

Position-Blind Ptychography: Viability of image reconstruction via data-driven variational inference

Simon Welker^{1,2,*}, Lorenz Kuger³, Tim Roith⁴, Berthy Feng^{5,6}, Martin Burger^{3,4}, Timo Gerkmann¹, and Henry Chapman^{2,7,8}

¹Department of Informatics, University of Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany

²Center for Free-Electron Laser Science CFEL, Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

³Department of Mathematics, Bundesstr. 55, University of Hamburg, 20146 Hamburg, Germany

⁴Helmholtz Imaging, Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

⁵Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA

⁶The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Cambridge, MA 02139, USA

⁷The Hamburg Center for Ultrafast Imaging, Luruper Chaussee 149, 22761 Hamburg, Germany

⁸Department of Physics, University of Hamburg, Notkestr. 9-11, 22607 Hamburg, Germany

*Corresponding author – simon.welker@uni-hamburg.de

October 1, 2025

Abstract

In this work, we present and investigate the novel blind inverse problem of *position-blind ptychography*, i.e., ptychographic phase retrieval without any knowledge of scan positions, which then must be recovered jointly with the image. The motivation for this problem comes from single-particle diffractive X-ray imaging, where particles in random orientations are illuminated and a set of diffraction patterns is collected. If one uses a highly focused X-ray beam, the measurements would also become sensitive to the beam positions relative to each particle and therefore ptychographic, but these positions are also unknown. We investigate the viability of image reconstruction in a simulated, simplified 2-D variant of this difficult problem, using variational inference with modern data-driven image priors in the form of score-based diffusion models. We find that, with the right illumination structure and a strong prior, one can achieve reliable and successful image reconstructions even under measurement noise, in all except the most difficult evaluated imaging scenario.

Keywords: single-particle imaging; ptychography; blind inverse problems; coherent diffractive imaging; diffusion models; data-driven regularization

AMS Subject Classification: 94A08, 68U10, 78A46, 68T07

1 Introduction

In the past two decades, there has been a strong push toward imaging ever smaller specimens such as nanoparticles, virus particles or even single proteins at X-ray free electron lasers (XFELs), and significant developments have been made on the experimental [3, 8, 14, 15, 31, 62] and algorithmic [4, 47, 53, 68, 70] fronts to realize these imaging modalities. To obtain measurable signals, intense femtosecond-duration X-ray pulses are used, which destroy the sample but only after the pulse has traversed the sample. single-particle diffractive imaging (SPI) therefore usually combines individual measurements from a stream of reproducible objects, each recorded in a random and unknown orientation. This approach promises benefits such as time-resolved, in-situ imaging of macromolecules such as proteins, including those that are not amenable to forming large crystals [54].

At the same time, the coherent diffractive imaging method of *ptychography* has shown remarkable success in microscopy with electron beams, optical light, and X-rays [57, 59]. Its advantage lies in the use of structured measurement redundancy by illuminating parts of an object in a scanning fashion and capturing multiple local diffraction patterns that can be merged into a single image with specialized algorithms [59]. This helps to avoid the need for prior information on the object under investigation and can achieve diffraction-limited resolution, even when various sources of experimental errors are present or the structure of the illuminating beam is unknown [59].

Recent innovations in X-ray optics can now achieve a beam focus with a spot size below three nanometers [5, 20], which is well within the size range of single biological macromolecules. Illuminating a single particle or nanocrystal with such a small beam would result in a ptychographic measurement, where only a part of the object is strongly illuminated for each diffraction pattern as the illuminating beam decays off the main beam spot. However, a significant advantage of the ptychographic measurement, the knowledge of the scan positions, is fully lost when imaging in the destructive regime of XFEL pulses. This leads to a blind inverse problem, where both the scan positions and the image of the object have to be recovered jointly. Well-established techniques [49, 74] and newer developments [21, 46] exist for the case of correcting local position errors, which is in a sense a semi-blind problem. For instance, Zhang et al. [74] find that their method, based on serial cross-correlation of objects in a modification of the ePIE algorithm [50], ceases to work well when the initial position error exceeds 20 pixels. The full position-blind problem has, to the best of our knowledge, not been investigated in prior works and presents a reconstruction task of high difficulty. This scenario is the subject of the present work.

Here, we perform a computational study on the viability of a simplified type of position-blind ptychographic imaging for small specimens such as single macromolecules. For simplicity and computational efficiency, we assume that the specimen is a thin sheet 2-D object of finite extent within the plane, and that the illuminating beam (also called the *probe*) is concentrated in a region roughly of the size of the specimen or smaller. Due to the increased difficulty caused by the loss of position information, we incorporate prior knowledge about the imaged object via diffusion models in order to facilitate the reconstruction process.

Generative diffusion models have strongly impacted the field of machine learning in the past few years, with widespread applications from unconditional and conditional image and audio generation [35, 45, 61, 65] to data-driven approaches for solving inverse problems [13, 40, 51, 75]. In this work, we use score-based generative models, a subclass of diffusion models formulated via a continuous-time diffusion process, as data-driven priors, which have recently begun to be employed in real imaging problems, for example in a Plug-and-Play (PnP)-based method for (non-blind) ptychography [18]. We compare these data-driven priors against the use of either no prior information or a simple model-based total variation (TV) prior. The comparison is carried out using two algorithmic

frameworks for variational inference [24, 51] suited for model-based or diffusion-based priors, as well as simpler optimization-based procedures.

2 Background on position-blind ptychography

In this section, we explain how a position-blind ptychographic reconstruction problem may arise in a potential SPI setup.

2.1 Phase Retrieval

In phase retrieval imaging problems, one seeks to recover the complex-valued image $x \in \mathcal{X} := \mathbb{C}^d$ from (noisy) intensity values y . This typically poses reconstruction problems of the form

$$y = |\mathcal{F}x|^2 + \varepsilon,$$

where $\mathcal{F}$ is a linear operator describing the light propagation in the measurement process and ε is measurement noise. For far-field data in applications like SPI, X-ray crystallography and ptychography, $\mathcal{F}$ is typically the Fourier transform, whereas for near-field measurements $\mathcal{F}$ is the Fresnel integral operator with an experiment-dependent defocus value. The image reconstruction task is ill-posed since the phase problem is non-linear and subject to several sources of measurement errors in practice. Robust and efficient reconstruction algorithms are subject to ongoing research within the mathematical and machine learning literature [12, 19, 22, 23, 26, 63].

2.2 Ptychography

Using modern X-ray sources at synchrotron facilities, coherent diffraction imaging (CDI) aims to solve the phase retrieval problem by reconstructing an image from the diffraction patterns y generated from a highly coherent X-ray beam illuminating the sample x . *Ptychography* [57, 59] is a special case of CDI allowing the reconstruction of high-resolution images from a collection of (far-field or near-field) diffraction patterns. It uses measurement redundancy by illuminating a sample multiple times at different positions, generating a set of diffraction patterns y_k with $k = 1, \dots, K$. Each pattern y_k is collected by strongly illuminating only a part of the object under investigation (real-space ptychography) or part of the diffraction space (Fourier ptychography) [59]. In real-space ptychography, the scan positions r_k are placed so that the illuminated parts overlap, and the resulting redundancy in the data mitigates the ill-posedness of the reconstruction problem.

Ptychography can also be interpreted as a specific type of *coded illumination* or *coded aperture* method, making links to other imaging disciplines [19], or as a variant of the *short-time Fourier transform phase retrieval* problem, with links to signal and audio processing [29, 30, 63].

Ptychography has been remarkably successful at retrieving high-resolution images of the object under investigation [59]. Moreover, it allows joint reconstruction of the object and the illuminating probe [59], an advantageous property since the latter is often only known approximately in phase retrieval tasks. It was also shown to be robust to various sources of measurement errors [59]. An important measurement error arises from imperfectly known scan positions. Such errors usually occur due to experimental factors such as imperfect scanning stages and thermal noise. Several prior works [21, 46, 49, 74] have proposed correction methods that have shown to work well as long as the initial estimates of the positions lie in a local vicinity of the true positions.

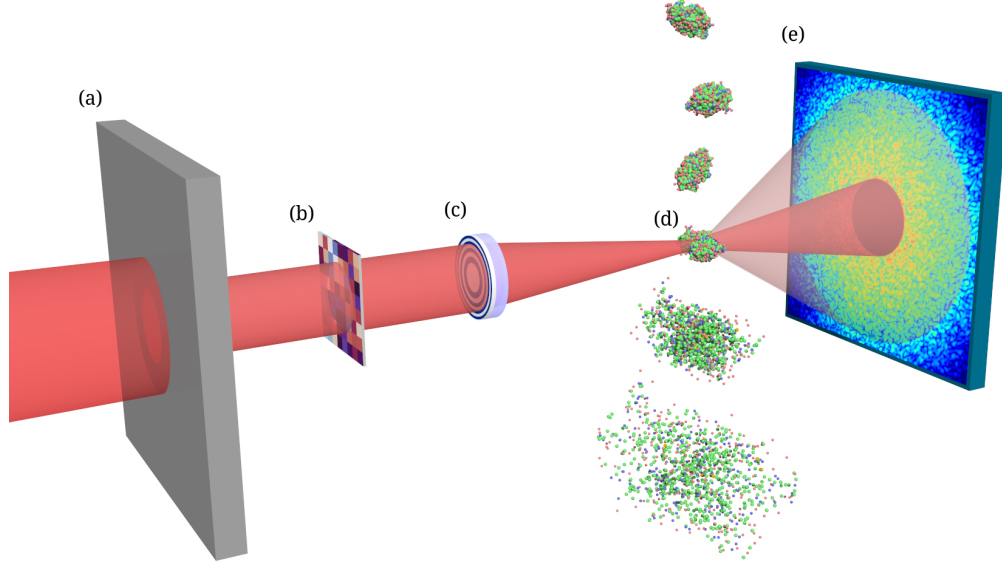


Figure 1: The ptychographic single-particle diffractive imaging (SPI) setup. Components from left to right: (a) a beam aperture, (b) an optional random phase mask, (c) a focusing optic, here illustrated as a Fresnel zone plate, (d) the interaction region at the beam focus, and (e) a detector. The photon beam is illustrated in transparent red. The particles move through the interaction region in an uncontrolled manner, and each particle generates a single diffraction pattern before disintegrating (*diffraction before destruction* [10, 54]). This makes the particle position and orientation relative to the beam unknown in every measurement.

2.3 Single-Particle Imaging and Ptychography

In this work, we investigate the extreme case of completely unknown scan positions. The underlying motivation comes from the methodology of single-particle diffractive imaging (SPI) [54]. Here, the imaged objects are micro- to nanometer-sized and do not reside on a well-controllable scanning stage, but are, e.g., in free flight within a *molecular beam* injected into the experimental chamber and exposed with a very short-duration X-ray pulse; see Fig. 1. Each object can only be exposed once, but it is assumed that all particles are identical in structure such that the reconstructed image x represents a single object. The SPI setup potentially enables time-resolved investigations of biochemical processes, but is experimentally limited by low probabilities that the randomly injected particles are hit by an X-ray pulse. The measurements further suffer from extremely low signal-to-noise ratio (SNR), since only a handful of photons are captured on the detector for each successful hit. We refer the interested reader to a review discussing these ideas and problems in more detail [11]. A way to increase the SNR is to use a highly focused beam with a focus size below 10 nm, which has recently become possible through advances in X-ray optics [20, 71, 72]. When these beams have a focus size smaller than the object under investigation, the measurements become ptychographic: each diffraction pattern encodes only a part of the object. While this increases the photon flux through the sample and hence the measurement SNR by allowing more photons to diffract, it also introduces severe measurement uncertainty since the “scan” positions r_k are unknown and must be recovered jointly with the object x . Since the measurement operator $\mathcal{A}_k$ depends on the scan positions r_k , the reconstruction problem is an instance of a *blind* inverse problem.

3 Reconstruction and Sampling Methods

In the following, we describe the theory behind the investigated methods for image reconstruction and sampling with and without score-based priors.

3.1 Score-Based Priors for Imaging Inverse Problems

In imaging inverse problems, we aim to recover a d -dimensional image $x \in \mathcal{X}$ from a measurement $y \in \mathcal{Y} := \mathbb{R}^m$, where the two are related via

$$y = \mathcal{A}(x) + \varepsilon. \quad (3.1)$$

For now, $\mathcal{A}$ is a general forward model describing the measurement acquisition and ε is measurement noise. Typically the inverse of $\mathcal{A}$ is discontinuous, making the solution of this problem ill-posed. The reconstruction of x from y therefore requires some form of regularization [6]. In order to quantify uncertainty in the reconstruction, it is further necessary not only to reconstruct a single solution to (3.1), but we have to make statements about the statistical properties of x . The distribution of possible solutions can be described using Bayesian inference [66]. In the Bayesian formulation, we recover the posterior distribution of the image x , the density of which is given by

$$p(x | y) = \frac{p(y | x)p(x)}{Z}. \quad (3.2)$$

Here $p(x)$ is the density function of the prior distribution that is imposed on the space of possible images, which acts as the regularization of the problem [66]. The term $p(y | x)$ is the measurement likelihood implied by (3.1). A standard point estimate for this posterior which can serve as an exemplary solution to the inverse problem is the maximum a posteriori (MAP) estimate $x^{\text{MAP}} = \arg \max_x p(x | y)$, but the posterior also allows to compute more advanced statistical properties like moments or confidence sets. In order to carry out these computations, we need to draw samples from the posterior. The density function $p(x | y)$, however, is generally intractable since its normalizing constant, the model evidence $Z = p(y) = \int p(y | x)p(x) dx$, is a high-dimensional integral and as such unknown. Algorithms therefore typically fall back to approximating the posterior by a simpler distribution with tractable density that is easy to sample from, or sample from the posterior using methods that do not require knowing Z , e.g., Markov chain Monte Carlo (MCMC) algorithms [38].

Inverse problems are termed *blind* if the forward model depends on an unknown parameter $\mathbf{r} \in \mathcal{R} \subset \mathbb{R}^K$. Instead of (3.1), the measurement in a blind problem is given by

$$y = \mathcal{A}(\mathbf{r}, x) + \varepsilon. \quad (3.3)$$

In the Bayesian setting, the model parameters can be treated in a similar way as the unknown: The likelihood term is now $p(y | x, \mathbf{r})$ and the posterior of both image and model parameter is given by

$$p(x, \mathbf{r} | y) = \frac{p(y | x, \mathbf{r})p(x, \mathbf{r})}{Z}. \quad (3.4)$$

Depending on the application, it can often be assumed that $x, \mathbf{r}$ are independent under the prior, so that $p(x, \mathbf{r}) = p(x)p(\mathbf{r})$ with respective prior distributions $p(x)$ and $p(\mathbf{r})$. Any inference task is now carried out with respect to the joint posterior $p(x, \mathbf{r} | y)$.

3.1.1 Learning Priors with Diffusion Models

Score-based diffusion models are a popular method for generative modeling due to their ability to learn complex distributions of training image datasets. In recent research, their adaptability to conditional/posterior distributions in inverse problems has been showcased with promising results [16]. The main concept behind the unconditional model is to transform an unknown distribution $p_0(x)$ of images to a normal distribution via a diffusion process. To draw new samples, one samples from the normal distribution and simulates an associated backward diffusion. The forward diffusion process is given by the stochastic differential equation (SDE)

$$dx_t = f(x_t, t) dt + g(t) dW_t, \quad t \in [0, T], \quad (3.5)$$

with $x_0 \sim p_0$, where W_t denotes standard Brownian motion. We will denote p_t for the distribution of x_t in (3.5) at time t . The drift $f(x, t)$, the diffusion coefficient $g(t)$ and the final time T are chosen such that p_t approximately equals an analytically tractable distribution π at the final time $t = T$, i.e. $p_T \approx \pi$.

In order to sample from p_0 , score-based models simulate the time-reversed SDE

$$dx_t = (f(x_t, t) - g(t)^2 \nabla_x \log p_t(x_t)) dt + g(t) d\bar{W}_t, \quad (3.6)$$

where $\bar{W}_t$ is a time-reversed Brownian motion. If the reverse SDE is initialized at time $t = T$ as $x_T \sim p_T$, then under mild conditions on the coefficients f, g , the process at time $t = 0$ obeys $x_0 \sim p_0$ [2]. We will compare results for two standard SDE choices. The first one is the variance preserving stochastic differential equation (VP-SDE) with $f(x_t, t) = -\frac{\beta(t)}{2}x$ and $g(t) = \sqrt{\beta(t)}$, with $\beta(t)$ linear and monotonically increasing so that p_t converges to the standard normal $\pi = \mathcal{N}(0, I)$. The second one is the variance-exploding variant (VE-SDE) with $f \equiv 0$ and $g(t)$ a monotonically increasing schedule, with $g(T)$ large enough such that $p_T \approx \mathcal{N}(0, g(T)^2 I)$. Samples from p_0 can thus be generated by sampling from the tractable distribution $x_T \sim \pi \approx p_T$ and simulating (3.6).

The practical difficulty of this approach lies in accurately approximating the score function $\nabla_x \log p_t(x_t)$ in (3.6). This is possible if we already have access to sufficient training data drawn from p_0 , since we can then train a score model $s_\theta(x_t, t) \approx \nabla_x \log p_t(x_t)$ which approximates the true score [36, 64]. The score model is parametrized by a network θ and trained using *denoising score matching*, see e.g. [69]:

$$\arg \min_{\theta} \int_0^T \lambda(t) \mathbb{E}_{(x_0, x_t) \sim p(x_0, x_t)} \left[\|s_\theta(x_t, t) - \nabla_x \log p(x_t | x_0)\|^2 \right] dt, \quad (3.7)$$

where $\lambda(t)$ is a weighting factor balancing the approximation quality at different time steps. Solving (3.7) requires estimating the expectation with respect to the joint distribution $p(x_0, x_t)$. If the drift coefficient f in (3.5) is affine linear, the forward in time conditional $p(x_t | x_0)$ is a normal distribution with a known closed-form mean and variance. Sample pairs (x_0, x_t) from $p(x_0, x_t) = p(x_0)p(x_t | x_0)$ can hence be easily generated by drawing x_0 from the training data $p(x_0)$ and generating x_t efficiently by calculating the closed-form mean and variance expressions and adding sampled Gaussian noise.

Once the score model is trained, new samples from p_0 can be generated by replacing $\nabla_x \log p_t(x_t)$ by $s_\theta(x, t)$ in (3.6) and then simulating the reverse SDE by discretizing it using, e.g., a standard Euler–Maruyama scheme [65] and an initialization $x_T \sim \pi$.

3.1.2 Sampling from a Bayesian Posterior

For unconditional sampling from p_0 , we can employ the training data to estimate the expectation in (3.7). Suppose we aim to sample from a posterior distribution instead, where the initial distribution

$p_0(x)$ in the SDE would be of the form $p(x|y)$ (3.2). In that case, sampling is no longer possible since there are no representative samples from the posterior to begin with. To circumvent this problem, most methods train a score model $s_\theta(x, t)$ on the image prior distribution $p(x)$. Crucially, since the prior distribution $p(x)$ is not a function of y , it allows the score model to be pre-trained offline before making any measurement.

Focusing for the moment on the non-blind setting (3.2), conditional sampling thus requires adjusting for a diffused likelihood term: If we want to simulate (3.6), where the target at time $t = 0$ is the posterior $p(x|y)$, the score is given by

$$\nabla_x \log p_t(x_t | y) = \nabla_x \log p_t(y | x_t) + \nabla_x \log p_t(x_t) \approx \nabla_x \log p_t(y | x_t) + s_\theta(x_t, t)$$

While the prior score $\nabla_x \log p_t$ can be efficiently approximated by the score model s_θ , the term $\nabla_x \log p_t(y | x_t)$ is generally intractable.

Some works have developed schemes to approximate $\nabla_x \log p_t(y | x_t)$, e.g., by building approximations based on the chain rule $p_t(y | x_t) = \mathbb{E}_{x_0} [p(y | x_0)p(x_0 | x_t)]$. One instance of these methods is diffusion posterior sampling (DPS) [13], and we refer to [16] for an overview of several other such algorithms.

Other works avoid approximating the intractable likelihood score by not simulating the reverse SDE (3.6) for the posterior at all. Instead, one can try to take a variational inference (VI) approach, which has been done for the RED-Diff method [51] and the works on principled score-based priors in [24, 25]. Consider the posterior $p(x|y) \propto p(y|x)p(x)$. One can use a learned approximation $\hat{p}_\theta(x)$ for the prior term $p(x)$, where $\hat{p}_\theta(x)$ is implicitly defined through a learned score-based prior $s_\theta(x, t)$. This defines an approximated, but still intractable posterior $\hat{p}_\theta(x|y) \propto p(y|x)\hat{p}_\theta(x)$, which can be modeled via VI by a tractable parametric distribution q_ϕ . To that end, one solves the optimization problem

$$\phi^* := \arg \min_{\phi} \{ \text{KL}(q_\phi \| \hat{p}_\theta(\cdot | y)) \}. \quad (3.8)$$

Depending on the chosen parametric class, the resulting approximate posterior q_ϕ can allow for direct sampling and density evaluations. The complexity of recovering ϕ^* is controlled by its dimensionality and the chosen parametric family of distributions. For instance, ϕ could consist of the mean and covariance parameters of a simple Gaussian or Gaussian mixture [7], or, allowing for more expressivity, ϕ could be the parameters of a neural network encoding a normalizing flow model [67]. Rewriting (3.8), those methods seek to recover

$$\begin{aligned} \phi^* &= \arg \min_{\phi} \{ \mathbb{E}_{x \sim q_\phi} [-\log \hat{p}_\theta(x, y) + \log q_\phi(x)] \} \\ &= \arg \min_{\phi} \{ \mathbb{E}_{x \sim q_\phi} [-\log p(y|x) - \log \hat{p}_\theta(x)] - \mathcal{H}(q_\phi) \}, \end{aligned} \quad (3.9)$$

where $\mathcal{H}$ denotes the entropy functional $\mathcal{H}(q) := -\mathbb{E}_{x \sim q} [\log q(x)]$.

3.1.3 RED-Diff

A simple variational distribution is a Gaussian with mean $\mu \in \mathbb{R}^d$ and isotropic covariance with a scalar $\sigma > 0$, i.e., $q_\phi = \mathcal{N}(\mu, \sigma^2 I)$, with $\phi = (\mu, \sigma)$. In [51], the authors prove that the VI objective (3.9) can then be written as

$$\arg \min_{\phi} \left\{ -\mathbb{E}_{x \sim q_\phi} [\log p(y|x)] + \int_0^T \omega(t) \mathbb{E}_{x \sim q_t(\cdot | y)} \left[\|\nabla_x \log q_t(x|y) - \nabla_x \log p_t(x)\|_2^2 \right] dt \right\}, \quad (3.10)$$

where $\omega(t)$ is a suitable weight, and $q_t = \mathcal{N}(\alpha(t)\mu, (\alpha(t)^2\sigma^2 + \sigma(t)^2)I)$ is the distribution that arises from simulating the forward SDE (3.5) with initial condition q_ϕ . The functions $\sigma(t), \alpha(t)$ are defined as $\sigma(t) = 1 - \exp(-\int_0^t \beta(s) ds)$ and $\alpha(t) = \sqrt{1 - \sigma(t)^2}$ (for VP-SDE) or $\sigma(t) = g(t)$ and $\alpha(t) \equiv 1$ (for VE-SDE), respectively. Since the diffused variational density q_t is available in closed form and $\nabla_x \log p_t(x)$ can be replaced by the trained score model $s_\theta(x, t)$, the terms under the integral can be evaluated efficiently. By modifying the weight $\omega(t)$ in the integral, the authors arrive at a loss function that bears similarities to the regularization by denoising (RED) approach to MAP estimation in inverse problems [60], despite it not being equal to the original VI loss anymore. Additionally, the authors assume for their numerical experiments that $\sigma \approx 0$, essentially fitting a point mass to the posterior and deviating from the Bayesian motivation of the VI approach. Despite these modifications, the method is reasonably fast and the reconstructed images of promising quality.

3.1.4 Variational inference with principled score-based priors

In two other works [24, 25], the VI loss is optimized without reweighting in time. Evaluating the objective (or its gradients) in (3.9) requires evaluating the prior log-density $\log \hat{p}_\theta(x)$ for unseen data x . The score model $s_\theta(x, t)$ is typically unstable around $t \approx 0$, but as previously shown by Song et al. [65], log-density values can instead be obtained by solving the initial value problem for the forward probability flow ordinary differential equation (ODE)

$$\frac{dx_t}{dt} = f(x_t, t) - \frac{1}{2}g(t)^2 \nabla_x \log p_t(x_t, t), \quad x_0 = x. \quad (3.11)$$

This generates the same dynamics of the distribution p_t as (3.5), since p_t solves both the continuity equation for (3.11) as well as the Fokker–Planck equation of (3.5). Replacing the score in (3.11) by the score model s_θ allows solving this ODE, giving an approximation $\log \hat{p}_\theta^{\text{ODE}} \approx \log \hat{p}_\theta$. The authors of [25] thus proposed to use $\log \hat{p}_\theta^{\text{ODE}}$ to evaluate the objective (3.9). As the numerical results of [25] showed, the resulting VI approach has a high computational cost, but generates very accurate approximations of the true posterior.

A follow-up paper [24] resolves the costly evaluation of $\log \hat{p}_\theta^{\text{ODE}}$ by instead employing the evidence lower bound (ELBO) surrogate $b_\theta^{\text{SDE}} \leq \log \hat{p}_\theta^{\text{ODE}}$; the exact definition of b_θ^{SDE} can be found in [24, p. 5]. For the readers convenience, it is repeated in Section B. After this modification, one instead solves

$$\tilde{\phi} := \arg \min_{\phi} \{ \mathbb{E}_{x \sim q_\phi} [-\log p(y|x) - b_\theta^{\text{SDE}}(x)] - \mathcal{H}(q_\phi) \}. \quad (3.12)$$

where the ELBO term b_θ^{SDE} can be estimated efficiently using Monte Carlo integration. Heuristically, the ELBO term induces a surrogate score-based prior (SSP) with density $\hat{p}_\theta^{\text{sur}} \propto \exp(b_\theta^{\text{SDE}})$. Empirically, it was shown that this surrogate prior is very effective, in the sense that for small dimensional examples $\tilde{\phi} \approx \phi^*$, i.e., the fitted distribution shows very good agreement with the variational approximation to the posterior. The reduced computational time of (3.12) allows to lift the problem dimension to realistic imaging sizes.

We now describe how we can extend this method to the case of blind inverse problems (3.4), using a standard variational Bayes approach for the joint posterior of image x and latent parameter $\mathbf{r}$. While we are specifically interested in solving the position-blind ptychography problem (4.2), the method we develop can be used for general blind imaging inverse problems.

3.2 Sampling from the Position-Blind Ptychography Posterior

Consider now a semi-blind or blind setting, where we need to reconstruct the posterior $p(x, \mathbf{r} | y)$ in the joint variable $(x, \mathbf{r})$, where x is an image and $\mathbf{r}$ the latent parameter. In our position-blind

ptychographic setup, $\mathbf{r}$ will be a vector containing the measurement positions.

3.2.1 Generalized VI Approach for Blind Problems

Using (3.4), we generalize the VI approach to the joint posterior $p(x, \mathbf{r} | y)$ in the blind case. Under the non-restrictive assumption of independent priors $\log p(x, \mathbf{r}) = \log \hat{p}_\theta(x) + \log p(\mathbf{r})$, the derivation of the VI objective (3.9) easily generalizes to

$$\phi^* = \arg \min_{\phi} \left\{ \mathbb{E}_{(x, \mathbf{r}) \sim q_\phi} [-\log p(y | x, \mathbf{r}) - \log \hat{p}_\theta(x) - \log p(\mathbf{r})] - \mathcal{H}(q_\phi) \right\} = \arg \min_{\phi} \mathcal{L}(\phi). \quad (3.13)$$

Note that the variational distribution q_ϕ is now the joint distribution of image and parameters. Depending on the forward model and dimensionality, the joint optimization problem can exhibit strong non-convexity and be computationally demanding. The problem can be provided with more structure by a standard mean-field assumption on the variational Bayesian posterior

$$q_\phi(x, \mathbf{r}) = q_\chi(x) q_\rho(\mathbf{r}), \quad (3.14)$$

where $\phi = (\chi, \rho)$, with χ encoding the image posterior and ρ the latent parameter posterior. Note that this comes at the cost of less accurately representing the correlation of uncertainties in image and parameters in q_ϕ , but is necessary to make the optimization feasible at scale. We note that a detailed analysis of this approximation is an important open question for future research, but is outside of the scope of this work. By inspecting the optimality conditions of $\min_{\phi} L(\phi) = \min_{\chi, \rho} L(\chi, \rho)$ with respect to χ and ρ separately, under the mean-field assumption, the posterior marginals obey the optimality conditions [7]

$$q_\chi \propto \exp \left(\mathbb{E}_{\mathbf{r} \sim q_\rho} [\log p(x | \mathbf{r}, y)] \right), \quad q_\rho \propto \exp \left(\mathbb{E}_{x \sim q_\chi} [\log p(\mathbf{r} | x, y)] \right).$$

This motivates the following alternating updates of the image and parameter distributions

$$\chi_{k+1} = \arg \min_{\chi} \mathcal{L}(\chi, \rho_k) = \arg \min_{\chi} \left\{ \mathbb{E}_{x \sim q_\chi} \mathbb{E}_{\mathbf{r} \sim q_{\rho_k}} [-\log p(y | x, \mathbf{r}) - \log \hat{p}_\theta(x)] - \mathcal{H}(q_\chi) \right\}, \quad (3.15a)$$

$$\rho_{k+1} = \arg \min_{\rho} \mathcal{L}(\chi_{k+1}, \rho) = \arg \min_{\rho} \left\{ \mathbb{E}_{x \sim q_{\chi_{k+1}}} \mathbb{E}_{\mathbf{r} \sim q_\rho} [-\log p(y | x, \mathbf{r}) - \log p(\mathbf{r})] - \mathcal{H}(q_\rho) \right\}. \quad (3.15b)$$

The technique of learning a score model using training data applies only to the image prior $\hat{p}_\theta(x)$, while the parameter prior $p(\mathbf{r})$ depends on knowledge in the specific application. For instance, in our ptychography setup, $p(\mathbf{r})$ could be chosen as a unimodal density (if we have an initial estimate of the measurement position, leading to a “semi-blind” problem) or constant (reflecting a uniform prior on the two-dimensional cell that the ptychographic measurements are restricted to, with no prior information imposed at all).

3.2.2 Blind RED-Diff

The RED-Diff method has been modified for tackling blind imaging inverse problems on the example of MRI with unknown off-resonance field map [1]. The authors there employed the same mean-field assumption (3.14) in order to separate optimization steps for the reconstructed image and parameters. As before, the objective for the image posterior can be rewritten using [51, Prop. 1] to arrive at a

loss similar to (3.10), but with the likelihood conditioned on the current parameter estimate, i.e., (3.15a) becomes

$$\chi_{k+1} = \arg \min_{\chi} \left\{ -\mathbb{E}_{x \sim q_{\chi}} \mathbb{E}_{\mathbf{r} \sim q_{\rho_k}} [\log p(y | x, \mathbf{r})] + \int_0^T \omega(t) \mathbb{E}_{x \sim q_t(\cdot | y)} [\|\nabla_x \log q_t(x | y) - \nabla_x \log p_t(x)\|_2^2] dt \right\}$$

Like [51], the work [1] then introduces time reweighting by replacing $\omega(t)$ by a different $\tilde{\omega}(t)$. The variational posterior q_{χ} is replaced by a point mass and the entropy term removed from the objective in order to arrive at an implementable scheme. We mention again that this essentially replaces the Bayesian character of the reconstructed quantity with an optimization scheme that rather resembles classical MAP computation. In our notation, the resulting image optimization step is

$$\begin{aligned} x_{k+1} &= \arg \min_x \left\{ -\log p(y | x, \mathbf{r}_k) + \int_0^T \tilde{\omega}(t) \mathbb{E}_{x_t \sim q_t(\cdot | y)} [\|s_{\theta}(x_t; t) - z\|_2^2] dt \right\} \\ &=: \arg \min_x \mathcal{L}^{\text{REDdiff}}(x, \mathbf{r}_k), \end{aligned} \quad (3.16)$$

where now $q_t = \mathcal{N}(\alpha(t)x, \sigma^2(t)I)$ and we abbreviated $z = -\frac{x_t - \alpha(t)x}{\sigma^2(t)}$. The time reweighting $\tilde{\omega}(t)$ is chosen such that $\tilde{\omega}(0) = 0$, since the gradients of the objective in (3.16) then allow the following form that does not require backpropagation through the score network

$$\nabla_x \mathcal{L}^{\text{REDdiff}}(x, \mathbf{r}_k) = -\nabla_x \log p(y | x, \mathbf{r}_k) - \int_0^T \tilde{\omega}(t) \mathbb{E}_{x_t \sim q_t(\cdot | y)} [s_{\theta}(x_t; t) - z] dt, \quad (3.17)$$

with a suitable weight $\tilde{\omega}(t)$; see [51, Prop. 2] for details. For the parameter update step (3.15b), the authors of [1] showed that assuming $p(\mathbf{r})$ and $q_{\rho}(\mathbf{r})$ are both Laplace distributions allows to obtain closed form representations of the relevant terms $p(\mathbf{r})$ and $\mathcal{H}(q_{\rho})$. Upon replacing q_{ρ} with a point mass on a single $\mathbf{r}$, the entropy is, however, dropped again in [1] and the parameter update becomes

$$\mathbf{r}_{k+1} = \arg \min_{\mathbf{r}} \{-\log p(y | x_{k+1}, \mathbf{r}) - \log p(\mathbf{r})\} =: \arg \min_{\mathbf{r}} \mathcal{L}^{\text{REDdiff}}(x_{k+1}, \mathbf{r}). \quad (3.18)$$

We line out the resulting algorithm with our notation in Algorithm 2.

3.2.3 Blind surrogate score-based prior (SSP) method

The scheme of [24, 25] can be generalized to the blind setting in a similar way. As before, the image step (3.15a) could be solved directly by deriving the density via the probability flow ODE. However, this becomes computationally infeasible in high dimensions, so the ELBO term $b_{\theta}^{\text{SDE}} \leq \log \hat{p}_{\theta}^{\text{ODE}}$ can be used instead, effectively approximating the learned image prior by the same surrogate prior as in (3.12). The image optimization step (3.15a) hence reads

$$\chi_{k+1} = \arg \min_{\chi} \mathcal{L}(\chi, \rho_k) = \arg \min_{\chi} \left\{ \mathbb{E}_{x \sim q_{\chi}} \mathbb{E}_{\mathbf{r} \sim q_{\rho_k}} [-\log p(y | x, \mathbf{r})(x) - b_{\theta}^{\text{SDE}}(x)] - \mathcal{H}(q_{\chi}) \right\} \quad (3.19)$$

For medium- to large-scale imaging problems, q_{χ} can be a Gaussian distribution with diagonal covariance – more complex models were computationally prohibitive in our setup; see also [24] for scaling experiments. We solve (3.15) using inner loops of stochastic gradient descent, where the partial derivatives $\nabla_{\chi} \mathcal{L}$ and $\nabla_{\rho} \mathcal{L}$ are approximated using Monte Carlo estimators of the expectations

Algorithm 1 Blind Variational Bayes reconstruction with surrogate prior

```

1: Initialize parameter distribution  $\rho^{(0)} = (\mu_{\mathbf{r}}^{(0)}, \Sigma_{\mathbf{r}}^{(0)})$ , image distribution  $\chi^{(0)} = (\mu_x^{(0)}, \Sigma_x^{(0)})$ ,  $l = 0$ ,
   data  $y \in \mathcal{Y}$ , forward model  $\mathcal{A} : \mathcal{R} \times \mathcal{X} \rightarrow \mathcal{Y}$ ,  $l = 0$ , maximum number of iterations  $l_{\max}$ , step
   size sequences  $(\tau^{(l,i)}), (\eta^{(l,i)})$ 
2: while  $l < l_{\max}$  and stopping criterion on  $\rho^{(l)} = (\mu_{\mathbf{r}}^{(l)}, \Sigma_{\mathbf{r}}^{(l)})$ ,  $\chi^{(l)} = (\mu_x^{(l)}, \Sigma_x^{(l)})$  is not satisfied do
3:    $\chi^{(l,0)} = \chi^{(l)}$ 
4:   for  $i \leftarrow 0, \dots, N_{\text{img}} - 1$  do ▷ Solve for image (3.19)
5:      $\chi^{(l,i+1)} = \chi^{(l,i)} - \tau^{(l,i)} \nabla_{\chi} \mathcal{L}(\chi^{(l,i)}, \rho^{(l)})$ 
6:   end for
7:    $\chi^{(l+1)} = \chi^{(l, N_{\text{img}})}$ 
8:    $\rho^{(l,0)} = \rho^{(l)}$ 
9:   for  $i \leftarrow 0, \dots, N_{\text{par}} - 1$  do ▷ Solve for parameters (3.15b)
10:     $\rho^{(l,i+1)} = \rho^{(l,i)} - \eta^{(l,i)} \nabla_{\rho} \mathcal{L}(\chi^{(l+1)}, \rho^{(l,i)})$ 
11:   end for
12:    $\rho^{(l+1)} = \rho^{(l, N_{\text{par}})}$ 
13:    $l \leftarrow l + 1$ 
14: end while
15: return image distribution  $q_{\chi^{(l)}}$ , parameter distribution  $q_{\rho^{(l)}}$ 

```

and the typical reparametrization trick [41] whenever q_{χ}, q_{ρ} are Gaussian. This sampling-based approach admits the use of a batch size $B \geq 1$ for each gradient evaluation, which can reduce the variance of estimated gradients. The surrogate prior term $b_{\theta}^{\text{SDE}}(x)$ can be implemented using a single forward-pass through s_{θ} ; see [24, sec. 4.2]. This leads to the variational Bayes approach summarized in Algorithm 1. For a clearer presentation, we formulate the updates as single gradient descent steps. However, in the practical implementation, lines 5 and 10 in Algorithm 1 are replaced by the Adam optimization scheme [42].

We also mention a connection to other previous work on blind inverse problems [27], which proposed an expectation maximization (EM) scheme to alternatingly optimize the image and a latent parameter. The EM algorithm naturally arises as a special case of the variational Bayes approach derived here, if we slightly abandon the Bayesian perspective and regard the latent parameter as having a true value $\mathbf{r}^*$. Equivalently, we can assume that q_{ρ} is a point mass in (3.15) and remove the entropy regularization. The EM iteration then reads as the alternating scheme:

$$\chi_{k+1} = \arg \min_{\chi} \left\{ \mathbb{E}_{x \sim q_{\chi}} \left[-\log p(y | x, \mathbf{r}_k) - b_{\theta}^{\text{SDE}}(x) \right] - \mathcal{H}(q_{\chi}) \right\}, \quad (3.20a)$$

$$\mathbf{r}_{k+1} = \arg \min_{\mathbf{r}} \left\{ \mathbb{E}_{x \sim q_{\chi_{k+1}}} \left[-\log p(y | x, \mathbf{r}) - \log p(\mathbf{r}) \right] \right\}. \quad (3.20b)$$

To solve the E-step (3.20a), we use a first-order optimization scheme and the learned surrogate prior. A closely related work [43] proposed a similar scheme with the same M-step for blind parameters, but used a likelihood-approximating sampling method like DPS to estimate the image posterior for a given parameter instance in the E-step.

3.2.4 Variational Minimization as Baseline

While Bayesian techniques like VI provide a powerful framework to recover not only a single image, but a whole distribution of images, they are also more computationally demanding. Classical, variational optimization algorithms are generally more economic and form a central methodology

in the ptychographic imaging literature [52]. As already noted, (blind) RED-Diff can be located between the two classes: It is motivated from a VI lens, but solves the imaging problem by fitting a single image—or, in distribution terms, a point mass, reflected by the fact that $\mathcal{L}^{\text{REDdiff}}$ contains no entropy terms. The iterative update of blind RED-Diff is ultimately an alternating first order descent method on the variational minimization (or MAP estimation) problem

$$\min_{x, \mathbf{r}} \{-\log p(y | x, \mathbf{r}) - \log p(x) - \log p(\mathbf{r})\}, \quad (3.21)$$

where the image prior is of the specific form $p(x)$ seen in Eq. (3.16). By replacing the image prior $p(x)$, we obtain other standard variational regularization formulations. In experiments, we will therefore compare the proposed SSP approach to RED-Diff, but also to simpler forms of (3.21) with cheaper, model-based regularization (as opposed to a data-driven score model), or no prior information at all. For the latter, completely omitting $p(x)$ leads to a maximum likelihood estimate of x , which we compute as before using first-order optimization methods. In the former case, one may use a typical hand-crafted image prior $p(x)$ that promotes features like sparsity in the image gradients or in a dictionary or frame like a wavelet basis.

4 Experiments

In the following, we describe the computational experiments we perform to arrive at our results and conclusions. For clarity, we preface this section with the following simplifying assumptions we make for computational viability in our simulated setting:

1. The illuminating wavefront (the probe function p) is fully known and constant in all measurements.
2. The measurement noise level is known and constant in all measurements.
3. The imaged object x is treated as infinitely thin sheet such that it can be treated as a 2-dimensional complex transmission function.
4. The object x is always oriented the same in every measurement, i.e., there are no unknown rotational parameters we need to recover.
5. The probe’s center is positioned somewhere on the object in every measurement; hence each measurement is significantly influenced by some portion of the object (no measurements of empty space).

Relaxations of these assumptions should be possible within our framework through extensions of our formalism, but are left for future work here due to the difficulty that this simplified setup already exhibits, and due to the likely need for parameter tuning in more complex settings.

4.1 Problem Setup

4.1.1 Forward Model

We treat here a simplified 2D far-field real-space ptychography problem. Let $x \in \mathbb{C}^{H \times W}$ be an image of the complex object transmission function which we want to recover,

$$x[h, w] = \exp(ikt \cdot (n(h/H, w/W) - 1)), \quad n(h', w') = 1 - \delta(h', w') + i\beta(h', w'), \quad (4.1)$$

where $n(h', w')$ is a complex-valued function describing the complex refractive index at normalized coordinates h', w' , with δ describing the local phase shift and β describing the local absorption induced by the object material. For simplicity we assume a unit wavenumber $k = 1$ and a unit thickness $t = 1$. We further assume that the image is square, i.e., $H = W$. Let y_k be the k -th measurement (diffraction pattern) at scan position r_k , $k = 1, \dots, K$. We model x and y_k to be related by the following differentiable forward operator:

$$y_k = |\mathcal{A}(r_k, x)|^2 + \varepsilon_k, \quad (4.2)$$

$$\mathcal{A}(r_k, x) = \mathcal{F}(p \odot \text{CROP}(S(r_k, x))), \quad (4.3)$$

$$S(r_k, x)[h, w] = \text{DFT}^{-1} \left\{ \text{DFT}\{\text{PAD}(x)\} \odot \exp \left(i \left(\Delta h_k \frac{2\pi h}{H} + \Delta w_k \frac{2\pi w}{W} \right) \right) \right\}, \quad (4.4)$$

where ε_k is the measurement noise of the k -th measurement, $\mathcal{F}$ models the wavefront propagation to the detector and DFT denotes the discrete Fourier transform, the probe array $p \in \mathbb{C}^{H_p \times W_p}$ models the complex-valued wavefield of the probe in the plane of the object and is applied through element-wise multiplication $\odot$, and $S(r_k, \cdot)$ is a shift operator depending on a two-dimensional shift $r_k = (\Delta h_k, \Delta w_k)$ which we refer to as the *scan position* in the following. The CROP operator crops its input to the same array size as the probe p before element-wise multiplication. The PAD operator pads its input with one full probe array size of empty space entries; see Section 4.1.2. In all experiments, we use an object array of size $x \in \mathbb{C}^{H \times W} := \mathbb{C}^{256 \times 256}$ and a probe array of size $p \in \mathbb{C}^{H_p \times W_p} := \mathbb{C}^{512 \times 512}$.

Since we assume a *far-field ptychography* problem, we set $\mathcal{F}$ to be the (discrete) Fourier transform. It may be interesting to investigate the case of Fresnel propagation (near-field ptychography), where position recovery could be simpler due to some real-space information being encoded in the diffraction patterns. However, we focus on the far-field case here, which should be more challenging due to a complete lack of position information in the measurements.

4.1.2 Scan Positions

We model the ground-truth positions to represent the center of each probe relative to the object coordinates, so if the probe is centered on some pixel of the object, it holds that $0 \leq \Delta h_k \leq H, 0 \leq \Delta w_k \leq W$. Under this assumption, we sample the ground-truth positions from a uniform distribution in the horizontal and vertical directions,

$$r_k \sim \mathcal{U}(0, H) \times \mathcal{U}(0, W). \quad (4.5)$$

The discrete Fourier transform (DFT) we use in S implicitly treats the input signal as periodic. We use this fact to make it impossible for estimated positions to escape into empty space, as $r_k + [aH, bW]$ is equivalent to r_k for all $a, b \in \mathbb{Z}$ under our DFT-based definition of S . Since the target image itself is however a single particle and thus aperiodic, we use the padding operator PAD, which pads x with one full probe array size of free space $(1 + 0i)$ entries. $S(r_k, \cdot)$ then effectively models shifting of an aperiodic object inside an indefinitely repeating reconstruction box. Furthermore, to avoid ringing artifacts in the image, we always round r_k to the nearest whole integer when evaluating the forward operator but keep the gradients as if we were using fractional shift positions.

4.1.3 Probe Functions

We simulate all probe functions $p \in \mathbb{C}^{512 \times 512}$ by taking the DFT of the aperture array $a \in \mathbb{C}^{512 \times 512}$ to propagate the wavefront to the focus. The aperture array a contains a centered circular aperture

with fractional diameter $d_{\text{ap}} \in (0, 1/2]$. In the aperture plane, we assume a wavefront of constant magnitude and a phase profile determined by a random Zernike polynomial [73] of order 4, with piston and tilt terms set to 0. We sample the random Zernike polynomial only once, shown in the leftmost column of Fig. 2. We optionally apply an additional block-wise random phase mask in the aperture plane through pixel-wise complex multiplication:

$$a_{\text{masked}}[h, w] = a \odot e^{i\Phi_{\text{mask}}}, \quad \Phi_{\text{mask}}[h, w] = M \left[\left\lfloor \frac{h}{b} \right\rfloor, \left\lfloor \frac{w}{b} \right\rfloor \right], \quad M \sim \mathcal{U}(0, 2\pi)^{\lceil h/b \rceil \times \lceil w/b \rceil} \quad (4.6)$$

where b determines the block size of the random mask in pixels. This construction is inspired by other works on randomized illumination [32, 44], which show more reliable reconstructions when using structured probes with high-frequency content. We will show in Section 5.2.1 that the additional structure from these phase masks is also helpful for position recovery in our position-blind setting. By default, we set $d_{\text{ap}} = \frac{1}{2}$, $b = 4$ in our experiments.

Since the probe is band-limited from the finite extent of the aperture and/or lens, it is in principle space-unlimited. Thus, to avoid the unrealistic case of the probe energy abruptly falling off to zero on some part of the object, we choose the probe array to be twice as large as the object array and set all ground-truth probe positions to be centered on some point on the object. Note that even though the probe array is larger in pixels than the object array, the measurements are ptychographic since the probe energy is concentrated in a smaller region, typically roughly the size of the object or smaller; see Fig. 2.

4.1.4 Noise Model

As the noise model in most of our experiments, we assume that the observation noise ε_k is independent and identically distributed Gaussian noise with mean zero and known variance σ_ε^2 for all $k = 1, \dots, K$. In later experiments, we also simulate more experimentally accurate noise by scaling the signal power of the probe function to match an assumed number of photons diffracted from an ideal non-absorbing object n_{phot} , and drawing y_k from a simulated Poisson distribution with mean $|\mathcal{A}(r_k, x)|^2$.

4.1.5 Simulated Measurements

To generate a simulated set of measurements $\{y_k\}_{k=1}^K$, we take x to be some simulated test object transmission function from our test dataset (see Section 4.3) and r_k to be 100 randomly sampled positions according to Eq. (4.5), compute each y_k according to Eq. (4.2), and then apply the chosen type of measurement noise. We use $K = 100$ measurement positions in all of our experiments.

Denoting the full set of scan positions by $\mathbf{r} = (r_1, \dots, r_K)$, the noise model leads to the following measurement likelihood under the Gaussian measurement noise model:

$$p(y | x, \mathbf{r}) \propto \exp \left(-\frac{1}{2\sigma_\varepsilon^2} \sum_k \left\| y_k - |\mathcal{A}(r_k, x)|^2 \right\|_2^2 \right), \quad (4.7)$$

whereas for Poisson measurement noise, we model the measurement likelihood using a Gaussian approximation to the Poisson distribution, leading to

$$p(y | x, \mathbf{r})_{\text{Poisson}} \propto \exp \left(-\sum_k \left\| \frac{1}{\sqrt{2y_k}} \odot (y_k - |\mathcal{A}(r_k, x)|^2) \right\|_2^2 \right), \quad (4.8)$$

where each pixel in y_k is treated as both the mean and the variance, and the square root is taken element-wise. An alternative would be to use $|\mathcal{A}(r_k, x)|^2$ as the variance of the approximating

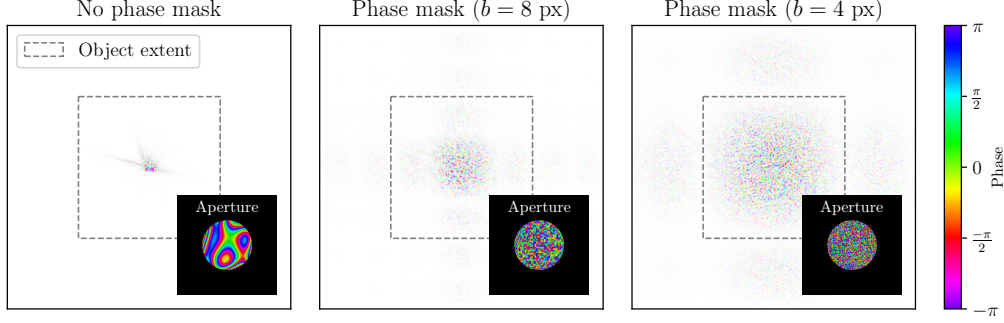


Figure 2: Comparison of several probe functions p used in this work. All three are based on the same random Zernike polynomial, with an optional random phase mask of block size b applied as indicated. The dashed white square shows the extent of the imaged object x for comparison. The diameter of the aperture is half a probe array size here, i.e., 256 pixels.

Gaussian instead of y_k , but in preliminary experiments we found this option to lead to less stable reconstructions and hence always use Eq. (4.8). In practice, within the expression $\frac{1}{2y_k}$ we clamp y_k to have a minimum value of 1 to avoid both division by zero and an overly high weighting of detector pixels with no measured photons.

4.2 Algorithmic Configuration

4.2.1 Variational Inference

For our variational inference methods and all choices of prior distribution, we approximate the posterior by the following variational distributions:

$$q_\chi(x|y) := \mathcal{N}_{\mathbb{C}}(x; \mu_\chi, \Sigma_\chi I), \mu_\chi \in \mathbb{C}^d, \quad \Sigma_\chi = \text{diag}(\sigma_\chi^2), \sigma_\chi \in \mathbb{R}_+^d, \quad (4.9)$$

$$q_\rho(r|y) := \mathcal{N}(r; \mu_\rho, \Sigma_\rho I), \mu_\rho \in \mathbb{R}^{K \times 2}, \Sigma_\rho = \text{diag}(\sigma_\rho^2), \sigma_\rho \in \mathbb{R}_+^K, \quad (4.10)$$

$$q_{\chi, \rho}(x, r|y) := q_\chi(x|y)q_\rho(r|y). \quad (4.11)$$

We assume complex isotropy of the image uncertainty in that the real and imaginary parts of each image pixel $x[h, w] \in \mathbb{C}$ share a single variance $\sigma_{\chi, hw}^2 \in \mathbb{R}$, and spatial isotropy of the shift uncertainty in that the horizontal and vertical components of each position $r_k \in \mathbb{R}^2$ share a single variance $\sigma_{\rho, k}^2 \in \mathbb{R}$. To enforce stability and positivity of σ_χ and σ_ρ , we reparametrize and minimize for $\log \sigma_\chi$, $\log \sigma_\rho$. We set the batch size of the Monte Carlo gradient estimators to $B = 4$; see Section 3.2.3.

For both the image and position parameters, we use the Adam optimizer [42], which has previously been established as a working method for (non-blind) ptychographic image reconstruction [33, 37, 39]. We set its momentum parameters to $\beta_1 = 0.9, \beta_2 = 0.999$, the default settings in the PyTorch package [55]. We run $N = 10,000$ outer steps unless otherwise noted, each with 1 inner image optimization step and 10 inner position optimization steps. For the position parameters we set the step sizes $\lambda_{\mu_\rho} = 10.0$ and $\lambda_{\sigma_\rho} = 0.01$, where we use a high learning rate of the means (λ_{μ_ρ}) due to the difficult position-dependent loss landscape; see Fig. 5. We keep λ_{σ_ρ} small to avoid a collapse of the position uncertainty. For the image parameters we set $\lambda_{\mu_\chi} = \lambda_{\sigma_\chi}$ to quickly arrive at an initial image estimate, and employ a falling cosine schedule from 0.1 to 0.001 between steps 4,000 and 6,000 to allow the image to stabilize and to improve the effectiveness of the high-variance prior term b_θ^{SDE} during the later optimization steps.

4.2.2 TV baseline details

Following the definition in [28], we employ the following variant of isotropic TV for complex valued images $x \in \mathbb{C}^{H \times W}$,

$$\text{TV}(x) := \sum_{h=1}^H \sum_{w=1}^W \sqrt{|x[h, w] - x[h, w+1]|^2 + |x[h, w] - x[h+1, w]|^2}, \quad (4.12)$$

where we set $x[H+1, w] = x[H, w]$ for all $w = 1, \dots, W$ and, respectively, $x[h, W+1] = x[h, W]$ for all $h = 1, \dots, H$. In order to circumvent the non-differentiability of the TV functional as defined above, we consider a smoothed version

$$\text{TV}_\alpha(x) := \sum_{h=1}^H \sum_{w=1}^W h_\alpha \left(\sqrt{|x[h, w] - x[h, w+1]|^2 + |x[h, w] - x[h+1, w]|^2} \right). \quad (4.13)$$

A typical choice for $h_\alpha : [0, \infty) \rightarrow [0, \infty)$ is the Huber regularizer (see, e.g., [9, Example 4.7]),

$$h_\alpha(t) := \begin{cases} \frac{t^2}{2\alpha} & \text{if } t \leq \alpha, \\ t - \frac{\alpha}{2} & \text{else,} \end{cases}$$

where $\alpha > 0$ is chosen small; for our experiments we set $\alpha = 10^{-5}$.

4.2.3 RED-Diff

The algorithm we use for blind RED-Diff in this setup is given in Algorithm 2. Following [51], in the notation of Eq. (3.17) we choose the inverse SNR weighting $\tilde{\omega}(t) = \lambda_{\text{RD}} \frac{\sigma(t)}{\alpha(t)}$. In [51, Prop. 2] the weight $\lambda_{\text{RD}} > 0$ may also depend on the observation noise σ_ε . In our experiments, we choose fixed step sizes $\tau^{(l,i)} = \tau = 0.1, \eta^{(l,i)} = \eta = 10$ in Algorithm 2.

4.2.4 Position log-barrier

In some of the scenarios we consider, such as phase-only objects, we observe that the loss landscape for the position recovery has an unfavorable structure, which leads to positions moving off the object rather towards reasonable estimates; see Section 5.2.4. To remedy this problem, we add a 2-dimensional log-barrier loss as a hand-crafted prior on the positions, with an empirically chosen weight λ_{pos} . Namely, we consider the following barrier function, $B(s) := -\log(1-s) - \log(1+s) + \iota_{\mathbb{R}^2 \setminus (-1,1)^2}(s)$, where for a set A , ι_A denotes the indicator function from convex analysis; see [58]. Our prior is then defined by an affine transformation T , which first shifts and scales each position r , i.e., $p(r) \propto \exp(-\lambda_{\text{pos}} B(T(r)))$. We define T such that the object extent within the domain is scaled to a smaller domain $(-a, a)^2 \subsetneq (-1, 1)^2$. This is done via

$$T(r) := \frac{r - m}{r_{\max} - r_{\min}},$$

where we set $r_{\max} = H + l, r_{\min} = H - l$ and $m = (r_{\max} + r_{\min})/2$ with $l = 20$, so l grants 20 pixels of leniency for the positions to lie slightly outside of the object extent. To avoid undefined gradients from the log-barrier loss, we clip the positions to lie inside the rectangle spanned by these edges before feeding them to the loss expression.

4.3 Dataset

Since there is not enough realistic data of X-ray wavefront modulation from single proteins or other single nanoscale particles readily available to train a deep generative model, we instead generate artificial complex-valued images from a large public image dataset. We choose the INRIA Aerial Image Labeling Dataset (AILD) [48] as a basis, since its images have detailed natural structures and high-frequency content but are nonetheless more predictable and of a simpler distribution than a generic large photograph dataset such as ImageNet [17]. This allows us to test the reconstruction methods in complex imaging scenarios. We then generate two types of complex-valued images based on random 256×256 crops of these RGB images, mapped to grayscale, with the following two procedures:

- (1) 80% of the time, to simulate the observation that the absorption image often has less structure than the phase image, we generate the image amplitudes from a Perona–Malik edge-preserving smoothing [56] of the grayscale input image. We use a uniformly random number of Perona–Malik iterations in $[30, 100]$ and a random κ parameter in $[0.03, 0.075]$. We scale the amplitudes by factors sampled from a log-normal distribution with $\mu = 0, \sigma = 0.25$. For the phase of the object, we use the grayscale image without any smoothing, mapping all grayscale values $v \in [0, 1]$ to a random phase of $4\pi\mathfrak{S}v + \mathfrak{s}$ where $\mathfrak{S} \sim \text{Beta}(3, 10), \mathfrak{s} \sim \mathcal{N}(0, \frac{\pi}{2})$ are sampled once for each image.
- (2) 20% of the time, to simulate non-absorbing objects, we generate phase-only images by setting all amplitudes to 1 and using the grayscale input image as the phase image. We map to a random phase range of $\mathfrak{S}v + \mathfrak{s}$ with $\mathfrak{S} \sim \mathcal{U}(\frac{\pi}{4}, 3\pi), \mathfrak{s} \sim \mathcal{U}(-\pi, \pi)$ where $v \in [0, 1]$ is again the input grayscale value.

In total, we generate 30,000 training images for training our score model from the training subset of AILD [48]. For evaluating the reconstruction methods, we generate 10 test images from the test subset of AILD, but here we only follow procedure (1) to generate test objects that exhibit both absorption and phase shifts. For our experiments where we consider phase-only objects (Section 5.2.4) we then generate phase-only variants of these same objects by setting their complex magnitude to 1 everywhere and keeping the phase.

4.4 Score model training

With the training dataset described in the previous section, we train a score model s_θ using the NCSN++ architecture [65] with a channel configuration of $[128, 128, 256, 256, 256, 256]$. As the diffusion process, we use the VP-SDE [35, 65] with $\beta_{\min} = 0.01, \beta_{\max} = 20$ and $t_\varepsilon = 0.001$. For training, we use the Adam optimizer [42] with a learning rate of 10^{-4} and an exponential moving average (EMA) weight smoothing with decay 0.999 [65], and train for 140 epochs.

4.5 Evaluation

To evaluate the quality of the reconstructions, we make use of three image metrics (aPSNR, aSSIM, cRMS), as well as one metric for evaluating position recovery which we call *posCorrect*. The metrics aPSNR and aSSIM are evaluated using only the object magnitudes

$$\text{aPSNR}(\hat{x}, x) := \text{PSNR}(\min(|\hat{x}|, 1), |x|), \quad \text{aSSIM}(\hat{x}, x) := \text{SSIM}(\min(|\hat{x}|, 1), |x|), \quad (4.14)$$

where x is the ground-truth image and $\hat{x}$ is a reconstructed image, and we clip the estimated magnitudes to $[0, 1]$ for evaluation to avoid large errors from isolated wrong pixels. We report

Non-blind scenario (baseline)			
Method / Metric	aPSNR $\uparrow$	cRMS $\downarrow$	aSSIM $\uparrow$
Optimization-based			
No prior	12.24 $\pm$ 0.99	0.55 $\pm$ 0.26	0.10 $\pm$ 0.05
H-TV prior ($\lambda = 0.1$)	23.96 $\pm$ 1.19	0.04 $\pm$ 0.01	0.51 $\pm$ 0.06
Variational Inference			
No prior	12.75 $\pm$ 0.87	0.48 $\pm$ 0.22	0.11 $\pm$ 0.05
H-TV prior ($\lambda = 5$)	23.69 $\pm$ 1.16	<u>0.04 $\pm$ 0.01</u>	0.49 $\pm$ 0.06
H-TV prior ($\lambda = 10$)	25.04 $\pm$ 2.14	0.04 $\pm$ 0.02	0.67 $\pm$ 0.04
H-TV prior ($\lambda = 20$)	24.49 $\pm$ 2.59	0.05 $\pm$ 0.03	0.78 $\pm$ 0.05
SSP	30.36 $\pm$ 2.39	0.02 $\pm$ 0.01	0.90 $\pm$ 0.04
RED-Diff ($\lambda_{RD} = 20$)	<u>29.41 $\pm$ 2.40</u>	0.02 $\pm$ 0.01	<u>0.87 $\pm$ 0.03</u>

Table 1: Metrics of reconstructions in the **non-blind** baseline setting, comparing different methods. Best in bold, second best underlined. H-TV represents the Huber-TV prior with weight λ , and SSP refers to the surrogate score-based prior method.

aPSNR values in dB, while aSSIM takes values in $[0, 1]$. Since aPSNR and aSSIM ignore errors in the phase structure, we include the complex-valued metric cRMS, introduced as E_o in [50], which is a normalized root-mean-square error metric with a complex-valued empirical scaling factor γ that corrects for a scale ambiguity and the global phase ambiguity inherent in phase retrieval problems:

$$\text{cRMS}(\hat{x}, x) = \frac{\sum_{h,w} |x[h, w] - \gamma \cdot \hat{x}[h, w]|}{\sum_{h,w} |x[h, w]|^2}, \quad \gamma = \frac{\sum_{h,w} x[h, w] \hat{x}[h, w]^*}{\sum_{h,w} |\hat{x}[h, w]|^2} \quad (4.15)$$

where $(\cdot)^*$ indicates the complex conjugate. Our position recovery metric *posCorrect* is defined as the number of estimated positions that are within a box of 3x3 pixels around their respective ground-truth positions. We argue that this small region of allowed error should be enough for further position refinement with well-established subpixel-capable methods, e.g., [74]. In our case where $K = 100$, posCorrect can directly be read as a percentage.

Our scenario of position-blind ptychography exhibits a global shift ambiguity in the reconstructed positions and image, since there is no absolute reference point for the positions. Therefore, before evaluating any metric, we run a simple greedy image registration procedure of the estimate $\hat{x}$ relative to the ground-truth x : we evaluate every shift between ± 20 pixels in both directions and choose the global shift with minimum error of the image magnitudes. We then translate the estimated positions and circularly shift the image according to this optimal shift, and use these shifted estimates for evaluation.

For all experiments, unless otherwise noted, we run every reconstruction method for all 10 test images (Section 4.3) with 3 repeats, each run with a different random seed, and for every metric we report the mean and standard deviation across these 30 runs.

5 Numerical Results

In this section, we present the numerical results of the reconstructions achieved by all evaluated methods and compare them. We evaluate **(1)** an optimization-based method with no prior or a Huber-TV prior (see Eq. (4.13)), **(2)** the proposed variational inference approach (Algorithm 1) under three choices of prior (no prior, Huber-TV prior, SSP using a score-based model s_θ) and **(3)** the blind RED-Diff method (Algorithm 2) using the same score-based model s_θ . In each table, “ $a \pm b$ ” indicates the empirical mean a and empirical standard deviation b of the respective metric,

Position-blind scenario				
Method / Metric	aPSNR $\uparrow$	cRMS $\downarrow$	aSSIM $\uparrow$	posCorrect $\uparrow$
Optimization-based				
No prior	14.19 $\pm$ 1.04	0.35 $\pm$ 0.15	0.11 $\pm$ 0.03	70.00 $\pm$ 19.06
H-TV prior ($\lambda = 0.1$)	19.08 $\pm$ 2.54	0.11 $\pm$ 0.04	0.37 $\pm$ 0.05	74.90 $\pm$ 12.76
Variational Inference				
No prior	15.68 $\pm$ 0.96	0.23 $\pm$ 0.08	0.16 $\pm$ 0.04	90.73 $\pm$ 6.14
H-TV prior ($\lambda = 5$)	23.35 $\pm$ 2.74	0.05 $\pm$ 0.03	0.65 $\pm$ 0.04	94.73 $\pm$ 4.29
SSP	25.34 $\pm$ 3.33	0.05 $\pm$ 0.03	0.85 $\pm$ 0.05	94.03 $\pm$ 5.24
SSP, pos. deltas	21.41 $\pm$ 5.39	0.20 $\pm$ 0.35	0.74 $\pm$ 0.17	65.07 $\pm$ 22.92
RED-Diff ($\lambda_{RD} = 20$)	24.17 $\pm$ 4.40	<u>0.09 $\pm$ 0.13</u>	<u>0.81 $\pm$ 0.09</u>	52.67 $\pm$ 23.55

Table 2: Metrics of reconstructions in the **position-blind** setting, comparing different methods. Best in bold, second best underlined. H-TV represents the Huber-TV prior with weight λ , and SSP refers to the surrogate score-based prior method.

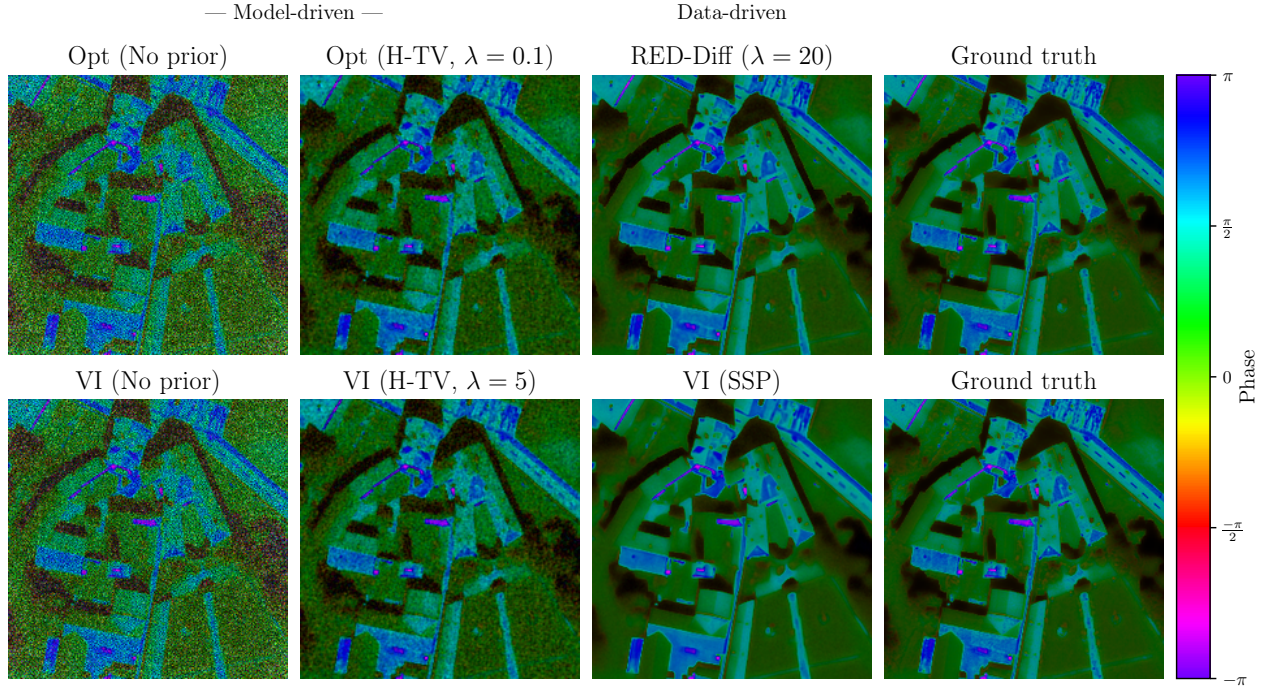
estimated using all 10 test objects, each with 3 independent reconstruction runs of the respective method.

5.1 Non-blind Baseline

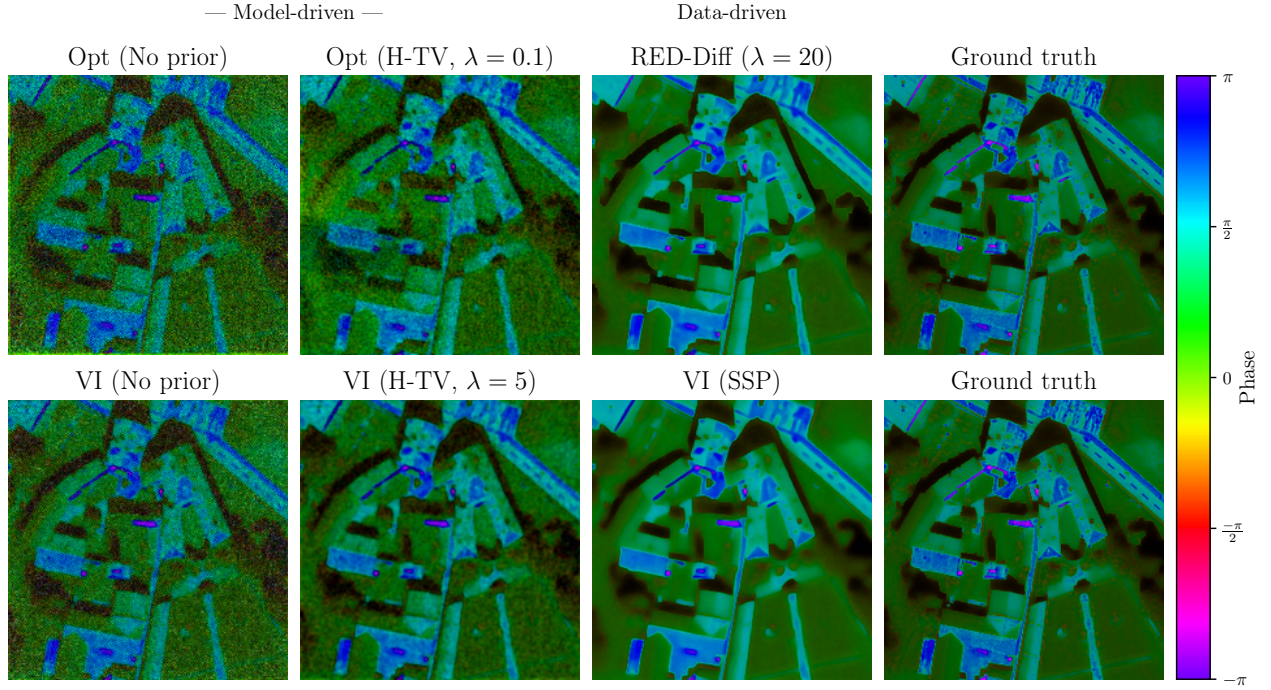
We first evaluate the described algorithms in the non-blind case. While this is not the main focus of the present work, we use this simpler case as a validation of our basic methodology and to gain an impression of the reconstruction quality that could be achievable under perfect position recovery. Here we use measurement noise of variance $\sigma_\epsilon^2 = 0.005$, which represents an average measurement signal-to-noise ratio (SNR) of 4.5 dB over all test objects.

In Table 1, we show the metric values of the compared methods. We can see that the variants without image priors only reach PSNR values around 12 dB, which can already be improved to around 25 dB by using a Huber-TV prior. Our variational method using the surrogate score-based prior (SSP) achieves a PSNR of 30 dB and SSIM of 0.90, closely followed by RED-Diff. This demonstrates the usefulness of the score-based data-driven priors for image quality in our problem. The optimization-based methods with no prior and a Huber-TV prior perform similarly to their VI counterparts here. For the weight parameter λ in the Huber-TV prior, we compare $\lambda \in \{5, 10, 20\}$, and find that larger λ achieve better aSSIM, though aPSNR already decreases again after $\lambda = 10$. Since we aim to reconstruct complex-valued images, we use $\lambda = 5$ for the VI method in all subsequent experiments, which achieves the best mean cRMS at the lowest standard deviation.

To illustrate the qualitative differences between the methods, in Fig. 3a we show the reconstruction with the best PSNR for each variational method. One can see that the reconstruction with no prior is heavily affected by the measurement noise, and the Huber-TV prior at $\lambda = 5$ only ameliorates this to some extent. H-TV with $\lambda = 10$ removes more noise but the image edges noticeably begin to blur. For a direct comparison, we show images with all three Huber-TV weights in Fig. 11. While RED-Diff seems to produce more fine details than SSP, these may be partly hallucinated, which is reflected in the slightly lower metric scores in Table 1. We note here that RED-Diff involves a weighting hyperparameter λ_{RD} for the score-based prior loss, the choice of which is ad-hoc and affects the image in a tradeoff between adherence to the measurement and adherence to the prior. RED-Diff therefore does not sample from the actual posterior induced by the likelihood and the prior, similar to issues of the related method DPS [13] as discussed in [25].



(a) Method and prior comparison in the **non-blind** setting



(b) Method comparison in the **blind** setting

Figure 3: Example image reconstructions of a single test object for the different optimization-based methods (*Opt*) and variational inference methods (*VI*), in (a) the non-blind baseline setting and (b) our position-blind setting. The images show complex magnitude as the brightness and complex phase as the hue.

Probe	aPSNR $\uparrow$	cRMS $\downarrow$	aSSIM $\uparrow$	posCorrect $\uparrow$
No phase mask				
$d_{\text{ap}} = 1/2$	10.53 ± 1.04	0.97 ± 0.31	0.25 ± 0.08	0.53 ± 0.68
$d_{\text{ap}} = 1/4$	11.32 ± 1.45	0.95 ± 0.37	0.23 ± 0.13	9.07 ± 10.13
$d_{\text{ap}} = 1/8$	21.16 ± 3.51	0.11 ± 0.07	0.68 ± 0.14	89.03 ± 7.85
Blockwise random phase mask, $d_{\text{ap}} = 1/2$				
$b = 32$	12.49 ± 3.19	0.81 ± 0.40	0.39 ± 0.17	15.10 ± 21.21
$b = 16$	18.77 ± 5.31	0.28 ± 0.38	0.65 ± 0.20	62.77 ± 31.23
$b = 8$	23.75 ± 4.94	0.16 ± 0.34	0.80 ± 0.14	85.97 ± 19.36
$b = 4$ (default)	25.34 ± 3.33	0.05 ± 0.03	0.85 ± 0.05	94.03 ± 5.24

Table 3: Metric comparison for the blind SSP method using different probe functions.

5.2 Blind case

Having established non-blind baseline performance, we now evaluate the methods in the fully position-blind case. Here, we now also report the “posCorrect” metric which indicates the percentage of correctly recovered positions; see Section 4.5. We use the same methods as in the previous Section 5.1, and also add a variant of our blind variational SSP method where the fitted position distributions are delta distributions, called “SSP, pos. deltas”. For this method the position estimation effectively turns into a pure optimization method while the image is still fitted as a variational distribution, similar to an EM scheme such as DeepGEM [27]. The goal of evaluating this method is to investigate the usefulness of the variational fitting of the positions in our SSP method.

The quantitative results are listed in Table 2. The SSP methods again achieve the best image metric results of all compared methods and recovers over 90% of the positions correctly, but regarding position recovery shows no advantage over using only a H-TV prior. This suggests that the score-based prior is helpful for retrieving better images, but not necessarily for easing the overall position-blind reconstruction problem. RED-Diff and our SSP variant with position deltas perform significantly worse at recovering the positions correctly at only 52-62% correctly recovered. Since RED-Diff also treats the fitted positions as fixed values rather than distributions (see Section 3), we conclude that the added randomness from fitting a random distribution on the positions is very helpful for position recovery. Nonetheless, both methods reach fair image metric values, suggesting that using only a subset of the measurements may be enough to retrieve a decent image in this specific simulated measurement setup.

One curious observation, when comparing the values for the non-blind case in Table 1 with the corresponding value for the blind case in Table 2, is that the methods with no prior or a Huber-TV prior to some extent reach better metrics in the blind case, in particular a better aSSIM. We found that this effect is due to the variations in the positions during reconstruction – both from the iterative updates and from the randomness of sampling from the variational distribution – which seem to act as an implicit smoothing prior on the image, resulting in better metrics in the presence of measurement noise.

5.2.1 Importance of the probe structure

Next we investigate the effect of the choice of probe function. First, we do not use a blockwise-random phase mask and only vary the aperture diameter d_{ap} , which inversely scales the probe size and therefore the measurement overlap. Then, for comparison, we add the blockwise-random phase mask to the largest aperture (smallest probe size) at $d_{\text{ap}} = 1/2$, with block sizes of $b \in \{4, 8, 16, 32\}$. We show the qualitative results visually in Fig. 4 and the quantitative results in Table 3. For completeness, we show all evaluated apertures and probe functions in the supplementary material;

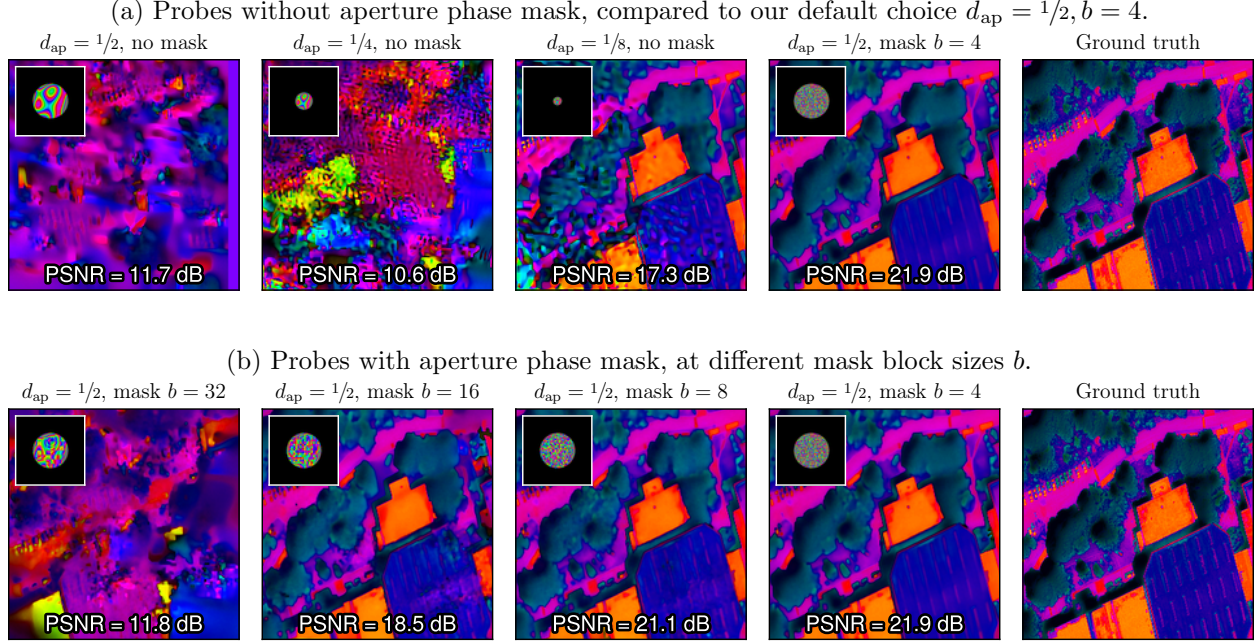


Figure 4: Reconstructed images for different probe functions with the SSP method. We compare (a) probes with different aperture diameters d_{ap} and (b) optional random aperture phase masks of different block sizes b . The insets show the aperture-plane wavefront generating each respective probe.

see Fig. 12.

Without a phase mask and $d_{\text{ap}} \in \{1/2, 1/4\}$, we recover almost no positions correctly and fail to produce usable images. While $d_{\text{ap}} = 1/8$ recovers 89% of positions correctly, its final images (Fig. 4a, center) are contaminated with low-frequency artifacts and the image metrics are subpar, which may be explained by the lack of high-frequency structure in the probe due to the low-pass from the small aperture. In comparison, adding block-wise random phase masks to the previously unusable aperture with $d_{\text{ap}} = 1/2$ leads to the best image quality we observed and also improves the position recovery, with a 25.3 dB aPSNR and 95% of positions recovered correctly for the block size $b = 4$. We further note that $b = 8$ still performs rather well compared to $b = 4$ despite illuminating a significantly smaller region of the object (see Fig. 2), again suggesting that probe structure is more helpful than probe size for our task.

These empirical observations are further corroborated by the loss landscapes shown in Fig. 5. In this figure, we compare three probe functions in the simplified task of recovering the position of a measurement when the ground-truth object is already known. We plot the sum of the squared errors between a noiseless measurement at the ground-truth position (0,0) and noiseless measurements at all other positions, showing the loss landscape of the likelihood term at all possible estimated positions. We can observe that the probe without the phase mask (Fig. 5, leftmost column) shows strong local minima separated by strong local maxima, and the loss landscape is much less convex than for the other two probes even around the true position (0,0).

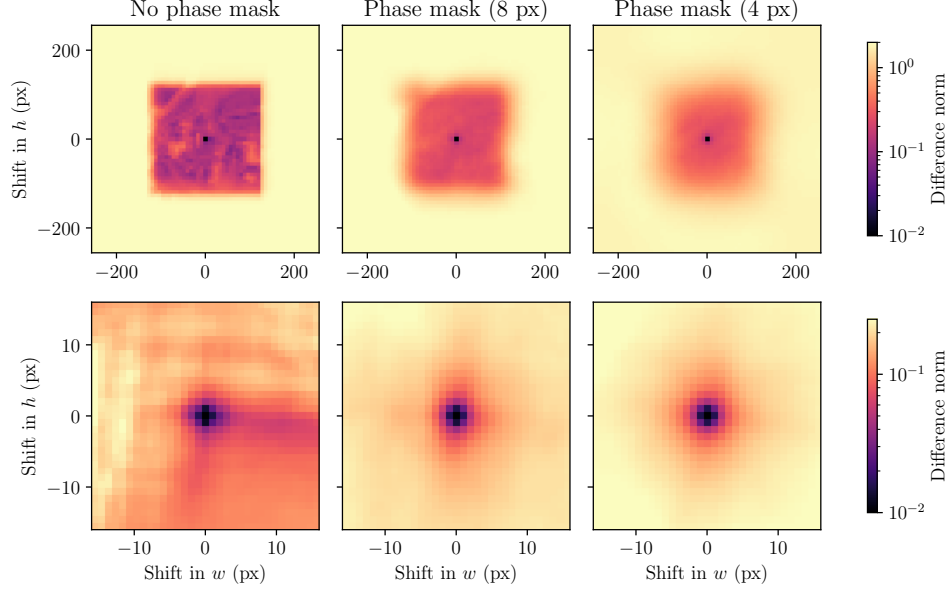


Figure 5: The loss landscape for an idealized position recovery problem. We plot summed squared errors between a noiseless measurement at the central position and simulated noiseless measurements at possible $(\Delta w, \Delta h)$ shifts relative to the center. We compare the three probe functions shown in Fig. 2 with $d_{\text{ap}} = 1/2$. The bottom row is zoomed in around $(0,0)$ and of higher resolution.

5.2.2 Different levels of measurement noise

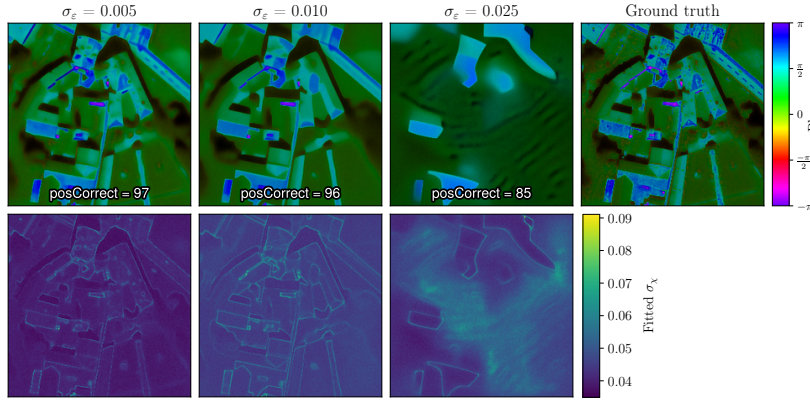
Here we analyze and compare the behavior of the blind SSP method, blind RED-Diff, and blind VI without an image prior when increasing the level (scale) of the Gaussian measurement noise σ_ϵ from the default $\sigma_\epsilon = 0.005$. We show quantitative results in Table 4, where we also list the corresponding measurement SNR, and qualitative results for the SSP method in Fig. 6a, where we also show the fitted per-pixel variance of the variational Gaussian that can inform uncertainty estimation. We can observe that **(1)** the data-driven priors allow for at least some image and position recovery even under -9.5 dB measurement noise, where the VI method without an image prior completely fails; **(2)** the SSP method performs best and most reliably under all noise levels and, surprisingly, recovers 69% of positions correctly even at the highest noise level; **(3)** the per-pixel uncertainty from SSP (Fig. 6) is, at least at the highest noise level, informative about regions that are heavily affected by artifacts. For a complete visual comparison of all methods under all measurement noise levels, see Fig. 13. We note here that the λ_{RD} parameter of RED-Diff could be tuned further in dependence of σ_ϵ in order to potentially improve the results, but we do not follow this here.

5.2.3 Poisson noise with SSP

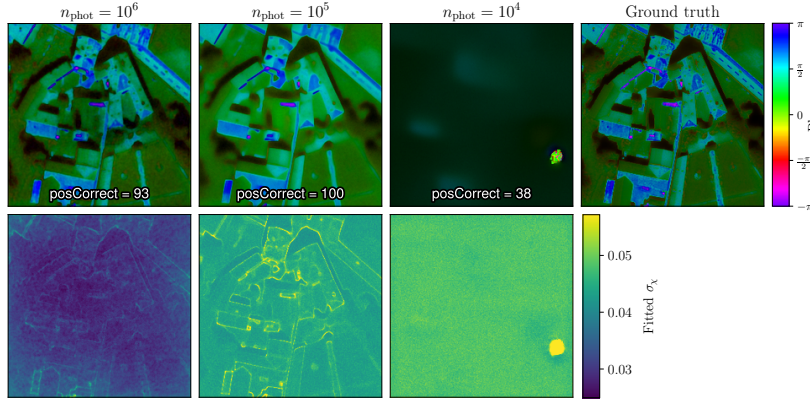
Here we evaluate the SSP method under different levels of Poissonian measurement noise, using the Gaussian approximation detailed in Section 4.1.4 and evaluating for different noise levels via different numbers of expected photons n_{phot} . We list the reconstruction metrics in Fig. 6c and show example images in Fig. 6b. We find that for $n_{\text{phot}} = 10^6$ and $n_{\text{phot}} = 10^5$ the method performs very well for position recovery and only somewhat worse for image recovery than in the Gaussian noise case with $\sigma_\epsilon = 0.005$. For $n_{\text{phot}} = 10^4$, the results degrade substantially, with the corresponding image only containing very coarse features of the ground truth and having a strong local artifact. Interestingly, at the same time, 37% of positions are still recovered correctly.

σ_ε	Meas. SNR	VI, SSP		RED-Diff ($\lambda_{RD} = 20$)		VI, no prior	
		cRMS ↓	pC ↑	cRMS ↓	pC ↑	cRMS ↓	pC ↑
0.005	4.47 dB	0.05±0.03	94.0±5.2	0.09±0.13	52.7±23.6	0.23±0.08	90.7±6.1
0.010	-1.55 dB	0.06±0.04	91.5±6.7	<u>0.21±0.38</u>	43.5±26.0	0.24±0.13	<u>74.4±12.5</u>
0.025	-9.51 dB	0.23±0.11	69.2±15.4	<u>0.30±0.20</u>	<u>16.2±11.9</u>	0.64±0.12	0.0±0.0

Table 4: Comparison of methods across Gaussian measurement noise levels in terms of the cRMS and posCorrect (here called pC for brevity) metrics. The best value in each row, i.e., for each noise level, is highlighted in bold; the second best is underlined. *Meas. SNR* indicates the average measurement signal-to-noise ratio corresponding to the respective noise level σ_ε .



(a) With **Gaussian** measurement noise at different noise levels σ_ε .



(b) With **Poissonian** measurement noise at different expected numbers of photons n_{phot} ; see Section 4.1.4.

n_{phot}	cRMS ↓	posCorrect ↑
10^4	0.66±0.13	37.33±11.89
10^5	0.13±0.04	96.17±6.79
10^6	0.08±0.05	94.20±7.86

(c) Metrics for SSP under different levels of **Poissonian** measurement noise with an expected number of photons n_{phot} .

Figure 6: Example reconstructions under different levels of (a) Gaussian and (b) Poissonian measurement noise, using the blind SSP method, and (c) corresponding metrics for Poisson noise. See Table 4 for the corresponding measurement SNRs. The bottom image rows show the fitted per-pixel standard deviation σ_χ of the Gaussian variational image distribution q_χ .

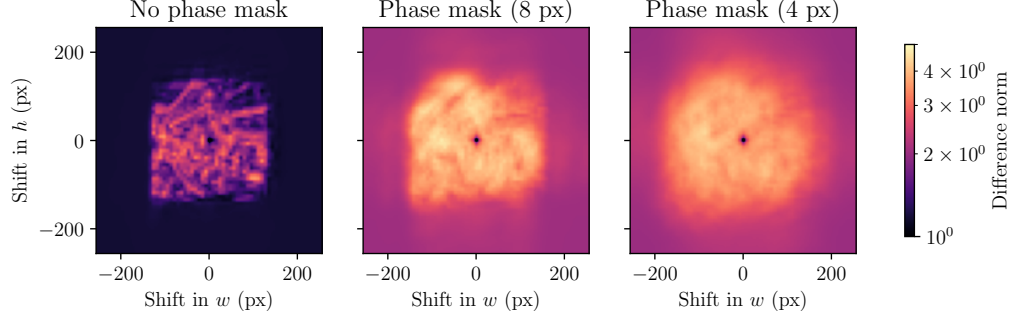


Figure 7: A position recovery loss landscape as in Fig. 5, but for a phase-only object (no absorption). The increased difficulty of this problem is evident from the broad regions of high loss around the global minimum and the broad regions of misleadingly low loss for positions lying outside of the object.

5.2.4 Phase-only objects

As a more difficult case, we now turn our attention to phase-only objects, that is, objects that are non-absorbing and only affect the phase of the incoming wavefront. We argue that this case is closer to real-world biological materials such as proteins, which typically consist only of light elements and thus have complex refractive indices $n = 1 - \delta + i\beta$ with $\delta \gg \beta$, $\delta \ll 1$ in the X-ray regime [34].

One problem we have to contend with in this scenario is the changed position recovery loss landscape, as illustrated in Fig. 7. This plot shows that, even with the added phase mask, the position loss landscape in this setting contains misleading broad regions of low loss for positions *off* the object and in fact has the highest loss values for most of the positions *on* the object, with only a small approximately convex region around the true position at (0,0). This is in contrast to the previous scenario with absorption (see Fig. 5), where the highest-loss regions were always those off the object. Without further constraints, we observed during initial reconstruction runs that the positions were frequently erroneously estimated to lie in the corners of the domain, i.e., maximally off the object, and we found successful image and position recovery to be unreliable.

We therefore added a log-barrier prior loss on the positions as described in Section 4.2, with the weight empirically set to 100. With this, we found a reliable reconstruction ability, with all objects and repeated runs leading to satisfactory image quality. In Fig. 8, we show successful example reconstructions for the standard probe $d_{\text{ap}} = 1/2, b = 4$ with the blind SSP method. Nonetheless, in terms of metrics, we find a worsened cRMS value of 0.06 ± 0.06 and a posCorrect value of only 61.60 ± 14.04 (cf. Table 2), showing the increased difficulty of this problem, particularly for position recovery.

5.2.5 Weak-phase objects with beamstop

Finally, we consider the most difficult scenario of position-blind ptychography of phase-only, weakly phase-distorting objects ($|\delta| \ll 1$), with a circular central beamstop in the measurements. The practical reasoning for using a central beamstop is that the direct portion of the beam has extremely high photon flux in XFEL experiments, which would lead to a measurement dynamic range that greatly exceeds what most detectors can handle, and may even damage or destroy the detector. We linearly rescale the phase shift δ (see Eq. (4.1)) of all phase-only test objects to the interval $[0, 10^{-3}]$. The beamstop blocks out the zero-order portion (and thus the full imaged aperture) of the diffracted beam on the detector, and is employed when faced with limited detector dynamic range and extreme source brightness such as for XFELs. To simulate the beamstop, we mask out

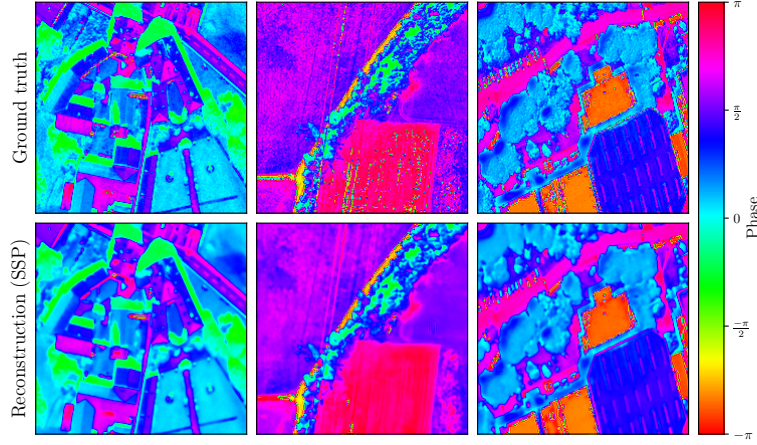


Figure 8: Example reconstructions of three phase-only test objects in the position-blind setting with the variational SSP method, when using a position log-barrier. Only the complex phase is shown as the hue.

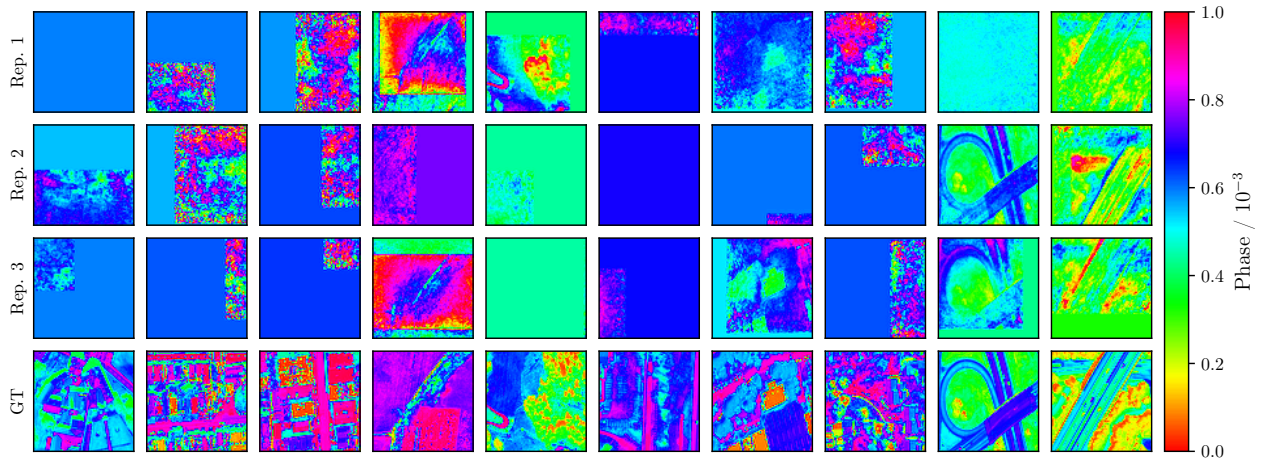


Figure 9: All attempted reconstructions (10 objects, 3 repeats) in the most difficult weak-phase phase-only object scenario with a measurement beamstop, where the maximum phase shift is 10^{-3} . The ground truth (GT) is shown in the bottom row. Only the complex phase is shown as the hue.

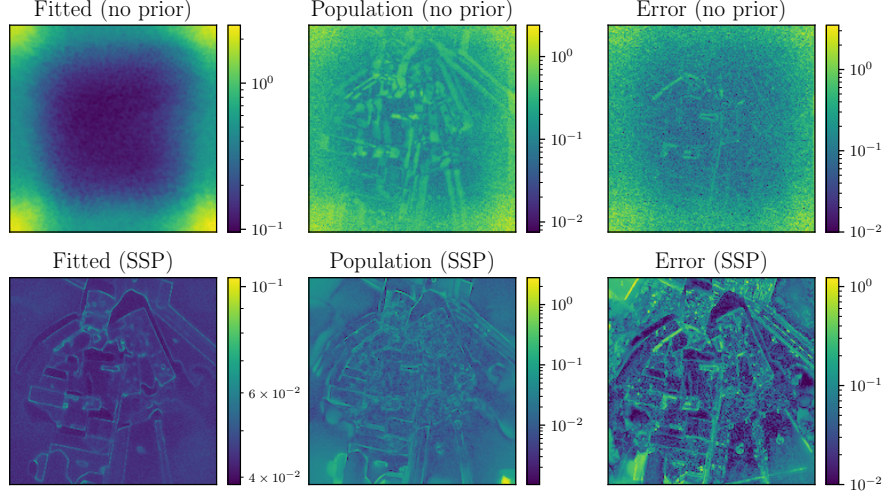


Figure 10: The per-pixel uncertainty without a prior and with a surrogate score-based prior (SSP), either estimated from the fitted diagonal covariance of the Gaussian distribution q_χ (*Fitted*) or estimated from a population of three independently fitted means (*Population*). For comparison, we also show the per-pixel to the ground-truth object, i.e., $|x - \hat{x}|$ (*Error*).

all pixels inside the circle that represent the aperture on the detector, i.e., all pixels that are within a circle of radius $(1 + \frac{d_{\text{ap}}}{2} \cdot 512)$ from the detector center, where we add another pixel to the radius to also block out high intensities that may occur at the edge of the imaged aperture. We inform our reconstruction method about this pixel-wise detector mask by ignoring all masked pixels when calculating the likelihood.

To block out fewer pixels from the measurement than for our default probe with $d_{\text{ap}} = 1/2$ where the imaged aperture covers half of each diffraction pattern, here we reduce the size of the aperture to $d_{\text{ap}} = 1/4$ but keep the random aperture-plane phase mask with $b = 4$. Furthermore, we decrease the measurement noise level to $\sigma_\varepsilon = 10^{-12}$ since the average intensity of pixels that are diffracted outside the region of the imaged aperture is much lower than those inside. We do not use a score-based prior here, since our neural network was not trained for images of weak-phase objects. For the algorithm (VI with no image prior), we set the image optimizer step sizes to $\lambda_{\mu_\chi} = \lambda_{\sigma_\chi} = 10^{-5}$ and use the position log-barrier (Section 4.2.4) with a weight of 10^6 . In this most difficult setting, we find that the probability of successful image recovery is very low, with frequent position misalignment of the entire object, but the method does occasionally achieve at least a partial reconstruction. This suggests that it is possible to further tune and improve our approach towards this scenario in the future. We show all result images in Fig. 9 for completeness.

5.3 Towards uncertainty quantification

A reliable estimate of uncertainty about each part of the reconstructed images would be helpful to inform experimental practice and scientific results. The variational approach with a Gaussian variational distribution yields a fitted per-pixel standard deviation, which may be used as a coarse approximation of a meaningful uncertainty estimate. Alternatively, the stochastic nature of our VI schemes also allows a straightforward population-based uncertainty estimate, by performing multiple independent reconstructions and constructing a per-pixel standard deviation map from this population.

To test both ideas, here we construct a set of measurements by restricting the simulated positions

to the central quarter of the image, which results in low coverage of the outer part of the object image in terms of information encoded in the collection of measurements. One would then hope that this lack of information about the outer part would be reflected in such uncertainty estimates. We compare SSP and VI with no image prior, and show both per-pixel uncertainty maps (fitted and population-based) in Fig. 10, in comparison to the actual per-pixel error to the ground truth image. We find that, without an image prior, the fitted uncertainty behaves as expected in the outer part of the image. With a data-driven prior, however, the fitted uncertainty does not reflect the low measurement coverage well, which may suggest an overly high confidence induced by the strong data-driven prior. The population estimate from SSP is somewhat more informative, but also does not clearly highlight the expected outer regions as uncertain.

6 Summary and Outlook

In this work, we investigate the novel blind inverse problem of position-blind ptychography for the first time, with possible applications in biological single-particle diffractive imaging (SPI) and ptychographic imaging under extreme movement of the sample. To attack this problem, we develop and demonstrate a Bayesian variational inference approach that can employ modern data-driven image priors and classic model-driven priors, and compare the developed method against another method from recent literature for data-driven solutions to blind inverse problems.

We evaluate our approach for increasingly difficult scenarios including phase-only objects, measurements with Poisson noise, and weak-phase phase-only objects under the presence of a beamstop, and show that we can reliably achieve successful reconstructions in all but the most difficult scenario. We investigate the underlying reasons for the difficulty of variants of our position-blind ptychography imaging problem, and propose remedies in the form of structured illumination and additional prior terms on the positions.

Since we treat only a simplified 2-D variant of the full 3-D position- and rotation-blind ptychography problem that would arise in SPI, we explicitly list our simplifying assumptions for clarity. We expect that it is possible for future works to build upon our Bayesian approach to include scenarios close to real-world SPI experiments towards 3D images with unknown positions *and* unknown rotations, unstable illumination (unknown or varying probe), and structured measurement noise, which we hope can eventually unlock this new imaging modality for real-world biological structure investigations.

Acknowledgments

MB, LK and TR acknowledge support from DESY (Hamburg, Germany), a member of the Helmholtz Association HGF. This research was supported in part through the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. This work was carried out while LK was a member of DESY. MB, TG, LK, TR and SW acknowledge funding by the German Federal Ministry of Research, Technology and Space (BMFT) under grant agreement No. 01IS24072A (COMFORT). MB, LK and TR acknowledge partial funding by the DAAD project 57698811 "Bayesian Computations for Large-scale (Nonlinear) Inverse Problems in Imaging".

References

- [1] Cagan Alkan et al. "Variational Diffusion Models for MRI Blind Inverse Problems". In: *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*. 2023 (cit. on pp. 9, 10).

- [2] Brian D.O. Anderson. “Reverse-time diffusion equation models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326. ISSN: 0304-4149 (cit. on p. 6).
- [3] Kartik Ayyer. “Reference-enhanced x-ray single-particle imaging”. In: *Optica* 7.6 (2020), pp. 593–601 (cit. on p. 2).
- [4] Kartik Ayyer et al. “Dragonfly: An Implementation of the Expand–Maximize–Compress Algorithm for Single-Particle Imaging”. In: *Journal of Applied Crystallography* 49.4 (2016), pp. 1320–1335 (cit. on p. 2).
- [5] Saša Bajt et al. “X-ray focusing with efficient high-NA multilayer Laue lenses”. In: *Light: Science & Applications* 7 (2018), 17162 EP - (cit. on p. 2).
- [6] Martin Benning and Martin Burger. “Modern regularization methods for inverse problems”. In: *Acta Numerica* 27 (2018), pp. 1–111 (cit. on p. 5).
- [7] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. ISSN: 1537-274X (cit. on pp. 7, 9).
- [8] M.J. Bogan et al. “Single Particle X-ray Diffractive Imaging”. In: *Nano Letters* 8.1 (2008), pp. 310–316 (cit. on p. 2).
- [9] Antonin Chambolle and Thomas Pock. “An introduction to continuous optimization for imaging”. In: *Acta Numerica* 25 (2016), pp. 161–319 (cit. on p. 16).
- [10] Henry N Chapman, Carl Caleman, and Nicusor Timneanu. “Diffraction before destruction”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1647 (2014), p. 20130313 (cit. on p. 4).
- [11] Henry N. Chapman. “X-Ray Free-Electron Lasers for the Structure and Dynamics of Macromolecules”. In: *Annual Review of Biochemistry* 88.1 (2019), pp. 35–58. ISSN: 0066-4154, 1545-4509 (cit. on p. 4).
- [12] Mathew J. Cherukara, Youssef S. G. Nashed, and Ross J. Harder. “Real-time coherent diffraction inversion using deep generative networks”. In: *Scientific Reports* 8.1 (2018). ISSN: 2045-2322 (cit. on p. 3).
- [13] Hyungjin Chung et al. “Diffusion Posterior Sampling for General Noisy Inverse Problems”. In: *The Eleventh International Conference on Learning Representations*. 2023 (cit. on pp. 2, 7, 19).
- [14] J. N. Clark et al. “Ultrafast Three-Dimensional Imaging of Lattice Dynamics in Individual Gold Nanocrystals”. In: *Science* 341.6141 (2013), pp. 56–59 (cit. on p. 2).
- [15] Alessandro Colombo et al. “Three-dimensional femtosecond snapshots of isolated faceted nanostructures”. In: *Science Advances* 9.8 (2023), eade5839 (cit. on p. 2).
- [16] Giannis Daras et al. *A Survey on Diffusion Models for Inverse Problems*. 2024. arXiv: [2410.00083 \[cs.LG\]](#) (cit. on pp. 6, 7).
- [17] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 17).
- [18] Alexander Denker et al. “Plug-and-Play Half-Quadratic Splitting for Ptychography”. In: *Scale Space and Variational Methods in Computer Vision*. Cham: Springer Nature Switzerland, 2025, pp. 269–281. ISBN: 978-3-031-92366-1 (cit. on p. 2).

- [19] Jonathan Dong et al. “Phase Retrieval: From Computational Imaging to Machine Learning: A tutorial”. In: *IEEE Signal Processing Magazine* 40.1 (2023), pp. 45–57. ISSN: 1558-0792 (cit. on p. 3).
- [20] J Lukas Dresselhaus et al. “X-ray focusing below 3 nm with aberration-corrected multilayer Laue lenses”. In: *Optics Express* 32.9 (2024), pp. 16004–16015 (cit. on pp. 2, 4).
- [21] Priya Dwivedi et al. “Lateral position correction in ptychography using the gradient of intensity patterns”. In: *Ultramicroscopy* 192 (2018), pp. 29–36. ISSN: 0304-3991 (cit. on pp. 2, 3).
- [22] Veit Elser. *Solving Problems with Projections: From Phase Retrieval to Packing*. Cambridge University Press, 2025 (cit. on p. 3).
- [23] Albert Fannjiang and Thomas Strohmer. “The numerics of phase retrieval”. In: *Acta Numerica* 29 (2020), pp. 125–228. ISSN: 1474-0508 (cit. on p. 3).
- [24] Berthy Feng and Katherine Bouman. “Variational Bayesian Imaging with an Efficient Surrogate Score-based Prior”. In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856 (cit. on pp. 3, 7, 8, 10, 11, 34).
- [25] Berthy T. Feng et al. “Score-Based Diffusion Models as Principled Priors for Inverse Imaging”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023 (cit. on pp. 7, 8, 10, 19).
- [26] J. R. Fienup. “Phase retrieval algorithms: a comparison”. In: *Applied Optics* 21.15 (1982), p. 2758. ISSN: 1539-4522 (cit. on p. 3).
- [27] Angela Gao et al. “DeepGEM: Generalized Expectation-Maximization for Blind Inversion”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 11592–11603 (cit. on pp. 11, 21).
- [28] Yunhui Gao and Liangcai Cao. “A complex constrained total variation image denoising algorithm with application to phase retrieval”. In: *arXiv preprint arXiv:2109.05496* (2021) (cit. on p. 16).
- [29] Timo Gerkmann, Martin Krawczyk, and Robert Rehr. “Phase estimation in speech enhancement – Unimportant, important, or impossible?”. In: *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. 2012, pp. 1–5 (cit. on p. 3).
- [30] Timo Gerkmann, Martin Krawczyk-Becker, and Jonathan Le Roux. “Phase processing for single-channel speech enhancement: History and recent advances”. In: *IEEE Signal Processing Magazine* 32.2 (2015), pp. 55–66 (cit. on p. 3).
- [31] Tais Gorkhover et al. “Femtosecond X-ray Fourier holography imaging of free-flying nanoparticles”. In: *Nature Photonics* 12.3 (2018), pp. 150–153 (cit. on p. 2).
- [32] Manuel Guizar-Sicairos et al. “Role of the illumination spatial-frequency spectrum for ptychography”. In: *Phys. Rev. B* 86 (10 2012), p. 100103 (cit. on p. 14).
- [33] Francesco Guzzi et al. “A modular software framework for the design and implementation of ptychography algorithms”. In: *PeerJ Computer Science* 8 (2022), e1036 (cit. on p. 15).
- [34] B.L. Henke, E.M. Gullikson, and J.C. Davis. “X-Ray Interactions: Photoabsorption, Scattering, Transmission, and Reflection at $E = 50\text{--}30,000$ eV, $Z = 1\text{--}92$ ”. In: *Atomic Data and Nuclear Data Tables* 54.2 (1993), pp. 181–342. ISSN: 0092-640X (cit. on p. 25).
- [35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851 (cit. on pp. 2, 17).

- [36] Aapo Hyvärinen. “Estimation of Non-Normalized Statistical Models by Score Matching”. In: *Journal of Machine Learning Research* 6.24 (2005), pp. 695–709 (cit. on p. 6).
- [37] Shaowei Jiang et al. “Solving Fourier ptychographic imaging problems via neural network modeling and TensorFlow”. In: *Biomed. Opt. Express* 9.7 (2018), pp. 3306–3319 (cit. on p. 15).
- [38] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Springer New York, 2005 (cit. on p. 5).
- [39] Saugat Kandel et al. “Using automatic differentiation as a general framework for ptychographic reconstruction”. In: *Opt. Express* 27.13 (2019), pp. 18653–18672 (cit. on p. 15).
- [40] Bahjat Kawar et al. “Denoising Diffusion Restoration Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 23593–23606 (cit. on p. 2).
- [41] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *International Conference on Learning Representations* (2014). arXiv preprint arXiv:1312.6114 (cit. on p. 11).
- [42] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. 2015 (cit. on pp. 11, 15, 17).
- [43] Charles Laroche, Andrés Almansa, and Eva Coupete. “Fast Diffusion EM: a diffusion model for blind inverse problems with application to deconvolution”. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024, pp. 5259–5269 (cit. on p. 11).
- [44] Abraham L. Levitan et al. “Single-frame far-field diffractive imaging with randomized illumination”. In: *Opt. Express* 28.25 (2020), pp. 37103–37117 (cit. on p. 14).
- [45] Haohe Liu et al. “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 21450–21474 (cit. on p. 2).
- [46] Li Liu et al. “An efficient and robust self-calibration algorithm for translation position errors in ptychography”. In: *IEEE Transactions on Instrumentation and Measurement* 73 (2024), pp. 1–12 (cit. on pp. 2, 3).
- [47] Ne-Te Duane Loh and Veit Elser. “Reconstruction Algorithm for Single-Particle Diffraction Imaging Experiments”. In: *Phys. Rev. E* 80.2 (2009), p. 026705 (cit. on p. 2).
- [48] Emmanuel Maggiori et al. “Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark”. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017 (cit. on p. 17).
- [49] AM Maiden et al. “An annealing algorithm to correct positioning errors in ptychography”. In: *Ultramicroscopy* 120 (2012), pp. 64–72 (cit. on pp. 2, 3).
- [50] Andrew M. Maiden and John M. Rodenburg. “An improved ptychographical phase retrieval algorithm for diffractive imaging”. In: *Ultramicroscopy* 109.10 (2009), pp. 1256–1262. ISSN: 0304-3991 (cit. on pp. 2, 18).
- [51] Morteza Mardani et al. “A Variational Perspective on Solving Inverse Problems with Diffusion Models”. In: *The Twelfth International Conference on Learning Representations*. 2024 (cit. on pp. 2, 3, 7, 9, 10, 16).

- [52] Oleh Melnyk. “Convergence properties of ePIE and stochastic gradient methods for blind ptychography”. In: *Analysis and Applications* (2025), pp. 1–49. ISSN: 1793-6861 (cit. on p. 12).
- [53] Benjamin Rainier Mobley. “Theoretical Techniques for the Recovery of Structural Information of Biomolecules Using Diffraction Data From X-ray Free-Electron Lasers”. PhD thesis. Arizona State University, 2024 (cit. on p. 2).
- [54] Richard Neutze et al. “Potential for Biomolecular Imaging with Femtosecond X-ray Pulses”. In: *Nature* 406.6797 (2000), pp. 752–757. ISSN: 0028-0836, 1476-4687 (cit. on pp. 2, 4).
- [55] Adam Paszke et al. “PyTorch: an imperative style, high-performance deep learning library”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on p. 15).
- [56] Pietro Perona and Jitendra Malik. “Scale-space and edge detection using anisotropic diffusion”. In: *IEEE Transactions on pattern analysis and machine intelligence* 12.7 (2002), pp. 629–639 (cit. on p. 17).
- [57] Franz Pfeiffer. “X-ray ptychography”. In: *Nature Photonics* 12.1 (2017), pp. 9–17 (cit. on pp. 2, 3).
- [58] R Tyrrell Rockafellar. *Convex analysis*. Vol. 28. Princeton university press, 1997 (cit. on p. 16).
- [59] John Rodenburg and Andrew Maiden. “Ptychography”. In: *Springer Handbook of Microscopy*. Springer International Publishing, 2019, pp. 819–904. ISBN: 9783030000691 (cit. on pp. 2, 3).
- [60] Yaniv Romano, Michael Elad, and Peyman Milanfar. “The Little Engine That Could: Regularization by Denoising (RED)”. In: *SIAM Journal on Imaging Sciences* 10.4 (2017), pp. 1804–1844. ISSN: 1936-4954 (cit. on p. 8).
- [61] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695 (cit. on p. 2).
- [62] M. Marvin Seibert et al. “Single mimivirus particles intercepted and imaged with an X-ray laser”. In: *Nature* 470.7332 (2011), pp. 78–81 (cit. on p. 2).
- [63] Yoav Shechtman et al. “Phase Retrieval with Application to Optical Imaging: A contemporary overview”. In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109. ISSN: 1053-5888 (cit. on p. 3).
- [64] Yang Song et al. “Maximum Likelihood Training of Score-Based Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 1415–1428 (cit. on p. 6).
- [65] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations*. 2021 (cit. on pp. 2, 6, 8, 17).
- [66] A. M. Stuart. “Inverse problems: A Bayesian perspective”. In: *Acta Numerica* 19 (2010), pp. 451–559 (cit. on p. 5).
- [67] He Sun and Katherine L. Bouman. “Deep Probabilistic Imaging: Uncertainty Quantification and Multi-modal Solution Characterization for Computational Imaging”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. Association for the Advancement of Artificial Intelligence (AAAI), 2021, pp. 2628–2637 (cit. on p. 7).
- [68] Miklós Tegze and Gábor Bortel. “Comparison of EMC and CM Methods for Orienting Diffraction Images in Single-Particle Imaging Experiments”. In: *IUCrJ* 8.6 (2021), pp. 980–991. ISSN: 2052-2525 (cit. on p. 2).

- [69] Pascal Vincent. “A connection between score matching and denoising autoencoders”. In: *Neural computation* 23.7 (2011), pp. 1661–1674 (cit. on p. 6).
- [70] August Wollter. “Phase Retrieval and Orientation Recovery in Single-Particle Coherent Diffractive Imaging: Background Noise and Biased Orientations”. Uppsala: Acta Universitatis Upsaliensis, 2024. 73 pp. (cit. on p. 2).
- [71] Jumpei Yamada et al. “Extreme Focusing of Hard X-ray Free-Electron Laser Pulses Enables 7 nm Focus Width and 10^{22} W cm $^{-2}$ Intensity”. In: *Nature Photonics* 18.7 (2024), pp. 685–690. ISSN: 1749-4885, 1749-4893 (cit. on p. 4).
- [72] Margarita Zakharova et al. “Focusing of X-ray free-electron laser pulses using multilayer Laue lenses”. In: *Opt. Express* 33.15 (2025), pp. 31884–31895 (cit. on p. 4).
- [73] Frits Zernike. “Beugungstheorie des Schneidenvorgangs und seiner verbesserten Form, der Phasenkontrastmethode”. de. In: *Physica* 1.7-12 (1934), pp. 689–704 (cit. on p. 14).
- [74] Fucai Zhang et al. “Translation position determination in ptychographic coherent diffraction imaging”. In: *Optics express* 21.11 (2013), pp. 13592–13606 (cit. on pp. 2, 3, 18).
- [75] Yuanzhi Zhu et al. “Denoising diffusion models for plug-and-play image restoration”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 1219–1229 (cit. on p. 2).

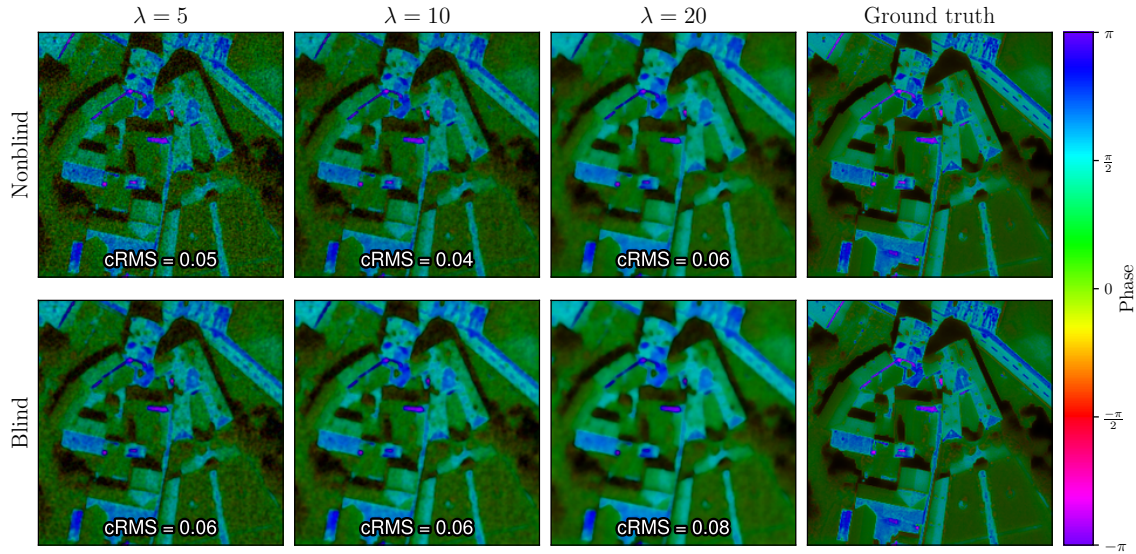


Figure 11: Comparison of VI reconstruction with the Huber-TV prior under different prior weights $\lambda \in 5, 10, 20$.

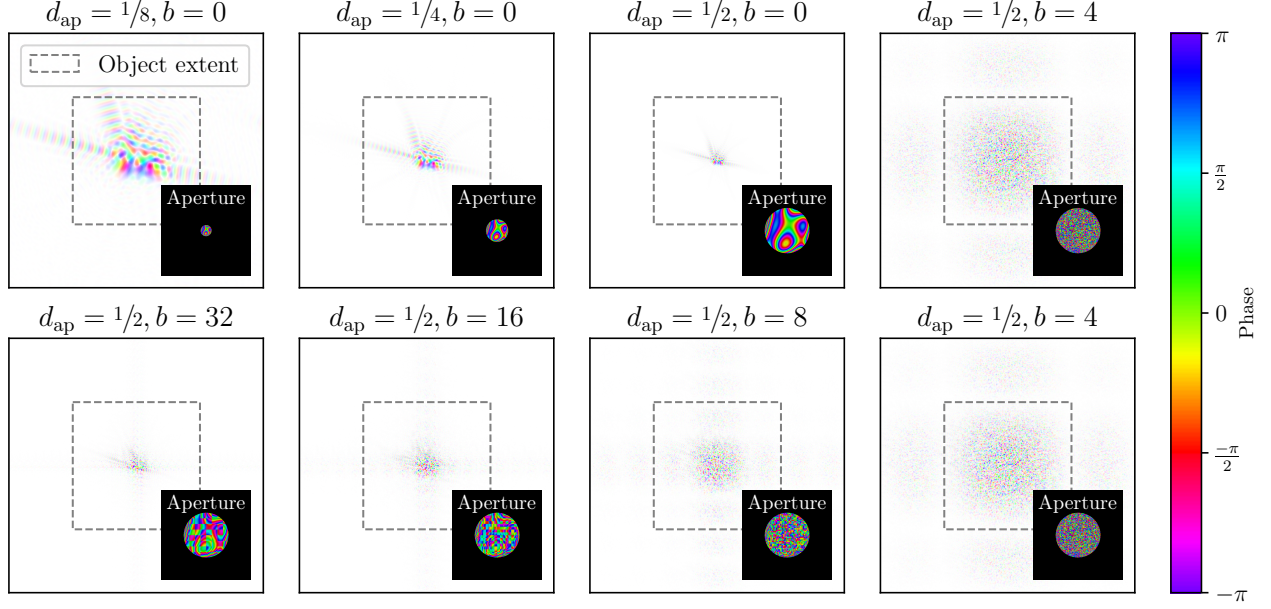


Figure 12: All probes and corresponding apertures used for evaluation in this work, in particular when comparing the effect of different probes. Our default choice throughout the work is the probe with $d_{\text{ap}} = 1/2, b = 4$, which is visually duplicated here for easier visual comparison and a cleaner layout. $b = 0$ indicates no phase mask was used for the respective probe.

A RED-diff Algorithm for Blind Inverse Problems

B Definition of the ELBO surrogate

We provide the definition of the ELBO surrogate b_{θ}^{SDE} , which is directly copied from [24, p. 5]. Namely, we define

$$b_{\theta}^{\text{SDE}}(x) = \mathbb{E}_{x \sim p_T(x|x_0)} [\log \pi(x)] - \frac{1}{2} \int_0^1 g(t)^2 h(t) dt$$

where

$$h(t) = \mathbb{E}_{x \sim p_t(x|x_0)} \left[\|s_{\theta}(x, t) - \nabla_x \log p(x|x_0)\|_2^2 - \|\nabla_x \log p_t(x|x_0)\|_2^2 - \frac{2}{g(t)^2} \nabla_x \cdot f(x, t) \right]$$

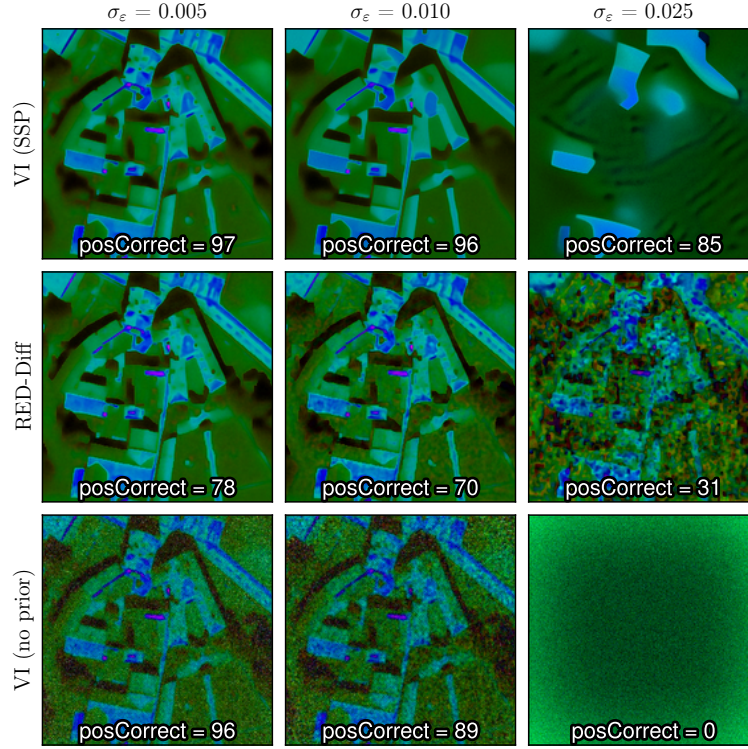


Figure 13: Reconstructed images of a test object from three methods (SSP, RED-Diff, VI without an image prior) under three increasing levels of measurement noise σ_ϵ .

Algorithm 2 RED-diff for blind inverse problems

- 1: Initialize parameter estimate $\mathbf{r}^{(0)}$, image estimate $x^{(0)}$, $l = 0$, data $y \in \mathcal{Y}$, forward model $A : \mathcal{R} \times \mathcal{X} \rightarrow \mathcal{Y}$, maximum number of iterations $l_{\max}$, step size sequences $(\tau^{(l,i)}), (\eta^{(l,i)})$
 - 2: **while** $l < l_{\max}$ and stopping criterion on $x^{(l)}, \mathbf{r}^{(l)}$ is not satisfied **do**
 - 3: $x^{(l,0)} = x^{(l)}$
 - 4: **for** $i \leftarrow 0, \dots, N_{\text{img}} - 1$ **do** ▷ Solve for image (3.16) using gradients (3.17)
 - 5: $x^{(l,i+1)} = x^{(l,i)} - \tau^{(l,i)} \nabla_x \mathcal{L}^{\text{REDdiff}}(x^{(l,i)}, \mathbf{r}^{(l)})$
 - 6: **end for**
 - 7: $x^{(l+1)} = x^{(l, N_{\text{img}})}$
 - 8: $\mathbf{r}^{(l,0)} = \mathbf{r}^{(l)}$
 - 9: **for** $i \leftarrow 0, \dots, N_{\text{par}} - 1$ **do** ▷ Solve for parameters (3.18)
 - 10: $\mathbf{r}^{(l,i+1)} = \mathbf{r}^{(l,i)} - \eta^{(l,i)} \nabla_{\mathbf{r}} \mathcal{L}^{\text{REDdiff}}(x^{(l+1)}, \mathbf{r}^{(l,i)})$
 - 11: **end for**
 - 12: $\mathbf{r}^{(l+1)} = \mathbf{r}^{(l, N_{\text{par}})}$
 - 13: $l \leftarrow l + 1$
 - 14: **end while**
 - 15: **return** image estimate $x^{(l)}$, parameter estimate $\mathbf{r}^{(l)}$
-