Towards single-shot coherent imaging via overlap-free ptychography

January 26, 2026

Abstract

Single-shot coherent imaging is important for XFEL science because it lowers photon dose, removes the overhead of overlapping scans, and enables real-time feedback. We present a *physics-constrained Deep Neural Network (DNN)* that delivers *overlap-free, single-shot* reconstructions in a Fresnel (near-field) CDI geometry and also accelerates conventional multi-shot ptychography. The DNN is *self-supervised*: it learns directly from raw diffraction patterns by enforcing a differentiable forward model of coherent scattering together with a Poisson photon-counting likelihood and a calibrated intensity scale. Two design choices enable robust experimental performance in both single- and multi-shot modes: (i) a *dual-resolution* decoder that prevents exit-wave truncation with realistic, extended probes, and (ii) real-space constraints through coordinate-based grouping that supports arbitrary scan geometries. On APS and LCLS data, the approach reconstructs orders of magnitude faster than iterative solvers, remains accurate at low counts ($\sim 10^4$ photons/frame) and zero overlap, achieves high fidelity with an order of magnitude less training data than supervised baselines, and generalizes across facilities without retraining. In the Fresnel regime we obtain stable *single-shot* reconstructions using only the probe curvature; that is, with no beam multiplexers or modulators. This unifies single-exposure Fresnel CDI and overlapped ptychography within one computational framework, enabling *high-resolution, dose-efficient operando* imaging at XFELs at substantially higher throughput than previously feasible.

1 Introduction

Modern light sources, such as fourth-generation synchrotrons and X-ray Free-Electron Lasers (XFELs), now generate coherent diffraction data at unprecedented rates, often reaching terabytes per second. This data deluge poses a significant challenge for Coherent Diffraction Imaging (CDI) techniques like ptychography, as traditional image reconstruction relies on computationally intensive iterative phase retrieval algorithms. An analytical bottleneck has emerged where offline processing time vastly exceeds data acquisition time. This disparity not only delays scientific insights but, more critically, precludes real-time feedback and on-the-fly experimental steering—capabilities essential for maximizing the efficiency and discovery potential of these facilities. Consequently, there is a pressing need for new reconstruction paradigms that deliver high-fidelity results at accelerated rates without sacrificing reliability or robustness.

Ptychographic coherent diffractive imaging is a cornerstone technique for x-ray nanoscale imaging at synchrotron and XFEL facilities [5]. However, classical iterative algorithms like the Ptychographic Iterative Engine (PIE) require $\sim 60-70\%$ scan overlap for robust convergence and can only process $\sim 0.1-1$ diffraction patterns per second on standard hardware [2, 9]. At high-repetition-rate sources (e.g., LCLS-II, MHz-class), data acquisition outpaces reconstruction by orders of magnitude [14]. Consequently, even highly-optimized GPU/HPC solvers require substantial infrastructure and struggle to provide the low-latency feedback needed for interactive experiments [10, 1].

To address this challenge, machine learning (ML) has been explored to accelerate reconstruction. Supervised convolutional networks, for instance, can achieve significant speedups[3]. However, their practical application is limited by poor generalization and resolution [8, 17, 12]. Moreover, these methods impose the onerous burden of generating large, curated datasets of ground-truth reconstructions for training. Our previous work showed that a Physics-Informed Neural Network (PINN) can achieve rapid, self-supervised reconstruction. However, this was demonstrated only on synthetic data with idealized probes and regular scan patterns.

Here, we extend the original framework with several improvements that enable robust reconstruction of actual experimental data. Specifically, we add support for long-tailed probe functions and arbitrary scan patterns.

We additionally explore this framework's capability for *single-shot* reconstructions—using only probe curvature—without beam multiplexers [13, 7] or downstream modulators [16, 4]. Unifying single-exposure Fresnel CDI and overlapped ptychography within one computational framework enables dose-efficient, high-throughput imaging and brings real-time feedback within reach at high-rate XFELs [14].

Inverse problem and constraints. Coherent diffractive imaging requires the recovery of a complex object from intensity-only diffraction measurements. Reconstruction methods enforce two complementary constraints: a reciprocal-space constraint requiring predicted intensities to match data (via a physics-based forward model), and a real-space constraint enforcing consistency between overlapping views. In our framework, the reciprocal-space constraint is enforced directly via a differentiable forward model and a Poisson likelihood. Real-space overlap is handled via a translation-aware merging operator. Crucially, this allows overlap to be treated as a flexible experimental parameter rather than a hard requirement; setting the group size to a single frame ($C_g = 1$) removes overlap constraints entirely, enabling "single-shot" reconstruction when the probe provides sufficient phase diversity.

2 Methods and Architecture

We train a physics-informed neural network (PINN) to perform self-supervised ptychographic reconstruction by composing a learned inverse map with a differentiable forward model of coherent scattering. This section consolidates the formulation, operators, normalization, network design (including extended-probe handling), and training objective.

2.1 Formulation and Forward Model

We learn an inverse map $G : X \to Y$ from diffraction space to real space and optimize it by composing with a differentiable forward model $F : Y \to X$. The overall autoencoder is $F \circ G$, trained to match measured diffraction statistics without ground-truth images.

Data model and notation. Each training sample comprises C_g diffraction amplitude images $\{x_k\}_{k=1}^{C_g}$ acquired at probe coordinates $\{\vec{r}_k\}_{k=1}^{C_g}$. The network $G(x, r)$ outputs C_g complex object patches $\{O_k\}_{k=1}^{C_g}$ on an $N \times N$ grid. We use:

- $\mathcal{T}_{\Delta\vec{r}}[\cdot]$: real-space translation by $\Delta\vec{r}$,
- $\mathrm{Pad}[\cdot]$: zero-padding to a canvas large enough to contain all translated patches,
- $\mathrm{Pad}_{N/4}[\cdot]$: zero-padding that embeds a central $N/2 \times N/2$ tile into an $N \times N$ grid,

- $\mathrm{Crop}_N[\cdot]$: center-cropping to $N \times N$,
- $\mathbf{1}$: an all-ones array of appropriate size,
- $\odot$: elementwise (Hadamard) product.

Constraint map (F_c): translation-aware merging. To enforce overlap consistency, per-patch reconstructions are merged in a translation-aligned frame:

$$O_{\text{region}}(\vec{r}) \;=\; \frac{\sum_{k=1}^{C_g} \mathcal{T}_{-\vec{r}_k}[\mathrm{Pad}(O_k)]}{\sum_{k=1}^{C_g} \mathcal{T}_{-\vec{r}_k}[\mathrm{Pad}(\mathbf{1})] + \epsilon}, \qquad \epsilon = 10^{-3}. \tag{1}$$

This "translational pooling" applies to arbitrary scan geometries.

Coordinate-aware grouping. Training groups are formed locally by nearest-neighbor sampling. For each anchor $\vec{r}_i$, let $\mathcal{N}_K(\vec{r}_i)$ be its K nearest distinct neighbors. A group $\mathcal{G}_{i,j}$ draws $C_g - 1$ neighbors uniformly without replacement:

$$\mathcal{G}_{i,j} = \{\vec{r}_i\} \cup S_{i,j}, \quad S_{i,j} \subset \mathcal{N}_K(\vec{r}_i),\ |S_{i,j}| = C_g - 1,$$

repeated n_{samples} times per anchor. If duplicate neighbor sets are disallowed, the effective number of distinct groups per anchor is

$$n_{\text{eff}} \;=\; \min\left(n_{\text{samples}},\ \binom{K}{C_g - 1}\right),$$

so the total number of training examples is $N_{\text{scan}} \times n_{\text{eff}}$, with the combinatorial upper bound $N_{\text{scan}}\binom{K}{C_g-1}$. Choosing $n_{\text{samples}} > 1$ augments the dataset through combinatorial re-grouping while preserving local spatial consistency.

Coordinates within each group are expressed in a stable local frame by re-centering to the group centroid

$$\vec{r}_{\text{global}} = \frac{1}{C_g}\sum_{k=1}^{C_g} \vec{r}_k, \qquad \vec{r}_k^{\,\text{rel}} = \vec{r}_k - \vec{r}_{\text{global}}.$$

Diffraction map (F_d): coherent scattering. Given O_{region}, the kth translated object patch and exit wave are

$$O'_k(\vec{r}) = \mathrm{Crop}_N\left[\mathcal{T}_{\vec{r}_k^{\,\text{rel}}}(O_{\text{region}})\right], \tag{2}$$

$$\Psi_k = \mathcal{F}\left\{O'_k(\vec{r}) \cdot P(\vec{r})\right\}, \tag{3}$$

where $P(\vec{r})$ is the (estimated) probe and $\mathcal{F}$ is the 2D Fourier transform. Predicted detector-plane amplitudes include a global intensity scale $e^{\alpha_{\log}}$ that links normalized network outputs to physical photon counts:

$$\hat{A}_k \;=\; |\Psi_k|\, e^{\alpha_{\log}}. \tag{4}$$

2.2 Data Preprocessing

A dataset consists of diffraction images from one or more objects measured with a fixed probe illumination P. After grouping images into overlapping samples (Section 2.1), we normalize the raw diffraction amplitudes to ensure numerical stability during training:

$$x_k \;=\; x'_k \cdot \sqrt{\frac{(N/2)^2}{\big\langle \sum_{i,j} |x'_{ij}|^2 \big\rangle}}, \tag{5}$$

where x' denotes raw measurements and the average is over all images in the dataset. This choice ensures order-unity activations in the neural network: by Parseval's theorem, unit-amplitude real-space objects produce diffraction power of approximately $N^2/4$, so this normalization maps experimental amplitude images to internal activations suited to gradient-based optimization.

Additionally, we introduce a trainable scalar $\alpha_{\log}$ that converts between the dimensionless internal model activations and per-pixel integrated amplitudes. As discussed in Section 2.4, the role of $\alpha_{\log}$ is to convert the output *intensity* into physical units of photons per pixel. The final, scaled, network input is $x_{\text{in}} = x \cdot e^{-\alpha_{\log}}$.

2.3 Neural Network Architecture

The inverse map G follows an encoder–decoder design (as in [6]), conditioned on $\{x_k\}_{k=1}^{C_g}$ and $\{\vec{r}_k^{\,\text{rel}}\}_{k=1}^{C_g}$, and outputs complex patches $\{O_k\}_{k=1}^{C_g}$. To respect oversampling while avoiding truncation artifacts from realistic probes with extended tails, the decoder allocates most capacity to the central, well-posed region and a lightweight continuation to the periphery.

Extended probe illumination (dual-resolution decoding). CNN architectures are limited to modest dimensions ($N \leq 128$) and we must furthermore restrict high-resolution reconstruction to the central $N/2 \times N/2$ region to satisfy oversampling conditions [11]. Probes with extended tails force inefficient use of this limited number of pixels because the real-space area brightly illuminated by the probe is small compared to the total probe area that must be represented to avoid truncation artifacts from non-zero amplitude at the edge of the real-space grid.

Consequently, given the modest magnitude of N, fully inscribing the probe—tails included—within the central $N/2 \times N/2$ pixels may require too much binning. This causes a dilemma: one must choose between truncation artifacts (and possible lack of convergence due to the associated physical inconsistency) and violation of the diffraction-space oversampling condition for coherent imaging.

We resolve this by reconstructing the object in high resolution in the central $N/2 \times N/2$ region of the real-space grid and low resolution in the periphery. Presuming the absence of high spatial frequency components in the probe tail, extending the probe times object reconstruction into the periphery does not compromise well-posedness of the inverse problem.

Concretely, we use most channels ($C-4$) of the penultimate decoder layer for the central region and the remaining 4 channels to coarsely reconstruct the periphery:

$$O_{\text{amp}} = \text{Pad}_{N/4}\big(\sigma_A(\text{Conv}(H_A^{\text{central}}))\big) \;+\; \sigma_A(\text{ConvUp}(H_A^{\text{border}})) \odot M_{\text{border}}, \tag{6}$$

$$O_{\text{phase}} = \text{Pad}_{N/4}\big(\pi \tanh(\text{Conv}(H_\phi^{\text{central}}))\big) \;+\; \pi \tanh(\text{ConvUp}(H_\phi^{\text{border}})) \odot M_{\text{border}}, \tag{7}$$

$$O_k = O_{\text{amp}} \cdot \exp\big(i\, O_{\text{phase}}\big), \tag{8}$$

where $H_{\{\cdot\}}^{\text{central}}$ (the first $C-4$ channels) targets the central region, $H_{\{\cdot\}}^{\text{border}}$ (the last 4 channels) produces a low-resolution continuation, and M_{border} is a binary mask that isolates the boundary contributions to the outer region. This modification avoids artifacts from truncation of the exit wave and enables stable reconstruction with experimentally realistic probes.

2.4 Training Objective and Optimization

Poisson negative log-likelihood. The training procedure optimizes the inverse map G using a negative log-likelihood loss under Poisson statistics:

$$\mathcal{L}_{\text{Poiss}} \;=\; -\sum_{k,i,j} \log f_{\text{Poiss}}(N_{kij}; \lambda_{kij}) \;=\; \sum_{k,i,j} \left(\lambda_{kij} - N_{kij}\, \log \lambda_{kij}\right), \tag{9}$$

where $N_{kij} = |x'_{kij}|^2$ is the measured photon count and $\lambda_{kij} = |\hat{A}_{kij}|^2$ is the predicted count.

Since the network operates on normalized inputs (Eq. 5) for numerical stability, a scale parameter $e^{\alpha_{\log}}$ bridges normalized and physical units. When the mean photon flux N_{photons} is known, we initialize:

$$e^{\alpha_{\log}} \;\leftarrow\; \frac{2\sqrt{N_{\text{photons}}}}{N}. \tag{10}$$

This ensures predicted intensities match measurement statistics. The parameter $e^{\alpha_{\log}}$ may be fixed or learned (see Table 3); learning it can absorb modest calibration errors.

Amplitude loss for unknown counts. For datasets lacking absolute photon counts, we resort to mean absolute error on normalized amplitudes:

$$\mathcal{L}_{\text{MAE}} = \sum_{k,i,j} \left| x_{kij} - \hat{A}_{kij} e^{-\alpha_{\log}} \right|.$$

Implementation notes. All operators in F_c and F_d are differentiable and implemented with padding-aware translations and FFT-based diffraction. Batching is performed over groups $\mathcal{G}_{i,j}$; nearest-neighbor sampling with $n_{\text{samples}} > 1$ provides dataset augmentation while preserving local spatial consistency. Default architectural and training hyperparameters are summarized in Table 3.

3 Results

We evaluate on experimental data from the Advanced Photon Source (APS) and Linac Coherent Light Source (LCLS), comparing against iterative solvers and supervised neural network baselines. We also provide quantitative comparisons based on reconstructions of simulated datasets, which allows a genuine comparison to ground truth for both classes of reconstruction methods (neural network and conventional / iterative). The fly64 dataset refers to the experimental 64×64 diffraction data (fly001_64_train*), whereas fly001-synthetic is a semi-synthetic dataset generated by forward-simulating diffraction from a conventional (TIKE) reconstruction of the fly001 experiment and then downsampling; it supplies the ground truth used in Table 1.

3.1 Reconstruction Quality

Figure **??** compares PtychoPINN and PtychoNN on experimental data from a Siemens Star pattern collected at APS (Velociprobe).

For evaluation, we use the top half of the Siemens star pattern for training and the bottom half for testing. PtychoPINN exhibits similar resolution in- and out-of-sample (Figure 3) independent of training set size. The supervised baseline tends to memorize the training samples but shows a considerable loss of image quality in the reconstruction of diffraction images unseen during training.

Semi-synthetic datasets derived from the conventionally-reconstructed object and probe are experimentally realistic and allow quantitative ground-truth comparisons. Here, PtychoPINN achieves an SSIM of 0.962 in phase reconstruction compared to 0.912 for the supervised-CNN baseline (Table 1).

Evaluation of both models on datasets of varying size (Figure 4) demonstrates their relative degrees of in-distribution generalization. At a training set size of 512 images—a stringent test—the supervised baseline (PtychoNN) fails to converge while PtychoPINN produces capable reconstructions.

Table 1: Reconstruction quality metrics at maximum training set size (16,384 images). Values shown are mean ± standard deviation across 5 trials. Best values per dataset are highlighted in green.

		PSNR (dB)		MS-SSIM	
Dataset	Method	Amplitude	Phase	Amplitude	Phase
fly001-synthetic	Baseline (supervised)	84.83 ± 0.23	68.62 ± 0.02	0.930 ± 0.002	0.912 ± 0.003
	PINN (self-supervised)	$\mathbf{85.53 \pm 0.02}$	$\mathbf{70.54 \pm 0.06}$	$\mathbf{0.955 \pm 0.001}$	$\mathbf{0.962 \pm 0.001}$

3.2 Overlap-Free Reconstruction

In overlap-free operation, we set the group size to a single diffraction frame ($C_g = 1$), removing overlap-based real-space consistency. Reconstruction then relies entirely on the diffraction likelihood and the known probe structure (defocused probe/Fresnel geometry). Figure 1 illustrates this single-frame mode compared with multi-position ptychography. While the overlap-free reconstruction shows a slight reduction in fidelity compared to the highly redundant ptychographic case, it successfully resolves the object features, enabling imaging without the need for redundant measurements. Table 2 provides a quantitative summary across overlap and probe-structuring variants.

Table 2: SIM-LINES-4X reconstruction metrics on the stitched test split. Ground truth uses the simulated object, cropped to the scanned region via scan-coordinate alignment. Test evaluation uses a random subsample (nsamples=1000, seed=7). Phase FRC50 is undefined because the ground-truth phase is constant zero.

Case	MAE (Amp)	MAE (Phase)	PSNR (dB) (Amp)	PSNR (dB) (Phase)	SSIM (Amp)
gs1 + idealized probe	0.274615	0.085438	55.702410	59.461522	0.620669
gs1 + custom probe	0.173138	0.047743	57.240780	59.372492	0.785203
gs2 + idealized probe	0.294089	0.027578	55.053092	59.844405	0.505578
gs2 + custom probe	0.188985	0.052160	56.732283	60.144200	0.712271

3.3 Data Efficiency

Figure 4 illustrates the reconstruction quality (phase SSIM) as a function of dataset size. The physics-informed framework maintains high fidelity (SSIM > 0.95) even when trained on as few as 512 diffraction patterns. In contrast, the supervised baseline degrades rapidly below 1024 samples. The horizontal shift between the curves indicates that PtychoPINN achieves comparable quality

using roughly an order of magnitude less training data than the supervised approach. This confirms that enforcing the diffraction forward model acts as a powerful regularizer, reducing the number of samples required to constrain the solution.

3.4 Out-of-distribution Generalization

To evaluate the model's out-of-distribution genrealization, we train on one dataset (APS-velociprobe) data and reconstruct a different one (LCLS-XPP) without retraining. Despite significant differences between the APS zone plate illumination and the LCLS compound refractive lens profile (Figure 2 legend), the physics-informed model successfully reconstructs the LCLS object features (Figure 2, 'PINN' column). The distribution shift causes significant phase distortion and loss of quantitative fidelity, but edge features are surprisingly well conserved.

3.5 Photon-Limited Performance

Figure **??** compares reconstructions trained with Poisson NLL versus MAE loss at low photon counts. The NLL-trained model preserves high-frequency features that MAE loses to noise, demonstrating the importance of correct statistical modeling for dose-efficient imaging.

3.6 Computational Performance

PtychoPINN processes approximately 2000 diffraction patterns per second on an NVIDIA RTX 3090, enabling real-time reconstruction at rates compatible with high-repetition-rate light sources.

4 Discussion

Physics-constrained learning

In classical ptychography, strong translational overlap provides real-space redundancy to make phase retrieval well-posed; but in principle, this redundancy can be derived either from either inter-measurement overlap or from the probe's real-space phase structure within a singular measurement (cite structured illumination microscopy and Fresnel cdi papers). Our framework exploits this by allowing the number of simulataneously-reconstructed images to be a configurable parameter of the model. This flexibility is useful in the Fresnel (curved-probe) regime, where we observe capable single-shot reconstructions. This contrasts with preexsitng ptychography solvers of both the physics-based and data-driven types. The former require overlap constraints by construction, whether or not a particular dataset requires them. The latter either formulate the optimization in the same way as conventional solvers (e.g. implicit neural representation-based networks for bespoke single-dataset reconstruction (cites)) or neglect overlap constraints entirely, leading to subpar imaging performance in the general case where overlaps may needed for the inverse problem to be well-posed (cites).

Dose efficiency

Our framework's Poisson NLL training objective matches actual photon statistics and improves dose efficiency relative to a standard amplitude mean absolute error (Fig. 5). For counts N and predicted rate λ, the per-pixel negative log-likelihood is (ommitting a zero-gradient constant $lam\hat{b}da!$) $L(\lambda) = \lambda - N \log \lambda$. This provides correct maximum-likelihood scoring of the model-predicted λ. In contrast, amplitude MAE provides the appropriate variance-stabilizing transformation in the

large-λ case but is a biased estimator at low intensities. Because the high-spatial frequency (i.e. large scattering angle) signal has relatively poor signal-to-noise ratios, an objective function that provides unbiased estimation at low photon counts – specifically, the pixel-wise Poisson NLL – is significantly more dose-efficient (Figs. 5, **??**). Our Poisson NLL training objective results in equal resolution at a factor of 10 less dosage, compared to MAE. (cite figure).

Single-shot Fresnel CDI

In Fresnel geometry, probe curvature acts as structured illumination that provides phase diversity sufficient for reconstruction without overlap constraints. We obtain stable single-shot reconstructions with modest quality loss relative to overlapped scans (Fig. 1). This unifies ptychography and single-shot CDI in one self-supervised solver. In contrast with conventional solvers for ptychochraphy, where overlap constraints are an algorithmic requirement, our framework can perform reconstruction with or without overlap constraints.

Our approach thus offers stable reconstructions of low-overlap fraction datasets where conventional solvers struggle (Figure cite figure) and is much more modest in its need for training data than conventional data-driven (i.e. neural network) approaches.

Generalization and overfitting: in-distribution vs out-of-distribution

We distinguish resistance to *overfitting* (a train–test gap under a within-dataset holdout) from *out-of-distribution (OOD) generalization* (performance under domain shift). On the Siemens-star *spatial holdout* (train: upper half; test: lower half), the supervised CNN attains high training quality but degrades on held-out positions, indicating overfitting (Figs. **??**, 3). The physics-informed model maintains comparable resolution in and out of sample (visually and by quantitative metrics; see Figs. **??**, 4). We attribute this generalization to the use of a diffraction space-only objective, which both enforces a better match to the data and guarantees insensitivity of the loss function to nuisance parameters, such as real-space global phase shift.

Data efficiency and inductive bias

PtychoPINN reaches a given SSIM with roughly an order of magnitude fewer training patterns than the supervised baseline (Fig. 4), such that even a few hundred training images are sufficient to extract coherent features. We previously posited that this gain in generalization is primarily a result of the unsupervised training procedure's better fit to the problem's physical invariants [6].

Practical performance

At $\sim 10^4$ photons/frame, Poisson NLL preserves high-q detail relative to MAE (Fig. 5, Fig. **??**). Inference runs at $\sim$2000 patterns/s on an RTX 3090, and stable reconstructions are obtained with a few hundred (Fig. 4). Resolution continues to improve with training set size, up to at least 16384 training images. In the high-dose regime, overlap-free reconstruction yields images with modest information loss compared to overlapped scans TODO check that these are the right figure references (Fig. 1, Table 2).

Limitations and extensions

We assume a pre-estimated probe; joint probe retrieval within the self-supervised loop is a natural extension. Extending to Bragg CDI and reflection geometries would require only forward-model and

sampling modifications. Probe drift and position errors are neglected; explicit stochastic position models or differentiable refinement could help when jitter dominates.

Conclusion

Coupling a Poisson photon-counting likelihood with a differentiable forward model yields dose efficiency, robustness to sparse/irregular acquisition and OOD shifts, and support for single-shot Fresnel CDI. Together with high throughput, these properties enable real-time coherent imaging at modern light sources.

Appendix A: Key Configuration Parameters

These parameters control critical aspects of the reconstruction process and should be tuned based on experimental conditions and computational constraints.

Table 3: Model parameters and their default values

Parameter	Default	Description
`N`	64	Patch dimension (pixels)
`C_g`	4	Number of patterns per group
`K`	7	Nearest neighbors for grouping
`nsamples`	1	Random samplings per scan point
`pad_object`	True	Enable adaptive boundary learning
`probe.mask`	True	Apply circular probe support
`gaussian_smoothing_sigma`	0.0	Probe boundary smoothing
`intensity_scale.trainable`	False	Learnable intensity scaling
`n_filters_scale`	1	Network width multiplier
`amp_activation`	sigmoid	Amplitude decoder activation
`offset`	4	Scan step size (pixels)
d	3-5	Encoder depth (resolution-dependent)
C	132	Total decoder channels (before split)
C_{latent}	128	Latent channels at bottleneck

Appendix B: Model Architecture

References

[1] Aswin Varghese Babu, Tao Zhou, Swarnendu Kandel, Vinayak Banakha, Kang Li, Sha Li, Junjing Deng, Youssef S. G. Nashed, Ross Harder, Mathew J. Cherukara, and Charudatta Phatak. Deep learning at the edge enables real-time streaming ptychographic imaging. *Nature Communications*, 14:7059, 2023.

[2] Oliver Bunk, Martin Dierolf, Søren Kynde, Ian Johnson, Olof Marti, and Franz Pfeiffer. Influence of the overlap parameter on the convergence of the ptychographical iterative engine. *Ultramicroscopy*, 108(5):481–487, 2008.

[3] Mathew J. Cherukara, Tao Zhou, Youssef S. G. Nashed, Alexander Hexemer, Ross Harder, Oleg G. Shpyrko, Jun Wang, Sunil K. Sinha, and Subramanian K. R. S. Sankaranarayanan. Ai-enabled high-resolution scanning coherent diffraction imaging. *Applied Physics Letters*, 117(4):044103, 2020.

[4] Xue Dong, Xingchen Pan, Cheng Liu, and Jianqiang Zhu. Single shot multi-wavelength phase retrieval with coherent modulation imaging. *Optics Letters*, 43(8):1762–1765, 2018.

[5] Manuel Guizar-Sicairos and Pierre Thibault. Ptychography: A solution to the phase problem. *Physics Today*, 74(9):42–48, 2021.

Table 4: Mathematical symbols and conceptual descriptions

Symbol	Type / Structure	Description
x'	Set of C_g real images	Raw diffraction patterns for one sample
x	Set of C_g real images	Normalized diffraction patterns for one sample
$\vec{r}_k$	2D Position Vector	Absolute scan position for the k-th image within a sample
$\vec{r}^{\,i}_{\text{global}}$	2D Position Vector	Centroid of a solution region (group of scans)
$\vec{r}_{\text{relative},k}$	2D Offset Vector	Relative scan offset within a solution region
$e^{\alpha_{\log}}$	Scalar (trainable)	Internal log-intensity scale parameter
N_{photons}	Scalar	Target average total photons per diffraction pattern
$P(\vec{r})$	$N \times N$ complex array	Effective probe function
O_k	$N \times N$ complex array	k-th object patch decoded by the network G
O_{region}	$M \times M$ complex array	Merged object representation for a solution region
O'_{k}	$N \times N$ complex array	Object patch extracted from O_{region} for forward model
Ψ_k	$N \times N$ complex array	Predicted complex wavefield at the detector
$\hat{A}_k$	$N \times N$ real array	Predicted final diffraction amplitude for one patch
λ_{ijk}	Scalar	Poisson rate parameter for a single pixel

N: patch dimension, C_g: patches per group, M: merged region size

[6] Oliver Hoidn, Aashwin Ananda Mishra, and Apurva Mehta. Physics constrained unsupervised deep learning for rapid, high resolution scanning coherent diffraction reconstruction. *Scientific Reports*, 13:22789, 2023.

[7] Konstantin Kharitonov, Masoud Mehrjoo, Mabel Ruiz-Lopez, Barbara Keitel, Svea Kreis, Seung-gi Gang, Rui Pan, Alessandro Marras, Jonathan Correa, Cornelia B. Wunderer, and Elke Plönjes. Single-shot ptychography at a soft x-ray free-electron laser. *Scientific Reports*, 12(1):14430, 2022.

[8] Andrew M. Maiden, Matthew J. Humphry, Matthew C. Sarahan, Bernd Kraus, and John M. Rodenburg. An annealing algorithm to correct positioning errors in ptychography. *Ultramicroscopy*, 120:64–72, 2012.

[9] Andrew M. Maiden and John M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009.

[10] Stefano Marchesini, Hari Krishnan, Benedikt J. Daurer, David A. Shapiro, Talita Perciano, James A. Sethian, and Filipe R. N. C. Maia. Sharp: a distributed gpu-based ptychographic solver. *Journal of Applied Crystallography*, 49(4):1245–1252, 2016.

[11] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999.

[12] Joachim P. Seifert, Zhaoning Chen, Myung-Jae Yoon, Andrew Ulvestad, Mathew J. Cherukara, Youssef S. G. Nashed, Savannah M. Wild, and Ross Harder. Maximum-likelihood ptychography in the presence of poisson–gaussian noise. *Optics Letters*, 48(18):4897–4900, 2023.

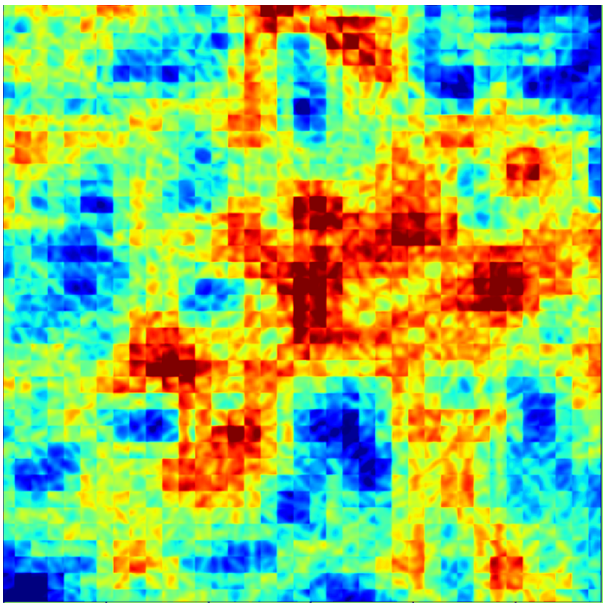
[13] Pavel Sidorenko and Oren Cohen. Single-shot ptychography. *Optica*, 3(1):9–14, 2016.

[14] SLAC National Accelerator Laboratory. LCLS-II-HE: Design and Performance. `https://lcls.slac.stanford.edu/lcls-ii-he/design-and-performance`, 2023. Accessed: 2025-08-14.

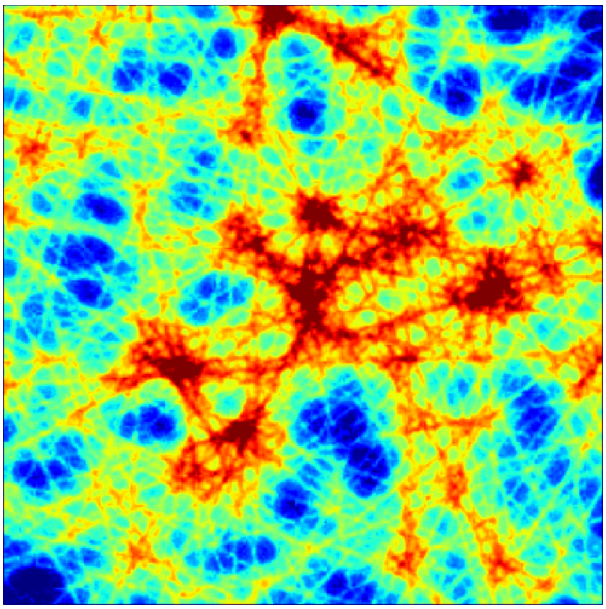
[15] Pierre Thibault and Manuel Guizar-Sicairos. Maximum-likelihood refinement for coherent diffractive imaging. *New Journal of Physics*, 14:063004, 2012.

[16] Fan Zhang, Irène Peterson, Joan Vila-Comamala, Ana Diaz, François Berenguer, Richard Bean, Bo Chen, Andreas Menzel, Ian K. Robinson, and John M. Rodenburg. Phase retrieval by coherent modulation imaging. *Nature Communications*, 7:13367, 2016.

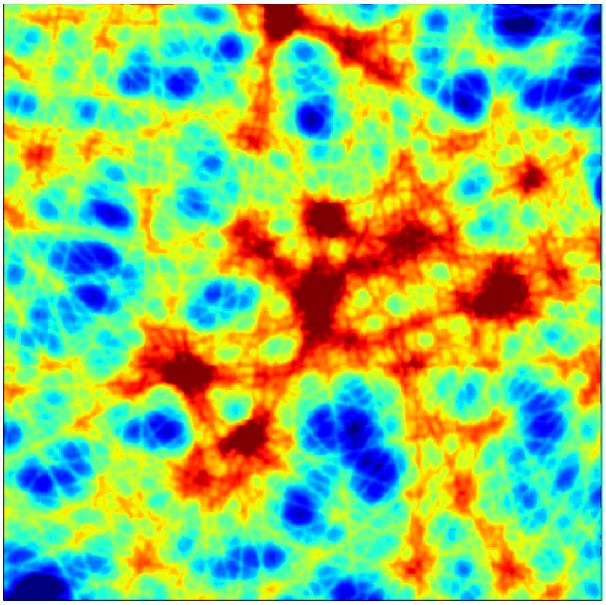
[17] Fei Zhang, Irène Peterson, Joan Vila-Comamala, Ana Diaz, François Berenguer, Richard Bean, Bo Chen, Andreas Menzel, Ian K. Robinson, and John M. Rodenburg. Translation position determination in ptychographic coherent diffraction imaging. *Optics Express*, 21(11):13592–13606, 2013.



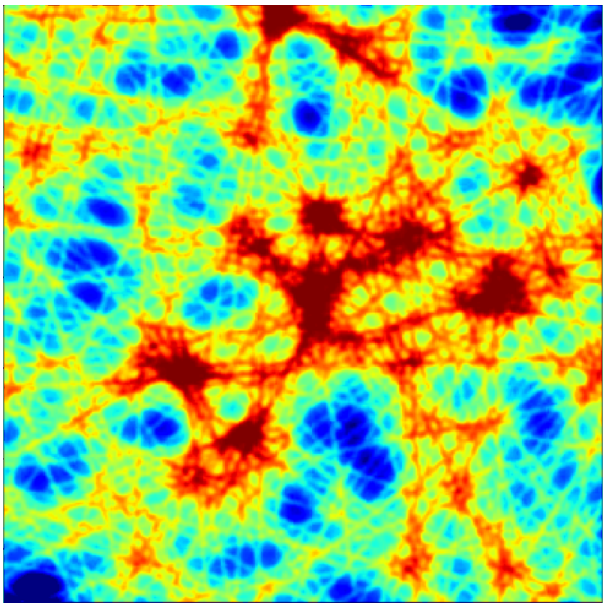
(a) Idealized — CDI



(b) Idealized — Ptycho



(c) Hybrid — CDI



(d) Hybrid — Ptycho

Figure 1: Reconstruction comparison. Rows: Idealized vs Hybrid probe. Columns: CDI vs Ptycho.

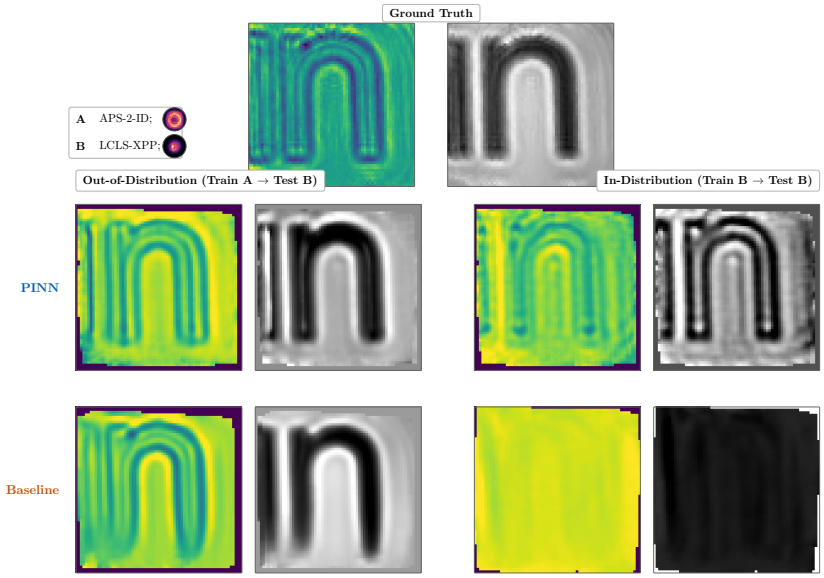
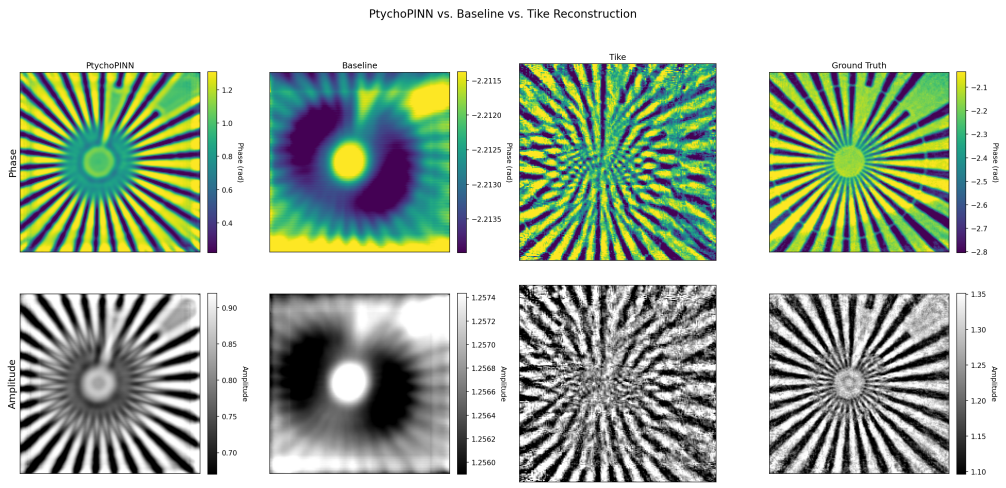
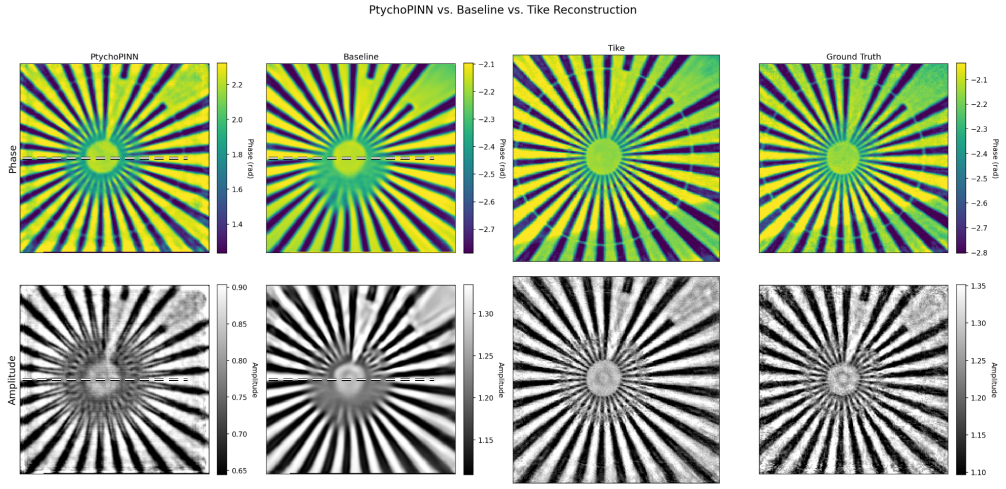


Figure 2: Comparison of methods across in- and out-of-distribution cases.



(a) 512 diffraction patterns of the Siemens star test pattern.



(b) 8192 diffraction patterns of the Siemens star test pattern.

Figure 3: Comparison of reconstruction quality with different numbers of diffraction patterns.

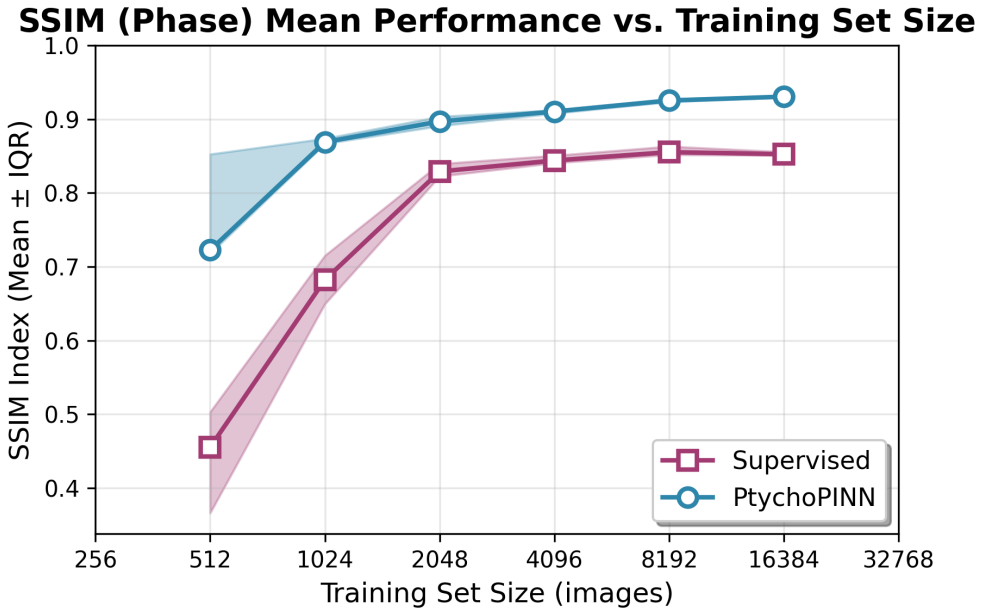


Figure 4: Structural similarity of PtychoPINN, conventional reconstruction (rPIE in Tike), and baseline as a function of training set size.

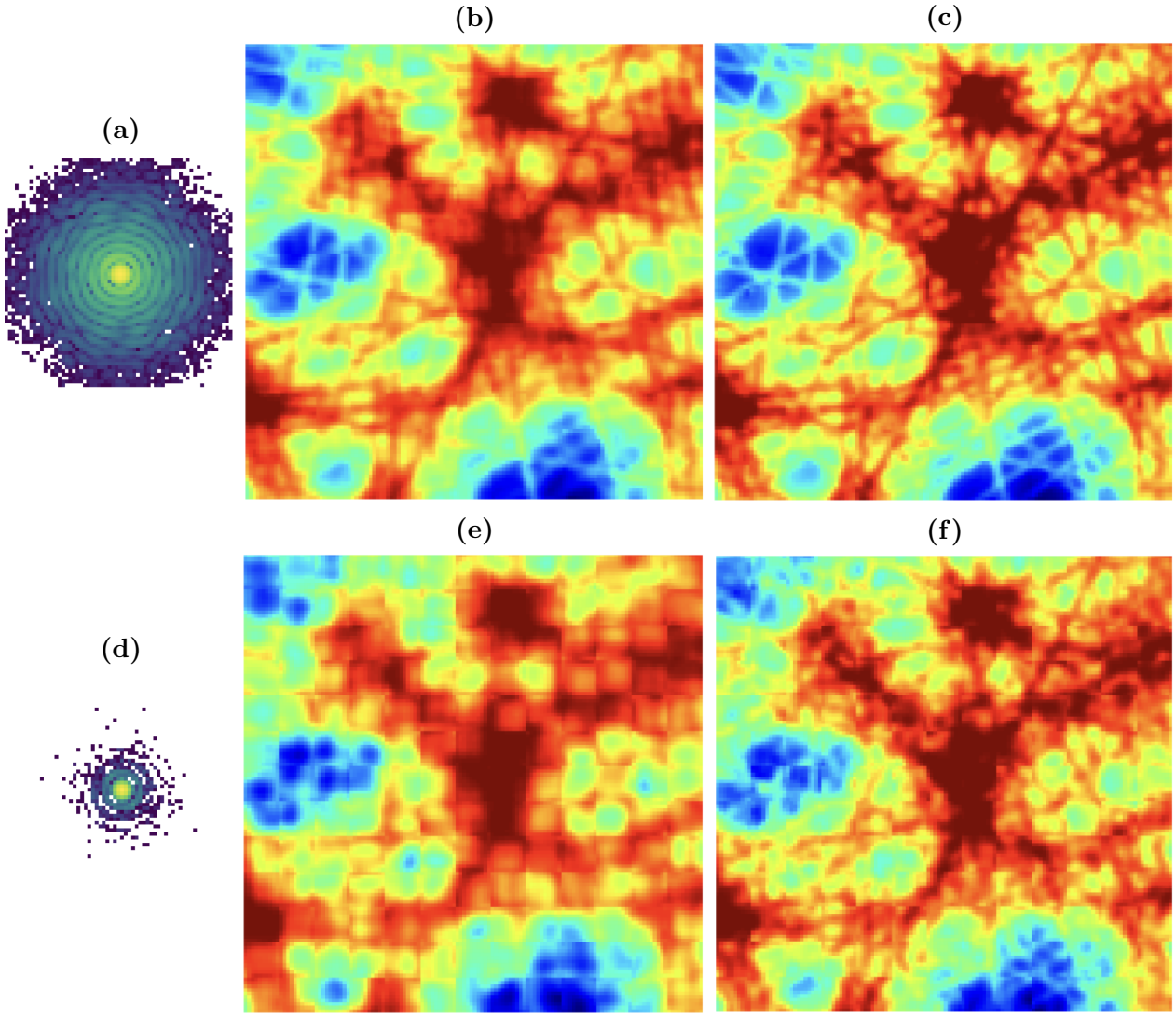


Figure 5: Resolution (FRC50) as a function of on-sample photon dose for two variants of the PtychoPINN framework trained with mean absolute error (MAE) and Poisson negative log likelihood (NLL) reconstruction penalties in the self-supervised loss function

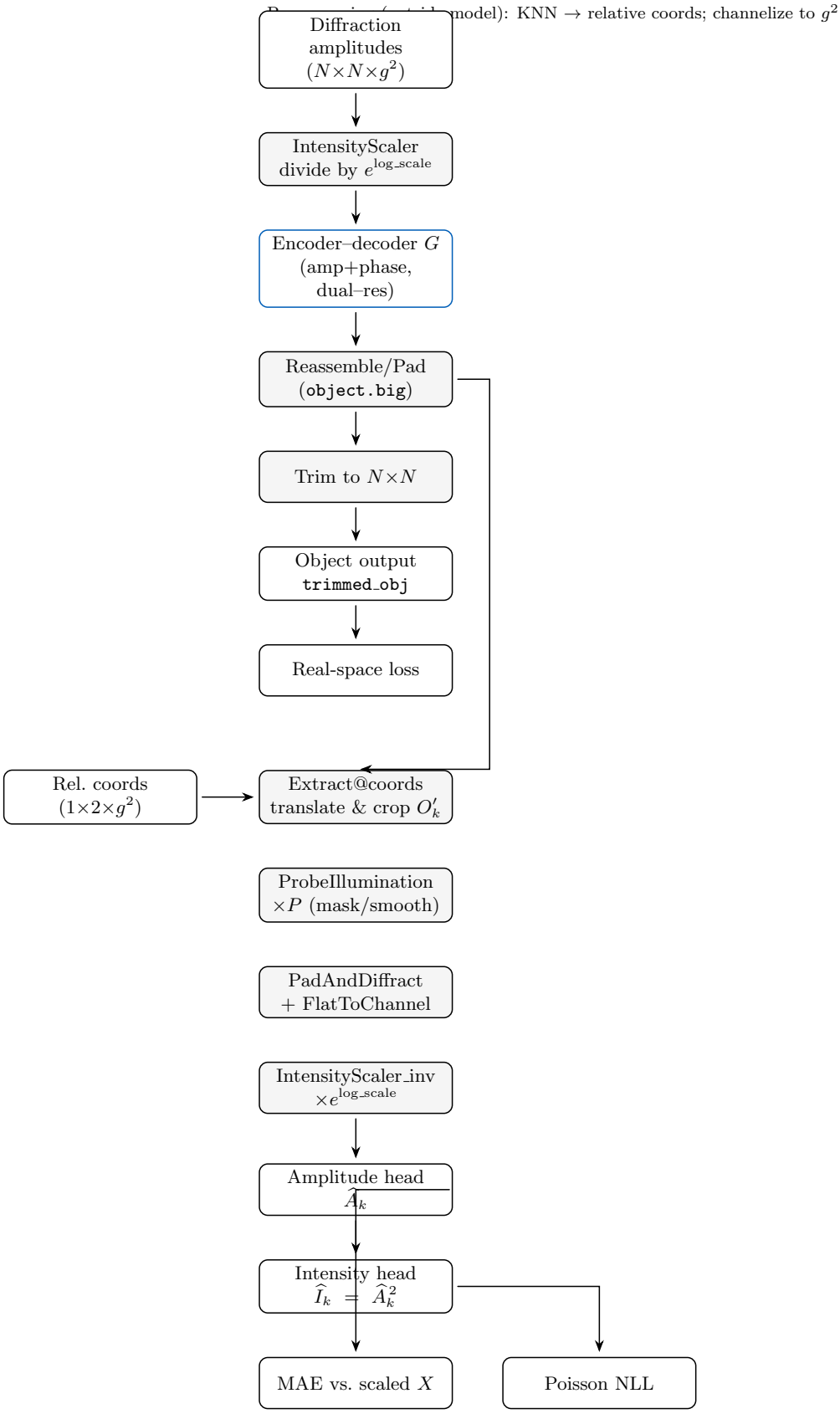


Figure 6: **PtychoPINN architecture (compact, vertical).** The model divides amplitudes by a learned scale, infers complex patches with a dual-resolution decoder, then forks: one path trims to an $N \times N$ object (real-space loss), the other feeds physics F_d (extract at coordinates, apply probe, FFT+channelize, rescale). Diffraction supervision uses an amplitude head (MAE) and an intensity head from squaring amplitudes (Poisson NLL). KNN grouping occurs outside the model.