# CrystalREPA: Transferring Physical Priors from Universal MLIPs to Crystal Generative Models

Chengqian Zhang[1,2] Yucheng Jin[1] Duo Zhang[1,2]
Tiejun Li[2,3,4,*] Han Wang[5,6,*]

[1] AI for Science Institute, Beijing, China
[2] Center for Data Science, Peking University, Beijing, China
[3] LMAM and School of Mathematical Sciences, Peking University, Beijing, China
[4] Center for Machine Learning Research, Peking University, Beijing, China
[5] National Key Laboratory of Computational Physics,
Institute of Applied Physics and Computational Mathematics, Beijing, China
[6] HEDPS, CAPT, College of Engineering, Peking University, Beijing, China

[*] Correspondence to: tieli@pku.edu.cn, wang_han@iapcm.ac.cn

Code: github.com/Chengqian-Zhang/CrystalREPA

# Overview

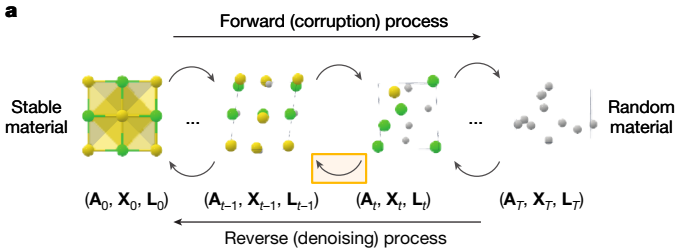
1 Background & Motivation

2 CrystalREPA

3 Results & Insights

# Crystal generation proposes new atomic structures

A crystal repeats a unit cell of atoms in space.




MatterGen: diffusion-based crystal generation.

Zeni et al., *Nature* (2025)

**What is generated?**

**Atom types**
Which elements?

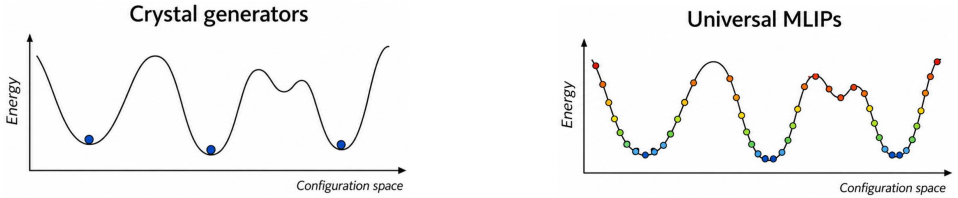
**Atom positions**
Where are the atoms?

**Cell geometry**
What cell shape and size?

**Materials discovery needs novel, energetically viable candidates.**

# The discovery goal and training signal differ

**Discovery goal:** new, stable crystals.




Training-data coverage (schematic)

### Structure reconstruction

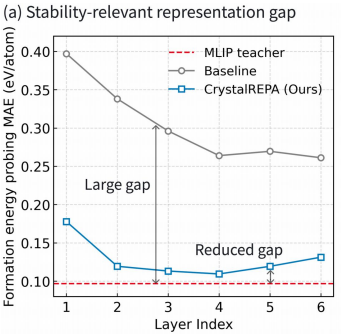
Recover atom types, positions, and cell geometry, typically without explicit energy or force targets.

### Energy and force supervision

Machine learning interatomic potentials (MLIPs) learn from energy and force labels across diverse configurations.

**MLIP representations could provide complementary supervision.**

# CrystalREPA narrows a representation gap




Formation-energy probe; lower error is better.

### A representation probe

Predict formation energy from representations of clean crystals.

### Before alignment

The baseline generator has higher prediction error than the MLIP teacher.

### With CrystalREPA

Lower probe error after representation alignment.

**CrystalREPA aligns generator representations with a frozen MLIP teacher.**

# Overview

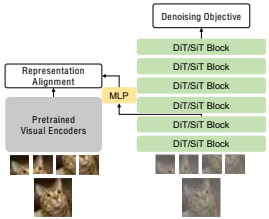
1 Background & Motivation

2 CrystalREPA

3 Results & Insights

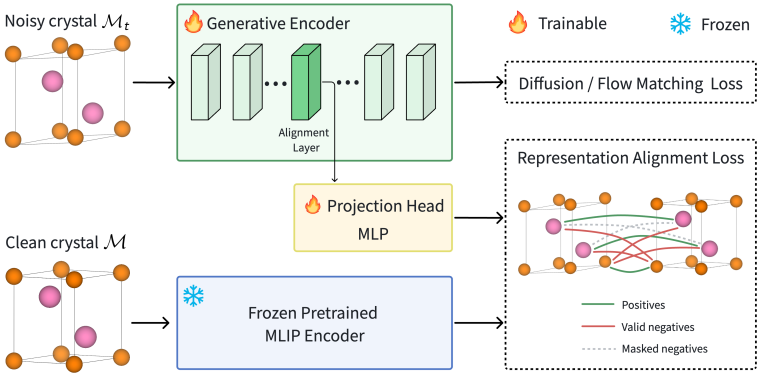
# Transfer pretrained representations during training

## Inspired by REPA



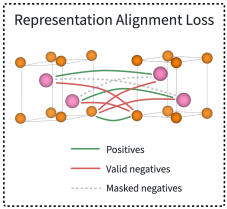

Yu et al., ICLR 2025
Representation Alignment for Generation

## CrystalREPA




**After offline precomputation: ∼1% extra training time per epoch; no extra inference cost.**

# Element-aware contrastive alignment respects chemistry




**Positive pairs:** the same atom in student and teacher.

**Negatives:** different elements; mask other same-element atoms.

$$M_{ij} = \begin{cases} 0, & i \neq j \text{ and } A_i = A_j, \\ 1, & \text{otherwise.} \end{cases}$$

$A_i$: element of atom $i$; diagonal positive pairs remain unmasked.

## Element-Aware InfoNCE (EA-NCE)

$$\mathcal{L}_{\mathrm{REPA}} = -\frac{1}{2N_B} \sum_{i=1}^{N_B} \left[ \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N_B} M_{ij} \exp(S_{ij}/\tau)} + \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N_B} M_{ji} \exp(S_{ji}/\tau)} \right].$$

$S_{ij} = \dfrac{z_i^\top \tilde{z}_j}{\|z_i\| \|\tilde{z}_j\|}$: student–teacher cosine similarity; $N_B$: atoms in the batch; $\tau$: temperature.

$$\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{REPA}} \mathcal{L}_{\mathrm{REPA}}$$

# Overview

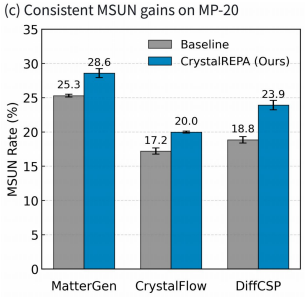
1 Background & Motivation

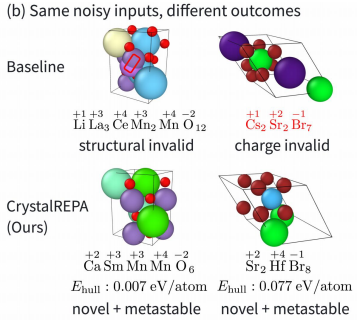
2 CrystalREPA

3 Results & Insights

# CrystalREPA improves useful crystal yield

MSUN rate: fraction of metastable (low-energy), unique, and novel candidates.




Dataset: MP-20 crystals (up to 20 atoms/cell).

Shared MLIP teacher: DPA-3.1-3M.

**MSUN gains across 3 generators, 10 MLIP teachers, and 2 crystal datasets.**

# DFT confirms gains in low-energy crystal yield

Density functional theory (DFT): 200 candidates per method.

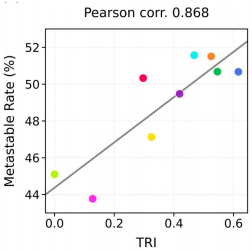
| Metric | Vanilla CrystalFlow | + CrystalREPA |
|---|---|---|
| Metastable (%) ↑ | 50.0 | **64.0** |
| MSUN (%) ↑ | 15.0 | **22.5** |
| Stable (%) ↑ | 11.0 | 11.0 |
| SUN: stable, unique, novel (%) ↑ | 1.5 | 1.5 |
| Mean energy above hull (eV/atom) ↓ | 0.149 | **0.114** |
| Structural displacement (RMSD) ↓ | 0.51 | **0.44** |
| Valid (%) ↑ | 89.5 | **90.0** |
| Unique (%) ↑ | 100.0 | 100.0 |
| Novel (%) ↑ | **61.5** | 55.5 |

**DFT relaxation starts from generated structures, with no prior MLIP relaxation.**

# Teacher representations matter beyond benchmark accuracy

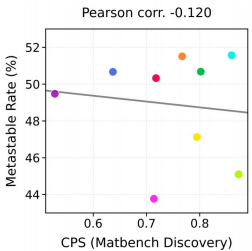
Each point is one of nine MLIP teachers.

**Representation resolution**




TRI: Total Resolution Index
Chemical + local-environment distinction

**Benchmark accuracy**




CPS: Combined Performance Score
Matbench Discovery benchmark

## Teacher selection

Inspect how well teacher representations distinguish atoms and their environments.

Thank you