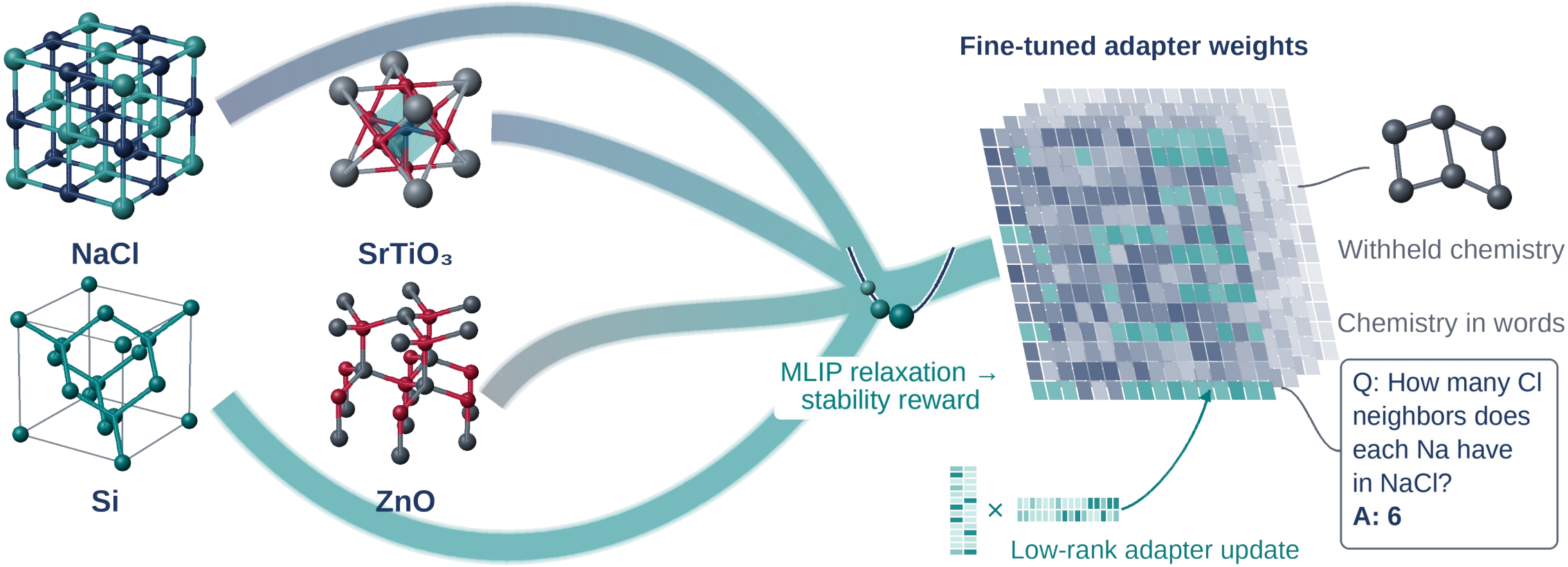
# Chemical intuition through physics practice

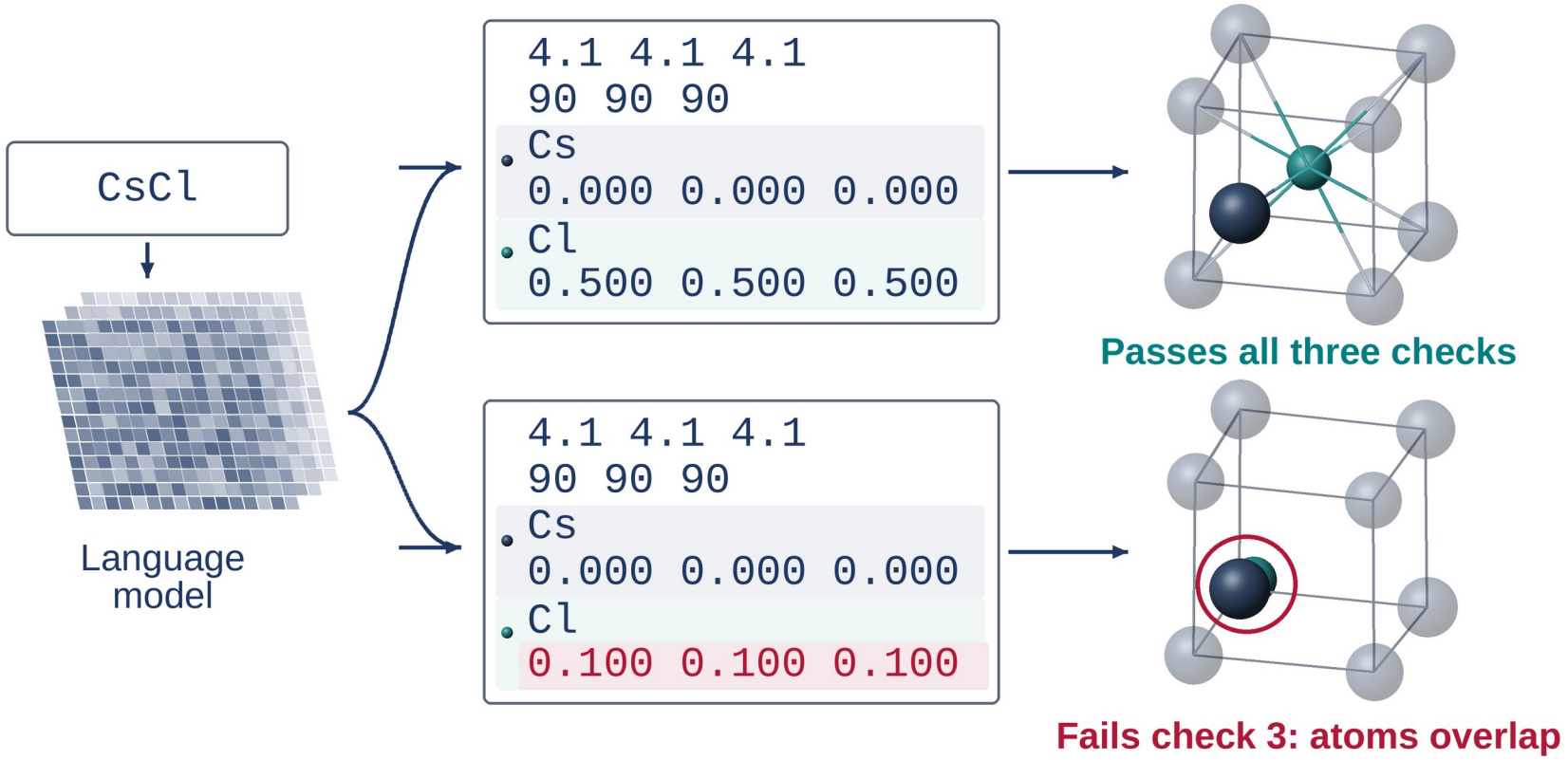
Sean Florez · CSCI 7000 · 7 October 2026




Crystal structures to scale; weights illustrative.

**Simulated stability rewards train a language model to propose stable crystals.**

# Writing a crystal as text




## Basic checks

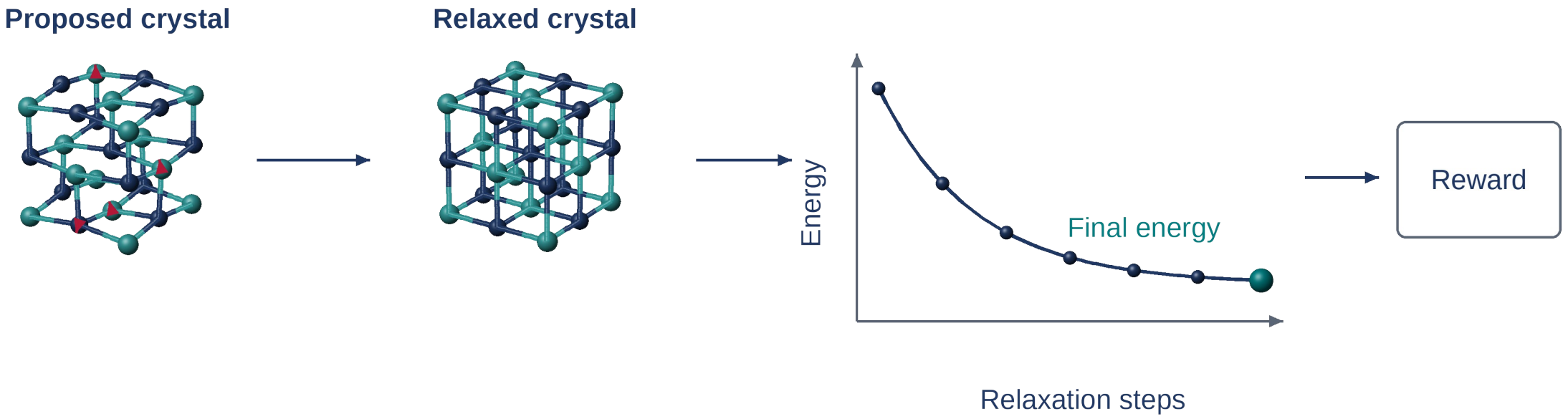
1. The text reads as a crystal

2. Elements and ratios match the formula

3. No two atoms overlap

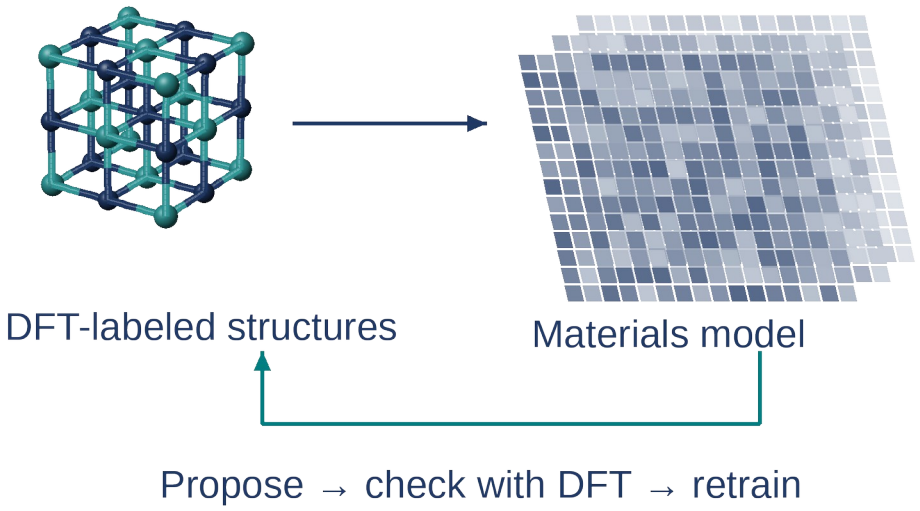
Text format: Gruver et al., 2024

# MLIP relaxation sets the reward

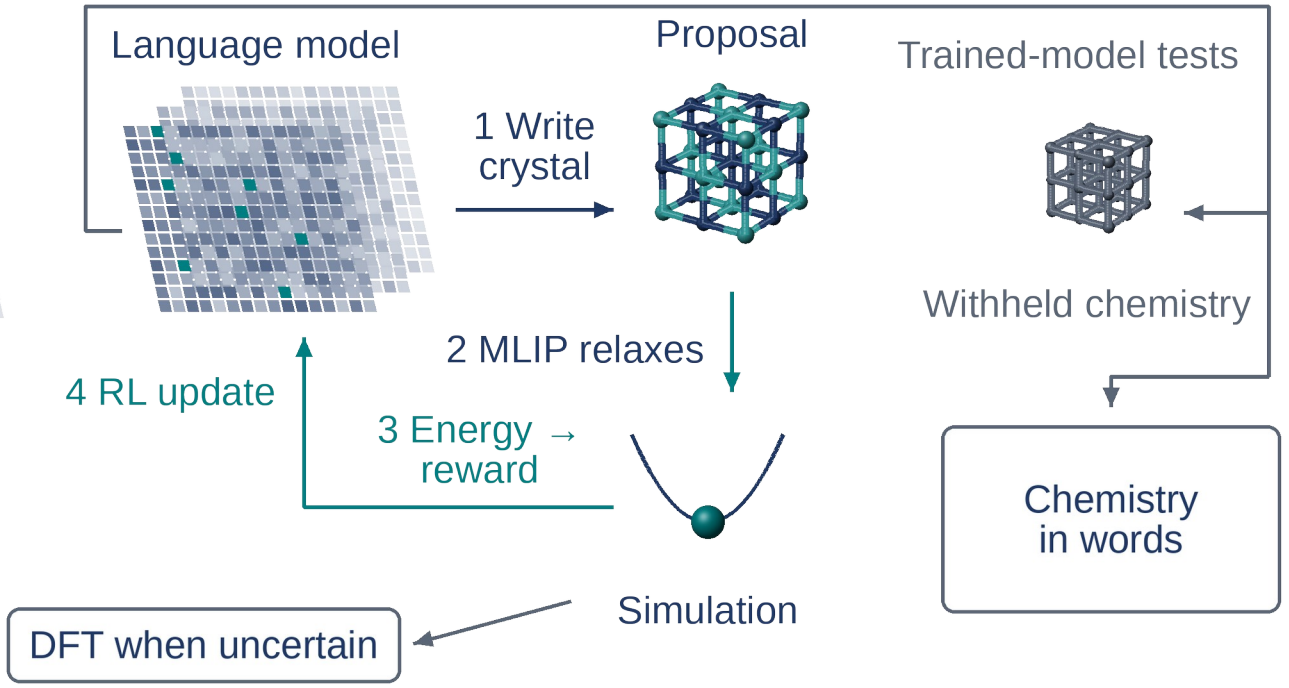


# Practicing against simulations

## Where we are



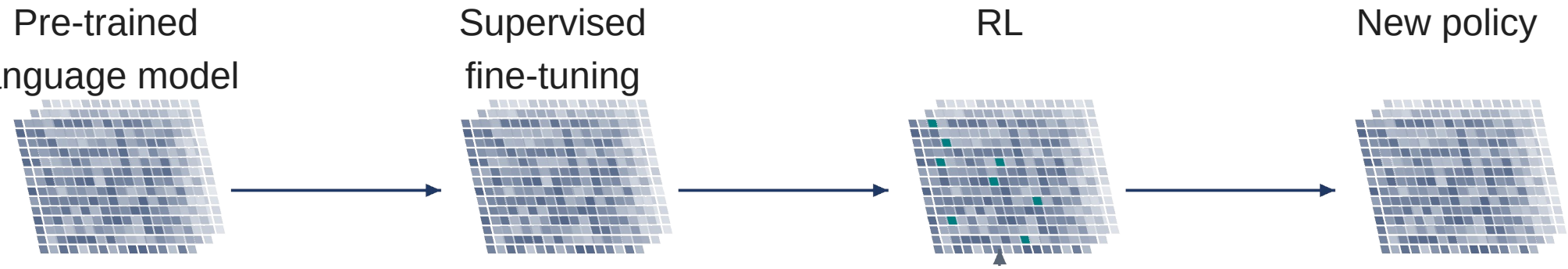

## Where we are going



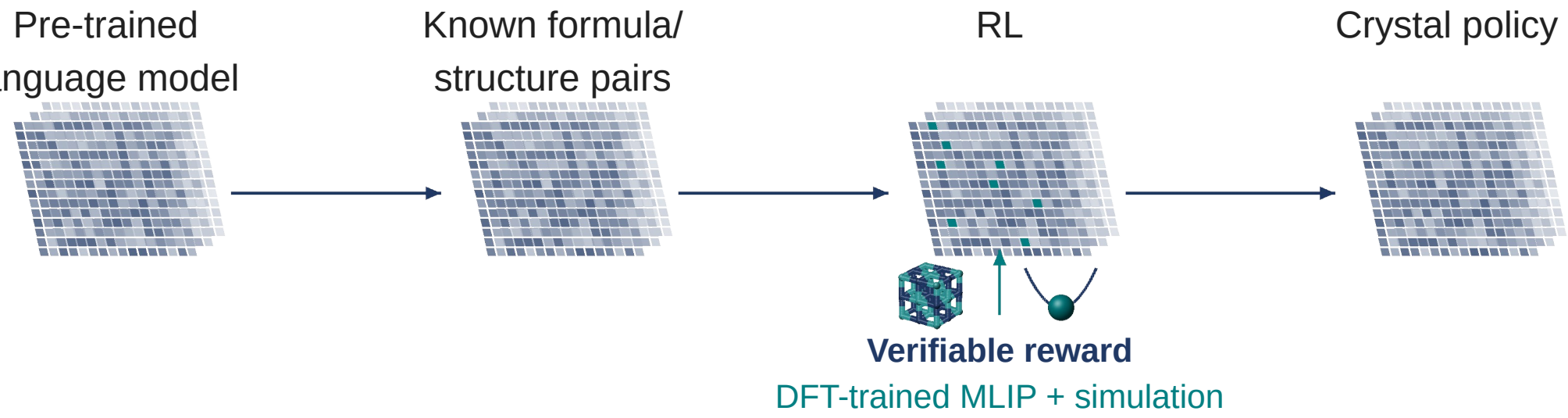

Florez, Presentation 1; Wu et al., 2026; Xu et al., 2026

# A physics reward replaces human preferences

## Human-feedback training




## Crystal training



# The reward and the update

**Reward**

$$r = -(E - E_c^*) \quad \text{if checks pass and relaxation succeeds}$$

$$r = -10 \quad \text{otherwise}$$

$E$ MLIP energy per atom, after relaxation

$E_c^*$ lowest known energy for formula c

**Update**

$$\theta \leftarrow \theta + \eta \nabla_\theta \left[ (r - \bar{r}_c) \log \pi_\theta(y \mid c) - \beta D(\pi_\theta \| \pi_0) \right]$$

$\theta$ adapter weights

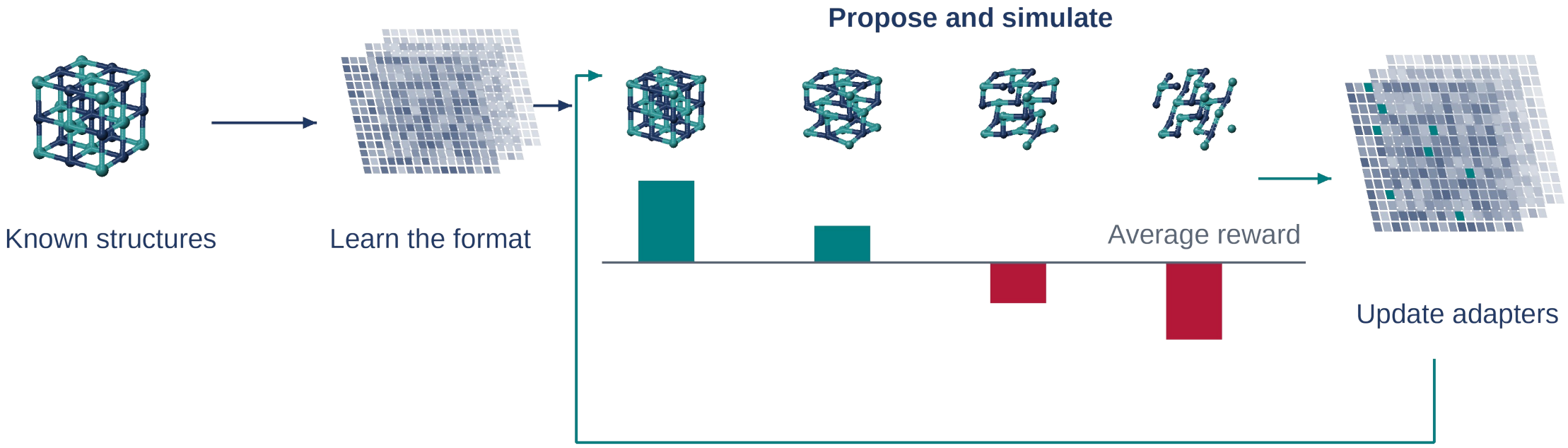
$y$ crystal text

$\bar{r}_c$ average reward for formula c

$\pi_0$ fine-tuned model, frozen

$\beta D$ drift penalty

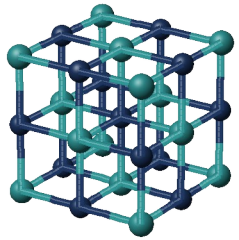
# The practice loop



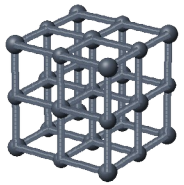

Gruver et al., 2024; Shao et al., 2024; Wu et al., 2026

# Three independent tests

Original model | Supervised fine-tuning only | Supervised fine-tuning + RL



**Practiced chemistry**



Chemistry withheld from RL

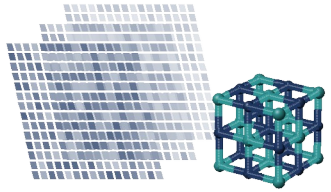
Chemistry question

**Chemistry in words**

Yue et al., 2025; Kirk et al., 2024; Yang et al., 2024; Cavanagh et al., 2026; Bran et al., 2026

# Research plan

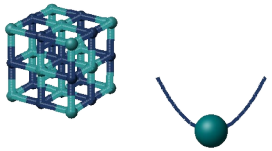
1



**Format baseline**

Valid crystal text

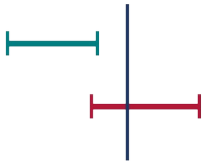
2



**Simulation reward**

Relaxed stability energy

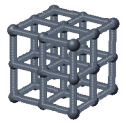
3



**Calibrated RL**

Uncertain cases to DFT

4




**Independent tests**

Structures and words

**Can physics rewards teach intuition that carries over to withheld chemistry and to chemistry in words?**

# Questions for discussion

1. What evidence would convince you that a model has learned chemical intuition rather than recalled known crystals?

2. When a learned simulator supplies the reward, how far should we trust it, and which crystals deserve DFT checks?

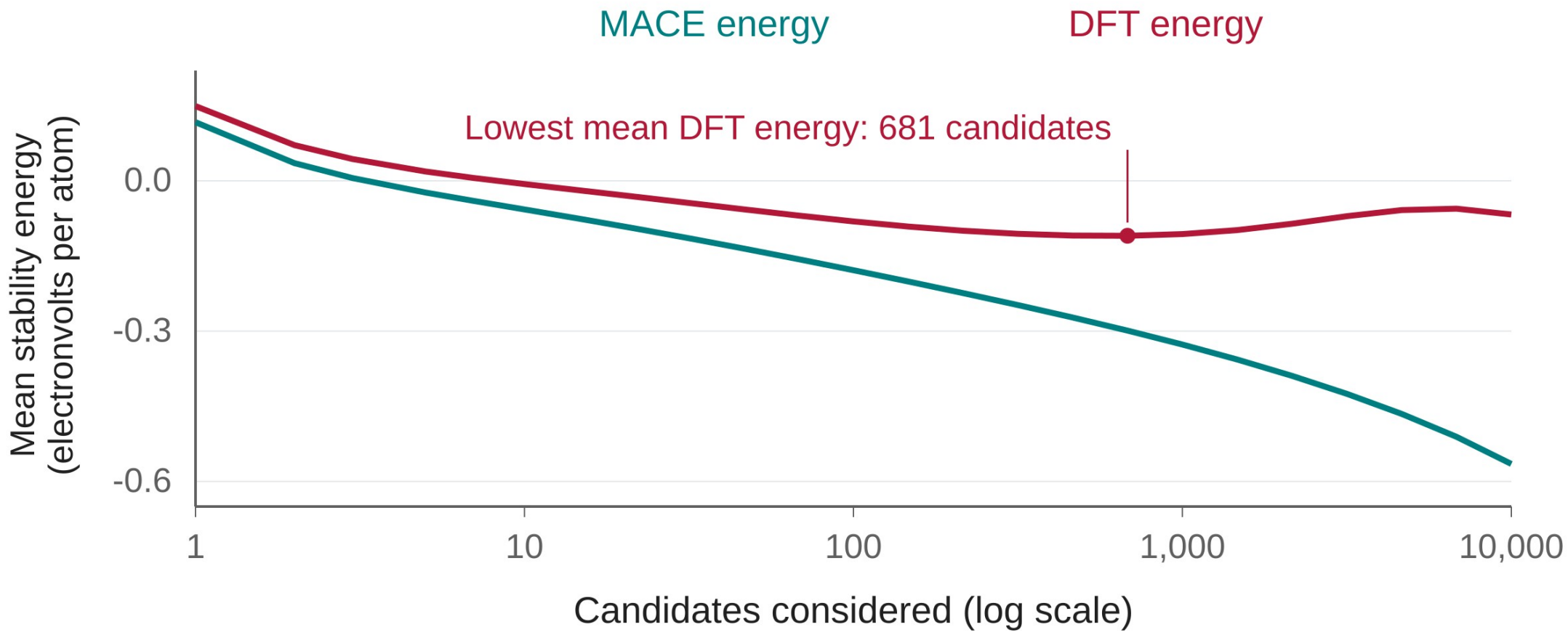
3. Should practice on crystal structures change how a model answers chemistry questions in words?

4. Does your field have a simulator or checker that could reward a model this way?

# Supplementary slides

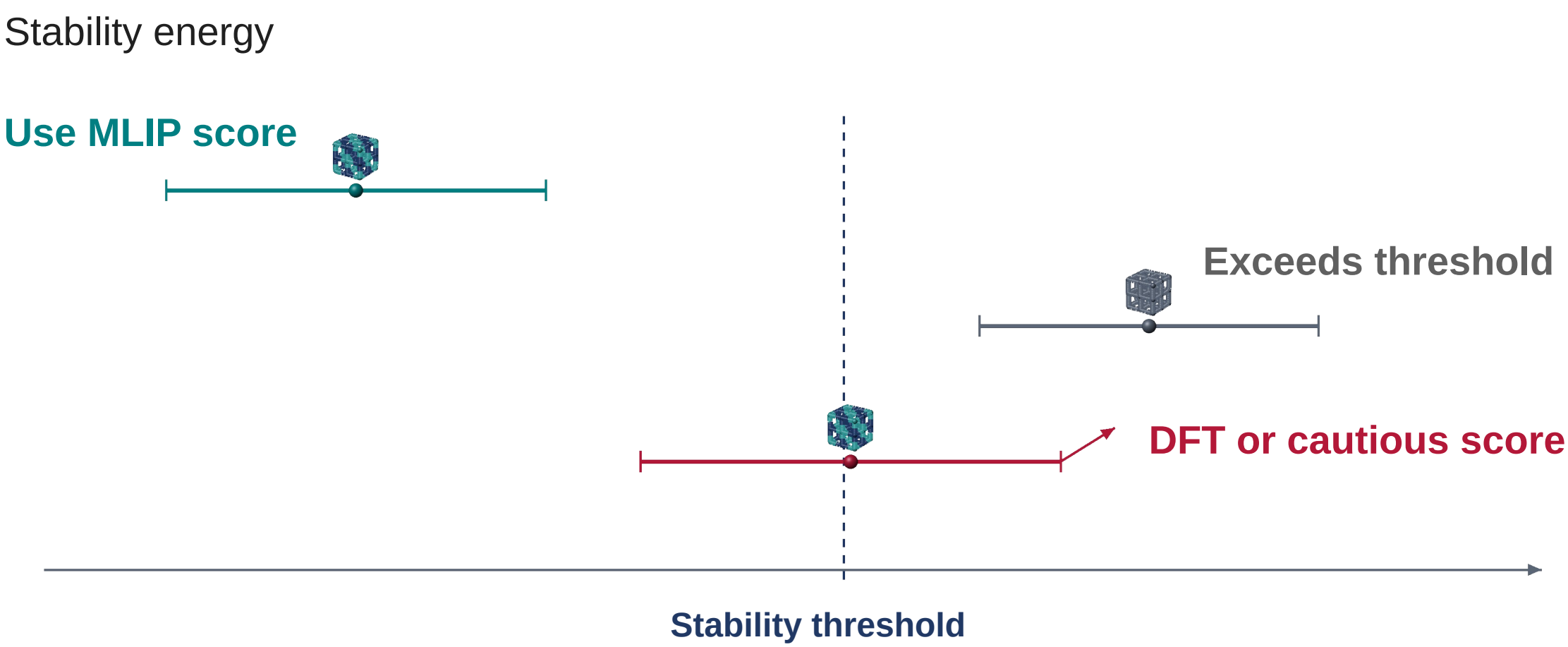
# MLIP and DFT energies diverge

Lanthanide and actinide compositions excluded




Florez analysis; Wang–Botti–Marques data; Matbench Discovery

# Uncertain crystals go to DFT




Angelopoulos & Bates, 2022; Fannjiang et al., 2022; Gui et al., 2024

# What prior work leaves open

| Ingredient | Precedent |
|---|---|
| Crystal RL | CrystalReasoner |
| Calibrated MLIP → DFT | Zaverkin et al., 2024 |
| Uncertain reward → stronger evaluator | Xu et al., 2025 |
| Conformal reward intervals | Pan et al., 2026; Lee et al., 2026 |

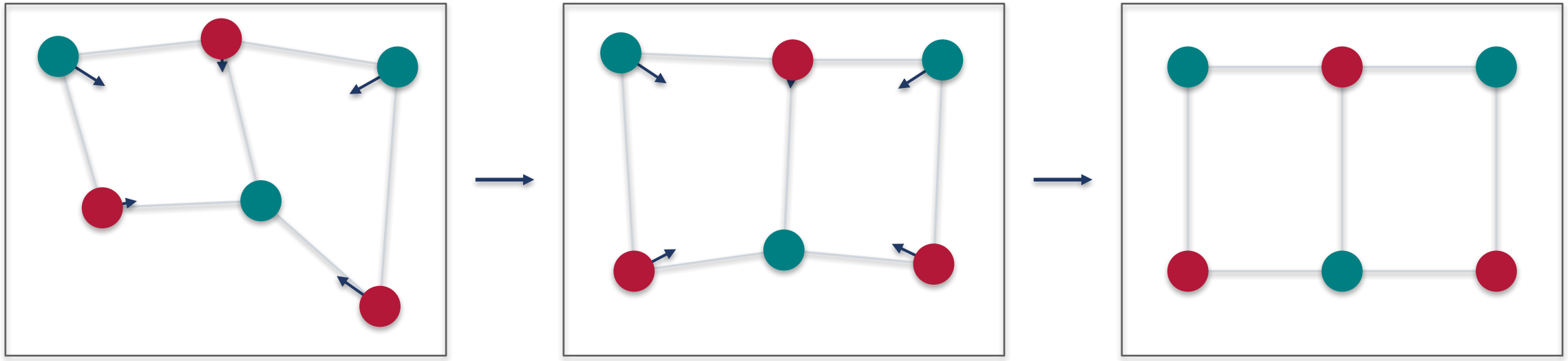
Wu et al., 2026; Zaverkin et al., 2024; Xu et al., 2025; Pan et al., 2026; Lee et al., 2026

# Relax, then score

## Proposed crystal

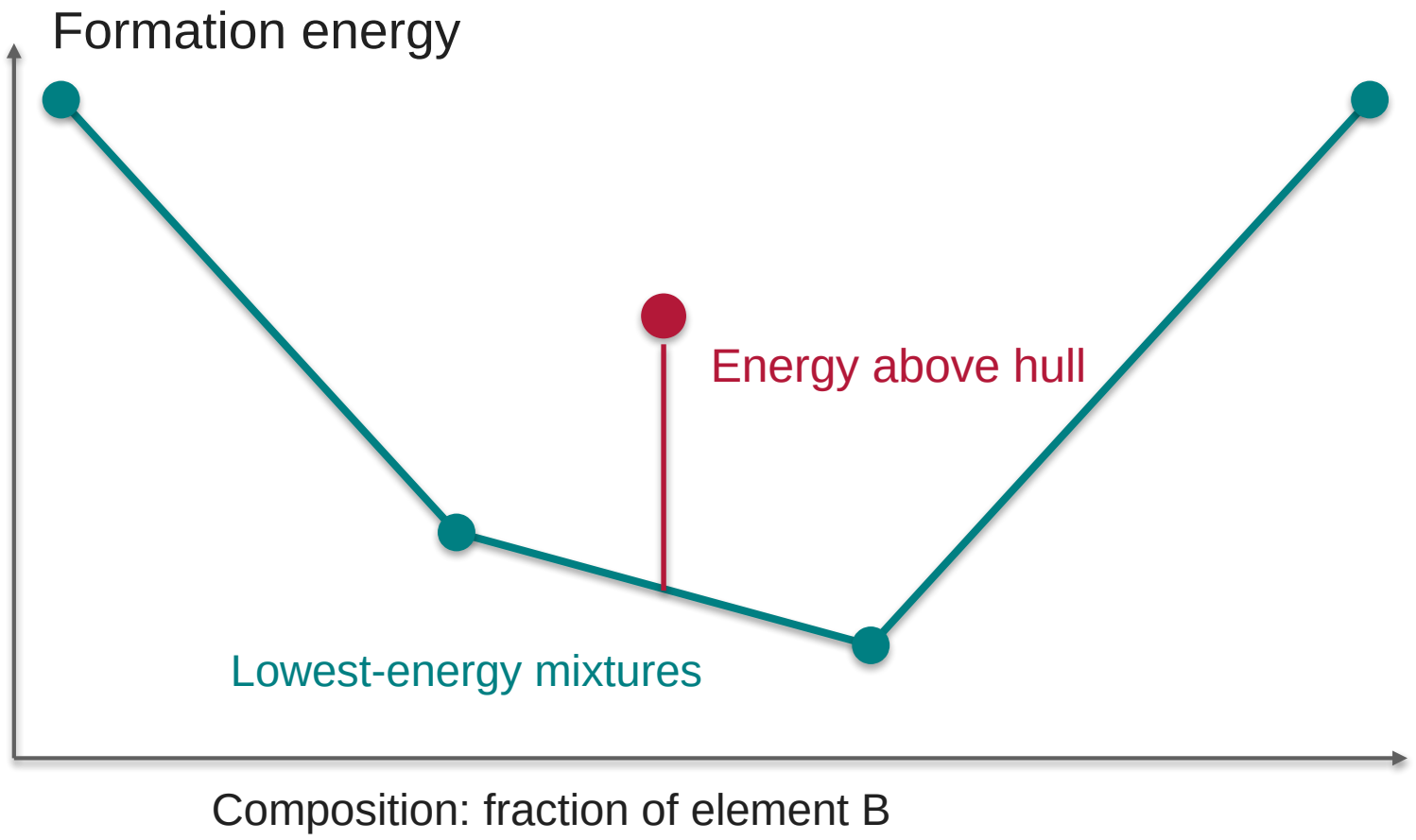
## Relaxing crystal

## Converged structure



Batatia et al., 2024; Wu et al., 2026

# Energy above the hull



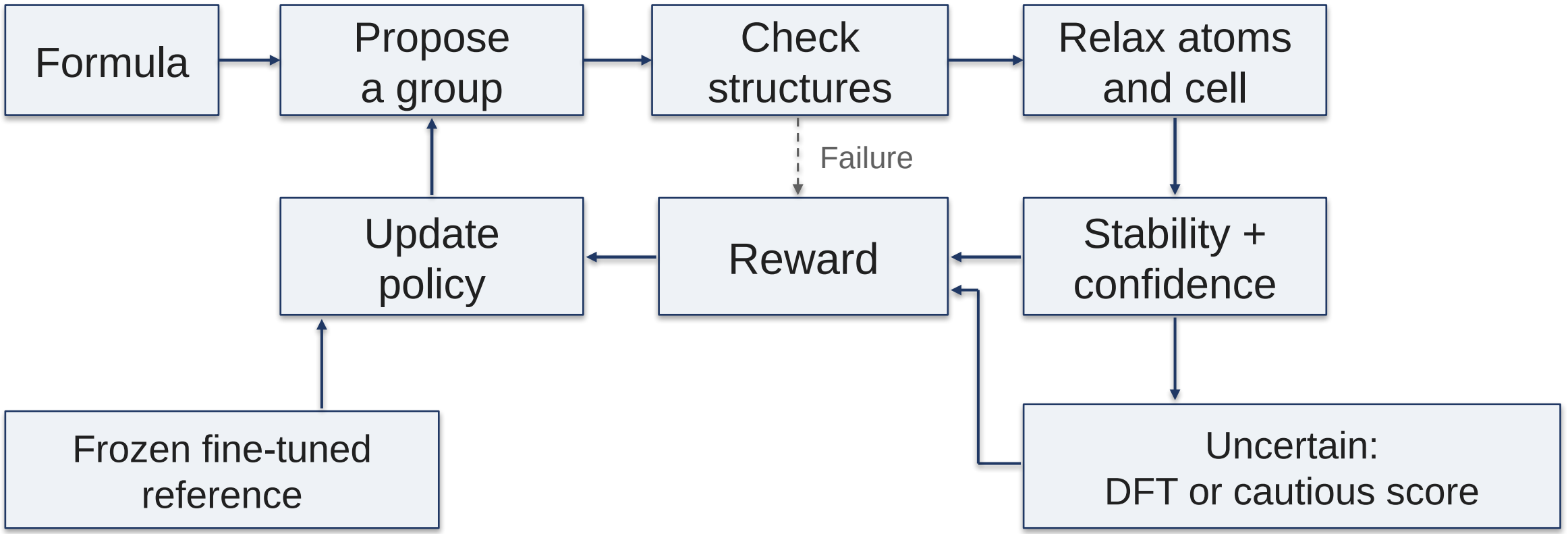

**On hull**
**0 eV/atom**

Near hull
up to 0.1 eV/atom

eV/atom = electronvolts per atom

Gruver et al., 2024, §5; Materials Project phase-diagram methodology

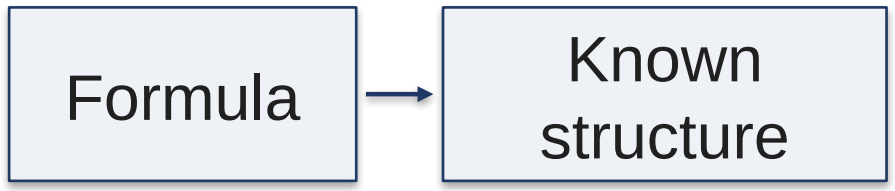
# The practice loop, in detail




Shao et al., 2024; Wu et al., 2026

# Fine-tuning first, then RL

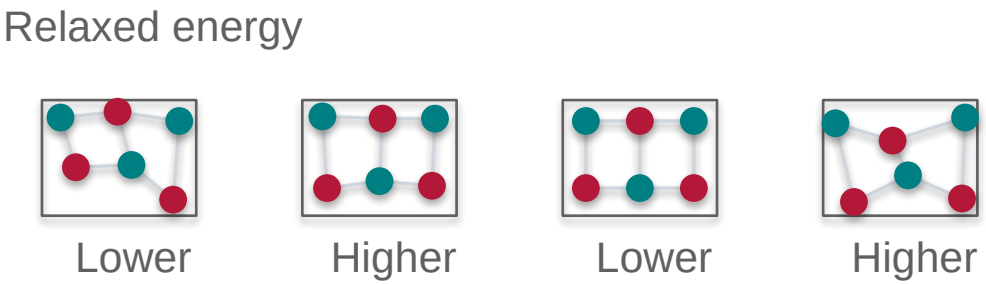
## Imitate known crystals




About 45,000 crystals overall

Qwen2.5-0.5B starting model

## Learn from simulated rewards




Group Relative Policy Optimization (GRPO)

$$A_i = (r_i - r_{mean}) / (s_r + \delta)$$

$r_i$: proposal reward; $r_{mean}$: group mean

$s_r$: group standard deviation; $\delta$: small positive constant

Gruver et al., 2024; Shao et al., 2024; DeepSeek-R1, 2025

# The full group-relative update

Maximize: $\min[\rho_{i,t} A_i, \text{clip}(\rho_{i,t}, 1 - \varepsilon_c, 1 + \varepsilon_c) A_i] - \beta d_{i,t}$

$\rho_{i,t}$: updated token probability divided by collection token probability

$A_i$: relative reward

$d_{i,t}$: Kullback–Leibler divergence from the frozen reference

$\beta$: reference-penalty weight; $\varepsilon_c$: clipping range

Reward: negative signed energy; invalid/failed output: −10

Shao et al., 2024; DeepSeek-R1, 2025; Hu et al., 2022

# Interval width can bias RL

## Calibrate normalized errors

Absolute energy error ÷ crystal difficulty → calibrated cutoff

Interval half-width = calibrated cutoff × crystal difficulty

Difficulty: disagreement among reward-side MACE models

Angelopoulos & Bates, 2022; Jin & Candès, 2023; Fannjiang et al., 2022

# What the tests measure

1. **Practiced chemistry**

   Stable, varied structures absent from training

2. **Chemistry withheld from RL**

   Exclude prompts and rewarded outputs

3. **Chemistry in words**

   Known answers; no question-answer training

**Compare checkpoints**

Original model

Supervised fine-tuning only

Supervised fine-tuning + RL

Yue et al., 2025; Kirk et al., 2024; Yang et al., 2024

# RL versus extra sampling

## Same evaluation budget

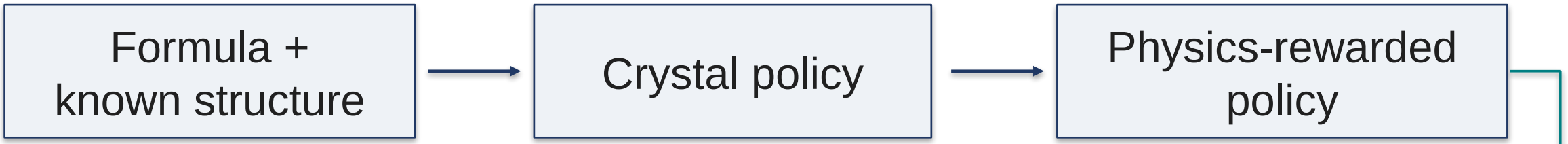
Model before RL
+ more attempts

Model after RL
+ fewer attempts

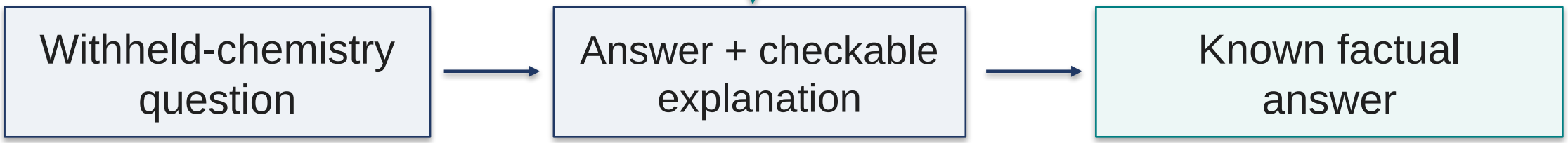
Yue et al., 2025; Liu et al., 2025; Florez selection analysis

# Transfer to chemistry in words

## Training




## Text evaluation




Cavanagh et al., 2026; Bran et al., 2026; Huan et al., 2025

# Why three checkpoints

## Earlier structure-to-text results

SmileyLlama: chemistry accuracy fell

MiST: text-ranking diagnostic improved

MiST = mid-stage scientific training

## Benchmark options

ChemBench-Mini: 236 questions

Materials science questions: 650

MMLU-Pro chemistry: 1,132

MMLU-Pro: Massive Multitask Language Understanding Pro

About 1,000 independent questions

80% power for an accuracy gain of 5 percentage points at 30% correctness disagreement (two-sided 5% significance)

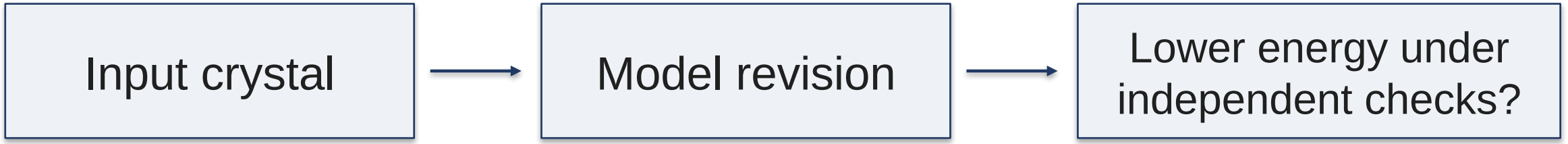
Cavanagh et al., 2026; Bran et al., 2026; ChemBench; MaScQA; MMLU-Pro; Florez analysis

# Measuring DFT and calibration costs

| Cost | What must be measured |
| --- | --- |
| Policy training and generation | Throughput and token lengths |
| Relaxation and difficulty models | All calls, including failures |
| DFT calibration and referrals | Code, runtime, reference phases |
| Independent evaluation | Random DFT sample + MatterSim |

Relaxation node-hours = relaxation count ÷ measured relaxations per node-hour

# Baseline: ordinary relaxation



# Sources

| | |
|---|---|
| **Earlier practice** | Florez, Presentation 1, 2 September 2026 |
| **Course flow** | Velasquez, Lecture 1; Ouyang et al., 2022 |
| **Crystal training** | Gruver et al., 2024; Wu et al., 2026; Xu et al., 2026 |
| **Interatomic models** | Batatia et al., 2024; Yang et al., 2024 |
| **Calibrated physics** | Zaverkin et al., 2024; Fannjiang et al., 2022 |
| **Intervals and selection** | Gui et al., 2024; Jin & Candès, 2023 |
| **Reward uncertainty** | Xu et al., 2025; Pan et al., 2026; Lee et al., 2026 |
| **RL comparisons** | Shao et al., 2024; Yue et al., 2025; Liu et al., 2025 |
| **Text transfer** | Cavanagh et al., 2026; Bran et al., 2026 |
| **Text retention** | Huan et al., 2025; Kirk et al., 2024 |
| **Existing evidence** | Florez analysis; Matbench Discovery data |
| **Chemical exclusions** | Periodic Labs report |
| **Text benchmarks** | ChemBench; MaScQA; MMLU-Pro |