FlowMM: Generating Materials with Riemannian Flow Matching

Benjamin Kurt Miller[1,2], **Ricky T. Q. Chen**[1], **Anuroop Sriram**[1], **Brandon M. Wood**[1]

[1]FAIR, Meta AI [2]University of Amsterdam

Summary

We introduce *FlowMM*, a pair of generative models for Crystal Structure Prediction (CSP) and De Novo Generation (DNG) that each jointly estimate symmetric distributions over fractional atomic coordinates and the unit cell (along with atomic types for DNG) in a single framework based on Riemannian Flow Matching. We train a Continuous Normalizing Flow with a finite time evolution and produce high-quality samples, as measured by standard metrics and thermodynamic stability, with significantly fewer integration steps than diffusion models. Our main contributions include:

i) We **generalize Riemannian Flow Matching to estimate a point cloud density that is invariant to translation with periodic boundary conditions**, a novel achievement for continuous normalizing flows, by proposing a new objective. With this step, it becomes possible to enforce isometric invariances inherent to the geometry of crystals as an inductive bias in the generative model.

ii) We **select a rotation invariant representation of the unit cell** and **choose a natural base distribution that samples plausible unit cells by design**. We find that this drastically simplifies fitting the lattice compared with a normal base distribution.

iii) We **choose a binary representation for the atom types that drastically reduces the dimensionality** compared with the simplex (one-hot). Our representation is $\lceil \log_2(100) \rceil = 7$ dimensions per atom, while the simplex requires 100 dimensions per atom. (Note that $\lceil \cdot \rceil$ denotes the ceiling function.)

iv) We **compare our method to diffusion model baselines** with extensive experiments on two realistic datasets and two simplified unit tests. Some of our results were state-of-the-art at time of submission.

Preliminaries

We review normalizing flows on manifolds of plausible materials that carry several symmetry properties.

Crystal representation

We represent a crystal $\boldsymbol{c} \coloneqq (\boldsymbol{a}, \boldsymbol{f}, \boldsymbol{l})$ as a tuple of atomic types, fractional coordinates, and unit cell.

Atom types $\boldsymbol{a} \in \mathcal{A}$ can be represented as either a tuple of n (a) h-dimensional one-hot vectors or (b) a $\{-1, 1\}$-bit representation of length $\lceil \log_2 h \rceil$. We choose (a) for CSP and (b) for DNG.

Lattice parameters $\boldsymbol{l} \in \mathcal{L}$ are a rotation invariant representation of the unit cell as a 6-tuple of three side lengths $a, b, c \in \mathbb{R}^+$ with units of Å and three internal angles $\alpha, \beta, \gamma \in [60°, 120°]$ in degrees.

Fractional coordinates $\boldsymbol{f} \in \mathcal{F} = [0,1)^{3 \times n}$ are a representation of the motif of atoms inside the unit cell. One can recover the Cartesian coordinates $\boldsymbol{x}$ within the unit cell with a matrix representation of the unit cell $\tilde{\boldsymbol{l}}$ and $\boldsymbol{x} = \tilde{\boldsymbol{l}} \boldsymbol{f}$. The volume of a unit cell $\mathrm{Vol}(\tilde{\boldsymbol{l}}) \coloneqq |\det \tilde{\boldsymbol{l}}|$ must be nonzero, implying that $\tilde{\boldsymbol{l}}$ is invertible.

Flow Matching on Manifolds

We are interested in working with flat, smooth, connected Riemannian manifolds $\mathcal{M}$ that "wrap-around."

Every $m \in \mathcal{M}$ has an associated *tangent space* $\mathcal{T}_m\mathcal{M}$ with an inner product $\langle u, v \rangle$ for $u, v \in \mathcal{T}_m\mathcal{M}$. These define minimum length curves (*geodesics*). Time-dependent vector fields on the manifold $u_t \in \mathcal{U}$ assign a vector to every time and point in the tangent bundle. We learn distributions by estimating elements of $\mathcal{U}$.

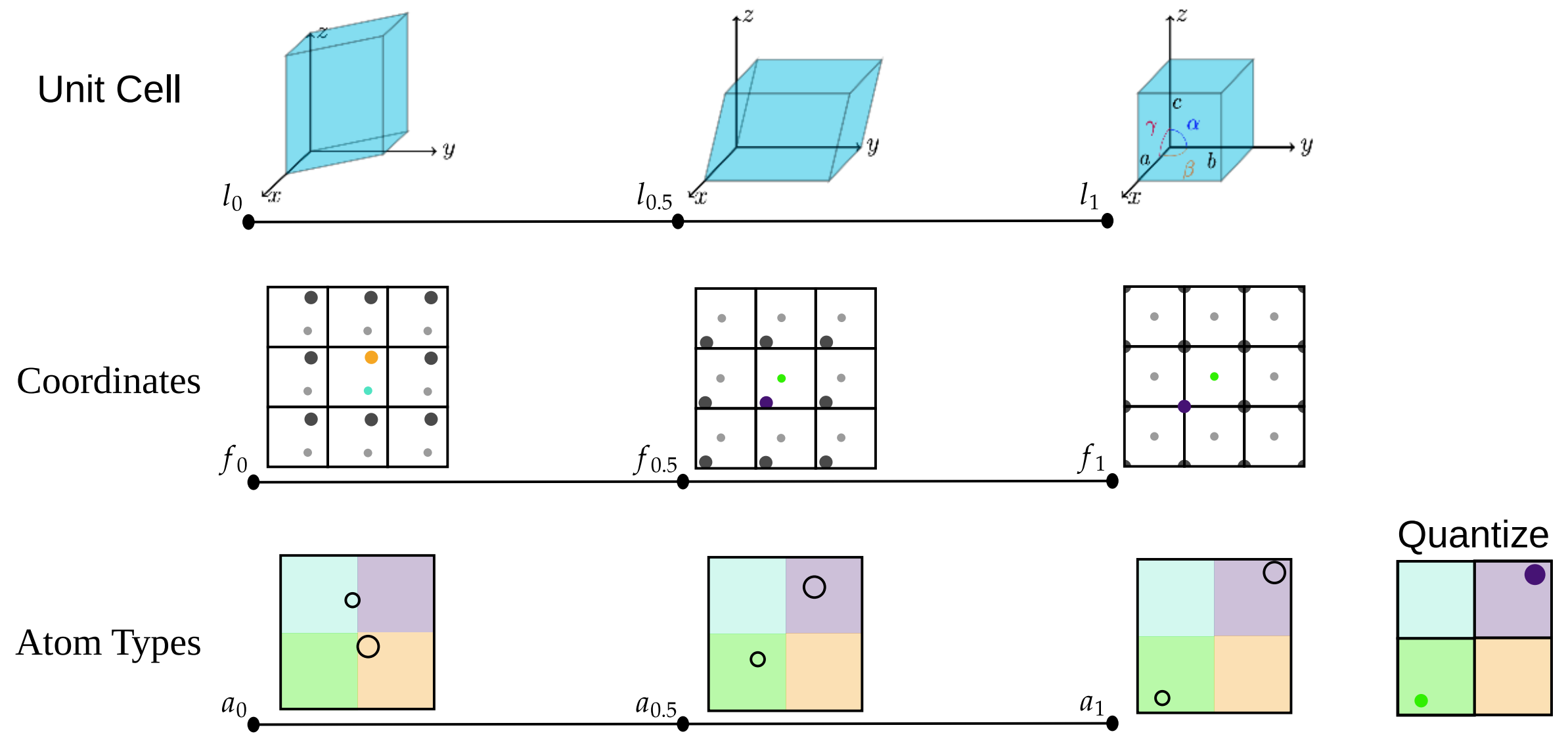
Probability densities on $\mathcal{M}$ are continuous functions $p \colon \mathcal{M} \to \mathbb{R}^+$ where $\int_{\mathcal{M}} p(m)\, dm = 1$ and $p \in \mathcal{P}$. We create new probability densities from existing ones by transforming them with a *flow* $\psi_t \colon [0,1] \times \mathcal{M} \to \mathcal{M}$, a time-dependent diffeomorphism solving: $\frac{d}{dt}\psi_t(m) = u_t(\psi_t(m))$, with initial conditions $\psi_0(m) = m_0$ and $u_t \in \mathcal{U}$. Continuous Normalizing Flows estimate the u_t that pushes base density p_0 forward to target density p_1.

Fitting a vector field $v_t^\theta \in \mathcal{U}$ with parameters θ requires regression on conditional vector fields u_t that are known *a priori* to generate p_t, on average. This is known as Flow Matching. The relevant objective is:

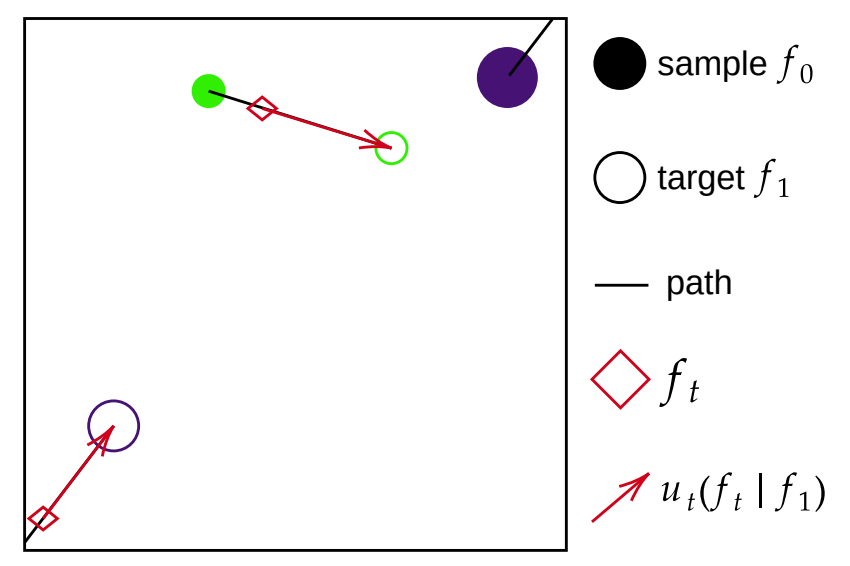
$$\mathfrak{L}(\theta) = \mathbb{E}_{t, p_1(m_1), p_t(m \mid m_1)} \| v_t^\theta(m) - u_t(m \mid m_1) \|^2,$$

where the $\|\cdot\|$ norm is induced by inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{T}_m\mathcal{M}$ and $t \sim$ Uniform$(0,1)$. At optimum, v_t^θ pushes p_0 forward to p_1. At inference, we sample p and propagate t from 0 to 1 using our estimated v_t^θ.

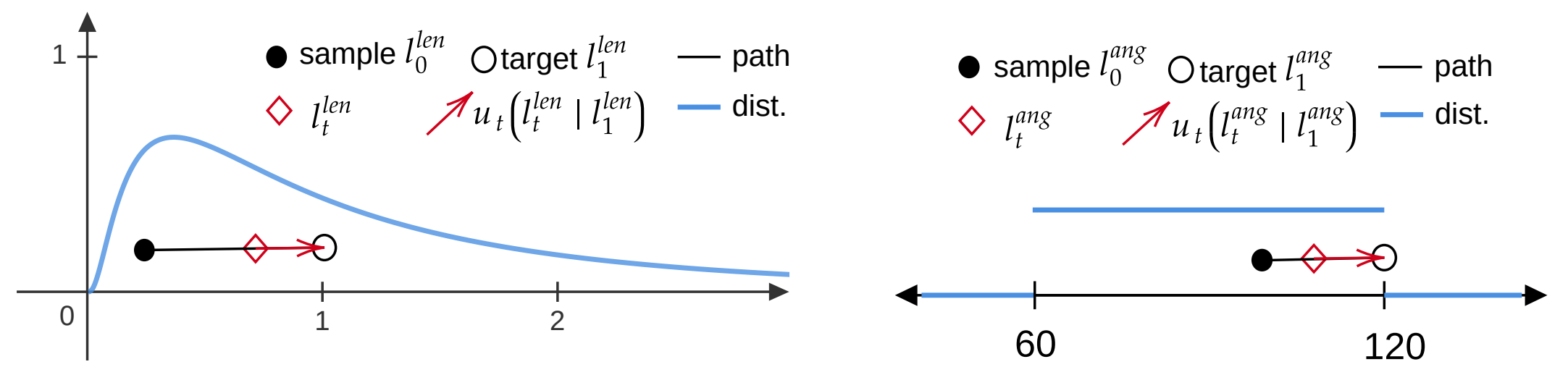
FlowMM



FlowMM is learned mapping from base distribution to target that optimizes a generalized version of Riemannian Flow Matching for labeled point clouds in periodic boundary conditions, chooses a base distribution that has support only on "physical" unit cells, and uses a graph neural network equivariant to isometries of crystals. In DNG, a binary representation reduces atomic type dimension. The learned density is invariant to isometries of crystals.



Hypothetical Riemannian Flow Matching regression target for fractional coordinates. The shortest path utilizes *periodic boundary conditions.*



Hypothetical Flow Matching regression target for lattice parameters (length & angles). The target is rotation and translation invariant. We chose (and ablated) custom base distributions for flow matching, and transformed the bounded target into "unconstrained space" using the logit function.

Lattice Parameters with Bespoke Base Distribution

Lattice parameters define a parallelepiped and are invariant to rotation. We select a positive base distribution (exponential) for lengths $\mathbb{R}^{+3}$, fit with maximum likelihood. The angles $[60, 120]^3$ often lie on the boundary. We send the boundaries to infinity with φ.

$$\text{logit}(\xi) \coloneqq \log \frac{\xi}{1-\xi}, \qquad \varphi(\eta) \coloneqq \text{logit}\left(\frac{\eta - 60}{120}\right),$$

$$\mathfrak{S}(\xi') = \frac{\exp(\xi')}{1 + \exp(\xi')}, \quad \varphi^{-1}(\eta') = 120\, \mathfrak{S}(\eta') + 60,$$

Generalized Riemannian Flow Matching for Point Clouds in Periodic Boundary Conditions

The supervision are conditional geodesic paths terminating at the target sample using the atom wise application of the $\log$ and $\exp$ maps.

That produces an equivariant–*not invariant*–target conditional vector field of $\frac{-\log_{\boldsymbol{f}_t}(\boldsymbol{f})}{1-t}$. **We remove the average torus translation** from $\boldsymbol{f}_1$ to $\boldsymbol{f}$, or "drift."

$$\exp_{f^i}(\dot{f}^i) \coloneqq f^i + \dot{f}^i - \lfloor f^i + \dot{f}^i \rfloor$$

$$\log_{f_0^i}(f_1^i) \coloneqq \frac{1}{2\pi} \operatorname{atan2}\left[\sin(\omega^i), \cos(\omega^i)\right]$$

$$\omega^i \coloneqq 2\pi(f_1^i - f_0^i)$$

$$\text{(ours)}\; u_t^{\mathcal{F}}(\boldsymbol{f} \mid \boldsymbol{f}_1) \coloneqq \log_{\boldsymbol{f}_1}(\boldsymbol{f}) - \frac{1}{n} \sum_{i=1}^{n} \log_{f_1^i}(f^i)$$

Analog Bits for Atom Types

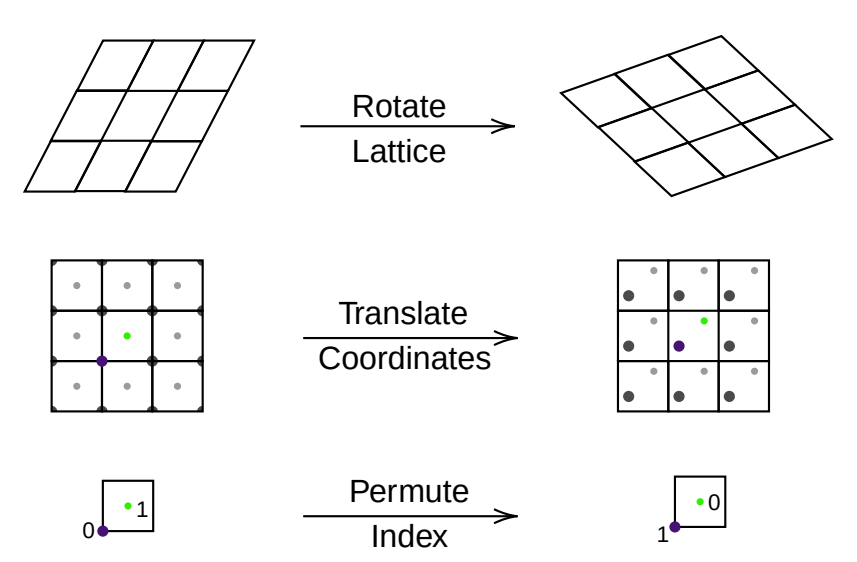
In CSP, the atomic types are only conditional information and may be considered a tuple of n, h-dimensional one-hot vectors. No density estimated.

In DNG, we learn a distribution over a categorical, binarized representation with length $\lceil \log_2 h \rceil$. The targets are made continuous and scaled and shifted to $\{-1., 1.\}^{\lceil \log_2 h \rceil}$. The flow transforms a multidimensional normal distribution to a continuous target vector. At inference time, estimates are discretized with $\text{sign} \colon \mathbb{R} \to \{-1, 1\}$. When $\lceil \log_2 h \rceil \neq \log_2 h$, we end up with "unused bits". We find that the model is able to learn to ignore these extra atom types in practice.

Neural Network

Our estimator and objective are invariant to:

- Rotation the unit cell.
- Translation of coordinates within unit cell.
- Permutation of index.



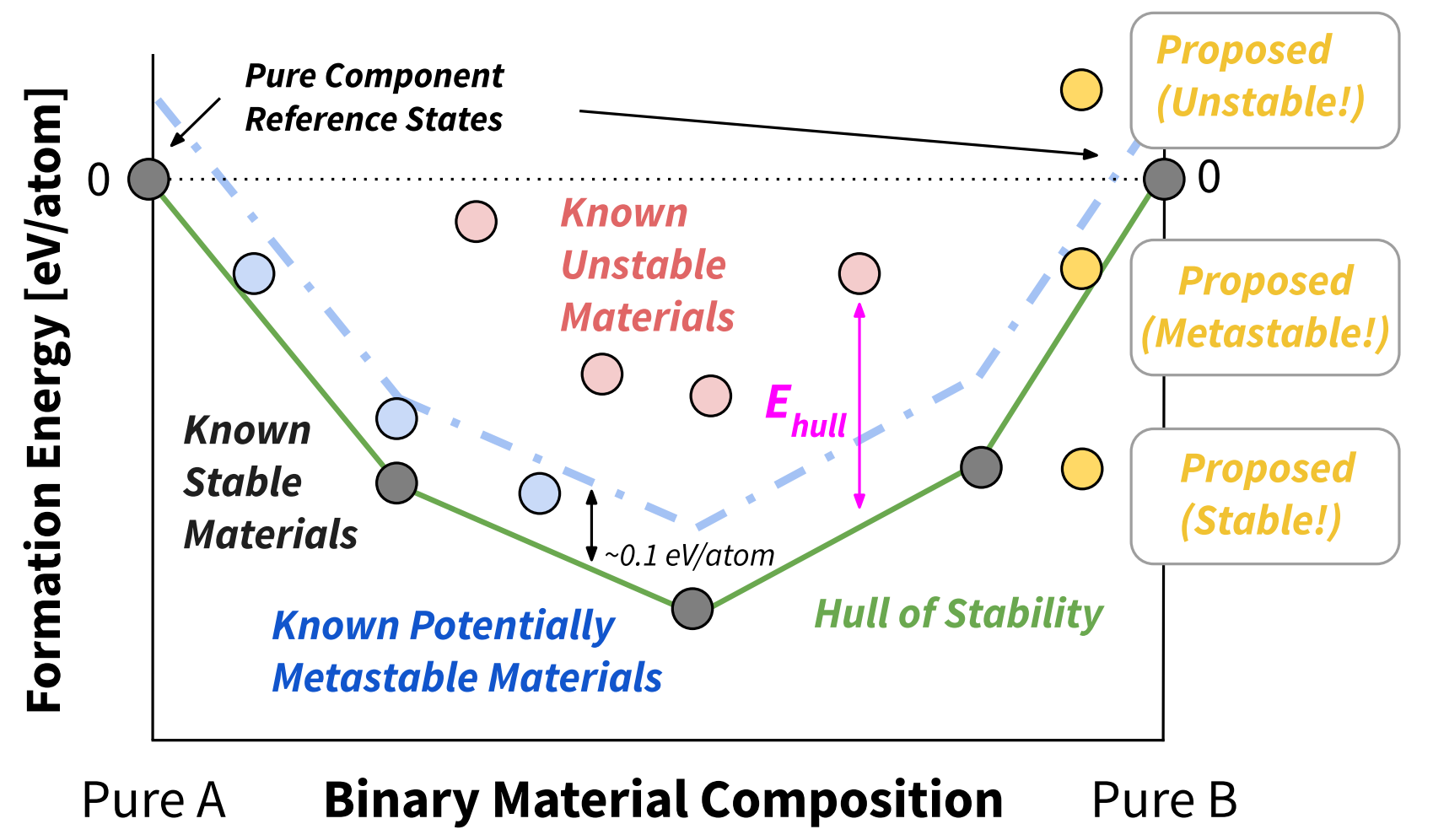
Edge features:

- Geodesic vec: SinusoidalEmbedding $\left(\log_{f^i}(f^j)\right)$
- Cosine of the angles between the Cartesian edge between atoms and the three lattice vectors $\frac{\tilde{\boldsymbol{l}}^T \tilde{\boldsymbol{l}} \boldsymbol{f}}{\|\tilde{\boldsymbol{l}}^T \tilde{\boldsymbol{l}} \boldsymbol{f}\|}$

Generating Stable Crystals

We want to generate crystals that are *stable*. Naively, stability is determined by a thermodynamic competition between a structure and competing alternatives. The known stable structures define a *convex hull* of stable compositions over the energy landscape. Structures extremely close to the hull, i.e. within $E_{\text{m}} \coloneqq 0.08$ eV/atom, we call *metastable*.

We determine the energy using a first-principles quantum mechanical method called density functional theory, which estimates the energy based on the electronic structure. Specifically, we use the default settings of the Materials Project (MP). Our convex hull references the MP database February 2023.



An explanation of the material composition convex hull.

Credit: Nate Gruver, et. al. Fine-Tuned Language Models... 2024.

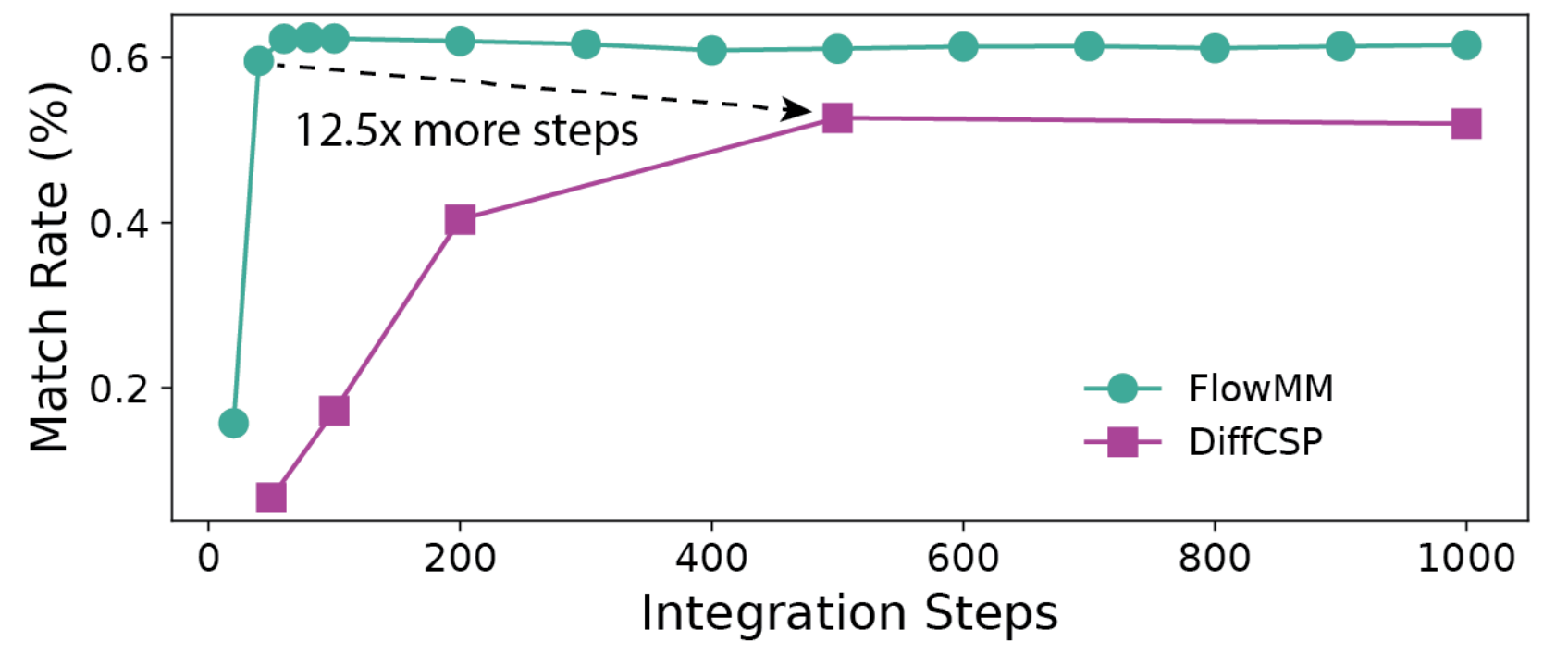
Crystal Structure Prediction

CSP aims to predict the distribution of metastable structures for given composition $q(\boldsymbol{f}, \boldsymbol{l}; \boldsymbol{a})$. Dataset:

$$\{\boldsymbol{f}' \in \mathcal{F}, \boldsymbol{l}' \in \mathcal{L} \mid E(\boldsymbol{a}_i, \boldsymbol{f}', \boldsymbol{l}') < E_{\text{m}}\}.$$

E_{m} is fixed by metastability, E is the single point energy prediction of density functional theory, $\boldsymbol{a}_i$ is the ith composition in the dataset.

Match Rate: The percentage of held-out data where $\texttt{StrutureMatcher}(\boldsymbol{c}, \boldsymbol{c}')$ returns a match (given only one FlowMM sample $\boldsymbol{c}' \sim p(\boldsymbol{f}, \boldsymbol{l} \mid \boldsymbol{a})$).
RMSE: average RMS dist. on *matching* samples.



Match rate as a function of number of integration steps on MP-20. FlowMM achieves a high maximum match rate and does so 450 steps earlier.

Results from crystal structure prediction on unit tests and realistic data sets.

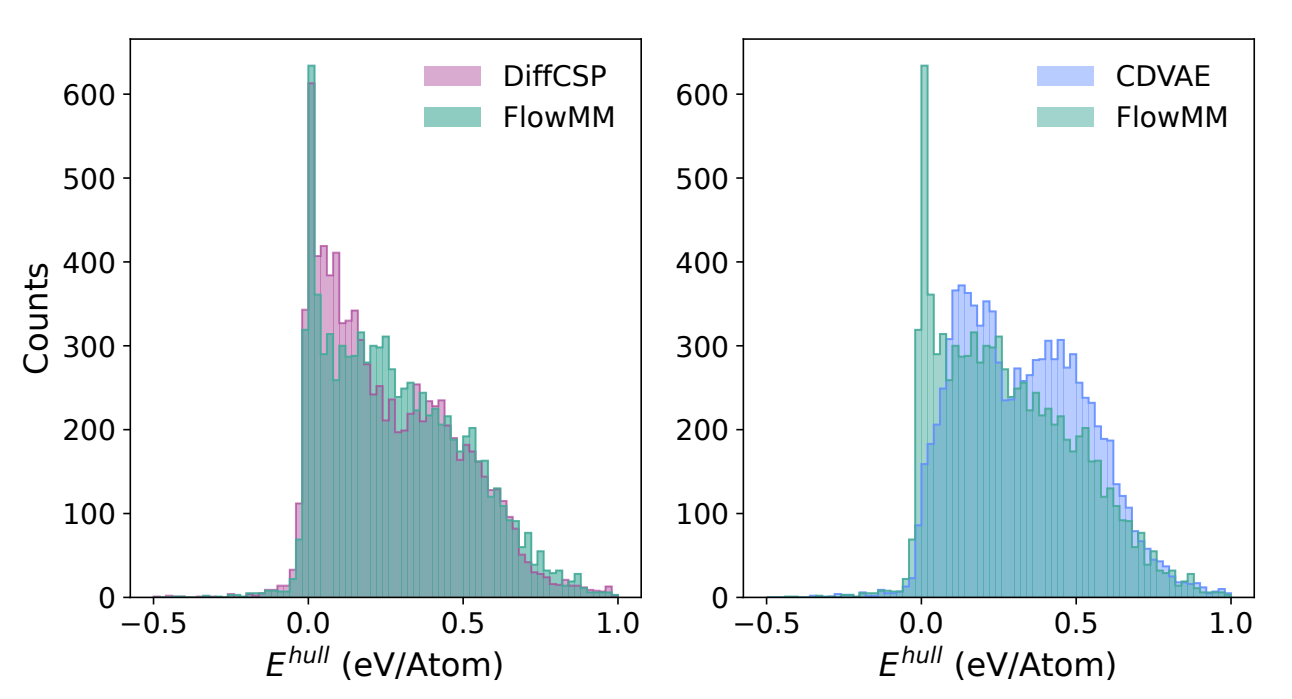
	Perov-5		Carbon-24		MP-20		MPTS-52	
	Match Rate (%) ↑	RMSE ↓	Match Rate (%) ↑	RMSE ↓	Match Rate (%) ↑	RMSE ↓	Match Rate (%) ↑	RMSE ↓
CDVAE	45.31	0.1138	17.09	0.2969	33.90	0.1045	5.34	0.2106
DiffCSP	52.02	**0.0760**	17.54	**0.2759**	51.49	0.0631	12.19	0.1786
FlowMM	**53.15**	0.0992	**23.47**	0.4122	**61.39**	**0.0566**	**17.54**	**0.1726**

De Novo Generation

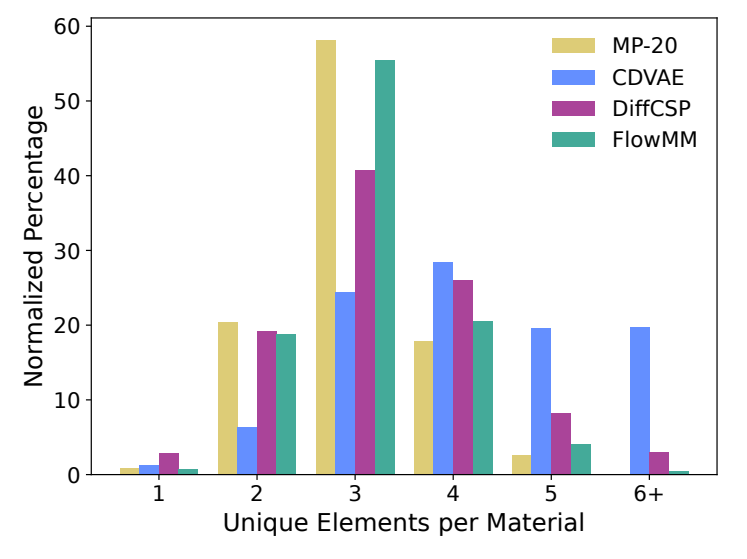
A goal of materials science is to discover stable and novel crystals. DNG aims to sample directly from a distribution of metastable materials $q(\boldsymbol{c})$, generating structure $\boldsymbol{f}, \boldsymbol{l}$ and composition $\boldsymbol{a}$. Dataset consist of metastable crystals:

$$\boldsymbol{a}_k, \boldsymbol{f}_k, \boldsymbol{l}_k \coloneqq \boldsymbol{c}_k \in \{\boldsymbol{c}' \in \mathcal{C} \mid E(\boldsymbol{c}') < E_{\text{m}}\}.$$

Compared with other methods, FlowMM accurately estimates the atomic density and the number of unique elements per crystal (N-ary / N_{el}). It also produces stable crystals with fewer integration steps.



Histogram of DFT relaxed E^{hull} for DiffCSP, CDVAE, and FlowMM. After relaxation on for all models, FlowMM generates lower energy structures compared to CDVAE and is competitive with DiffCSP.



Histogram of N-ary / N_{el} for the MP-20 distribution and the generative models. Compared to FlowMM, CDVAE and DiffCSP generate too many materials with N-ary ≥ 5.

Results from De Novo generation on the MP-20 dataset.

Method	Integration	Validity (%) ↑		Coverage (%) ↑		Property ↓		Stability Rate[†] (%) ↑	Cost ↓	S.U.N. Rate ↑	S.U.N. Cost ↓
	Steps	Structural	Composition	Recall	Precision	wdist (ρ)	wdist (N_{el})	MP-2023	Steps/Stable[†]	MP-2023	Steps/S.U.N.
CDVAE	5000	**100.00**	**86.70**	99.15	99.49	0.688	0.278	1.57	31.85	1.43	34.97
DiffCSP	1000	**100.00**	83.25	**99.71**	**99.76**	0.350	0.125	**5.06**	1.98	**3.34**	2.99
FlowMM	250	96.58	83.47	99.48	99.65	**0.261**	**0.107**	4.32	**0.58**	2.38	**1.05**
	500	96.86	83.24	99.38	99.63	**0.075**	**0.079**	4.19	**1.19**	2.45	**2.04**
	750	96.78	83.08	99.64	99.63	**0.281**	**0.097**	4.14	**1.81**	2.22	3.38
	1000	96.85	83.19	99.49	99.58	**0.239**	**0.083**	4.65	2.15	2.34	4.27

Stable[†] implies $E^{hull} < 0.0$& N-ary ≥ 2.