

ED-CSP: Crystal Structure Prediction from Electron Diffraction

Germain Poloudenny^{1,2}, Yaël Frégier¹, Arnaud Demortière^{2,3,4}

¹Laboratoire de Mathématiques de Lens (LML), UR 2462, Université d'Artois, Lens, France

²Laboratoire de Réactivité et de Chimie des Solides (LRCS), CNRS UMR 7314, UPJV, Amiens, France

³Réseau sur le Stockage Electrochimique de l'Energie (RS2E), CNRS FR 3459, Amiens, France

⁴ALISTORE-European Research Institute, CNRS FR 3104, Amiens, France

Abstract

Recovering a periodic 3D crystal structure from sparse, unindexed detector-plane observations is a challenging generative inverse problem. Prior electron diffraction (ED) learning methods largely predict crystallographic labels, reconstruct structures from indexed reflections, or retrieve structures from finite libraries. In this paper, we consider the task of crystal structure prediction from known composition and atom count, together with multiple detector-plane ED spot sets, and introduce ED-CSP, a machine learning model that combines a relational set encoder, permutation-invariant multi-view aggregation, and a periodic flow generator to generate the lattice and fractional atomic coordinates. To train ED-CSP, we construct Electron Diffraction Crystal Structures (ED-CS), a 4.85-million-structure resource with simulated multi-view ED, deduplicated across seven materials repositories and filtered to exclude CHILI-100K matches under a fixed structural matcher. On 2,075 held-out CHILI-100K materials, CHILI-only ED-CSP achieves a structural match rate (MR) of 57.5 % at five candidates per query (MR@5), versus 52.9 % for PXRGen, a state-of-the-art crystal structure prediction (CSP) model conditioned on powder X-ray diffraction (PXRD); both use the same periodic-generator architecture but modality-specific encoders and inputs. To demonstrate that increasing the dataset size improves model performance, we warm-start the full model from a separate one-million-structure precursor and show that this raises MR@5 to 66.3 %. On 1,024 queries whose reduced formulas are absent from the train/validation retrieval library, this registry-1M-initialized model retains 53.5 % MR@5, demonstrating recovery where exact-formula lookup has no candidate. For the same model, replacing target ED observations with those from a non-isomorphic same-formula donor on 67 queries lowers mean MR@5 by 22.1 percentage points across five generation seeds, providing evidence of query-specific diffraction use. ED-CSP and ED-CS provide a controlled benchmark for generative inference from sparse simulated ED and future experimental-transfer studies.

Introduction

Many crystalline materials cannot be grown as crystals large enough for conventional single-crystal X-ray diffraction, whereas ED can collect structural signal from individual nanocrystals (Gemmi et al. 2019; Unge et al. 2025). Powder X-ray diffraction (PXRD) remains broadly accessible for bulk powders composed of randomly oriented crystallites,

making the two modalities complementary rather than interchangeable (Segal et al. 2026; Gemmi et al. 2019). Unlike the orientation-averaged PXRD profile, each ED view samples an oriented region of reciprocal space and preserves detector-plane relationships among scattering vectors. Both 3D ED and scanning diffraction experiments produce orientation-dependent reciprocal-space observations, although their acquisition geometries differ from the discrete simulated views studied here (Gemmi et al. 2019; Savitzky et al. 2021). This orientation dependence motivates a multi-view inverse problem in which geometric information is available while simulation-to-experiment transfer and incomplete orientation coverage remain explicit limitations. Figure 1 contrasts these observation regimes.

Diffraction-conditioned CSP is now established for PXRD through contrastive pretraining, diffusion and flow models, and autoregressive generation (Lai et al. 2025; Guo et al. 2025; Li et al. 2025; Johansen et al. 2026; Segal et al. 2026). In ED, however, machine learning has mainly predicted crystallographic labels or retrieved structures from finite databases (Gleason et al. 2024; Nathani et al. 2026; Peng et al. 2026). These tasks show that sparse ED patterns contain learnable structural signal, but they stop short of predicting lattice and fractional atomic coordinates from unindexed multi-view spot lists given composition. ED-CSP targets this gap with composition-conditioned crystal-geometry prediction.

ED-CSP encodes each view with a shared relational spot encoder, aggregates information across views, and jointly optimizes the resulting representation with a periodic flow generator. Known composition fixes the atom types and count, while the ED branch receives detector-plane spot coordinates and intensities.

On CHILI-100K, we use a common held-out split to evaluate signal use, registry-scale transfer, finite-library coverage, same-dataset PXRD generation, and indexed-reflection reconstruction (Friis-Jensen et al. 2024). ED input interventions and a converged composition-only reference support signal use beyond composition; registry scaling improves held-out recovery, and library-coverage stratification separates generation from finite-database lookup. The study uses simulated detector-plane ED with known composition.

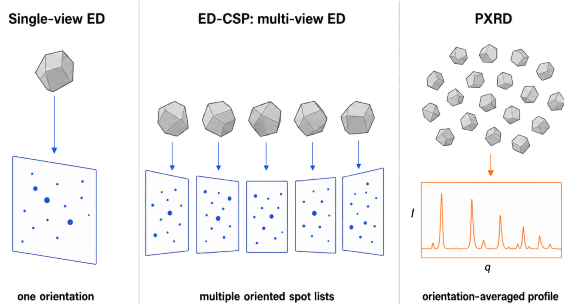


Figure 1: **Diffraction observation regimes.** ED-CSP conditions on discrete sparse ED views from multiple orientations, whereas PXRD aggregates randomly oriented crystal-lites into a one-dimensional radial profile.

Contributions. This paper makes the following focused contributions:

- A formulation of composition-conditioned CSP from unindexed sparse multi-view ED, instantiated by ED-CSP.
- Controlled ED-input interventions and a converged composition-only reference testing use of the ED conditioning signal.
- ED-CS, a CHILI-disjoint corpus of 4,852,131 deduplicated structures with simulated multi-view ED, together with registry-scale transfer and comparisons across generation, retrieval, and indexed reconstruction.

Related Work

Diffraction-conditioned generation. PXRD-conditioned CSP has progressed from establishing diffraction as a generative condition to testing which auxiliary information is available at inference and whether diffraction resolves structural ambiguity. XtalNet combines contrastive PXRD-structure pretraining with equivariant generation; PXRD-net and PXRDGen condition diffusion or flow generators on formula and PXRD, with PXRDGen additionally supporting lattice inference and Rietveld refinement; deCIFer generates crystallographic information file (CIF) sequences autoregressively (Lai et al. 2025; Li et al. 2025; Johansen et al. 2026; Guo et al. 2025). PXRDGen evaluates its contrastively pretrained XRD encoders through retrieval and then uses them, either frozen or trainable, to condition structure generation (Li et al. 2025). More recent systems emphasize experimental transfer and varying chemical or crystallographic inputs: XRDSol receives stoichiometry and unit-cell parameters, RealPXRD-Solver supports lattice-conditioned and lattice-free inference after large-scale simulated pre-training, and XRDiff evaluates full and partial composition with composition-grouped polymorph splits (Yu et al. 2026; Li et al. 2026; Segal et al. 2026). Outside diffraction-conditioned CSP, Atomistic Language Models couple a language backbone to an atomistic diffusion decoder and report strong composition-conditioned crystal recovery, emphasizing the importance of isolating diffraction-specific gains from

learned structural priors (Edamadaka et al. 2026). ED-CSP addresses the complementary setting of sparse multi-view ED spot lists.

Sparse multi-view ED learning. ED representation learning provides the closest architectural precedent. RF-ED predicts crystal systems, space groups, and lattice parameters from one or more simulated two-dimensional ED patterns, while PE-AG-GMoE processes Bragg spots as variable-size relational sets and aggregates predictions across orientations for crystallographic classification (Gleason et al. 2024; Nathani et al. 2026). Recent multi-view selected-area electron diffraction (SAED) learning fuses two views for symmetry prediction and formula-constrained retrieval from a finite structure database (Peng et al. 2026). ED-CSP changes the output from labels or database identities to lattice and fractional atomic coordinates given composition.

Indexed ED solution and refinement. Learned ED inverse methods also operate after crystallographic preprocessing. GraPhAI phases indexed three-dimensional reflection amplitudes, while hybrid physics-ML refinement optimizes an existing structural model against integrated per-HKL intensities using differentiable dynamical simulation (Melgalvis and Rekis 2026; Malik et al. 2026). Conventional continuous-rotation 3D ED likewise combines indexing and integration with structure solution and refinement (Klar et al. 2023). These methods complement ED-CSP but solve phasing or refinement after indexing, rather than generation from unindexed detector-plane spot lists.

Method

Problem Setting and Inputs

Given a composition A and K ED views $S_{1:K}$, ED-CSP models candidate lattices L and fractional coordinates F through $p_\theta(L, F \mid A, S_{1:K})$. Each view is a variable-length spot list $S_k = \{(q_x, q_y, \log(1 + I))_j\}_j$, padded only at batching time. The composition fixes the atom types and atom count, while the ED branch receives only the sampled detector-plane spot lists. Indexed Miller labels, zone-axis vectors, and crystallographic labels are not provided to the ED branch. Figure 2 summarizes the resulting pipeline.

Sparse Multi-View ED Encoder

The CHILI-100K ED-CSP runs reported here use an ED encoder adapted from the EDiffCrystals PE-AG-GMoE sparse diffraction backbone (Nathani et al. 2026). Each ED view is encoded by a shared PE-AG-GMoE-style graph-attention module over the raw py4DSTEM detector-plane spot list. The resulting per-view representations are then aggregated across the sampled ED views with a mean-max pooling head and projected into the conditioning space of the generator.

Training and Initialization

The ED encoder is optimized jointly with the periodic generator in all reported ED-CSP runs. For the CHILI-only checkpoint, its initial weights come from a separate ED-structure contrastive model that aligns paired multi-view observations and crystal structures in a normalized embedding space; the

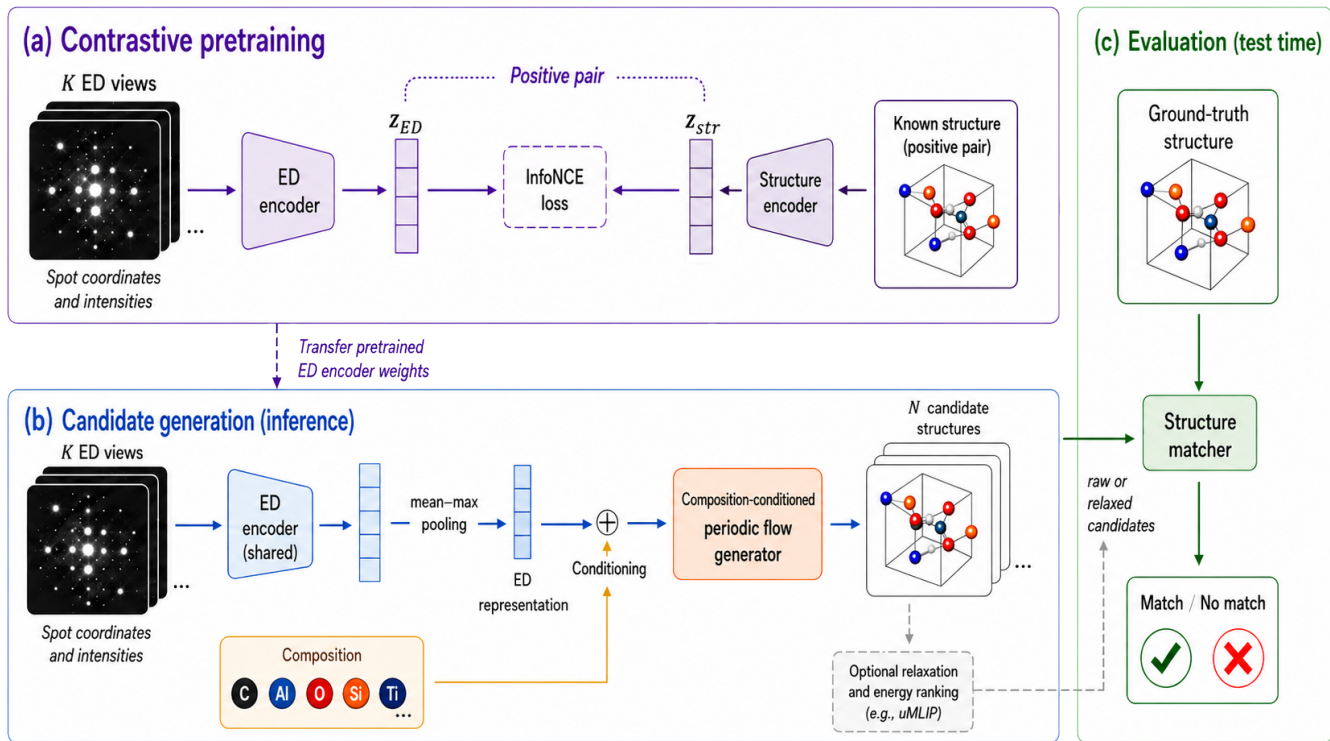


Figure 2: **ED-CSP pretraining, inference, and optional post-processing.** The optional uMLIP branch receives generated candidates but neither ED observations nor ground truth; structure matching is evaluation only.

transferred ED encoder remains trainable and the structure encoder is discarded. Following the PXRGen training design, this stage serves two roles: ED-to-structure retrieval benchmarks the learned diffraction representation, and the pretrained ED weights initialize the conditioning encoder before generator training (Li et al. 2025). It is particularly useful for screening encoder and view-aggregation choices: retrieval isolates the ED representation from composition conditioning and requires neither periodic-generator optimization nor iterative structure sampling. We evaluate the first role directly; without a matched randomly initialized generator, the downstream contribution of the second is not isolated. The registry-1M experiment instead warm-starts the complete registry-trained ED-CSP model before CHILI fine-tuning, so its gain measures full-model transfer rather than ED-CL alone. This completed one-million-structure precursor predates the final ED-CS filtering and is therefore reported separately from the 4.85-million-structure corpus.

Periodic Flow Generation

A six-layer CSPNet-style periodic graph decoder (Jiao et al. 2023; Li et al. 2025) conditions on flow time, the current lattice and fractional coordinates, and the aggregated ED state to predict lattice and periodic-coordinate vector fields. Training minimizes their weighted mean-squared errors, $\mathcal{L} = \mathcal{L}_{\text{lat}} + 100\mathcal{L}_{\text{coord}}$. ED-CSP and PXRGen share the same CSPFlow/CSPNet structure generator but use modality-specific encoders; their comparison therefore

evaluates the complete ED- and PXRGen-conditioned systems rather than isolating diffraction modality alone.

Experimental Setup

Datasets and ED Simulation

We use CHILI-100K, a KDD graph-ML benchmark derived from experimentally determined inorganic structures (Friis-Jensen et al. 2024). We additionally construct Electron Diffraction Crystal Structures (ED-CS) v1, a frozen snapshot comprising 4,852,131 structures selected from AFLOW, Alexandria, the Crystallography Open Database, GNoME, Materials Project, OQMD, and JARVIS-DFT (Curtarolo et al. 2012; Gražulis et al. 2012; Schmidt et al. 2024; Merchant et al. 2023; Jain et al. 2013; Saal et al. 2013; Choudhary et al. 2020). ED-CS v1 is a curated construction snapshot, not an exhaustive mirror of any upstream repository; each source contribution is defined by its versioned eligibility, simulation, deduplication, and exclusion manifests, and later additions require a separately versioned expansion. The v1 candidate snapshot contains entries with at most 100 sites; we merge canonical identifiers, deduplicate same-formula structures with StructureMatcher, require a certified payload with at least ten valid simulated views, and exclude every match to CHILI-100K under the same strict matcher settings. The resulting certificate contains zero CHILI matches; source counts and the full construction record are reported in the supplementary material. For every retained identifier, a reconstruction registry links the stored canonical

Table 1: ED-CS construction record; bold marks the final v1 snapshot.

Stage	Structures
Canonical entries in v1 candidate snapshot	5,330,947
After global structural deduplication	5,324,221
Entries with at least 10 certified ED views	4,852,708
After CHILI exclusion (ED-CS)	4,852,131

ED-CS v1 is a frozen curated snapshot, not an exhaustive export of its upstream repositories. The final strict StructureMatcher certificate reports zero CHILI matches and zero matcher errors.

structure, exact orientations, and ED arrays to record-level hashes; all 4,852,131 entries pass this technical completeness check. The Code and Data Supplement provides the construction code, provenance and integrity schemas, and a source-stratified subset of 256 real ED-CS records; source-specific redistribution terms for the full staged corpus are detailed in the supplementary material. Table 1 records the attrition at each certified construction stage.

For each retained structure, we precompute dynamical ED spot patterns using the py4DSTEM simulation pipeline and parameters adopted by EDiffCrystals (Savitzky et al. 2021; Gleason et al. 2024; Nathani et al. 2026), at 300.00 keV electron energy, 20.00 nm thickness, and a reciprocal-space cutoff of $2.00 \text{ \AA}^{-1}$. This multi-orientation protocol follows recent ML electron-diffraction benchmarks based on py4DSTEM or Bloch-wave point-list patterns (Gleason et al. 2024; Nathani et al. 2026). Simulation starts from ten random orientations. A throughput-oriented online policy extends a case to at most 100 views only when its ten-view pilot runtime falls below a break-even threshold estimated from recent simulation throughput and extension cost; the complete rule is given in the supplementary material. A structure is retained only when at least ten valid views are available. Each accepted view has at least ten spots before retaining its 16 strongest intensities. The completed million-structure transfer curriculum uses an earlier pool drawn from Materials Project, the Crystallography Open Database, and Alexandria while retaining source provenance; it is a precursor rather than a claimed subset of the final ED-CS corpus. The main CHILI benchmark uses 16,611 training structures and the same 2,075 ED-valid held-out queries across its reported comparisons. ED-CSP samples ten views per structure and retains the 16 strongest spots per view.

Training Protocol

The CHILI-100K protocol is frozen before model evaluation. ED-CSP is trained on the CHILI-100K train split with Adam optimization, gradient clipping, and mixed precision. Validation match rate with one candidate selects the reported ED-CSP checkpoints, with the policy fixed before test-set evaluation.

Baselines and Comparison Settings

We evaluate three comparison settings with distinct inputs. PXRDGen, XRDSol, and deCIFer provide same-dataset

diffraction-conditioned generative comparisons; ED library matching measures finite-database retrieval with explicit coverage; and Superflip/EDMA measures reconstruction when simulator-indexed reflections are supplied (Palatinus and Chapuis 2007; Li et al. 2025; Yu et al. 2026; Johansen et al. 2026). For the matched deCIFer adaptation, we train from scratch on the same 16,611 CHILI structures, provide only composition and a clean PXRD profile at inference, and select the checkpoint by validation loss before test evaluation. For the XRDSol adaptation, we use the same CHILI split, its published 1,000-step training budget, five full 1,000-step diffusion samples, and its native PXRD, composition, and ground-truth unit-cell inputs.

The retrieval library is restricted to train and validation materials; test structures are excluded. It searches all 18,688 ED-valid train/validation reference structures and returns five nearest neighbors. Radial retrieval compares normalized reciprocal-radius histograms; canonical-Chamfer retrieval compares sparse two-dimensional spot sets after radial pre-filtering by matching each query view to its closest library view. Formula-aware variants restrict candidates by anonymous formula, chemical system, or exact reduced formula before ranking. If no train/validation candidate passes a formula filter, that protocol has no valid candidate for the query.

The Palatinus-style control exports simulator-indexed reflection lists to Superflip/EDMA and uses five fixed reconstruction restarts. It is evaluated separately from methods that receive detector-plane spot lists.

Evaluation Protocol

Structure generation is evaluated with the shared structural matcher implemented through pymatgen (Ong et al. 2013), with the same matcher settings for ED-CSP, library retrieval, and all evaluations. We report match rate (MR): the fraction of test materials for which at least one candidate structure matches the ground truth under this fixed matcher. For CHILI-100K, ED-CSP, PXRDGen, and deCIFer are reported at both MR@1 and MR@5; XRDSol is reported at MR@5 over five independent samples; ED library matching is reported at MR@1 and MR@5 for the exact-formula control and at MR@5 for the unfiltered full-coverage control; Superflip/EDMA is reported as MR@5 over five fixed restarts. We compute each confidence interval (CI) by non-parametric bootstrap over held-out materials and assess paired significance with a sign-flip test on the per-query ED-CSP-minus-library outcomes.

Results

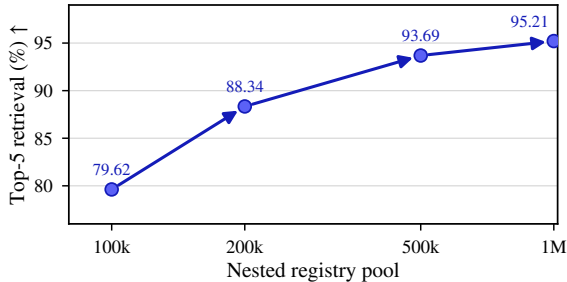
Table 2 summarizes the CHILI-100K benchmark. Its blocks answer different questions under a common test split and matcher; they are not an input-equivalent leaderboard.

Registry Scaling and Signal Use

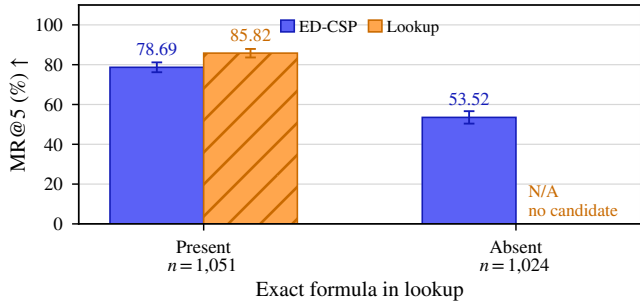
Full-model registry-1M initialization is strongest on the aligned ED-valid split: MR@1 reaches 51.66 % versus 42.12 %, and MR@5 reaches 66.27 % versus 57.49 % for CHILI-only initialization. The paired gains are 9.54 percentage points (95% CI [7.57, 11.52]) at MR@1 and 8.78 per-

Table 2: CHILI-100K benchmark on the same 2,075 held-out queries. Coverage is the fraction of queries for which the method has a valid input candidate set; dashes denote unmeasured metrics. Rows are grouped by their available inputs and are not a single input-equivalent leaderboard. Bold marks the strongest measured MR within the first two ED-CSP/generator blocks; control rows are not ranked.

Method / condition	Information available	Coverage (%)	MR@1 $\uparrow$ (%)	MR@5 $\uparrow$ (%)
<i>Registry-scaled ED-CSP and signal ablations</i>				
ED-CSP + registry-1M (ours)	py4DSTEM ED spots + composition	100.00	51.66	66.27
Zero ED spots	ED spots removed (test-time intervention)	100.00	17.35	–
<i>CHILI-only generator comparison</i>				
ED-CSP	py4DSTEM ED spots + composition	100.00	42.12	57.49
PXRGen	PXRD + composition	100.00	34.02	52.92
XRDSol	PXRD + composition + ground-truth unit cell	100.00	–	16.00
deCIFer	PXRD + composition	100.00	23.28	34.99
<i>Crystallographic and retrieval controls</i>				
Superflip/EDMA	Indexed ED reflections (hidden HKL)	99.76	–	10.51
ED library, exact formula	Train/validation ED neighbors + composition	50.65	42.60	43.47
ED library, unfiltered	Train/validation ED neighbors	100.00	–	18.31



(a) Sequential nested-pool curriculum



(b) Recovery beyond lookup coverage

Figure 3: **Effect of pretraining scale and exact-formula retrieval availability.** (a) Sequential ED-to-structure stages: each nested-pool checkpoint resumes its converged predecessor, so the points are not independent fits. (b) Registry-initialized ED-CSP recovery with and without an exact-formula train/validation candidate; exact-formula lookup has no candidate in the absent stratum. Error bars are query-bootstrap 95% CIs.

centage points ([7.04, 10.51]) at MR@5. Registry-1M ED-CSP achieves 66.27% MR@5 for the designated headline seed; across three inference seeds, mean MR@5 is 66.28%, with a sample standard deviation of 0.12 percentage points. We retain the reference CHILI-100K split for comparability; a post-hoc sensitivity analysis excluding 888 queries with a strict StructureMatcher near-duplicate still ranks registry-1M ED-CSP first at 38.84%/53.92%, versus 27.38%/42.46% for CHILI-only ED-CSP and 23.67%/38.75% for PXRGen. The separate representation diagnostic in Figure 3a follows a progressive nested-pool curriculum: each scale resumes the converged checkpoint from the preceding scale, expands the training pool, and continues until validation retrieval plateaus. Its monotonic gains show that adding registry structures improves representation retrieval along this curriculum. Figure 3b separately evaluates full-model ED-CSP transfer.

The full-split signal ablation changes only the ED input. Removing the ED spots reduces MR@1 from 51.66% to 17.35%, a paired drop of 34.31 percentage points (95% CI [32.14, 36.48]). On the stricter 67-query subset with a non-isomorphic same-formula donor, swapping in donor ED reduces mean MR@5 by 22.09 percentage points across five generation seeds (95% CI [10.45, 34.03]). A separately optimized converged composition-only reference reaches 50.94% MR@5, compared with 57.49% for CHILI-only ED-CSP on the same queries and candidate budget; because their optimization and sampling layouts differ, this 6.55-point gap is descriptive rather than a paired causal estimate. Together, the converged reference and input interventions show that ED-CSP uses ED beyond composition and learned priors on the full split, and query-specific ED on the same-formula donor subset.

Input Sensitivity

In a separate single-seed evaluation, we apply paired corruptions to the cached CHILI-only ED inputs while fixing the checkpoint, queries, compositions, matcher, five-candidate budget, and inference setting. These interventions measure sensitivity to perturbed model inputs, not transfer to a new physical simulation regime. Table 3 shows a graded response to missing spots: 10% dropout lowers MR@5 by 2.07 percentage points, while 25% lowers it by 7.71 percentage points. Intensity noise at $\sigma = 0.50$ produces a 2.02-percentage-point decrease, whereas perturbing stored view angles by up to 5° has no resolved effect under this cached-input protocol.

A separate single-seed detector-frame intervention starts from the registry-1M checkpoint and fine-tunes with random shared in-plane rotations. It raises MR@5 under shared and independent rotations by 7.28 and 5.93 percentage points, leaving a 1.35-percentage-point gap to clean inputs in both cases, at a 2.31-percentage-point clean-input cost (Supplementary Material).

Benchmark Comparisons

The exact-formula library control reaches 42.60 % MR@1 and 43.47 % MR@5 at 50.65 % coverage, whereas unfiltered full-coverage retrieval reaches 18.31 % MR@5. The apparent strength of formula-filtered lookup is therefore tied to analogue availability: among the 1,051 queries with a same-formula train/validation candidate, retrieval reaches 85.82 % MR@5 versus 78.69 % for registry-pretrained ED-CSP; on the remaining 1,024 queries it has no candidate, while ED-CSP reaches 53.52 %. This stratification separates phase lookup from out-of-library generation rather than averaging the two regimes into a misleading leaderboard. Using exact-formula retrieval when it has coverage and ED-CSP otherwise reaches 69.88 % MR@5 at the same five-candidate budget, a paired gain of 3.61 percentage points over ED-CSP (95% CI [2.17, 5.06]).

The indexed-reflection Superflip/EDMA control reaches 10.51 % MR@5, with 99.76 % valid-CIF query coverage and 96.40 % valid-candidate coverage. It consumes simulator-indexed reflections rather than detector-plane spots; its role is to measure reconstruction performance when indexed reflections are provided.

Under CHILI-only training and identical query IDs and candidate budgets, ED-CSP reaches 42.12 % MR@1 and 57.49 % MR@5. PXRDGen reaches 34.02 % and 52.92 %, respectively. At MR@5, ED-CSP alone solves 241 queries, PXRDGen alone solves 146, both solve 952, and both miss 736. The matched autoregressive deCIFer adaptation reaches 23.28 % MR@1 and 34.99 % MR@5. ED-CSP’s paired gains over deCIFer are 18.84 percentage points (95% CI [16.48, 21.16]) at MR@1 and 22.51 percentage points ([20.19, 24.82]) at MR@5; at MR@5, ED-CSP alone solves 584 queries and deCIFer alone solves 117. XRDsSol, retrained on the same CHILI split and additionally given the ground-truth unit cell, reaches 16.00 % MR@5; CHILI-only ED-CSP’s paired gain is 41.49 percentage points (95% CI [39.18, 43.81]). These paired results compare specific systems rather than establish intrinsic ED superiority: ED-CSP

receives multiple ED spot-list views through its encoder; PXRDGen uses a one-dimensional powder profile with a convolutional neural network (CNN) encoder but shares ED-CSP’s CSPFlow/CSPNet generator; XRDsSol additionally receives the unit cell; and deCIFer uses an autoregressive CIF-generation architecture.

Multi-View Diagnostic

The supplementary train-time view-count diagnostic keeps a 100-view simulation pool fixed and retraining an ED-structure retrieval encoder for each input count. Increasing the consumed views from one to twenty improves Top-5 ED-to-structure retrieval from 1.95 % to 13.38 %, supporting multi-view representation learning in this encoder-level benchmark without claiming a downstream generation optimum.

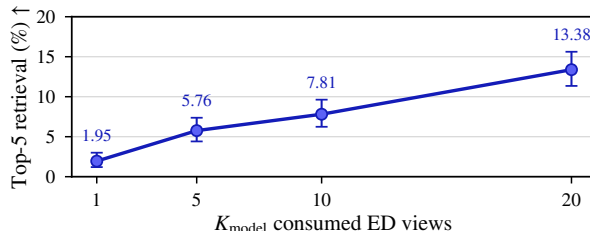


Figure 4: **Train-time K_{model} ablation.** Single-seed checkpoints; error bars are exact binomial 95% CIs over 1,024 queries and exclude training-run variation.

Post-Generation Relaxation

We test whether a target-free interatomic potential can stabilize and rank a frozen five-candidate CHILI-only ED-CSP payload. ORB-v3 (Rhodes et al. 2025), MACE-MPA-0 (Batatia et al. 2025), and eSEN-30M-OAM (Fu et al. 2025; Barroso-Luque et al. 2024) perform up to 100 FIRE steps, while CHGNet (Deng et al. 2023) performs 30 relaxation steps; all four relax the cell and rank candidates by final energy per atom. None of the potentials receives the ED observations or ground-truth structure.

ORB-v3 improves top-1 recovery by 13.16 percentage points (95% CI [11.37, 14.94]) and the relaxed candidate-pool MR@5 by 4.63 percentage points ([3.42, 5.88]). MACE-MPA-0 provides a near-identical independent check, with gains of 13.20 and 4.53 percentage points ([11.47, 14.99] and [3.33, 5.73]), respectively; ORB-v3 remains the validation-selected potential. eSEN-30M-OAM independently gives gains of 12.87 and 4.10 percentage points ([11.08, 14.65] and [2.89, 5.35]), respectively, without exceeding MACE-MPA-0 or ORB-v3. CHGNet independently gives gains of 10.65 and 3.28 percentage points ([8.96, 12.34] and [2.27, 4.34]), respectively. The separately sampled frozen payload supplies its own raw baseline, which differs slightly from the designated headline evaluation. The gains show that both candidate ordering and local geometry limit recovery, while the 61.69 % relaxed-pool ceiling leaves substantial room for ED-aware refinement rather than energy-only post-processing.

Table 3: CHILI-100K sensitivity to cached ED input corruptions.

ED input condition	MR@5 $\uparrow$ (% , 95% CI)	Matches	Queries	Delta vs. clean (95% CI)
Clean ED	57.40 [55.28, 59.52]	1,191	2,075	0.00
10% spot dropout	55.33 [53.20, 57.45]	1,148	2,075	-2.07 [-3.37, -0.82]
25% spot dropout	49.69 [47.52, 51.81]	1,031	2,075	-7.71 [-9.25, -6.22]
Intensity noise $\sigma = 0.25$	56.43 [54.31, 58.55]	1,171	2,075	-0.96 [-2.07, +0.14]
Intensity noise $\sigma = 0.50$	55.37 [53.25, 57.49]	1,149	2,075	-2.02 [-3.28, -0.82]

Single-seed cached-input evaluation. All deltas are paired percentage-point changes from this table’s 57.40% clean row, not from the 57.49% primary evaluation.

Table 4: Post-generation relaxation and energy ranking on CHILI-100K using a separately sampled, fixed candidate set. Deltas are paired gains over that set in percentage points; bold marks the validation-selected potential.

Condition	Top-1 $\uparrow$	Δ	Pool MR@5 $\uparrow$	Δ
Raw ED-CSP	42.31	–	57.06	–
ORB-v3	55.47	+13.16	61.69	+4.63
MACE-MPA-0	55.52	+13.20	61.59	+4.53
eSEN-30M-OAM	55.18	+12.87	61.16	+4.10
CHGNet	52.96	+10.65	60.34	+3.28

Every paired delta uses the same fixed candidate set, sampled separately from the candidates used for the 42.12%/57.49% primary evaluation.

Discussion and Limitations

The signal-use interventions show that the generator responds to query-specific diffraction geometry rather than treating ED as an optional auxiliary input. Together with the coverage-stratified retrieval results, this supports generative recovery as a complement to analogue lookup and indexed-reflection workflows.

The candidate analyses identify generation quality and selection as immediate bottlenecks: relaxation and energy ranking improve top-1 recovery, yet the remaining candidate-pool ceiling indicates room for ED-aware refinement. A natural next step is differentiable dynamical Bloch-wave refinement of generated candidates against observed ED views, jointly regularized by crystallographic or learned energy priors (Malik et al. 2026). Beginning with indexed, orientation-aware upper bounds, this would provide an ED-specific post-generation analogue to Rietveld refinement while modeling thickness-dependent intensities. The current ED-CSP generator consumes ten ED views, while the encoder-level diagnostic in Figure 4 improves Top-5 retrieval from 7.81 % at $K_{\text{model}} = 10$ to 13.38 % at $K_{\text{model}} = 20$, suggesting that the present input regime does not saturate multi-view representation learning. Evaluating larger view sets in full generator training is therefore a promising direction. ED-CS makes the scale and provenance of this direction explicit, while its simulation cost motivates adaptive allocation of orientations. The registry results also motivate continuing to expand the unique-structure pool, which already spans millions of structures, while reducing its ED simulation cost. Rather than assigning every structure the same simulation

budget, future work should test whether Krivovichev-style Shannon structural complexity (Krivovichev 2014) can help estimate the number of informative orientations required per structure. The principal external-validity gap remains the transition from calibrated simulated point lists with known composition to experimental 3D ED, where detector calibration, uncertain spot finding, background, missing reflections, thickness variation, indexing, and expert refinement all affect the observed data.

Conclusion

ED-CSP predicts lattices and fractional atomic coordinates from known composition and sparse simulated multi-view ED. The benchmark separates generation from finite-library retrieval and indexed-reflection preprocessing, while registry-scale full-model transfer further improves recovery. It provides a reproducible basis for developing ED-consistent candidate ranking and refinement.

Acknowledgments

The authors gratefully acknowledge GENCI/IDRIS for providing high-performance computing resources on the Jean Zay supercomputer, which supported the computational aspects of this work. This work was also supported by MAIA (“Maitrise des Applications de l’IA”) project from alliance A2U (Université d’Artois, UPJV et ULCO) and Hauts-de-France (HdF) region.

References

- Barroso-Luque, L.; Shuaibi, M.; Fu, X.; Wood, B. M.; Dzamba, M.; Gao, M.; Rizvi, A.; Zitnick, C. L.; and Ulissi, Z. W. 2024. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. *arXiv preprint arXiv:2410.12771*.
- Batatia, I.; et al. 2025. A Foundation Model for Atomistic Materials Chemistry. *The Journal of Chemical Physics*, 163(18): 184110.
- Choudhary, K.; Garrity, K. F.; Reid, A. C. E.; DeCost, B.; Biacchi, A. J.; et al. 2020. The Joint Automated Repository for Various Integrated Simulations (JARVIS) for Data-Driven Materials Design. *npj Computational Materials*, 6: 173.
- Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; and Levy, O. 2012. AFLOWLIB.ORG: A Distributed Materials Properties

- Repository from High-Throughput Ab Initio Calculations. *Computational Materials Science*, 58: 227–235.
- Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C. J.; and Ceder, G. 2023. CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling. *Nature Machine Intelligence*, 5(9): 1031–1041.
- Edamadaka, S.; Ramesh, K.; Li, J.; and Gómez-Bombarelli, R. 2026. Atomistic Language Models Understand and Generate Materials. *arXiv preprint arXiv:2606.21395*.
- Friis-Jensen, U.; Johansen, F. L.; Anker, A. S.; Dam, E. B.; Jensen, K. M. Ø.; and Selvan, R. 2024. CHILI: Chemically-Informed Large-Scale Inorganic Nanomaterials Dataset for Advancing Graph Machine Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4962–4973. ACM.
- Fu, X.; Wood, B. M.; Barroso-Luque, L.; Levine, D. S.; Gao, M.; Dzamba, M.; and Zitnick, C. L. 2025. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction. *arXiv preprint arXiv:2502.12147*.
- Gemmi, M.; Mugnaioli, E.; Gorelik, T. E.; Kolb, U.; Palatinus, L.; Boullay, P.; Hovmöller, S.; and Abrahams, J. P. 2019. 3D Electron Diffraction: The Nanocrystallography Revolution. *ACS Central Science*, 5(8): 1315–1329.
- Gleason, S. P.; Rakowski, A.; Ribet, S. M.; Zeltmann, S. E.; Savitzky, B. H.; Henderson, M.; Ciston, J.; and Ophus, C. 2024. Random Forest Prediction of Crystal Structure from Electron Diffraction Patterns Incorporating Multiple Scattering. *Physical Review Materials*, 8: 093802.
- Grāzulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; and Le Bail, A. 2012. Crystallography Open Database (COD): An Open-Access Collection of Crystal Structures and Platform for World-Wide Collaboration. *Nucleic Acids Research*, 40(D1): D420–D427.
- Guo, G.; Saidi, T. L.; Terban, M. W.; Valsecchi, M.; Billinge, S. J. L.; and Lipson, H. 2025. Ab Initio Structure Solutions from Nanocrystalline Powder Diffraction Data via Diffusion Models. *Nature Materials*, 24: 1726–1734.
- Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; and Persson, K. A. 2013. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials*, 1(1): 011002.
- Jiao, R.; Huang, W.; Lin, P.; Han, J.; Chen, P.; Lu, Y.; and Liu, Y. 2023. Crystal Structure Prediction by Joint Equivariant Diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 17464–17497.
- Johansen, F. L.; Friis-Jensen, U.; Dam, E. B.; Jensen, K. M. Ø.; Mercado, R.; and Selvan, R. 2026. deCIFer: Crystal Structure Prediction from Powder Diffraction Data Using Autoregressive Language Models. *Transactions on Machine Learning Research*.
- Klar, P. B.; Krysiak, Y.; Xu, H.; Steciuk, G.; Cho, J.; Zou, X.; and Palatinus, L. 2023. Accurate Structure Models and Absolute Configuration Determination Using Dynamical Effects in Continuous-Rotation 3D Electron Diffraction Data. *Nature Chemistry*, 15: 848–855.
- Krivovichev, S. V. 2014. Which Inorganic Structures Are the Most Complex? *Angewandte Chemie International Edition*, 53(3): 654–661.
- Lai, Q.; Xu, F.; Yao, L.; Gao, Z.; Liu, S.; Wang, H.; Lu, S.; He, D.; Wang, L.; Zhang, L.; Wang, C.; and Ke, G. 2025. End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction. *Advanced Science*, 12(8): 2410722.
- Li, Q.; Guo, M.; Jiao, R.; Gao, J.; Xu, F.; Xue, H.; Zhang, W.; Huang, W.; Yan, J.; Zhang, L.; Wang, C.; Yan, Z.; Ke, G.; E, W.; Tang, Z.; Jin, S.; and Yao, L. 2026. Experimental Powder X-ray Diffraction Crystal Structure Determination with RealPXRD-Solver. *arXiv preprint arXiv:2603.00965*.
- Li, Q.; Jiao, R.; Wu, L.; Zhu, T.; Huang, W.; Jin, S.; Liu, Y.; Weng, H.; and Chen, X. 2025. Powder Diffraction Crystal Structure Determination Using Generative Models. *Nature Communications*, 16: 7428.
- Malik, S. A.; Doherty, T. A. S.; Colmey, B.; Roberts, S. J.; Gal, Y.; and Midgley, P. A. 2026. Hybrid Physics-Machine Learning Models for Quantitative Electron Diffraction Refinements. *Nature Communications*, 17: 5056.
- Melgalvis, D. M.; and Rekiş, T. 2026. GraPhAI: Neural Networks for Solving Centrosymmetric Crystal Structures. *Journal of the American Chemical Society*, 148(27): 28754–28763.
- Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; and Cubuk, E. D. 2023. Scaling Deep Learning for Materials Discovery. *Nature*, 624: 80–85.
- Nathani, A.; McCray, A. R. C.; Liu, Y.; Ding, H.; Kazempoor, P.; Xu, S.; Ophus, C.; and Ghamarian, I. 2026. Accelerating Electron Diffraction Analysis Using Graph Neural Networks and Attention Mechanisms. *npj Computational Materials*, 12: 56.
- Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; and Ceder, G. 2013. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Computational Materials Science*, 68: 314–319.
- Palatinus, L.; and Chapuis, G. 2007. SUPERFLIP—A Computer Program for the Solution of Crystal Structures by Charge Flipping in Arbitrary Dimensions. *Journal of Applied Crystallography*, 40(4): 786–790.
- Peng, Q.; Han, X.; Wang, Z.; Hong, Y.; Meng, F.; Zhang, Z.; Zhang, J.; and Zhao, X. 2026. Diffraction-Native Representation Learning for Automatic Symmetry Recognition and Structural Retrieval. *Advanced Functional Materials*, 36(50): e75949.
- Rhodes, B.; Vandenhaute, S.; Šimkus, V.; Gin, J.; Godwin, J.; Duignan, T.; and Neumann, M. 2025. Orb-v3: Atomistic Simulation at Scale. *arXiv preprint arXiv:2504.06231*.
- Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; and Wolverton, C. 2013. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65: 1501–1509.
- Savitzky, B. H.; Zeltmann, S. E.; Hughes, L. A.; Brown, H. G.; Zhao, S.; Pelz, P. M.; Pekin, T. C.; Barnard, E. S.;

Donohue, J.; Rangel DaCosta, L.; Kennedy, E.; Xie, Y.; Janish, M. T.; Schneider, M. M.; Herring, P.; Gopal, C.; Anapolsky, A.; Dhall, R.; Bustillo, K. C.; Ercius, P.; Scott, M. C.; Ciston, J.; Minor, A. M.; and Ophus, C. 2021. py4DSTEM: A Software Package for Four-Dimensional Scanning Transmission Electron Microscopy Data Analysis. *Microscopy and Microanalysis*, 27(4): 712–743.

Schmidt, J.; Cerqueira, T. F. T.; Romero, A. H.; Loew, A.; Jäger, F.; Wang, H.-C.; Botti, S.; and Marques, M. A. L. 2024. Improving Machine-Learning Models in Materials Science through Large Datasets. *Materials Today Physics*, 48: 101560.

Segal, N.; Li, M.; Miller, B. K.; and Gómez-Bombarelli, R. 2026. XRDiff: Crystal Structure Prediction from Powder X-Ray Diffraction Data Using Diffusion Models. *arXiv preprint arXiv:2606.14003*.

Unge, J.; Nannenga, B. L.; Oliver, A. G.; and Gonen, T. 2025. Standards for MicroED. *Acta Crystallographica Section C*, 81: 376–390.

Yu, D.; Zhu, Z.; Leng, F.; and Zhu, Y. 2026. Equivariant Diffusion Solution for Inorganic Crystal Structure Determination from Powder X-ray Diffraction Data. *Nature Communications*, 17: 3274.