

Robust Indexing for Challenging Serial X-ray Diffraction Patterns

Marc M Nasser ^{a,b}, Frédéric Poitevin ^b, and Kevin M Dalton ^b

^aElectrical Engineering, Stanford University

^bLCLS Data Systems, SLAC National Accelerator Laboratory

Abstract

Serial crystallography experiments routinely produce thousands of diffraction patterns from crystals in random orientations. To turn this stream of images into a usable dataset, each pattern must be indexed (having its crystal orientation recovered from a small set of measured reflections) before integration and merging can proceed. In practice, diffraction patterns can contain only a small number of reliable peaks, are contaminated by background and spuriously detected reflections, or arise from crystals with highly skewed unit cells. These factors make indexing unstable precisely in the small- N regime.

We introduce a robust indexing algorithm tailored to this setting. We formulate indexing as a symmetry-aware lattice decoding problem and design a loss that explicitly incorporates lattice symmetries while trimming outlier peaks that are inconsistent with any plausible orientation. We combine this objective with a reciprocal-space basis reparameterization that stabilizes decoding for skewed or poorly conditioned lattices, and we develop a dedicated small- N objective mode that couples refined peak scoring with a method to recover orientations from very few reflections. The resulting method can be implemented in a memory-efficient way and used as a robust indexer.

We evaluate our approach on three protein datasets from the Coherent X-ray Imaging Data Bank collected at XFEL facilities, using identical preprocessing and unit-cell information across methods. Across all datasets, our algorithm matches or outperforms established indexers such as XGANDALF and TORO, with particularly large gains for patterns with few indexed peaks and for crystals with skewed unit cells. While slower, our algorithm is extremely memory-efficient, and its structure readily enables high-degree parallelization on CPUs or larger batch sizes on GPUs. Our results show that carefully exploiting lattice structure, symmetry, and small- N -aware search yields substantial improvements in indexing robustness.

Keywords: Serial crystallography; Indexing; Orientation determination; Diffraction patterns; Small- N regime; Skewed lattices; Lattice symmetry; Memory-efficient algorithms

1 Introduction

1.1 Motivation and Background

Serial crystallography has become a central approach for determining macromolecular structures at both X-ray free-electron laser (XFEL) facilities. In serial experiments, a large number of diffraction patterns are collected from many microcrystals delivered in a continuous stream. Each pattern typically samples a different random crystal orientation and contains only a subset of all possible reflections. To reconstruct a full reciprocal-space dataset, each frame must be *indexed*: one must determine the crystal orientation and assign Miller indices to the observed peaks. Indexing accuracy directly affects downstream integration, merging, and phasing, and is therefore a crucial component of modern serial crystallography pipelines.

As such, it is important to robustly and efficiently index large numbers of sparsely populated, noisy diffraction patterns. Existing indexers have made substantial progress towards this goal, particularly in computational speed, yet a nontrivial fraction of frames in many datasets remain unindexed, reducing completeness and the quality of the final merged structures.

In this work we focus on the common scenario where approximate unit-cell parameters are known and supplied to the indexer. This matches the standard usage in serial crystallography pipelines such as CrystFEL/indexamajig (White *et al.*, 2016), where a unit-cell file is typically provided alongside the detector geometry and other experimental parameters (White *et al.*, 2016).

1.2 Indexing Challenges in Serial Crystallography

Several factors make indexing particularly challenging in serial crystallography:

- **Sparse peak sets.** Many diffraction patterns contain only a small number of measured reflections, for example from weakly diffracting samples or partially illuminated crystals. In this low- N regime, existing indexers typically experience a marked decline in performance compared to patterns with larger numbers of reflections, as their methods are sensitive to per-peak assignment errors.
- **High noise and false peaks.** Background scattering, detector artifacts, and peak-finding errors introduce outliers that can mislead least-squares or grid-based search methods, especially since loss landscape is highly non-convex.
- **Skewed or ill-conditioned reciprocal bases.** Conventional unit-cell bases may lead to poorly conditioned integer decoding, especially in low-symmetry systems or poorly chosen primitive cells.

- **Throughput and deployment constraints.** High repetition-rate XFELs and fast serial synchrotron beamlines generate data at rates that challenge both online (real-time) and offline analysis pipelines. Indexers must balance robustness against computational cost, memory usage, and latency.

These difficulties motivate the development of indexing methods that (i) remain robust for small- N patterns, (ii) handle noise and outliers effectively, (iii) leverage lattice symmetries and basis reduction for stable integer decoding, and (iv) explore the orientation space efficiently, balancing speed and memory, making them usable across both XFEL and synchrotron serial crystallography experiments.

1.3 Related Work

Several indexing algorithms have been developed for serial crystallography, with PinkIndexer (Gevorkov *et al.*, 2020) and XGANDALF (Gevorkov *et al.*, 2019) among the most widely used.

PinkIndexer formulates indexing as a geometric constraint problem and was one of the first methods tailored to snapshot diffraction patterns (including pink-beam X-rays and electron diffraction) rather than conventional rotation-series data. PinkIndexer assumes that approximate unit-cell parameters are known in advance.

XGANDALF introduced a modern formulation of indexing for serial crystallography based on an extended gradient-descent algorithm for lattice finding. It constructs a continuous score function from proximity to families of lattice planes and performs fast gradient-based refinement, making it well suited for real-time feedback and widely adopted in CrystFEL (White *et al.*, 2016) pipelines. XGANDALF can operate both with a known unit cell and in a mode where the lattice parameters are inferred from the data, i.e., the unit cell is optional rather than strictly required. However, improvements can be made in terms of robustness over several factors (noise, false peaks, and small number of reflections present).

More recently, TORO (Gasparotto *et al.*, 2024) introduced a PyTorch-based indexing framework that combines relaxed Laue score functions with robust optimization techniques, including least-trimmed-squares-style estimation and residual-threshold annealing, together with an speed efficient orientation-space exploration strategy, although memory heavy. Its advantage comes in the low number of iterations required for convergence. TORO runs on GPUs, CPUs and other accelerators supported by PyTorch, making it an attractive candidate for real-time serial crystallography pipelines while matching or surpassing the accuracy of established indexers such as XGANDALF and MOSFLM on their tested benchmarks. In the experiments reported, TORO is used with the unit-cell parameters provided to the indexer; the authors note that the method could in principle be extended to also guess an initial cell, but this is not explored in their benchmarks.

Despite these advances, neither XGANDALF nor TORO explicitly addresses issues arising from poorly conditioned reciprocal bases, nor do they exploit geometric tools such as reciprocal-basis

reduction and Babai’s nearest-plane decoder. Their objectives are designed around relaxed Laue equations and fast online processing, which naturally prioritizes speed over small- N -specific robustness. These gaps motivate the development of indexing methods that remain accurate under small number of reflections present, noise, and basis distortion, and that can also maintain efficient memory usage if needed, as pursued in the present work, where we assume that approximate unit-cell parameters are known.

Our Contributions

In this work we introduce a robust indexing algorithm for serial crystallography with the following main contributions:

- We formulate indexing as a symmetry-aware lattice decoding problem and design a robust orientation-dependent loss that explicitly incorporates lattice symmetries and outlier trimming.
- We combine this loss with a reciprocal-space basis reduction and a nearest-plane integer decoder, leading to stable indexing even for skewed unit cells.
- We develop a dedicated small- N regime, using a refined per-peak objective and local search, that substantially improves indexing rates for patterns with very few reflections.
- We show that the method can be implemented in a memory-efficient way on GPUs and evaluate it on three XFEL datasets, where it consistently matches or outperforms state-of-the-art indexers (TORO and XGANDALF), with particularly large gains in the small- N regime.

This positions our method as a complementary tool to existing indexers: highly robust and memory-efficient, particularly effective for images with a small number of reflections, and for skewed unit cells.

2 Indexing Problem

The goal of crystallographic indexing is to determine the orientation of a crystal and assign Miller indices to the observed diffraction peaks. In serial crystallography, each diffraction pattern contains a sparse set of reciprocal-space vectors

$$\mathcal{Q} = \{q_1, \dots, q_N\} \subset \mathbb{R}^3,$$

corresponding to detector-localized peak positions mapped into laboratory coordinates. Given the experimental geometry and approximate unit-cell parameters, the reciprocal lattice is described by a basis matrix

$$B \in \mathbb{R}^{3 \times 3}$$

(typically chosen as the inverse transpose of a direct-space basis). For a rotation matrix $R \in SO(3)$, the corresponding reciprocal lattice in the laboratory frame is generated by RB .

Indexing can therefore be formulated as the problem of finding an orientation R such that each observed q_n is well explained by some Miller index $h_n \in \mathbb{Z}^3$ through the relation

$$q_n \approx RBh_n.$$

The assignment of (h_n) is implicit: for a fixed R , candidate indices are obtained by projecting each q_n onto the lattice spanned by RB and applying an integer-decoding rule (rounding, nearest-plane search, or similar). Thus, the core task is to identify the rotation R for which these decoded indices agree with the underlying crystal lattice structure.

It is important to note that the solution (R, H) is not unique: the reciprocal lattice is invariant under the crystal's symmetry group. If S is any symmetry operation, then $RBh = R(SB)(S^{-1}h)$ generates the same set of reciprocal-lattice vectors. Therefore, indexing is defined only up to the symmetry group, and any valid orientation R must be interpreted modulo these symmetry-equivalent transformations.

2.1 Indexing as an Optimization Problem

In mathematical optimization terms, the indexing task can be expressed as a joint inference problem over the crystal orientation and the corresponding Miller indices. Given observed reciprocal-space vectors $\mathcal{Q} = \{Q_1, \dots, Q_N\} \subset \mathbb{R}^3$, $Q \in \mathbb{R}^{3 \times N}$ and a reciprocal basis B derived from the known unit-cell parameters, the ideal orientation $R \in SO(3)$ and integer triples $H_i \in \mathbb{Z}^3$ should satisfy

$$Q_i \approx RBH_i.$$

for Q_i the reciprocal-space vector of a true reflection.

A general formulation of the indexing problem is therefore

$$(R^*, H^*) = \arg \min_{R \in SO(3), H \in \mathbb{Z}^{3 \times N}} L(Q, RBH), \quad (1)$$

where L is any suitable discrepancy measure between the observed peak positions $\mathcal{Q}$ and the lattice points predicted by the oriented reciprocal basis RBH .

or an analogous robust variant. Because crystallographic lattices are invariant under the Laue symmetry group $\mathcal{G}$, solutions are not unique: (R, H) and $(RS, S^{-1}H)$ represent the same physical orientation for any $S \in \mathcal{G}$.

This abstract formulation highlights the intrinsic structure of the indexing problem: a highly non-convex search over rotations coupled with discrete integer assignments, defined only up to lattice symmetry.

3 Methods

3.1 Algorithm Overview

Given reciprocal-space vectors $\{Q_i\}_{i=1}^N$, a reciprocal basis B , and unit-cell parameters, the indexer seeks an orientation $R^* \in SO(3)$ and Miller indices $\{H_i^*\}_{i=1}^N$ such that $Q_i \approx R^* B H_i^*$ for as many peaks as possible. We precompute a reduced reciprocal basis and symmetry operators, define a robust orientation-dependent loss, minimize this loss over $SO(3)$ using differential evolution (DE), and finally decode the Miller indices using a nearest-lattice procedure based on Babai’s Nearest Plane Algorithm (Section 3.3) (with a local-search for patterns with small number of reflections). The steps are summarized in Algorithm 1.

Algorithm 1 Robust symmetry-aware indexing (high-level)

Require: Reciprocal vectors $\{Q_i\}_{i=1}^N$, basis B , trim fraction κ , small- N threshold N_{small} , DE population size P , mutation scale F , crossover probability p_{cx} , maximum iterations G_{max} , refinement checkpoint interval T_{ref}

Ensure: Orientation $R^* \in SO(3)$, indices $\{H_i^*\}_{i=1}^N \subset \mathbb{Z}^3$

1: **Precomputation:**

 Perform reciprocal basis reduction (see Section 3.4) to obtain $B_{\text{red}} = BU$ with $U \in SL(3, \mathbb{Z})$

 Compute $B_{\text{red}} = Q_B R_B$ and symmetry operators for the lattice (in reciprocal space)

2: Represent each candidate crystal orientation by a unit quaternion $q \in S^3$, with $R(q) \in SO(3)$ the corresponding rotation matrix

3: Initialize the DE population by sampling quaternions uniformly on S^3 (see Section 3.2)

4: **Define orientation loss** (see Section 3.5):

5: **if** $N \leq N_{\text{small}}$ **then**

$\triangleright$ e.g. $N_{\text{small}} = 25$

6: $L(q) \leftarrow L_{\text{small}}(R(q); Q, B_{\text{red}}, Q_B, R_B, \text{symmetry}, \kappa)$

7: **else**

8: $L(q) \leftarrow L_{\text{big}}(R(q); Q, B_{\text{red}}, Q_B, R_B, \text{symmetry}, \kappa)$

9: **end if**

10: **Run DE with occasional local refinement** (see Section 3.2 and 3.6):

11: **while** DE not converged **do**

12: Perform one DE generation on the quaternion population
 (mutation, crossover, normalization)

13: Evaluate the loss $L(q)$ for all population members

14: **if** generation index is a refinement checkpoint **then**

15: $q_{\text{DE}} \leftarrow \arg \min_q L(q)$ in the current population

16: $R_{\text{DE}} \leftarrow R(q_{\text{DE}})$

17: Locally refine R_{DE} (Babai/Local-Search + Procrustes) to obtain R_{ref} (see Section 3.3)

18: $R_{\text{best}} \leftarrow \arg \min_{R \in \{R_{\text{best}}, R_{\text{ref}}\}} L(R)$

19: **end if**

20: **end while**

21: **Select final orientation:**

22: $R^*(q^*) \leftarrow$ best orientation ($\arg \min L(q)$) among DE and shadow-refined candidates

23: **Final indexing:**

24: **for** $i = 1, \dots, N$ **do**

25: $y_i \leftarrow Q_B^\top R^* Q_i$

26: Decode reduced index $H_{\text{red},i}$ by Babai / Local-search in the reduced basis (with symmetry)

27: Map back to original indices: $H_i^* \leftarrow U H_{\text{red},i}$

28: **end for**

29: **return** $R^*, \{H_i^*\}_{i=1}^N$

3.2 Differential Evolution

Differential evolution (DE) (Storn & Price, 1997) is a population-based stochastic optimizer for real-valued, nonconvex objectives. Given a loss $L(\theta)$ with $\theta \in \mathbb{R}^d$, DE maintains a population

$$\Theta^{(g)} = \{\theta_1^{(g)}, \dots, \theta_P^{(g)}\}$$

at generation g , and iteratively produces a new population by three steps for each index $i \in \{1, \dots, P\}$:

Mutation. Select distinct indices $r_1, r_2 \neq i$ uniformly at random without replacement from the population (excluding i), and let $\theta_{\text{best}}^{(g)}$ denote the current best individual. In the **best-1-bin** scheme we use, the mutant vector is

$$v_i^{(g)} = \theta_{\text{best}}^{(g)} + F(\theta_{r_1}^{(g)} - \theta_{r_2}^{(g)}), \quad r_1, r_2 \neq i, \ r_1 \neq r_2.$$

where $F > 0$ is a mutation scale factor. The parameter F controls the exploration step size in orientation space: larger values promote broader exploration, while smaller values bias the search toward local refinement. In our experiments we follow standard DE practice and sample F uniformly in the interval $[0.5, 1]$ for each generation.

Crossover. A trial vector $u_i^{(g)}$ is formed by mixing coordinates from $v_i^{(g)}$ and $\theta_i^{(g)}$. In binomial crossover, for each component $k \in \{1, \dots, d\}$,

$$u_{i,k}^{(g)} = \begin{cases} v_{i,k}^{(g)}, & \text{with probability } C_r, \\ \theta_{i,k}^{(g)}, & \text{otherwise,} \end{cases}$$

with crossover rate $C_r \in [0, 1]$. In practice we use the standard DE convention of setting C_r to be 0.7, which balances exploration and stability

Selection. The trial vector competes with the parent:

$$\theta_i^{(g+1)} = \begin{cases} u_i^{(g)}, & \text{if } L(u_i^{(g)}) \leq L(\theta_i^{(g)}), \\ \theta_i^{(g)}, & \text{otherwise.} \end{cases}$$

Thus DE performs a stochastic, *elitist* search over $\mathbb{R}^d$, in the sense that the current best solution is always preserved in the population, and is an efficient method for non-convex, non-smooth landscapes that can escape local minima.

Application to orientation search. In our indexer, the parameter θ is a unit quaternion $q \in S^3 \subset \mathbb{R}^4$ representing an orientation $R(q) \in SO(3)$.

Note that unit quaternions provide a double cover of $SO(3)$, since q and $-q$ represent the same rotation. Because our loss depends only on $R(q)$, this redundancy has no effect on the optimization. In practice, keeping both q and $-q$ in the DE population is harmless: mutation, crossover, and selection operate directly on the quaternion parameters, so the two representatives typically follow different stochastic trajectories, providing additional exploration at negligible cost.

We apply DE to the quaternion coordinates using the loss

$$L(q) \in \{L_{\text{big}}(R(q)), L_{\text{small}}(R(q))\}$$

defined in Section 3.5, depending on the number of observed reflections, N . After each mutation and crossover step we renormalize,

$$q \leftarrow \frac{q}{\|q\|_2},$$

to keep all population members on the unit 3-sphere.

DE is responsible for exploring the orientation space, while the Babai step provides accurate lattice decoding for each proposed orientation so that the loss can be evaluated reliably.

Retries and memory considerations. Indexing pipelines such as CRYSTFEL commonly rely on multiple retries with different seeds. DE is especially compatible with this practice, since different initializations or random seeds lead to distinct stochastic explorations of orientation space. Our implementation also maintains a very small memory footprint, as DE does not require large population size since it is more dependent on number of iterations. In practice, DE does not require a large population, since increasing the number of generations often yields similar improvements in solution quality. This well-known trade-off allows small populations to converge reliably in many settings, including ours, and keeps the memory footprint minimal. In practice, it is often recommended to set the population size between $5d$ - $10d$, where d is the number of dimensions of candidates.

3.3 Babai Rounding

Given a full-rank lattice basis $A \in \mathbb{R}^{3 \times 3}$, the associated lattice is

$$\Lambda(A) := \{Ah : h \in \mathbb{Z}^3\}.$$

The *closest vector problem* (CVP) asks, for a point $x \in \mathbb{R}^3$, to find the lattice point Ah closest to x in Euclidean distance. Babai’s nearest-plane algorithm (Babai, 1986) is a classic and efficient approximation: it takes x and produces an integer vector $\hat{h}$ by rounding in a suitable orthonormal coordinate system.

Babai in QR coordinates. Let $A = QR$ be a QR factorization with $Q \in O(3)$ orthogonal and R upper triangular with $R_{ii} > 0$. For $x \in \mathbb{R}^3$ define

$$y := Q^\top x.$$

In the noiseless case $x = Ah$ with $h \in \mathbb{Z}^3$ gives $y = Rh$. Babai's decoder computes an estimate $\hat{h} \in \mathbb{Z}^3$ via backward substitution:

$$\hat{h}_i = \text{round}\left(\frac{y_i - \sum_{j>i} R_{ij} \hat{h}_j}{R_{ii}}\right), \quad i = 3, 2, 1,$$

and returns the lattice point $\hat{x} := A\hat{h}$.

We choose Babai rounding over the more common least-squares-plus-rounding approach (i.e., solving $\min_h \|Ah - x\|$ and then rounding h to the nearest integer vector), as it offers clear advantages, as discussed in Section 3.7.1.

Use in our indexer. In our setting $A = B_{\text{red}} = BU$ is the reduced reciprocal basis (Section 3.4), with QR factorization

$$B_{\text{red}} = Q_B R_B.$$

For a fixed orientation $R \in SO(3)$ and peak Q_i , we work with

$$y_i := Q_B^\top R^\top Q_i = R_B h_i + \varepsilon_i,$$

where $h_i \in \mathbb{Z}^3$ are the (unknown) Miller indices in the reduced basis and ε_i is noise. Applying Babai's rule with R_B yields an estimate

$$\hat{H}_i^{\text{red}}(R) \in \mathbb{Z}^3,$$

and we map back to the original reciprocal basis via $H_i(R) := U \hat{H}_i^{\text{red}}(R)$. This Babai step is the core nearest-lattice decoder used in our objectives, with a local integer search around $\hat{H}_i^{\text{red}}(R)$ in the small- N case (Section 3.5).

3.4 Reciprocal Basis Reduction

Introduction. The reciprocal lattice is

$$\mathcal{L}^* = \{Bh : h \in \mathbb{Z}^3\},$$

where h are Miller indices and $B \in \mathbb{R}^{3 \times 3}$ is a reciprocal basis. We replace B by a lattice-equivalent *reduced* basis

$$B_{\text{red}} := BU, \quad U \in SL(3, \mathbb{Z}),$$

where $SL(3, \mathbb{Z})$ denotes the group of 3×3 integer matrices with determinant 1 (unimodular integer changes of basis). Such a change preserves the lattice $\mathcal{L}^*$ but re-labels Miller indices. Reduced bases are chosen to avoid very short and highly correlated basis vectors: they are typically closer to orthogonal and better balanced in length, which improves the conditioning of lattice decoding. As we show in Section 3.7.2, in the reduced coordinates Babai’s nearest-plane decoder succeeds whenever the noise falls inside an axis-aligned box whose side lengths are given by the Gram-Schmidt lengths of the basis; enlarging these lengths, especially in the weakest directions, enlarges the success region and tends to increase the probability of correct indexing.

Index mapping, QR, and orientation. Unimodularity preserves the lattice and maps indices via

$$H = U H_{\text{red}}, \quad H_{\text{red}} = U^{-1} H,$$

so that indexing in the reduced basis (H_{red}) is equivalent to indexing in the original basis (H). Compute a QR decomposition of B_{red} :

$$B_{\text{red}} = Q_B R_B, \quad Q_B \in O(3), \quad R_B \text{ upper triangular with } (R_B)_{ii} > 0,$$

where $O(3)$ is the group of orthogonal 3×3 matrices (rotations and reflections), and the diagonal entries $(R_B)_{ii}$ are the Gram-Schmidt lengths of the reduced basis vectors. For a candidate orientation $R \in SO(3)$ (the group of 3×3 rotation matrices), we rotate the measured reciprocal vectors Q_i into the reduced, orthonormal coordinates via

$$y_i := Q_B^\top R^\top Q_i.$$

In these coordinates the noiseless model is

$$y_i = R_B H_{\text{red}, i},$$

so each measurement lies exactly on the lattice generated by R_B when the Miller indices are correct. Babai (nearest-plane) decoding then reduces to triangular back-substitution using R_B .

3.5 Objective Function For a Given Orientation

For a given orientation $R \in SO(3)$ the indexer should reward configurations where many observed peaks $\{Q_i\}_{i=1}^N$ lie close to reciprocal-lattice points of the form

$$Q \approx RBH, \quad H \in \mathbb{Z}^3,$$

up to lattice symmetry. We therefore define an orientation-dependent loss $L(R)$ that penalizes the distance from each rotated peak to its nearest symmetry-equivalent lattice point, and then aggregates these distances in a robust way. The same construction is used for the quaternion

parameterization q via $R = R(q)$.

Symmetry-aware per-peak residual. Let $\{S_\ell\}_{\ell=1}^L$ denote the reciprocal-space rotation matrices representing the point-group (lattice) symmetries of the unit cell. For a candidate orientation R and peak Q_i , we first rotate into the crystal frame and then apply each symmetry:

$$v_{i,\ell}(R) := S_\ell^\top R^\top Q_i.$$

Expressing the residuals in the reduced QR coordinates proceeds in two steps. First, for each symmetry operator S_ℓ we define

$$v_{i,\ell}(R) = S_\ell^\top R^\top Q_i.$$

Next, projecting into the reduced coordinates gives

$$y_{i,\ell}(R) := Q_B^\top v_{i,\ell}(R) = Q_B^\top S_\ell^\top R^\top Q_i = R_B H_{i,\ell}^* + \varepsilon_{i,\ell}.$$

for some (unknown) integer $H_{i,\ell}^* \in \mathbb{Z}^3$ and noise $\varepsilon_{i,\ell}$. We decode $H_{i,\ell}^*$ by applying Babai's nearest-plane algorithm in the reduced basis:

$$\hat{H}_{i,\ell}^{\text{red}}(R) := \text{Babai}(R_B^{-1} y_{i,\ell}(R)) \text{ (nearest-plane lattice decoding in the } R_B \text{ basis)} \in \mathbb{Z}^3,$$

and map back to the original reciprocal basis via $H_{i,\ell}(R) := U \hat{H}_{i,\ell}^{\text{red}}(R)$. The corresponding predicted reciprocal-space vector in the lab frame is

$$\hat{Q}_{i,\ell}(R) := R S_\ell B H_{i,\ell}(R).$$

We define the symmetry-aware squared residual for peak i as

$$r_i^2(R) := \min_{\ell \in \{1, \dots, L\}} \|Q_i - \hat{Q}_{i,\ell}(R)\|_2^2,$$

and write $r_i(R) = \sqrt{r_i^2(R)}$. Intuitively, $r_i(R)$ is the best fit to peak Q_i obtainable by applying any lattice symmetry followed by nearest-lattice decoding in the reduced basis.

Robust aggregation. Given $\{r_i(R)\}_{i=1}^N$ for a fixed R , we want a scalar loss $L(R)$ that is robust to false peaks and noise. We use two closely related objectives, depending on the number of peaks N : for $N > N_{\text{small}}$ we use a trimmed mean of squared residuals with a self-calibrating scale estimate; for $N \leq N_{\text{small}}$ we use a softened nearest-lattice objective that more carefully resolves individual peaks before trimming. In both cases a trim fraction $\kappa \in (0, 1]$ controls the fraction of peaks that we want to be explained by a given orientation. The higher the estimated fraction of false peaks, the lower the value of κ should be.

3.5.1 Large- N loss L_{big}

For larger peak sets we rely on the fact that, for a good orientation R , a nontrivial fraction of peaks cluster near the lattice while the remainder are outliers. We therefore aggregate the symmetry-aware residuals $\{r_i^2(R)\}_{i=1}^N$ using a two-stage robust procedure: trimming followed by winsorization.

Trimming and winsorization. Let $r_i^2(R)$ denote the squared residual for peak i as defined above, and let

$$h := \lceil \kappa N \rceil$$

for a user-chosen trim fraction $\kappa \in (0, 1]$. We first select the h peaks with the smallest squared residuals,

$$\mathcal{I}_\kappa(R) := \arg \min_{I \subset \{1, \dots, N\}, |I|=h} \sum_{i \in I} r_i^2(R),$$

which corresponds to trimming away the worst $(1 - \kappa)N$ peaks. On this trimmed set $\mathcal{I}_\kappa(R)$ we compute an empirical upper cutoff as a high quantile of the retained residuals. Specifically, letting

$$\{s_j(R)\}_{j=1}^h := \{r_i^2(R) : i \in \mathcal{I}_\kappa(R)\}$$

be the multiset of trimmed squared residuals, we define

$$t^2(R) := \text{Quantile}_{0.90}(\{s_j(R)\}_{j=1}^h),$$

the empirical 90th percentile of the h smallest squared residuals. We then winsorize the trimmed residuals by clipping any $s_j(R)$ that exceeds $t^2(R)$:

$$\tilde{s}_j(R) := \min\{s_j(R), t^2(R)\}, \quad j = 1, \dots, h.$$

Intuitively, the trimming step enforces that any plausible orientation must explain at least a κ fraction of peaks with relatively small residuals, while the winsorization step limits the influence of the worst points even among those retained as inliers. This construction does not assume any specific parametric form for the residual distribution; it uses only order statistics.

Definition of L_{big} . The large- N loss is defined as the mean of the winsorized squared residuals on the trimmed set,

$$L_{\text{big}}(R) := \frac{1}{h} \sum_{j=1}^h \tilde{s}_j(R).$$

Equivalently,

$$L_{\text{big}}(R) = \frac{1}{|\mathcal{I}_\kappa(R)|} \sum_{i \in \mathcal{I}_\kappa(R)} \min\{r_i^2(R), t^2(R)\}.$$

The additional winsorization makes the loss less sensitive to a small number of unusually large

residuals even among the κN best peaks. In practice this combination provides a high degree of robustness to false positives and noise when many reflections are available.

3.5.2 Small- N loss L_{small}

When the number of peaks N is small (e.g. $N \leq N_{\text{small}}$), pure nearest-plane decoding can be sensitive: a single rounding error on one reflection may flip its Miller indices, and with few peaks such misassignments are not averaged out (see Section 3.7.3). To mitigate this, we augment Babai with a small local search in index space around each candidate label before trimming.

Local integer neighborhood. For each orientation R and symmetry operator S_ℓ , Babai applied in the reduced basis produces a baseline label

$$H_{i,\ell}^{\text{Babai}}(R) \in \mathbb{Z}^3$$

for peak Q_i . We then consider a small cubic neighborhood of integer offsets

$$\Delta \in \{-\delta, \dots, \delta\}^3 \subset \mathbb{Z}^3,$$

where $\delta \in \mathbb{N}$ is a small radius (in our implementation, $\delta = 1$ to account for rounding error, so there are $K = (2\delta + 1)^3 = 27$ offsets). For each offset Δ , we form the candidate index

$$H_{i,\ell,\Delta}(R) := H_{i,\ell}^{\text{Babai}}(R) + \Delta$$

and the corresponding predicted reciprocal-space vector in the lab frame

$$\hat{Q}_{i,\ell,\Delta}(R) := R S_\ell B H_{i,\ell,\Delta}(R).$$

The associated squared residual is

$$\rho_{i,\ell,\Delta}^2(R) := \|Q_i - \hat{Q}_{i,\ell,\Delta}(R)\|_2^2.$$

Per-peak residual for small N . For each peak i we take the best fit over both symmetry and local integer offsets:

$$r_i^2(R) := \min_{\ell \in \{1, \dots, L\}} \min_{\Delta \in \{-\delta, \dots, \delta\}^3} \rho_{i,\ell,\Delta}^2(R).$$

Equivalently, $r_i^2(R)$ is the squared distance from Q_i to the closest lattice point reachable by changing the Babai label by at most δ in each Miller index, after accounting for lattice symmetries.

Trimming and definition of L_{small} . Given the per-peak residuals $\{r_i^2(R)\}_{i=1}^N$ for a fixed orientation, we apply the same least-trimmed-squares idea as in the large- N case, but without

additional winsorization. Let

$$h := \lceil \kappa N \rceil,$$

and let $\mathcal{I}_\kappa^{\text{small}}(R)$ be the indices of the h smallest squared residuals:

$$\mathcal{I}_\kappa^{\text{small}}(R) := \arg \min_{I \subset \{1, \dots, N\}, |I|=h} \sum_{i \in I} r_i^2(R).$$

The small- N loss is then defined as

$$L_{\text{small}}(R) := \frac{1}{h} \sum_{i \in \mathcal{I}_\kappa^{\text{small}}(R)} r_i^2(R).$$

Thus L_{small} is an LTS objective built on top of a more careful, locally refined per-peak residual: for each reflection, we allow the Miller indices to move within a small cube around Babai's guess and pick the best nearby lattice point. This strictly enlarges the region in which a reflection is assigned its correct indices compared to pure Babai decoding, which is especially important when only a small number of peaks are available and individual rounding errors cannot be averaged out.

3.6 Callback Function and Local Refinement

The DE optimizer in Section 3.2 explores the orientation space globally, but its population is constrained to a finite step size and may approach a good basin only coarsely. To improve accuracy once promising orientations have been found, we interleave DE with a local refinement step implemented as a callback: at prescribed generations, the current best orientation is extracted, locally refined using lattice information (Babai decoding and Procrustes alignment), and the refined orientation is injected back into the population if it improves the loss.

Inlier selection and relabeling. Given a candidate orientation R and per-peak residuals $r_i^2(R)$ (Section 3.5), we first identify a set of inlier reflections $\mathcal{I}_\kappa(R)$ of size $h = \lceil \kappa N \rceil$ by keeping the h smallest residuals (large- N case) or using the corresponding small- N trimmed set. For each $i \in \mathcal{I}_\kappa(R)$ we then recompute Miller indices using Babai decoding (and local search when $N \leq N_{\text{small}}$):

$$y_i := Q_B^\top R^\top Q_i, \quad \hat{H}_i^{\text{red}}(R) := \text{Babai/Local}(R_B, y_i), \quad \hat{H}_i(R) := U \hat{H}_i^{\text{red}}(R).$$

The corresponding lattice predictions in the lab frame are

$$\hat{Q}_i(R) := B \hat{H}_i(R).$$

Orthogonal Procrustes refinement. Conditioned on the (re)estimated indices $\{\hat{H}_i(R)\}_{i \in \mathcal{I}_\kappa(R)}$, we refine the orientation by solving an orthogonal Procrustes problem on the inlier set:

$$R_{\text{ref}} := \arg \min_{R \in SO(3)} \sum_{i \in \mathcal{I}_\kappa(R)} \|Q_i - R \hat{Q}_i(R)\|_2^2.$$

Let

$$C := \sum_{i \in \mathcal{I}_\kappa(R)} Q_i \hat{Q}_i(R)^\top \in \mathbb{R}^{3 \times 3}$$

be the 3×3 cross-covariance matrix between observed and predicted reciprocal-space vectors. Writing the singular value decomposition as $C = U_C \Sigma V_C^\top$, the unique Procrustes solution in $SO(3)$ is

$$R_{\text{ref}} = U_C D V_C^\top, \quad D := \text{diag}(1, 1, \det(U_C V_C^\top)).$$

This R_{ref} is the least-squares optimal rotation mapping the predicted lattice vectors $\hat{Q}_i(R)$ to the observed peaks Q_i over the inlier set.

Callback logic and early stopping. In practice we expose this refinement as a callback from the DE optimizer: at each refinement checkpoint we take the current best quaternion q_{DE} , form $R_{\text{DE}} = R(q_{\text{DE}})$, compute the refined orientation R_{ref} as above, and, if $L(R_{\text{ref}}) < L(R_{\text{DE}})$, replace the best individual in the population by R_{ref} . The optimization is stopped either when DE’s built-in convergence criterion on the population (controlled by its tolerance parameter) is satisfied, or when a stagnation criterion is met (the best loss has not improved by more than a small threshold over a fixed number of generations).

3.7 Proofs and Theoretical Justification

3.7.1 Babai Advantages Over Least-Squares Rounding

A very common heuristic for integer decoding is *least-squares plus rounding*: given a basis A and a point x , one solves the unconstrained least-squares problem

$$h_{\text{LS}} := \arg \min_{h \in \mathbb{R}^n} \|Ah - x\|_2^2 = A^+ x,$$

and then rounds h_{LS} coordinatewise to the nearest integer vector. This procedure can perform well when A is close to orthogonal, but in general it admits no known dimension-only worst-case approximation guarantee for the closest vector problem (CVP): for sufficiently skew bases, the distance from the LS+round solution to the lattice can be made arbitrarily larger than the true CVP distance (see, e.g., the discussion in (Agrell *et al.*, 2002)).

Babai’s nearest-plane algorithm (Babai, 1986), applied in QR coordinates, *does* come with such a guarantee. For any lattice $\Lambda(A)$ in dimension n and any target $t \in \mathbb{R}^n$, the Babai point x_{Babai}

satisfies

$$\|x_{\text{Babai}} - t\|_2 \leq \gamma_n \text{dist}(t, \Lambda(A)),$$

for some constant γ_n depending only on n and the quality of the basis. For example, using an LLL-reduced basis one obtains a bound of the form $\gamma_n \leq 2^{n/2}$, with significantly smaller factors available for stronger reductions (Agrell *et al.*, 2002). Since we always apply Babai in a reduced reciprocal basis B_{red} (Section 3.4), our decoder inherits a bounded worst-case CVP error factor, whereas LS+rounding in the original reciprocal basis does not.

Additionally, if we model the noise in the QR coordinates by a Gaussian integer least-squares model

$$y = Ah + v, \quad h \in \mathbb{Z}^n, \quad v \sim \mathcal{N}(0, \sigma^2 I),$$

then results from the box-constrained OILS (optimal integer least squares) literature (“box-constrained” meaning the integer vector h is assumed a priori to lie in a finite hyper-rectangular index set) give a complementary probabilistic comparison: when h is (approximately) uniform over this bounded index set, the success probability of the box-constrained rounding detector is always no larger than that of the corresponding Babai detector for the same A and σ (Wen & Yin, 2013). Under this simple stochastic model, Babai decoding is therefore at least as reliable as LS+rounding in terms of the probability of assigning the correct integer vector, in addition to enjoying the worst-case geometric bounds above.

3.7.2 Basis Reduction Advantage with Babai

Setup (reduced coordinates). Fix a unimodular $U \in SL(3, \mathbb{Z})$ and set $B_{\text{red}} := BU$. Let $B_{\text{red}} = Q_B R_B$ be a QR factorization with $Q_B \in O(3)$ and R_B upper-triangular with $(R_B)_{ii} > 0$. For a single measured reflection Q_i , we model it in the reduced basis as

$$x = B_{\text{red}} h + \eta, \quad h \in \mathbb{Z}^3,$$

where h are the (unknown) Miller indices and η is measurement noise in reciprocal space. Premultiplying by $Q_B^\top$ gives

$$y := Q_B^\top x = R_B h + \varepsilon, \quad \varepsilon := Q_B^\top \eta. \quad (\text{noise in orthonormal QR coordinates})$$

In these coordinates, the indexing problem for one reflection is: given y , recover the correct integer vector h .

Decoder (Babai / nearest-plane). Babai’s nearest-plane decoder in these coordinates is

$$\hat{h}_i^{\text{Babai}} = \text{round}\left(\frac{y_i - \sum_{j>i} (R_B)_{ij} \hat{h}_j^{\text{Babai}}}{(R_B)_{ii}}\right),$$

applied successively for $i = 3, 2, 1$. This is the standard lattice-decoding rule for an upper-triangular generator matrix. Define the success set

$$\mathcal{S}_{\text{Babai}} = \{\varepsilon : \hat{h}^{\text{Babai}}(y) = h \text{ when } y = R_B h + \varepsilon\},$$

and, for a given noise distribution of ε , the per-reflection success probability

$$P_{\text{succ}}(R_B) = \Pr(\varepsilon \in \mathcal{S}_{\text{Babai}}).$$

In crystallographic terms, $P_{\text{succ}}(R_B)$ is the probability that a single noisy reciprocal-space peak is assigned the correct Miller indices when expressed in the reduced basis.

Theorem (Babai success region and Gram-Schmidt diagonals). Define the axis-aligned box

$$\mathcal{B}(R_B) = \prod_{i=1}^3 \left(-\frac{1}{2}(R_B)_{ii}, \frac{1}{2}(R_B)_{ii} \right).$$

(A) Characterization of the success region. Babai's decoder succeeds if and only if the noise coordinates satisfy

$$|\varepsilon_i| < \frac{1}{2}(R_B)_{ii} \quad \text{for all } i = 1, 2, 3.$$

Equivalently,

$$\mathcal{S}_{\text{Babai}} = \mathcal{B}(R_B) \quad \Rightarrow \quad P_{\text{succ}}(R_B) = \Pr(\varepsilon \in \mathcal{B}(R_B)).$$

Intuition. Decoding proceeds from $i = 3$ up to 1. Once the higher indices are decoded correctly, the i -th step rounds

$$\frac{y_i - \sum_{j>i} (R_B)_{ij} h_j}{(R_B)_{ii}} = h_i + \frac{\varepsilon_i}{(R_B)_{ii}}.$$

If $|\varepsilon_i| < (R_B)_{ii}/2$, the nearest integer is the true h_i ; conversely, if Babai returns the correct h_i , then necessarily $|\varepsilon_i| < (R_B)_{ii}/2$. Requiring this for all i yields the box $\mathcal{B}(R_B)$.

(B) Enlarging the Gram-Schmidt diagonals increases P_{succ} . Fix the noise distribution of ε in the reduced coordinates. Let R_B and R'_B be upper-triangular matrices with positive diagonals, and suppose

$$(R'_B)_{ii} \geq (R_B)_{ii} \quad \text{for all } i = 1, 2, 3.$$

Then Babai's success probability cannot decrease:

$$P_{\text{succ}}(R'_B) = \Pr(\varepsilon \in \mathcal{B}(R'_B)) \geq \Pr(\varepsilon \in \mathcal{B}(R_B)) = P_{\text{succ}}(R_B),$$

because enlarging each diagonal enlarges the box,

$$\mathcal{B}(R_B) \subseteq \mathcal{B}(R'_B),$$

and hence increases its measure under any fixed noise distribution for ε .

Relation to reciprocal-basis reduction. For a fixed lattice, unimodular changes of basis preserve the lattice volume, $\det B_{\text{red}} = \prod_i (R_B)_{ii}$, so it is impossible to enlarge all Gram–Schmidt lengths simultaneously via basis changes alone. In crystallographic practice, however, conventional reciprocal cells for high-symmetry systems (cubic, tetragonal, orthorhombic, hexagonal, trigonal) are already close to reduced; LLL/KZ reduction (Wen & Chang, 2017) typically returns the same (or close to) basis up to permutations and sign changes, so $P_{\text{succ}}(R_B)$ is essentially unchanged in these cases. For low-symmetry or skew cells (e.g., triclinic/monoclinic, or poorly chosen primitive bases), reciprocal-lattice reduction algorithms such as LLL or KZ are explicitly designed to remove nearly dependent and highly correlated basis vectors: they tend to increase the smallest Gram–Schmidt lengths and produce a more balanced, less skew basis while keeping the volume fixed. Since $P_{\text{succ}}(R_B)$ is coordinatewise non-decreasing in the diagonals $(R_B)_{ii}$, improving these “weak directions” enlarges the Babai success region and, under the continuous and moderately anisotropic noise distributions for ε that arise from detector geometry and peak integration in serial crystallography, tends to maintain or improve the overall success probability of Babai decoding. In particular, this translates into a higher chance of assigning correct Miller indices to noisy peaks, and provides a strong practical justification for applying reciprocal-basis reduction before Babai decoding in our indexing pipeline.

3.7.3 Small- N Local Search Justification and Advantage

For a fixed orientation R and basis R_B , let $Z_i \in \{0, 1\}$ indicate whether Babai decoding assigns the correct Miller indices to reflection i , with $\Pr(Z_i = 1) = p$ under an i.i.d. noise model in the QR coordinates. The empirical fraction of correctly decoded reflections is

$$\hat{p}_N := \frac{1}{N} \sum_{i=1}^N Z_i, \quad \mathbb{E}[\hat{p}_N] = p, \quad \text{Var}(\hat{p}_N) = \frac{p(1-p)}{N}.$$

By the weak law of large numbers, $\hat{p}_N \rightarrow p$ in probability as $N \rightarrow \infty$, so for large N the effect of individual Babai errors on the orientation score is likely averaged out after being handled by trimming. When N is small, however, $\text{Var}(\hat{p}_N)$ is large and a single decoding error changes $\hat{p}_N$ by $1/N$ (e.g. 0.2 for $N = 5$), making the loss $L(R)$ much more sensitive to per-reflection rounding mistakes.

Our small- N loss L_{small} mitigates this by augmenting Babai with a local integer neighborhood $\Delta \in \{-\delta, \dots, \delta\}^3$, $\delta = 1$ around the Babai label. Denote by E_{Babai} the event that pure Babai yields the correct h_i , and by E_{local} the event that the correct h_i lies within this neighborhood and is selected by the local search. By construction (and under the mild assumption that the correct lattice point is the unique best fit within this neighborhood), we have

$$E_{\text{Babai}} \subseteq E_{\text{local}},$$

and therefore

$$p' := \Pr(E_{\text{local}}) \geq \Pr(E_{\text{Babai}}) = p.$$

Thus the local search typically enlarges the per-reflection success region, increasing the probability that peaks are correctly indexed and stabilizing the orientation score in the small- N regime.

4 Results

In this section we evaluate the proposed indexing algorithm on publicly available serial femtosecond crystallography datasets from the Coherent X-ray Imaging Data Bank (CXIDB) (Maia, 2012). We focus on protein systems collected at XFEL facilities, spanning different beamlines and detector geometries. For each dataset, all indexers are run using identical CrystFEL settings (White *et al.*, 2016), with retries, and the unit-cell parameters are provided to every method, as our algorithm requires it. We enable the default validation checks in `indexamajig`.

4.1 Datasets

We benchmark our method on three CXIDB entries: 61, 62, and 83.

CXIDB 61: POMGnT1 stem domain. Entry 61 contains a serial femtosecond crystallography dataset of the stem domain of human POMGnT1, collected at SACLA (beamline BL3) with a photon energy of 13.0 keV (0.954 Å) (Yamashita *et al.*, 2017). The dataset was originally acquired for experimental phase determination using SeMet or Hg derivatization, and is distributed in CXIDB with the corresponding MPCCD detector geometry and associated PDB structures (5XFC, 5XFD, 5XFE).

CXIDB 62: ACG (Agrocybe cylindracea galectin). Entry 62 provides a second SFX dataset from the same study (Yamashita *et al.*, 2017), comprising diffraction images of ACG (Agrocybe cylindracea galectin) collected at SACLA BL3 under nominally identical beam conditions (13.0 keV, 0.954 Å). As for CXIDB 61, the data are supplied with facility and refined CrystFEL geometry files and share the same associated PDB entries (5XFC, 5XFD, 5XFE).

CXIDB 83: β -lactamase at megahertz repetition rate. Entry 83 corresponds to one of the datasets from the “Megahertz serial crystallography” experiment at the European XFEL (Wiedorn *et al.*, 2018). It contains diffraction patterns of a complex of β -lactamase collected at the SPB/SFX beamline with a photon energy of 9.30 keV (1.33 Å) at megahertz repetition rates, and is distributed together with detector geometry and the associated PDB structures (6GTH, 6FTR). Compared to the SACLA datasets, CXIDB 83 features different detector layout, noise characteristics, and pulse structure, providing a complementary test case for robust indexing.

4.2 Indexing rate

Table 1 presents the overall indexing rate (in %) for the three CXIDB datasets when using our method, TORO (Gasparotto *et al.*, 2024), and XGANDALF (Gevorkov *et al.*, 2019) within CrystFEL (White *et al.*, 2016). Across all systems, our indexer gives the highest indexing rate: it is slightly better than XGANDALF on CXIDB 61, ahead by a few percent of both TORO and XGANDALF on CXIDB 62, and slightly better than TORO on CXIDB 83.

The main advantage/use case of our algorithm is shown in 4.2.1, where we focus on the small N patterns.

Table 1: Indexing rate (in %) on all diffraction patterns for the three CXIDB datasets. The total number of processed patterns is given in parentheses.

Dataset (patterns)	Ours	TORO	XGANDALF
CXIDB 61 (35 294)	90.6	88.2	89.7
CXIDB 62 (28 835)	95.2	93.0	92.6
CXIDB 83 (17 043)	91.5	90.6	85.3

A detailed step-by-step description of the experimental procedure to obtain indexing rates, including the parameters used in CrystFEL, will be provided in the Appendix

4.2.1 Indexing rate for $N \leq 25$

To highlight performance in the challenging small- N regime, we restrict the analysis to patterns with at most 25 detected reflections. Table 2 summarizes the indexing rate on this subset. With $N \leq 25$, our method indexes substantially more patterns than TORO and XGANDALF on all three datasets. In contrast to the baselines, which suffer a marked drop in indexing rate for sparse patterns, our indexer remains close to its full-dataset performance, underscoring the effectiveness of the small- N loss and local refinement.

Table 2: Indexing rate (in %) restricted to patterns with at most 25 reflections ($N \leq 25$). The number of such patterns is given in parentheses.

Dataset ($N \leq 25$ patterns)	Ours	TORO	XGANDALF
CXIDB 61 (2 657)	97.2	79.6	85.9
CXIDB 62 (3 881)	95.9	74.2	72.3
CXIDB 83 (2 059)	89.2	76.3	83.2

4.2.2 Indexing Rate vs. Number of Reflections

Figures 1a–1c show the indexing rate as a function of the number of detected reflections N for CXIDB 61, 62, and 83. Across all datasets, the three indexers converge to similar performance once a sufficiently large number of reflections is available, except for CXIDB 83m where XGandalf

underperforms across most N . However, in the sparse-reflection regime (small N), our method consistently maintains a clear advantage: it yields higher indexing rates for small N on all datasets, and remains competitive for larger N . These results reinforce the trends observed in Table 2, confirming that the proposed small- N loss and refinement steps provide a decisive benefit when only a few reflections are present, while remaining competitive for well-populated patterns.

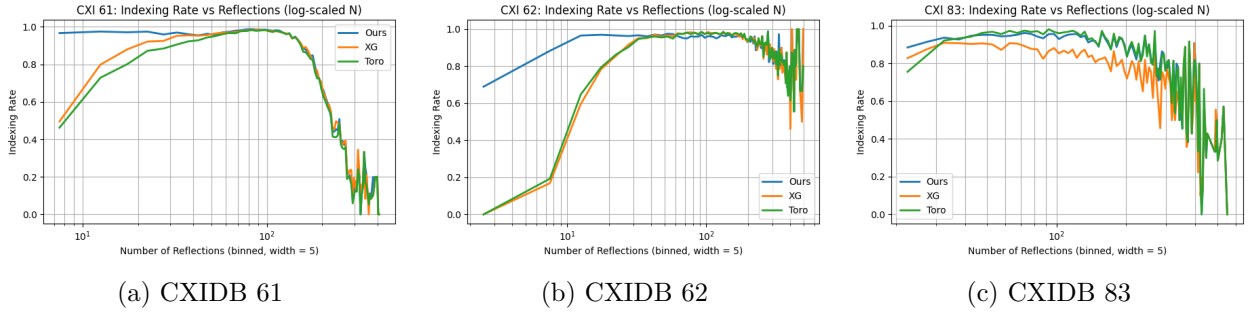


Figure 1: Indexing rate vs. number of reflections for CXIDB 61, 62, and 83.

4.3 Runtime and Memory Usage

To assess computational efficiency, we measured the wall-clock time and GPU memory consumption when indexing batches of diffraction patterns. Runtime was recorded using Python’s `time` module, while GPU memory was measured through PyTorch’s built-in memory profiling utilities. All GPU experiments were performed on the same NVIDIA A100 under identical software and memory constraints, ensuring that the comparison between indexers reflects algorithmic differences rather than hardware variations. Tables 3 and 4 report the runtime per batch (with the batch size indicated) together with the peak GPU memory allocated and reserved during execution. Additional implementation details on the measurement setup will be provided in Appendix.

Table 3: Runtime per batch and peak GPU memory usage for 100 random CXI62 patterns with $N = 33$ reflections.

HW	Indexer	Batch	Time / batch (s)	GPU mem. (alloc/res) [GB]
AMD EPYC (CPU)	TORO	1	0.37 ± 0.05	N/A
AMD EPYC (CPU)	Ours	1	1.19 ± 0.57	N/A
NVIDIA A100 (GPU)	TORO	100	0.56 ± 0.09	11.69 / 22.51
NVIDIA A100 (GPU)	Ours	100	7.06 ± 0.09	0.41 / 0.55

Table 4: Runtime per batch and peak GPU memory usage for 100 random CXI62 patterns with $N = 15$ reflections.

HW	Indexer	Batch	Time / batch (s)	GPU mem. (alloc/res) [GB]
AMD EPYC (CPU)	TORO	1	0.26 ± 0.05	N/A
AMD EPYC (CPU)	Ours	1	3.51 ± 1.64	N/A
NVIDIA A100 (GPU)	TORO	100	0.16 ± 0.03	5.34 / 10.26
NVIDIA A100 (GPU)	Ours	100	10.34 ± 0.03	1.21 / 1.34

Across GPU experiments, our indexer is markedly more memory-efficient: for identical batch sizes, it requires orders of magnitude less GPU memory than TORO. However, in terms of raw speed, TORO is consistently much faster, typically by at least one order of magnitude on both CPU and GPU. This difference is expected and arises from several characteristics of our optimization procedure:

1. **Iteration-heavy optimization.** Differential Evolution (DE) requires many population updates, each involving several function evaluations. Consequently, even modest per-iteration cost accumulates substantially over the optimization trajectory.
2. **Exploration-first orientation search.** We use DE to explore the orientation space. While DE provides strong robustness to noise, helps escape local minima, and has low memory requirements on GPU, its crossover–mutation cycle introduces non-negligible computational overhead. These operations are inexpensive individually, but the cumulative effect over many iterations leads to higher runtime.
3. **Additional local search for small- N patterns.** For patterns with $N \leq 25$, we apply a small local-search step. As established in Section 3.7.3, this refinement is critical for small- N robustness but introduces extra computational cost. The increase in runtime is justified: Table 2 and Figures 1a–1c show that our method dramatically outperforms other indexers in the low-reflection regime, where noise and ambiguity are typically highest.

Notably, because DE is stochastic, it also benefits from multiple retries: rerunning the population with different random initializations can lead to a total different exploration strategy of the space, which is particularly valuable for noisy datasets with narrow valleys.

From a deployment perspective, current real-time pipelines at XFEL facilities typically process diffraction patterns sequentially with tight per-pattern latency constraints. In this online setting, our present implementation which relies on DE iterations and an additional refinement step for small- N patterns is better suited to offline reprocessing than to live feedback, especially in memory-constrained environments, and when the number of small- N images is dominant and noisy. **Nevertheless**, because most of the computational overhead comes from DE’s exploration of the orientation space and the need for many iterations to converge, a promising direction for future work, particularly if one wishes to prioritize speed over memory usage such as in online deployment, is to replace our

Differential Evolution–based exploration of the orientation space with a faster sampling strategy inspired by TORO (Gasparotto *et al.*, 2024), while retaining the objective functions introduced in this work.

Overall, although TORO achieves superior runtime on both CPUs and GPUs, our method provides competitive throughput on modern GPUs and maintains extremely low memory usage due to the population-based nature of DE, which trades memory for iterative refinement. Given the significantly higher indexing rates we obtain, especially for small- N patterns, the additional runtime is often justified and acceptable for offline processing. In addition, the much smaller memory footprint means that, on GPUs, substantially larger batch sizes can be used on the same hardware, which can narrow the effective gap in frames processed per second compared to TORO. Similarly, on multi-core CPUs, the lower per-process memory usage allows more workers to be run in parallel before exhausting system memory, further improving throughput.

5 Discussion and Future Work

5.1 Advantages

Our approach offers several key strengths. First, the use of Differential Evolution (DE), combined with our objective functions and Babai-based evaluation, yields strong robustness to noise and maintains high accuracy even in the small- N regime (with the help of the local-search), where most indexers degrade.

Second, using Babai’s nearest-plane algorithm instead of least-squares rounding can improve the reliability of Miller index assignment, especially for skewed or ill-conditioned unit cells. This is further reinforced by reciprocal-space basis reduction, which produces a more orthogonal basis and improves the success probability of Babai’s nearest-plane algorithm as seen in [3.7.1](#).

Finally, because DE is population-based and does not rely on dense orientation grids or large precomputed structures, the method is extremely memory-efficient on GPUs, often an order of magnitude lighter than TORO for identical batch sizes. This low memory usage can be traded for higher parallelism, either through larger GPU batch sizes or more concurrent CPU workers, which helps tighten the effective speed gap relative to TORO despite higher per-batch runtime.

These properties make the algorithm particularly attractive for indexing of noisy datasets and scenarios dominated by low- N patterns, especially since DE can benefit from multiple stochastic retries.

5.2 Limitations

The main limitation of the current implementation is runtime. DE requires many generations and repeated objective evaluations to sufficiently explore the orientation space, and the additional refinement step used for small- N patterns further increases the computational cost. As a result, the

method is considerably slower than TORO and XGANDALF on both CPUs and GPUs and is not suited for online deployment, where low latency per diffraction pattern is essential.

5.3 Applications and Integration

Given its accuracy and low memory requirements, the method is well aligned with offline reprocessing pipelines, where full datasets are available on disk and batching can be exploited to achieve high throughput. In contrast, real-time pipelines process images sequentially and require rapid per-pattern indexing, making our current DE-based approach less suitable for live feedback.

Future Work. A promising direction for future work is to accelerate orientation and make it suitable for online pipelines by replacing DE with a more speed-efficient sampling strategy, such as the basis-orientation search used in TORO (Gasparotto *et al.*, 2024), while preserving the robustness advantages of our objective functions. Such a hybrid design could potentially bridge the gap between the small- N performance of our method and the low-latency requirements of online processing, enabling deployment in both offline and live-crystallography contexts.

6 Conclusion

We introduced a new crystallographic indexing algorithm that combines Differential Evolution-based orientation search, a symmetry-aware objective function, Babai-based integer assignment, and reciprocal-space basis reduction. Our method assumes known/estimated unit-cell parameters. Across three XFEL datasets, the method achieves consistently higher indexing rates than existing indexers, particularly for small- N patterns where methods such as TORO and XGANDALF experience significant degradation. This robustness arises from the combination of our symmetry-aware loss, the use of Babai’s nearest-plane algorithm in a reduced reciprocal basis, and a dedicated local search applied when only a few reflections are available.

Although the algorithm is slower than existing methods due to the iterative nature of DE, it remains highly memory-efficient on modern GPUs and is well suited for offline reprocessing, particularly for noisy datasets and cases dominated by sparse diffraction patterns. The improved accuracy, especially for small N , can offset the additional runtime in practical scenarios. Moreover, the markedly lower memory footprint means that larger GPU batches and more parallel CPU workers can be employed on the same hardware, so that, in practice, the effective indexing throughput can approach that of but more memory-intensive methods such as TORO.

Future work involves making the indexer more suitable for online pipelines by replacing DE with a faster orientation-sampling strategy, similar to the one employed in TORO, offering a promising direction for reducing runtime while preserving our advantages in our objective functions and reductions. Such approach could offer variants of our algorithm for both offline and online pipelines, bridging the gap between high accuracy at sparse and noisy reflection patterns and the low-latency

requirements of real-time serial crystallography.

Acknowledgements

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility using NERSC award [ABC]-ERCAP [12345]. This work used the resources of the SLAC Shared Science Data Facility (S3DF) at SLAC National Accelerator Laboratory. S3DF is a shared High-Performance Computing facility, operated by SLAC, that supports the scientific and data-intensive computing needs of all experimental facilities and programs of the SLAC National Accelerator Laboratory. SLAC is operated by Stanford University for the U.S. Department of Energy’s Office of Science.

Funding

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Contract No. DEAC02-76SF00515.

Conflicts of interest: The authors declare no conflicts of interest.

Data availability: The datasets used in this study are publicly available through the Coherent X-ray Imaging Data Bank (CXIDB). Code and implementation details will be released at a later date.

References

- Agrell, E., Eriksson, T., Vardy, A. & Zeger, K. (2002). *IEEE Transactions on Information Theory*, **48**(8).
- Babai, L. (1986). *Combinatorica*, **6**(1).
- Gasparotto, P., Barba, L., Stadler, H.-C., Assmann, G., Mendonça, H., Ashton, A. W., Janousch, M., Leonarski, F. & Béjar, B. (2024). *Journal of Applied Crystallography*, **57**(4).
- Gevorkov, Y., Barty, A., Mariani, V., Yefanov, O., Brehm, W., Tolstikova, A., Chervinskii, F., Chapman, H. N. & White, T. A. (2020). *Acta Crystallographica Section A: Foundations and Advances*, **76**(2), 121–131.
- Gevorkov, Y., Yefanov, O., Barty, A., White, T. A., Tolstikova, A., Wiedorn, M. O., Meents, A., Grigat, R.-R., Chapman, H. N. & Brehm, W. (2019). *Acta Crystallographica Section A: Foundations and Advances*, **75**(5).
- Maia, F. R. N. C. (2012). *Nature Methods*, **9**(9).
- Storn, R. & Price, K. (1997). *Journal of Global Optimization*, **11**(4).
- Wen, J. & Chang, X.-W. (2017). *arXiv preprint arXiv:1702.08152*.
<https://arxiv.org/abs/1702.08152>
- Wen, Z. & Yin, W. (2013). *Mathematical Programming*, **142**(1–2).
- White, T. A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K. R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashita, K., Yoon, C. H., Diederichs, K. & Chapman, H. N. (2016). *Journal of Applied Crystallography*, **49**(2).
- Wiedorn, M. O., Oberthür, D., Bean, R., Schubert, R., Werner, N., Abbey, B., Aepfelbacher, A. *et al.* (2018). *Nature Communications*, **9**(1).
- Yamashita, K., Kuwabara, N., Nakane, T., Murai, T., Mizohata, E., Sugahara, M., Takegawa, S.-i., Ohsaki, Y., Takagi, K., Noguchi, H., Masuda, T., Kato, K., Nango, E., Tono, K., Joti, Y., Kameshima, T., Yabashi, M. & Yamamoto, M. (2017). *IUCrJ*, **4**(5).