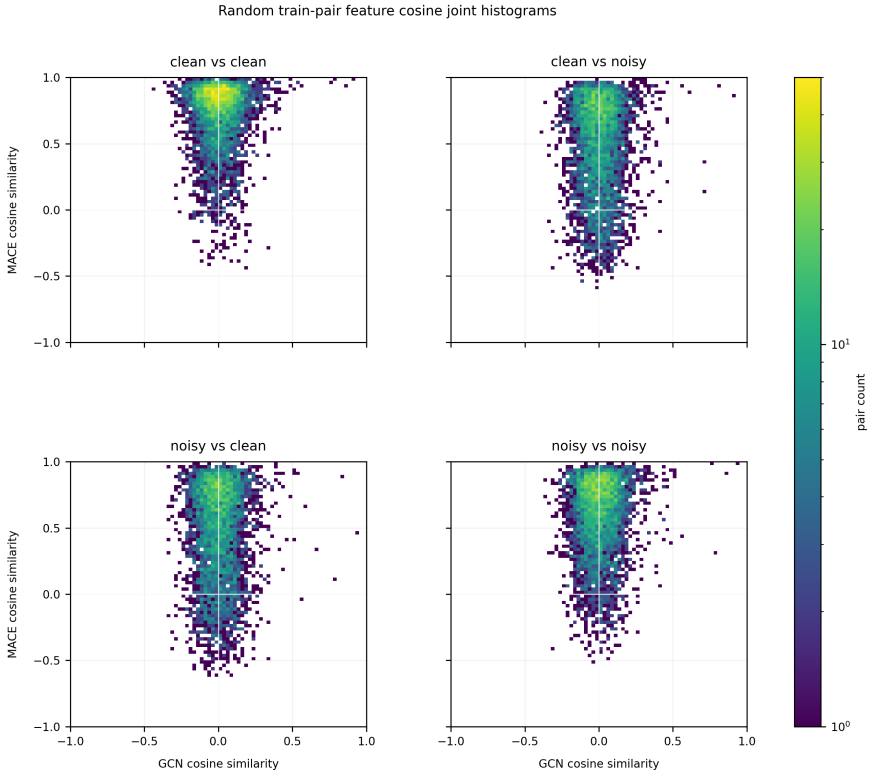


Figure 9: Joint histograms of cosine similarities between random pairs of clean or noisy data for the two featurizers.

E On the implicit stability of MACE features

A natural question that arises is what notion of stability the MACE features really encode. We call this *implicit* stability. Of course, due to mean pooling and the random projections, exact node-level information is lost. Since MACE predicts energies and forces, it is still reasonable to expect that it captures these even in a condensed feature space. We test this with the following experiments on formation energies.

MP20 Sweep: We test how well we can predict formation energies of MP20 from the mean-pooled MACE feature projections and train a simple multilayer perceptron (MLP) to predict the formation energies using the MAE as the loss function. The results are shown in Fig. 10 for a varying number of projections of the mean-pooled invariant MACE features as input for the MLP. We can see that all of the variants capture some notion of stability and greatly improve with training. More projections also improve the information encoded. Therefore, Fig. 10 shows that mean pooled MACE feature projections are a reasonable proxy for estimating formation energy, and possibly also local stability.

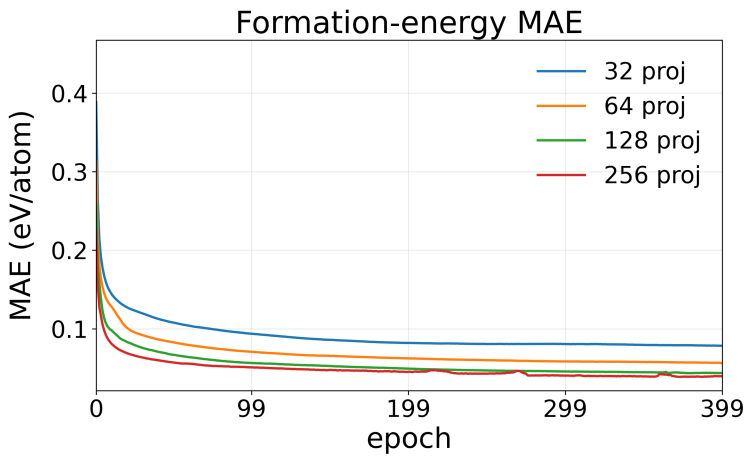
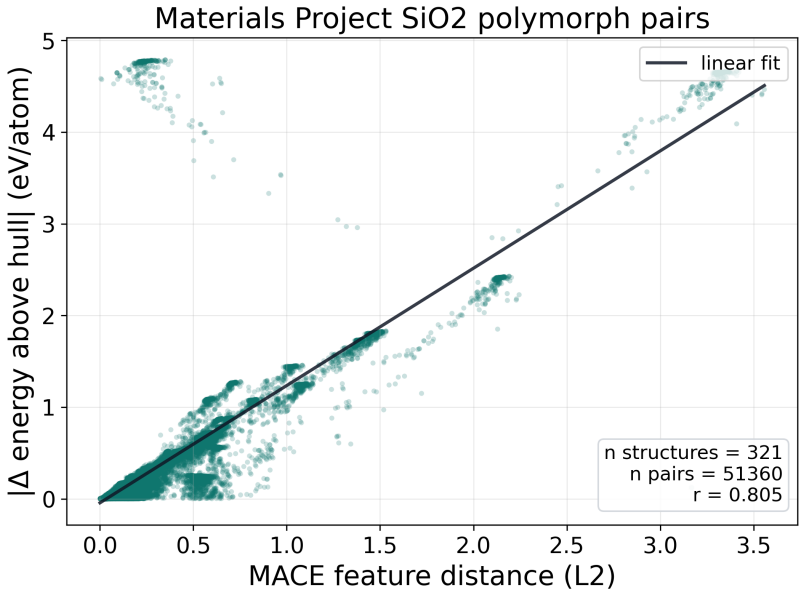


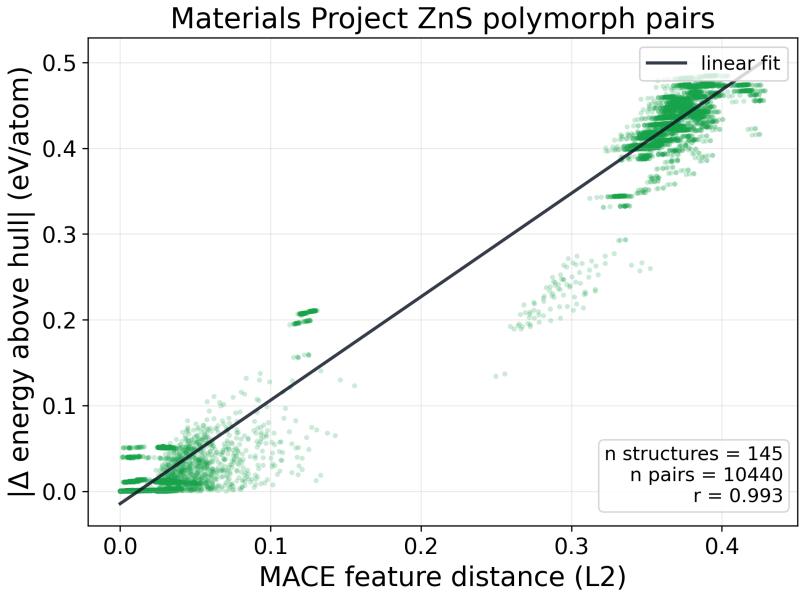
Figure 10: MAE of formation energy per atom over epochs for different lengths of mean pooled MACE feature projections.

Experiments on Polymorphs: We continue our investigation by analyzing the formation energies of polymorphs (i.e., compositions where different crystal structures exist). With this test, we want to check whether the differences in their energy above the hull (essentially, their differences are equivalent to the formation energy differences for the polymorphs) can be sufficiently separated by the MACE featurizer. We check all polymorphs of SiO_2 and ZnS in the MP dataset (Jain et al., 2013). We specifically choose polymorphs because they have the same composition, which we can therefore exclude as a confounder. In the results shown in Fig. 11, both investigated polymorphs exhibit a near linear correlation between their energy above hull differences with the distance between their respective MACE feature projections and although several outliers are visible in the plots, a general trend persists. This trend is unsurprising, as the readout phase for the invariant features in the first layers of the MACE models uses a linear readout. The experiment underscores our general thesis that pooled MACE feature projections are still a useful proxy for precise stability information. Nevertheless, there may be several outliers, as shown in the figure.

Overall, we observe a high correlation between MACE feature distances and polymorph energy distances, as well as MACE cosine similarities that differ systematically between random pairs of (meta)stable structures and pairs of structures, for which the fractional coordinates were altered by inducing random mean Gaussian noise. Both results demonstrate that MACE features carry implicit stability information.



(a) Hull-energy difference versus MACE feature L2 difference for SiO_2 polymorphs.



(b) Hull-energy difference versus MACE feature L2 difference for ZnS polymorphs.

Figure 11: Polymorph plots showing hull-energy differences as a function of MACE feature L2 differences.

F HYPERPARAMETER SELECTION

Here, we provide a precise description of how the hyperparameters are chosen in our CFTD framework.

The input to the hyperparameter selection is as follows:

1. The training data set
2. The validation data set
3. The β-Goldilocks percentage, i.e., the relative number of points that are categorized into the Goldilocks zone.

From there on, we calculate the featurized versions of the training and the validation samples with respect to the two featurizers F^{id}, F^{mace}, and find the optimal coupling with respect to the coupling $C = \frac{1}{2}C^{id} + \frac{1}{2}C^{mace}$. Then, we assign that $\frac{1-\beta}{2}$ mass should belong to the "memorized" and $\frac{1-\beta}{2}$ belongs to the "low quality" zone.

This allows us to enumerate the coupled cost. Now, we set $\tau_{mem} = \text{Quantile}_{\frac{1-\beta}{2}}(C^{id})$ and $\tau_{qual} = \text{Quantile}_{1-\frac{1-\beta}{2}}(C^{mace})$ with respect to the OT plan π between train and validation, which just says that the thresholds are chosen such that only $\frac{1-\beta}{2}$ of the mass are categorized below τ_{mem} or above τ_{qual}. Further, during the calculation of CFTD, we choose these regions to be mutually exclusive, favoring a categorization as memorized via setting the "low quality ones" as the ones with C^{mace} above τ_{qual} *that are not categorized as memorized*.

Now, we can calculate the memorized value $\sum_{i,j} \pi_{i,j}(\tau_{mem} - C^{id}_{i,j})$ as well as the "low quality" value $\sum_{i,j} \pi_{i,j}(C^{mace}_{i,j} - \tau_{qual})$ and choose M as the fraction of "low quality" value over memorized, i.e., scaling the values to match on the validation set. We remark that this is purely heuristic and intended to retain interpretability. A promising future direction is to improve this with concrete data-labeling or explicit data-perturbation procedures.

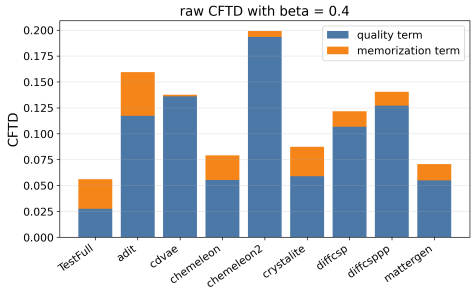
G ABLATION STUDY ON GOLDILOCKS ZONE β

Here, we ablate the influence of the size of the Goldilocks zone. This is controlled by the parameter β. The higher β is, the less quality and memorization errors are penalized. Hence, we need to ablate whether our conclusions, i.e., the internal rankings of the models, stay consistent with varying β. For this, we run model ablation and comparison experiments for $\beta \in \{0.4, 0.6, 0.8\}$ on the raw and relaxed versions of the generated materials. In Fig. 12, we can see that the general trend of the performance persists very well across different β, indicating that the results by our CFTD are highly robust. In particular, the strongest and weakest models remain clearly separated, and only minor ranking changes occur among models with similar CFTD values.

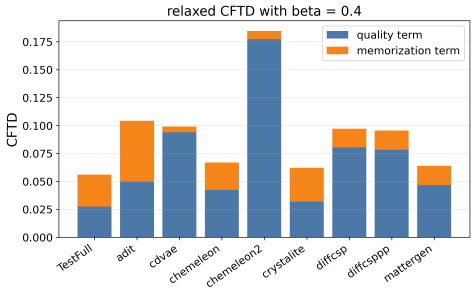
H ABLATION ON THE BALANCE FACTOR M

Similarly, we ablate the results with respect to the factor M. Usually, it is chosen that, on a calibration set (e.g., the validation set), both the quality and memorization penalties are balanced. However, this can be changed according to the type of model that is supposed to be favored. For instance, one could weigh memorization less than quality, favoring more conservative, training-like models, or penalize it more heavily. Therefore, we ablate on calibrating M such that both terms are balanced, and then set $M_{new} \in \{0.5M, M, 1.5M\}$.

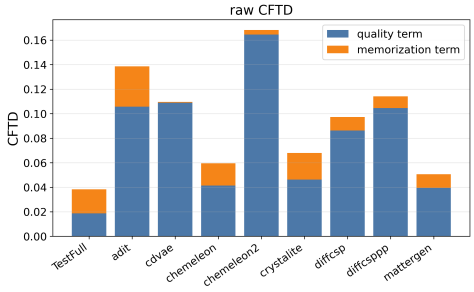
The results on the model rankings are shown in Fig. 13, where the relative ranking does not change for unrelaxed models, but it does for relaxed ones. In particular, Crystalite benefits strongly from a lower memorization penalty. For $M_{new} = 0.5M$, it becomes the best generated model after relaxation. Also ADiT's ranking benefits from a lower memorization penalty. Generally, M is a tuning knob for the quality-novelty profile, and hence rankings are not expected to be stable.



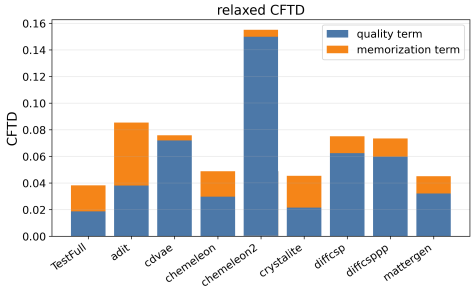
(a) CFTD for unrelaxed models $\beta = 0.4$.



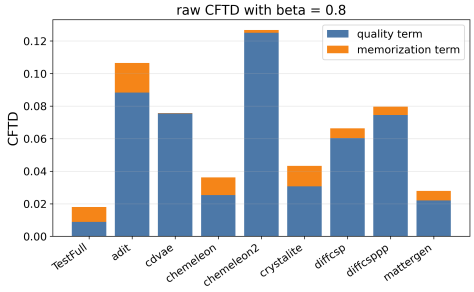
(b) CFTD for relaxed models $\beta = 0.4$.



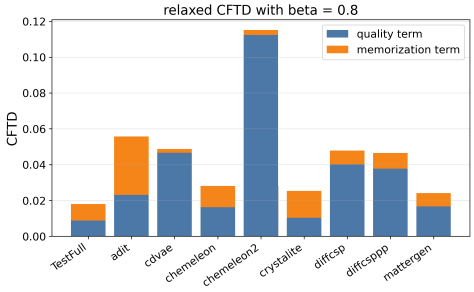
(c) CFTD for unrelaxed models $\beta = 0.6$.



(d) CFTD for relaxed models $\beta = 0.6$.

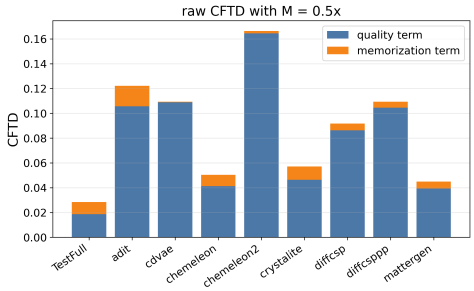


(e) CFTD for unrelaxed models $\beta = 0.8$.

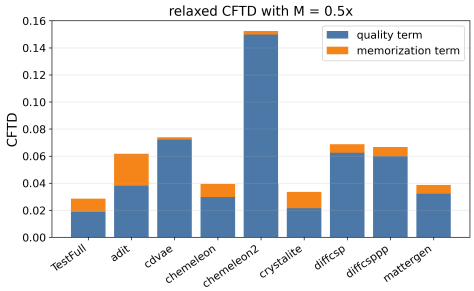


(f) CFTD for relaxed models $\beta = 0.8$.

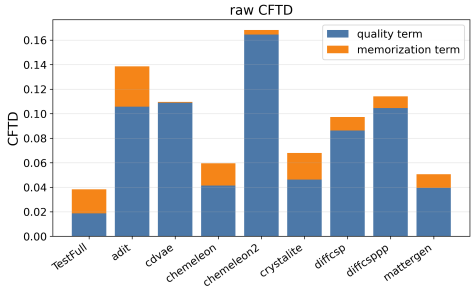
Figure 12: Comparison of common material generative models: relaxed and unrelaxed for varying β.



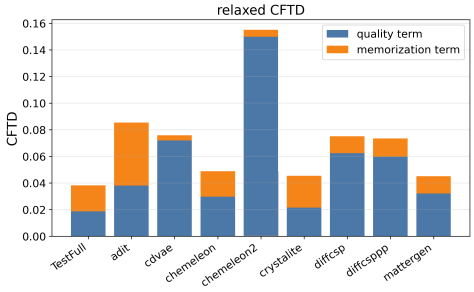
(a) CFTD for unrelaxed models $M_{new} = 0.5M$.



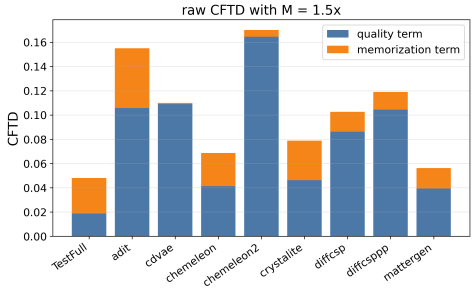
(b) CFTD for relaxed models $M_{new} = 0.5M$.



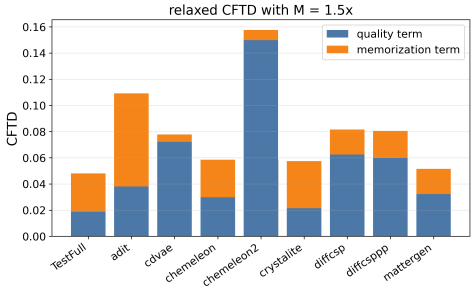
(c) CFTD for unrelaxed models $M_{new} = 1M$.



(d) CFTD for relaxed models $M_{new} = 1M$.

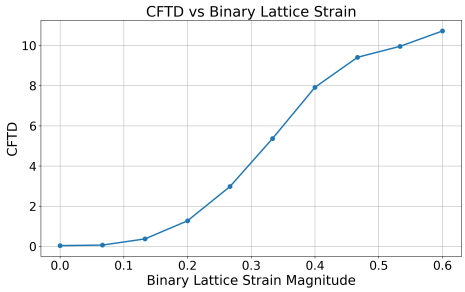


(e) CFTD for unrelaxed models $M_{new} = 1.5M$.

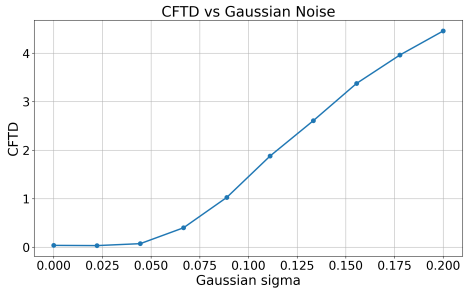


(f) CFTD for relaxed models $M_{new} = 1.5M$.

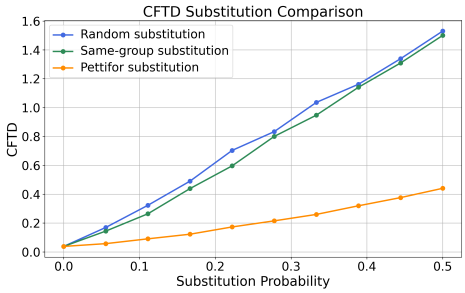
Figure 13: Comparison of common material generative models: relaxed and unrelaxed for varying M.



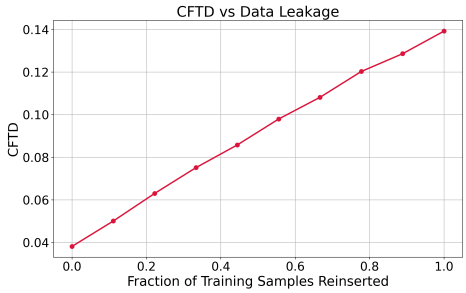
(a) CFTD over lattice strains.



(b) CFTD over Gaussian noise.



(c) CFTD over substitutions.



(d) CFTD over train leakage.

Figure 14: CFTD behavior across different perturbation settings on Perov-5.

I Toy Experiment on Perov-5

To show that the toy experiments from Fig. 2 in extend to larger training sets and different structures, we repeat them for the Perov-5 dataset (Castelli et al., 2012a;b; Xie et al., 2022). It contains materials that share a similar structure but different compositions (over 18,000 in total), making it structurally different from MP20. We repeat the same deformations as before and also test for data leakage.

Note that we reused the Identity featurizer that was trained on MP20, as well as the same MACE featurizer. The results can be seen in Fig. 14. Note that for this experiment, we chose modified Pettifor nearest neighbors from the MP20 set, not only from the atoms contained in Perov-5. A close investigation reveals that for the first addition of Gaussian noise (0.022) to the atomic positions, the CFTD dips slightly, which indicates that for such low numbers of noise, memorization decreases more quickly than quality penalization kicks in.

As for MP20, the plots of the individual experiments reveal large differences in the absolute values of the new metric. Artificial crystal degeneration is penalized more strongly than data leakage, and it responds particularly sensitively to lattice strain and changes in atomic positions. This is again different from the behavior observed in TNovD. We argue that this result comes from the inclusion of the coarse MACE featurizer, which enables the new CFTD to now accurately capture the quality of a crystal structure. The much lower values observed in the data leakage experiment result from our implementation of tuning the hyperparameter M. The optimal value of M is determined by aligning the naturally lower values of the memorization regime with the values of the quality regime on the same scale. This adjustment is performed on a split of MP20 train and validation set and, since the quality regime in this scenario has lower CFTD values, M will scale the memorization regime accordingly.

J RANKING OF EXISTING GENERATIVE MODELS

Here, we comment in more detail on the rankings of the models and how they compare to the rankings implied by continuous SUN (Negishi et al., 2026). For this comparison, we dismiss Crystalite as it was not available in their study.

Also, they ran their results only on unrelaxed models and we therefore compare the rankings only for those. Table 1 shows that our method largely agrees, with only minor differences between MatterGen and the test set, and DiffCSP++. The only major disagreement is the ranking of Chemeleon2. Note that their ranking changes significantly depending on the choice of the tunable weight hyperparameters for stability (w_S), uniqueness (w_U), and novelty (w_N). In their study, they show results for varying stability weights w_S. For instance, with $w_S = 0$, CDVAE is ranked 2nd, since novelty is overvalued and it is very novel by generating low-quality structures.

Table 1: Rankings taken from the cSUN study (Negishi et al., 2025) with $w_S = 1$ (SUN with elm+am) and our CFTD ranking for $\beta = 0.6$ and calibrated M, so that quality and memorization are balanced.

Model	**cSUN Rank**	**CFTD Rank**
MP20 Test Set	3	1
MatterGen	2	2
Chemeleon	4	3
DiffCSP	5	4
CDVAE	6	5
DiffCSP++	7	6
ADiT	8	7
Chemeleon2	1	8

K MACE CONDITIONAL MODEL DESCRIPTION

Atom type sampling We use a transformer trained with a log-likelihood objective. We model the sequence of atom types as $(a_1, .., a_n)$ with n as the number of atoms in a unit cell. Each atom type takes a value in the discrete set $a_i \in \{1, ..., K\}$, where K is the number of distinct atom types we wish to model (i.e., atom types present in the dataset). Sequence modeling is therefore hard, since we have K^n different possible sequences. As the atom types are categorical, we use the framework of discrete flow matching (Gat et al., 2024). Within this framework, we are given training samples $A_1 \sim P(A|n, h)$, where $A = (a_1, ..., a_n)$ is a random variable drawn according to the law (or dataset). We start with a fully masked atom-type assignment $A_0 = (m, ..., m)$, where m is a mask token. Then, we interpolate between the fully masked A_0 and the unmasked vector A_1 of atom types, by setting $A_t = (1 - B)\, A_0 + B\, A_1$, where B is a vector of independent Bernoulli random variables, where an entry is one with the probability $t \in [0, 1]$. Then, similar to flow matching, we learn a conditional probability on the categorical simplex Δ^{K-1}, i.e., probabilities on the atom types for each atom given the partially unmasked structure, which we denote by $q_\theta(A_1|A_t, t)$. We train this via forward KL, i.e., maximize the cross entropy

$$\mathbb{E}_{a_1 \sim p, t \in U([0,1])}[-\log q_\theta(a_1|a_t, t)].$$

Here, in practice, we condition on all relevant variables, i.e., the number of atoms and the pooled MACE feature projections as additional conditioning.

After training our categorical distribution on $\{1, ..., K\}^n$, we sample following the masked diffusion procedure from Sahoo et al. (2024). From the definition, it follows for each component of the path at time t that $p(A_t^i = m) = 1 - t$ and correspondingly $p(A_t \neq m) = t$. Then, for sampling, the probability that a previously masked token A_{t_1} is unmasked at a later time A_{t_2} is given by $p(A_s = m|A_t = m) = \frac{1-s}{1-t}$, and hence $p(A_s \neq m|A_t = m) = \frac{s-t}{1-t}$. Following this, we start with $s_0 = (m, ..., m)$ and then for each time step of $T = (t_1, ..., t_T)$ steps, then for each position in s_t we do nothing if it is already unmasked or we unmask with probability $\frac{t_{k+1}-t_k}{1-t_k}$. This matches

the conditional formulas given by the path. Furthermore, as t approaches 1, all tokens become unmasked.

Lattice sampling For lattice and position sampling, we mainly use continuous-flow matching. In contrast to discrete-flow matching, we here interpolate continuously in the space of random vectors, not in the space of measures. We follow mainly the presentation in Liu et al. (2023). Assume we have already drawn the positions X_1. Flow matching interpolates between an auxiliary latent $X_0 \sim \mathcal{N}(0, I)$ and $X_t = t\, X_1 + (1-t)X_0$ to learn its velocity $v_t(X_t) \approx X_1 - X_0$. It can be shown that considering the Ordinary Differential Equation (ODE) $dX_t = v_t(X_t)$ with initial $X_0 \sim \mathcal{N}(0, I)$ yields (approximate) samples from the positions.

For lattice generation, we follow Jiao et al. (2024) and its proposed polar decomposition. We apply this decomposition to the lattice matrix L, which is given by the lattice vectors. With real volume, this is invertible, and we can find a decomposition $L = Q\exp(S)$ with an orthogonal Q and symmetric S. By utilizing the space of symmetric matrices as a vector space, we can find a basis and decompose $S = \sum_{i=1}^{6} k_i B_i$, where B_i is the basis and k_i are the coefficients. The basis matrices are taken from Jiao et al. (2024). Since the space of 3×3 symmetric matrices are six-dimensional, we choose the following basis:

$$B_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$B_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \quad B_6 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For the flow matching onto the lattice coefficients k_i, we need to convert the lattice representation $(a, b, c, \alpha, \beta, \gamma)$ into the spaces of the k_i. From the $(a, b, c, \alpha, \beta, \gamma)$ we can compute the metric matrix $G = LL^T$. From this, we can find an eigendecomposition, $G = V\Lambda V^T$, where Λ is the diagonal matrix of eigenvalues and V are the eigenvectors. Now the matrix $S = V\log(\Lambda^{1/2})V^T$ satisfies $\exp(S)\exp(S)^T = G$, since S is symmetric itself. Now we convert these matrix entries into the representation with respect to the k-coefficients. It is easy to see that $k_1 = S_{0,1}$, $k_2 = S_{0,2}$ and $k_3 = S_{1,2}$, since B_1 is the only basis matrix with non zero entry at $(0, 1)$ (and similar for the other two). For the remaining k, we need to solve a linear system. For $(0, 0)$ we have $S_{0,0} = k_4 + k_5 + k_6$, and further $S_{1,1} = -k_4 + k_5 + k_6$ and $S_{2,2} = -2k_5 + k_6$. It is simple to check that $k_4 = \frac{1}{2}(S_{0,0} - S_{1,1})$, $k_5 = \frac{1}{6}(S_{0,0} + S_{1,1} - 2S_{2,2})$ and $k_6 = \frac{1}{3}(S_{0,0} + S_{1,1} + S_{2,2})$ solve this. Now we can get from $(a, b, c, \alpha, \beta, \gamma)$ to the coefficients, and use Euclidean flow matching with a latent $Z \sim \mathcal{N}(0, I)$ to the k_i, from which we can build a representation of the lattice matrix S.

Position sampling For the positions, we also employ continuous-flow matching conditioned on the pooled MACE feature projections, the lattice, and atom types. We adopt an adapted version of an equivariant GNN (Satorras et al., 2021) and modify it into a non-equivariant architecture. We do this by adding directional and moment information (Novikov et al., 2020) that is passed through a multilayer perceptron (MLPs) and FiLM-style conditioning (Perez et al., 2018). The main technical difficulty is that material generation introduces a periodic structure, so we cannot naively use flow matching in the Euclidean space. Let $\mathbb{T}_n^3$ denote n atoms (points) living each in the 3-dimensional torus space with the periodic structure. Further, we make the assumption that $[x + \delta] = [x_1 + \delta, ..., x_n + \delta] = [x]$, i.e., we have translational invariance for $x \in \mathbb{R}^{n,3}$ and $\delta \in \mathbb{R}^3$. We perform flow matching on the fractional coordinates. Assume we have $X_0 \in \mathbb{T}_n^3$ as a latent random vector, and $X_1 \in \mathbb{T}_n^3$.

The general idea of flow matching is now to regress onto the velocity $X_1 - X_0$. However, $X_1 - X_0$ is in general not in $[0, 1]^{n,3}$ due to the periodic boundary conditions, so we need to be careful there. For an element in $\mathbb{T}_n^3$, there are many different elements representing $[X_1 - X_0]$, namely $X_1 - X_0 + l + \delta$ with $l \in \mathbb{Z}^{n,3}$ and $\delta \in \mathbb{R}^3$. In particular, since flow matching is defined to live on the tangent space (Wu et al., 2026; Sriram et al., 2024), we use the representative given by $P(\mathrm{wrap}(X_1 - X_0))$, where $\mathrm{wrap}(x) = (x + 0.5 \bmod 1) - 0.5$ and $(P(x))_i = x_i - \sum_{k=1}^{n} \frac{1}{n} x_k$. This can be seen either by Riemannian geometry (Chen & Lipman, 2024) or by choosing the representative $\Delta = \mathrm{argmin}_{l,\delta} \|x_1 - x_0 + l + \delta\|^2$.

Similarly, we choose the congruent path representative $\mathrm{frac}(X_0 + t\Delta)$, whose derivative (up to the discontinuity points of frac) is Δ.

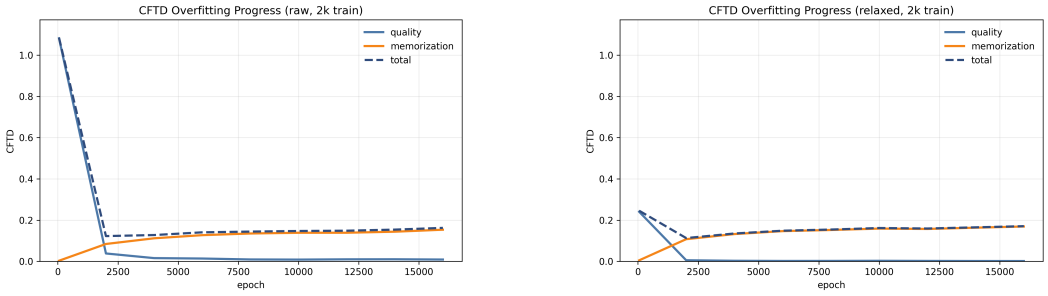
Then we regress our velocity onto

$$L_{pos}(v_\theta) = \mathbb{E}_{x_0,x_1,t}\|v_\theta(t, x_t) - \Delta\|^2.$$

This also has neat differential-geometric interpretations (Chen & Lipman, 2024).

L Detection of memorization

To judge both CFTD and the consistency of our generator, we first train our model on a small subset of 2,000 materials from the MP20 dataset (Xie et al., 2022). With so few samples for training, we expect our model to memorize to some degree. We train it for 16,000 epochs and validate the generated structures, both relaxed and unrelaxed, every 2,000 steps using the CFTD metric. The CFTD allows an insightful comparison of quality coverage and novelty, which we visualize in Fig. 15. We see that both unrelaxed (Fig. 15 (a)) and relaxed (Fig. 15 (b)) runs have a minimal CFTD at epoch 2000. We see that both runs overfit heavily from epoch 5000 on; however, the memorization increases more (and more strongly) for the relaxed version.. The quality error is closer to zero for the relaxed version. These results show a vital advantage of CFTD for evaluating the training of material generative models: it can be easily combined with relaxation procedures, unlike, e.g., the validation loss. Increased memorization is a clear indicator of overfitting, and since this might only be visible after relaxation (as seen by comparing Fig. 15 (a) and (b)), CFTD can identify such behavior more easily than other metrics.



(a) CFTD over epochs (unrelaxed).

(b) CFTD over epochs (relaxed).

Figure 15: Comparison of our proposed material generator trained on an MP20 subset of 2,000 materials only.

M Possible Extensions

M.1 Hierarchical Generation

One of the motivating pillars of this work is to strip away complexity from generating composition via LLM ideas or positions via GNNs. The first one seems particularly prone to overfitting, whereas the latter often struggles with finding stable or space group diverse structures, as indicated by the results in Betala et al. (2026) (see also their leaderboard). Conveniently, the pooled MACE features projections now live in a fixed-dimensional continuous space, where we can apply ideas from standard generative modeling in Euclidean space. In particular, this enables us to learn $p(n, h)$ via a generative model, for instance, a transformer using the flow matching framework. Then, we could pursue precisely as in our material generator, and conditionally sample atom types, lattice and positions, therefore being a full "hierarchical" MACE conditional generative model.

In particular, this allows us to sample "new" MACE features, therefore bypassing the composition-overfitting problem. However, a priori, the MACE feature space does not seem to have any particular structure. Many of our choices, such as mean-pooling, are arbitrary. This suggests a deeper analysis of how to compress the information in the MACE features while making it as "learnable" as possible

by a generative model. We leave this for future work, but preliminary results suggest that we can achieve a high novelty generator with a similar MSUN rate.

Further, another direction to explore is using *atom-level* features of the MACE embeddings. We did not pursue this since OT deliberately requires a fixed dimension. However, for material generation, finer-grained information might be helpful.

M.2 Conditional Generation

In this paper, we were mainly concerned with unconditional material generative models, but both the metric idea and the generative model can be extended for conditional generators, albeit extending the metric is not trivial and might need ideas like conditional Wasserstein-like couplings (Chemseddine et al., 2025). We want to focus on extending a MACE generator to a conditional one. Indeed, we could subsample or steer a MACE generator (say a generative model in the MACE feature space) to sample only from $p(h|c)$. This is possible as soon as we have paired data (h_i, c_i) and would avoid training another conditional generative model that takes the condition as input for a factorization. This might provide a significant speed-up as soon as the condition is identifiable enough using the MACE features. The experiments in Appendix D.1 suggest that MACE features indeed encode important material aspects.