High-Dimensional Uncertainty Quantification with Deep Data-Driven AI Priors

Jason D. McEwen[1,2], Tobías I. Liaudat[3]

[1]Mullard Space Science Laboratory (MSSL), University College London (UCL), Dorking RH5 6NT, UK (jason.mcewen@ucl.ac.uk)

[2]Alan Turing Institute, London NW1 2DB, UK

[3]IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette Cedex, France (tobias.liaudat@cea.fr)

Abstract

High-dimensional inverse problems are central to modern scientific discovery, yet the transition to the exascale era presents significant computational challenges. As instruments like the Square Kilometre Array (SKA) produce data at unprecedented volumes, traditional Bayesian inference via Markov chain Monte Carlo (MCMC) becomes computationally infeasible. While deep learning offers a powerful alternative for reconstructing complex signals, standard black-box models often lack the necessary physical consistency and rigorous uncertainty quantification required for scientific analysis. We synthesize disparate trends in the literature to chart a path toward a unified framework for high-dimensional uncertainty quantification with deep data-driven AI priors. We evaluate state-of-the-art methods against the competing demands of computational efficiency, robustness, reconstruction fidelity, and statistical reliability, finding that no single existing method simultaneously satisfies all these criteria. Consequently, we advocate for a trustworthy framework that integrates efficient physics-informed architectures to ensure data consistency, expressive deep data-driven artificial intelligence (AI) priors to capture complex signal structure, and scalable uncertainty quantification strategies. Crucially, we highlight the use of post-hoc calibration via conformal prediction to bridge the reliability gap, transforming heuristic uncertainty estimates into rigorous statistical bounds. We conclude that combining physics-informed generative unrolled networks with conformal calibration offers a promising path toward robust, scalable, and scientifically reliable imaging in the exascale regime.

1. INTRODUCTION

High-dimensional inverse problems are ubiquitous in modern science, from medical imaging to astrophysics. As we enter the "exascale" era, instruments are producing data at unprecedented volumes and resolutions. A prime example is the Square Kilometre Array (SKA), which will generate petabytes of visibility data to reconstruct gigapixel images of the radio sky [40]. In these regimes, the sheer scale of both the data volumes and parameter spaces, and the cost of the measurement operator, renders traditional computational methods inadequate. We require new methods that are computationally efficient, physics-informed, expressive, and able to quantify uncertainties.

We focus on the ubiquitous setting of linear inverse problems of the form $y = \Phi x + n$, where $y \in \mathbb{R}^M$ represents the observed data of dimension M, $x \in \mathbb{R}^N$ is the underlying signal to be reconstructed, of dimension N, $\Phi : \mathbb{R}^N \to \mathbb{R}^M$ is the measurement operator, and $n \in \mathbb{R}^M$ denotes noise (extensions to non-linear problems are also possible). Prominent examples of linear inverse problems include radio interferometric imaging and weak gravitational lensing mass-mapping in astronomy, alongside magnetic resonance imaging (MRI) and computed tomography (CT) in medical imaging.

In many scientific applications, such as those mentioned above, inverse problems are ill-conditioned and ill-posed, in the sense of Hadamard [19]. That is, a solution may not exist, or may not be unique, or may not be stable with respect to the data. Consequently, we must inject prior information to regularize the problem. Moreover, quantifying uncertainty is thus of critical importance. In fields like astronomy and medical imaging, a single "best guess" point estimate is insufficient. Scientific inquiry requires quantifying the reliability of reconstructed structures to determine whether a feature is a physical reality or an artifact of the reconstruction. These dual

requirements of *regularization for stability* and *uncertainty quantification for scientific rigour*, strongly motivate the *Bayesian inference paradigm*, where the goal is not merely a single image but the characterization of the full posterior distribution $p(x|y)$.

Ideally, one would explore the full posterior distribution using Markov chain Monte Carlo (MCMC) sampling [e.g. 30, 28]. However, for modern high-dimensional problems the computational cost of standard MCMC is prohibitive due to three compounding factors: (i) the high-dimensional parameter space, where the number of parameters N of x is very large (e.g., gigapixel images); (ii) the large data volume, where the number of measurements M of y is massive (e.g., petabytes of visibility data); and (iii) the costly measurement operator Φ, which is computationally expensive to evaluate (e.g., non-uniform FFTs with modelling to account for wide-field effects and non-coplanar baselines in radio interferometry [34, 35]). For instance, radio interferometric imaging with the SKA hits all three bottlenecks simultaneously, rendering standard MCMC approaches computationally infeasible.

As instruments become more powerful, capturing higher resolutions and more complex structures, the simple, hand-crafted priors of the past (e.g., sparsity, wavelets, total variation) are no longer sufficient. Reconstructing these complex images requires more expressive data-driven priors, a task where artificial intelligence (AI) excels. Deep data-driven priors offer superior reconstruction fidelity compared to classical methods, capturing intricate signal structures that analytical priors fail to model.

To address the challenges of modern, highly computationally demanding inverse problems, which are exacerbated at the exascale as demonstrated by the SKA, we advocate for approaches that satisfy four simultaneous criteria:

1. **Computationally Efficient:** Methods must scale to high dimensions, typically relying on optimization rather than full MCMC.
2. **Physics-Informed:** The forward operator Φ must be explicitly integrated to ensure data consistency and physical plausibility, allowing methods to generalize across varying measurement configurations.
3. **Expressive Data-Driven AI Priors:** Leverage deep learning to enhance reconstruction fidelity beyond hand-crafted priors such as, e.g., wavelet sparsity.
4. **Quantified Uncertainties:** Go beyond point estimates to provide rigorous error bars or confidence intervals that are reliable.

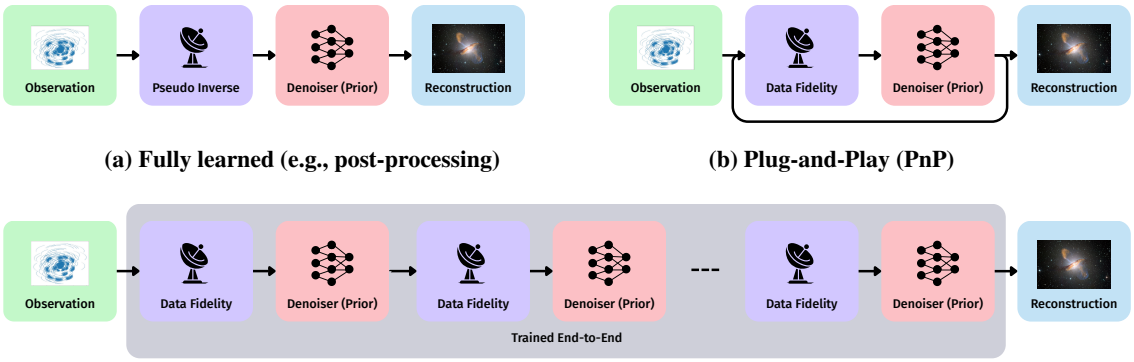
In this article we review the state-of-the-art through the lens of these four competing requirements. To date, no single method appears to simultaneously satisfy all these criteria. Consequently, this article serves not merely as a review but as a synthesis of disparate trends in the literature, charting a path toward a unified framework for high-dimensional uncertainty quantification with deep data-driven AI priors. We begin by reviewing reconstruction methods to solve the inverse problem (Section 2), proceed to discuss scalable uncertainty quantification strategies (Section 3), and then address the critical issue of trustworthiness via coverage testing and calibration (Section 4). Finally, concluding remarks are given in Section 5.

2. RECONSTRUCTION: SOLVING THE INVERSE PROBLEM

We review contemporary data-driven approaches for solving high-dimensional inverse problems, focusing initially on methods designed to recover accurate point estimates of the underlying signal. We trace the evolution of these techniques from purely data-driven "black-box" models to hybrid architectures that explicitly integrate the physical measurement operator. By combining the expressivity of deep learning with the robustness of physical models, these strategies aim to overcome the limitations of classical regularization while scaling to the data volumes of the exascale era. A diagrammatic overview of the various approaches is illustrated in Figure 1.

2.1 Fully Learned Reconstruction (Pre-/Post-Processing)

The most straightforward application of deep learning to inverse problems is to learn a direct mapping from the observed data y, or more commonly a proxy such as the "dirty" image $x_{\text{dirty}} = \Phi^\dagger y$, to the underlying signal x [1, 21]. Architectures like the U-Net [38] are typically employed to approximate the inverse mapping $x_{\text{dirty}} \mapsto \hat{x}$, effectively treating the reconstruction as a learned denoising or artifact-removal task. This approach is computationally efficient, often requiring only a single network pass and one application of the adjoint operator $\Phi^\dagger$, making



(a) Fully learned (e.g., post-processing) **(b) Plug-and-Play (PnP)**

(c) Unrolled

Figure 1: Schematic overview of deep learning-based imaging methods for inverse problems. (a) Fully learned reconstruction: *A neural network (e.g., U-Net) maps a proxy reconstruction (such as the dirty image, $\Phi^\dagger y$) to the estimated signal $\hat{x}$, with the measurement operator used only implicitly during training.* ***(b) Plug-and-Play (PnP):*** *Iterative algorithms alternate between explicit physics-based data consistency and a learned denoiser, treating the network as a proximal operator replacement.* ***(c) Unrolled:*** *The iterations of an optimization algorithm are "unfolded" into a fixed-depth neural network, often allowing parameters and proximal operators to be learned from data, and explicitly incorporating the measurement operator at each layer.* ***Generative extensions:*** *Generative models (e.g., GANs, diffusion models) can be incorporated for all approaches, i.e., post hoc (direct mapping from the dirty image or intermediate reconstructions to posterior samples), as the denoising step in PnP, or within unrolled architectures (e.g., generative GU-Nets), combining physical data consistency with expressive learned priors.*

it highly attractive for real-time applications like the SPIDER instrument [33, 25] or large-scale surveys like those expected with the SKA [26].

However, these methods face significant limitations. First, they are generally restricted to the specific training conditions; a network trained on a fixed telescope configuration (e.g., a specific uv-coverage) may fail to generalize to observations with different sampling patterns, leading to poor performance when the forward operator varies [26]. Second, because the measurement operator Φ is not explicitly incorporated into the inference step (only implicitly during training), there is no guarantee of data consistency; the reconstructed $\hat{x}$ may not satisfy $y \approx \Phi\hat{x}$. This "black box" nature can result in hallucinations or the removal of faint physical structures not well-represented in the training set.

2.2 Plug-and-Play (PnP)

Plug-and-Play (PnP) approaches offer a flexible framework that bridges the gap between model-based optimization and deep learning. In traditional iterative algorithms, such as the Alternating Direction Method of Multipliers (ADMM) or Forward-Backward Splitting [13], the solution is found by alternating between a data-fidelity step and a regularization step. The PnP approach replaces the proximal operator associated with the explicit prior (e.g., total variation or ℓ_1 sparsity) with a learned off-the-shelf denoiser, such as a deep neural network [45]. This substitution effectively injects a learned data distribution as the prior without requiring an explicit analytical formulation.

A primary benefit of the PnP framework is the decoupling of the prior from the forward measurement operator. Because the denoiser is trained independently of the specific inverse problem, it remains robust to changes in the physics of the instrument, such as varying telescope configurations in radio interferometry. This contrasts with fully learned reconstruction methods, which often fail to generalize when the measurement operator deviates from the training distribution. Furthermore, recent theoretical developments have addressed the stability of these heuristic methods [32]. By constraining the deep data-driven prior, for instance, by enforcing a Lipschitz constant strictly less than unity via spectral normalization, it is possible to prove fixed-point convergence for the iterative scheme [39].

Despite these advantages, PnP methods face a significant computational bottleneck. The iterative nature of the algorithm typically requires hundreds or even thousands of iterations to reach convergence [43, Table 1]. Since each iteration involves applying the forward measurement operator and its adjoint, due to the likelihood gradient

calculation, the total computational cost can be prohibitive for high-dimensional exascale problems where the operator itself is expensive to evaluate.

2.3 Unrolled architectures

Unrolled networks represent a hybrid approach that unfolds the iterations of a classical optimization algorithm, such as ADMM or Forward-Backward Splitting, into a fixed number of layers in a deep neural network [18]. By interpreting the iteration index as a layer index, these architectures can be trained end-to-end, allowing the network to learn optimal step sizes and regularization parameters (or even the proximal operators themselves) from data. However, these approaches face significant challenges: they typically require the measurement operator to be differentiable, which may not always be available for complex instrument models, and they necessitate multiple expensive evaluations of the measurement operator for a single network pass during training, resulting in high computational and memory costs.

To overcome the significant computational cost of standard unrolled approaches, a Gradient U-Net (GU-Net) can be considered, where the measurement operator is effectively applied at different resolutions inside the U-Net architecture [25, 26]. This multiscale integration significantly reduces the computational burden, as only two full-resolution measurement operator applications are required during the network pass. At each layer, the network ingests not only the current image representation but also the gradient of the data fidelity term, ensuring data consistency is actively enforced.

2.4 Generative AI Extensions

Generative models offer a powerful mechanism to enhance reconstruction fidelity by leveraging deep learning to capture the complex, non-Gaussian statistics of the underlying signal. Unlike simple analytical priors, generative models, such as Generative Adversarial Networks (GANs) [16, 5] or diffusion models [20, 41], learn an expressive distribution from high-quality training data, enabling the recovery of intricate structures that traditional methods often smooth over. While their primary advantage here is the superior quality of the point estimates they can produce, these models also naturally support the generation of multiple samples, offering a potential route to uncertainty quantification (as discussed further in Section 3).

Generative Post-Processing. The most direct extension of learned reconstruction is to condition a generative model, such as a conditional GAN [1], on the dirty image or a preliminary reconstruction. In this framework, the generator learns a mapping from the observed data and a latent noise vector to a sample from the posterior distribution. For example, Whitney et al. [46] demonstrate this approach for weak gravitational lensing mass-mapping, using a conditional GAN to produce posterior samples that capture non-Gaussian statistics. To prevent mode collapse, a common failure mode where the generator ignores the latent code and produces a single deterministic output, which can be particularly problematic in the context of inverse problems, regularization strategies are essential. The regularization strategy of Bendel et al. [7] is adopted by Whitney et al. [46], which explicitly penalizes the lack of diversity in generated samples, ensuring that the first and second moments of the posterior are correctly recovered in idealised settings. While computationally efficient, post-processing methods based on generative models may be less robust to shifts in the measurement operator, as the physics is not explicitly encoded in the generation process.

Generative Plug-and-Play. A more physically rigorous approach integrates generative models as priors within iterative schemes. Diffusion models, or score-based generative models, have emerged as powerful data-driven priors. By learning the score function $\nabla_x \log p(x)$ of the signal class, these models can be combined with the likelihood score $\nabla_x \log p(y|x)$ to sample from the approximate posterior [12, 14] following Langevin dynamics. This "Generative PnP" approach has been successfully applied to radio interferometry [15] and mass-mapping [36]. These methods offer high reconstruction quality and can handle complex, non-linear measurement operators. However, they inherit the computational cost of iterative schemes, often requiring thousands of likelihood gradient evaluations to generate a single sample, which can be prohibitive for computationally demanding problems.

Generative Unrolled. To bridge the gap between the efficiency of direct mapping and the rigor of iterative sampling, recent work has proposed integrating generative models into unrolled architectures. A prime example is the RI-GAN framework [27], which builds a conditional GAN upon a Gradient Unrolled Network (GU-Net) [25, 26] and again adopts the regularization strategy of Bendel et al. [7]. In this architecture, the generator is not a standard U-Net but a physics-informed unrolled network that explicitly incorporates the measurement operator (or its approximation) at multiple scales. This design ensures data consistency by construction while leveraging the

adversarial training to capture complex signal priors. The result is a method that is both expressive and extremely fast, capable of generating independent posterior samples with only a few evaluations of the forward operator.

3. UNCERTAINTIES: SCALABLE UNCERTAINTY QUANTIFICATION

Having reviewed methods for point estimation, we now turn to the critical challenge of quantifying uncertainty in high-dimensional inverse problems. In scientific applications, a single reconstructed image is often insufficient; rigorous error bars are required to distinguish physical signals from reconstruction artifacts. We review strategies that scale to the exascale regime, ranging from fast heuristic approximations, to approximate posterior sampling, to methods that exploit symmetry.

3.1 Learned Summary Statistics

A computationally efficient approach to uncertainty quantification involves training deep neural networks to output pixel-wise summary statistics directly. This strategy treats uncertainty estimation as a supervised learning problem, employing distinct loss functions to capture different heuristic notions of uncertainty.

One common method is to regress the magnitude of the residual, where the network learns to predict the absolute error $|y - \hat{y}|$ at each pixel [3]. This provides a direct estimate of the reconstruction error but typically yields symmetric intervals that may not capture complex, asymmetric posterior distributions . Alternatively, one can adopt a parametric approach, such as modeling each pixel as a Gaussian distribution [29]. In this case, the network outputs both a mean and a standard deviation, trained by minimizing the negative Gaussian log-likelihood. The problem can also be formulated as a classification task, where the network predicts a softmax distribution over discrete pixel value bins, effectively learning a histogram at each pixel [8]. Another robust approach is pixel-wise quantile regression [23, 22, 37], where the network estimates specific lower and upper quantiles (e.g., for a 90% prediction interval) by minimizing the pinball loss.

While these methods are attractive for their speed, requiring only a single forward pass during inference, they remain inherently heuristic. The predicted standard deviations or quantile intervals are not guaranteed to be statistically valid, particularly when the test data distribution shifts from the training set. Consequently, these learned statistics should be viewed as heuristic signals of uncertainty that require subsequent calibration to provide rigorous coverage guarantees.

3.2 Exploiting Convexity

Before the advent of deep learning, significant progress was made in uncertainty quantification by exploiting the mathematical properties of convex optimization with analytical priors. While full posterior sampling via MCMC provides a gold standard, as discussed it is often computationally prohibitive for high-dimensional imaging [9]. An alternative strategy, established by Pereyra [31], leverages the geometric properties of high-dimensional probability distributions. Pereyra [31] showed that for log-concave posterior distributions, typically resulting from convex priors such as ℓ_1 sparsity, the probability mass concentrates heavily around the mode. Consequently, the maximum a posteriori (MAP) estimate acts as an accurate surrogate for the posterior mean, and rigorous highest posterior density (HPD) credible regions can be approximated directly from the MAP solution using theoretical concentration bounds. Cai et al. [10] leveraged this result to also construct local credible intervals and successfully applied the framework to radio interferometry using sparse wavelet priors, demonstrating that rigorous error bars and hypothesis tests could be computed orders of magnitude faster than by MCMC.

To extend this rigorous and efficient uncertainty quantification framework to the data-driven AI setting presents challenges. Standard deep learning priors (e.g., denoisers or GANs) generally define non-convex potentials, breaking the log-concavity guarantee required for the concentration of measure results to hold. To bridge this gap, one requires a learned prior that is both expressive, convex and also exhibits an explicit potential. The QuantifAI approach [24] extends this framework to the data-driven AI setting by parameterizing the regularization potential with an input convex neural network (ICNN) [17]. By ensuring the learned prior is convex, and assuming a log-concave likelihood (e.g., Gaussian noise with a linear measurement operator), the resulting posterior is guaranteed to be log-concave. This allows the concentration of measure theory from Pereyra [31] to be applied to a deep learning model.

In this framework, uncertainty is quantified through hypothesis testing, where one can statistically test whether

specific structures in the reconstructed image are significant or consistent with noise, and local credible intervals. One can compute rigorous Bayesian error bars for individual pixels (or superpixels) without sampling. By defining the local interval as the range of values a pixel can take while the full image remains within the global HPD credible region, the problem is recast as a constrained optimization task [10, 24]. This enables the computation of pixel-wise uncertainties at multiple scales with orders of magnitude fewer likelihood evaluations than MCMC sampling.

The primary advantages of this method are its scalability and theoretical guarantees. It provides theoretically grounded Bayesian uncertainties with the speed of convex optimization, while also guaranteeing convergence of the underlying optimization algorithms. However, the requirement of convexity inevitably limits the expressivity of the prior compared to non-convex generative models. Moreover, the approach is restricted to forms of uncertainty quantification that can be derived from HPD regions, precluding access to the full posterior distribution. This represents a clear trade-off between theoretical rigour on the one hand, and representational power and flexibility on the other.

3.3 Generative Posterior Sampling

While QuantifAI provides a rigorous and scalable path to uncertainty quantification by exploiting convexity, it inevitably restricts the expressivity of the prior and limits the forms of uncertainty that can be quantified. To capture more complex, non-convex signal structures, and provide more flexible uncertainty quantification, we must turn to generative models. The goal here shifts from finding a single best reconstruction estimate that is supplemented with various quantified uncertainties, to instead generating samples from the full posterior distribution $p(x|y)$. This allows for the exploration of multimodal distributions and the characterization of complex uncertainties that simple summary statistics cannot capture.

Conditional Generative Adversarial Networks. A powerful approach to achieve fast posterior sampling is to train a conditional GAN, as discussed above. In this framework, the generator learns a mapping from the observed data y (or a proxy like the dirty image) and a latent noise vector z to a sample $\hat{x} \sim p(x|y)$. By sampling different z, one can rapidly generate multiple independent realizations of the signal that capture the learned posterior distribution.

This strategy can be deployed in two primary ways, as discussed briefly above. First, as a *stochastic post-processor*, where the model refines a preliminary reconstruction. For instance, Whitney et al. [46] apply this to weak gravitational lensing mass-mapping (MM-GAN), demonstrating the ability to recover non-Gaussian statistics effectively and extremely computationally efficiently (an approximate posterior sample can be generated in less than a second [46], whereas an alternative diffusion approach requires of order 10 minutes to generate a single posterior sample [36]). Second, and more robustly, the generative model can be integrated into an *unrolled physics-informed architecture*. The RI-GAN framework of Mars et al. [27] exemplifies this by building a conditional GAN on top of a Gradient U-Net (GU-Net). Here, the generator explicitly incorporates the measurement operator at multiple scales, ensuring that the generated samples are not only realistic but also consistent with the observed data.

A historic challenge with GANs is mode collapse, where the generator ignores the latent code and produces deterministic outputs. To address this, Bendel et al. [7] propose a regularized conditional GAN that includes specific penalties to enforce diversity. Crucially, they provide theoretical guarantees that their formulation recovers the correct first and second moments (mean and variance) of the posterior distribution in the idealised Gaussian setting, offering a degree of statistical rigour often missing in adversarial methods. This regularization approach is integrated in both the MM-GAN [46] and RI-GAN [27] frameworks discussed above.

Diffusion Models. Diffusion models, or score-based generative models, represent the current state-of-the-art in terms of generative image fidelity. By learning the score function of the data distribution, they allow for sampling from the posterior via Langevin dynamics [12, 36, 15, 14]. While these methods produce samples of exceptional fidelity and can handle complex non-linear inverse problems, they are computationally expensive, typically requiring hundreds to thousands of iterations for a single sample. This makes them less suitable for the real-time or highly computationally demanding processing required by exascale experiments compared to the single-pass efficiency of GANs. Furthermore, rigorous posterior sampling faces the challenge of an intractable likelihood [12, 14]. The reverse diffusion process requires the score of the likelihood $\nabla_{x_t} \log p(y|x_t)$ at each noise level t, but the measurement model is defined only for the clean signal x_0. Evaluating $p(y|x_t)$ formally requires marginalising over all possible clean images x_0, which is computationally infeasible. To circumvent this, methods typically employ approximations that replace the integral with a likelihood evaluated at a denoised estimate $\hat{x}_0(x_t)$, introducing a trade-off between theoretical exactness and tractability [12, 14].

3.4 Exploiting Symmetry

A fundamentally different approach to uncertainty quantification is proposed by Tachella and Pereyra [42], which avoids the need for explicit Bayesian modeling or posterior sampling entirely. Instead, this method, termed the *equivariant bootstrap*, leverages the inherent symmetries of the signal class to construct a frequentist confidence region. It is a "method agnostic" wrapper that can be applied to any reconstruction algorithm, whether a simple unregularized inverse or a complex deep learning model, to provide rigorous high-dimensional error bars.

The core concept relies on group invariance. In many imaging problems, the set of plausible signals is invariant under a group of transformations G, such as rotations or translations (e.g., a rotated image of the sky is still a valid image of the sky). However, the measurement operator Φ is typically *not* equivariant with respect to these transformations; for instance, a radio interferometer observes different spatial frequencies as the sky rotates relative to the baseline distribution. The equivariant bootstrap exploits this property to probe the nullspace of the operator.

The procedure generates bootstrap samples by transforming the data and estimates. Specifically, for a given reconstruction $\hat{x}(y)$, one draws a random transformation $g \in G$ and generates a synthetic measurement $\tilde{y}_g = \Phi T_g \hat{x}(y) + n$, where T_g is the operator corresponding to the transformation g and n is a new noise realisation. The reconstruction method is then applied to this synthetic data to obtain $\hat{x}(\tilde{y}_g)$, which is subsequently inversely transformed: $\tilde{x} = T_{g^{-1}} \hat{x}(\tilde{y}_g)$. The variation in these bootstrapped samples $\{\tilde{x}\}$ provides a proxy for the estimation error. For example, Cherif et al. [11] have effectively applied the equivariant bootstrap in a simplified radio interferometric imaging setting, demonstrating its ability to quantify uncertainty in this challenging setting.

Standard parametric bootstrapping typically underestimates uncertainty in ill-posed problems because the estimator $\hat{x}(y)$ is often biased towards the subspace where Φ is well-conditioned, failing to explore the nullspace. By introducing the group transformation, the equivariant bootstrap effectively "rotates" the problem, forcing the measurement operator to sample different components of the signal, thereby mitigating this bias.

The primary benefit of this approach is its ability to produce accurate, high-dimensional confidence regions without the computational burden of MCMC or the training complexity of generative models. It is also applicable in unsupervised settings where ground truth data is unavailable. However, the method relies on the existence of a known symmetry group for the signal class, and its effectiveness depends on the interplay between these symmetries and the measurement operator; ideally, the operator should not be equivariant to the group transformations to maximize nullspace exploration.

4. TRUSTWORTHINESS: COVERAGE TESTING AND CALIBRATION

As we transition to using deep data-driven priors in high-stakes scientific applications, a critical question arises: can we trust the "black box"? While the methods discussed in Section 3 offer mechanisms to quantify uncertainty, ranging from heuristic statistics to full approximate posterior sampling, the mere production of a probability distribution does not guarantee its validity. Many techniques rely on heuristics, such as learned summary statistics, or fail to provide comprehensive theoretical guarantees. For example, regularized conditional GANs offer assurances primarily within idealized Gaussian settings, while diffusion posterior sampling contends with intractable likelihoods. Conversely, approaches that do secure rigorous guarantees, like the QuantifAI framework, typically require restricting the flexibility of the prior or the forms of uncertainty quantification available. In order to adopt more the expressive and flexible uncertainty quantification approaches discussed, we need a mechanism to assess the trustworthiness of our models and to ensure that our uncertainty estimates are reliable. This brings us to the crucial final stage of our proposed framework: rigorous coverage testing and calibration.

4.1 The Reliability Gap

There exists a fundamental tension between the Bayesian interpretation of probability as a measure of subjective belief and the frequentist requirement for long-run frequency guarantees. In an ideal world where the model perfectly matches reality (the "M-complete" setting), Bayesian credible regions naturally possess frequentist coverage. However, in practice, our models, especially deep generative priors, are approximations.

Recent empirical studies have highlighted a significant "reliability gap" in modern Bayesian imaging. Thong et al. [44] conducted an extensive evaluation of state-of-the-art Bayesian methods, assessing whether the reported probabilities are meaningful under replication. Their findings are striking: methods that achieve the highest reconstruction fidelity, such as diffusion models, can be dangerously overconfident. For instance, credible regions that

claimed to capture 99% of the probability mass were found to contain the ground truth in less than 2% of trials [44]. Conversely, simpler methods like empirical Bayes with total variation priors, while producing lower fidelity images, were found to be conservative, often over-estimating uncertainty. This reveals a disconnect between *estimation accuracy* (e.g., PSNR) and *uncertainty quantification reliability*, underscoring the danger of assuming that better images imply better error bars.

Consequently, to deploy these powerful methods in practice, we must adopt a two-stage trustworthiness protocol: first, we must rigorously *test* the coverage probabilities of our method to diagnose any reliability gaps; second, we must *calibrate* the uncertainties to ensure they deliver the advertised frequentist coverage.

4.2 Coverage Testing

To bridge the reliability gap, we must rigorously test whether our Bayesian credible regions are statistically valid. Coverage testing evaluates this by checking the empirical frequency with which the ground truth lies within the predicted credible regions over a large set of test examples.

Marginal Coverage. Marginal coverage assesses reliability at the level of individual pixels or parameters. For an image reconstruction task, this involves checking whether the true value of a pixel falls within its predicted $1-\alpha$ credible interval $(1-\alpha)\%$ of the time, averaged over all pixels and all test images. Angelopoulos et al. [3] formalize this for image-to-image regression, advocating for methods that control the risk (e.g., the expected fraction of uncovered pixels) at a user-specified level (as discussed further in the next subsection). This approach provides fine-grained, interpretable error bars for local features and is often computationally straightforward to evaluate. However, by focusing on individual pixels, it may not explicitly capture the joint statistics of the full posterior, meaning that while each pixel is well-calibrated in isolation, the coherent spatial structures in the sampled images might not reflect the true global uncertainty.

Global Coverage. Global coverage testing, conversely, assesses whether the entire ground truth image lies within a high-dimensional credible region C_α (such as the HPD region) with probability $1-\alpha$. Thong et al. [44] demonstrate that this provides a holistic test of the posterior geometry, sensitive to correlations that marginal checks might miss. While this offers a rigorous check of the full high-dimensional distribution, global metrics can be difficult to interpret physically—a failure in global coverage does not necessarily pinpoint which features are unreliable. Furthermore, passing a marginal coverage test does not guarantee global coverage, and vice versa; thus, these two perspectives offer complementary, rather than competing, views on trustworthiness.

4.3 Calibration with Conformal Prediction

Once coverage testing has diagnosed the reliability gap, the final step is to fix it. This is the domain of *calibration*. While we cannot easily force a deep generative model to learn the perfect posterior, we can apply post-hoc corrections to its uncertainty estimates to ensure they satisfy frequentist guarantees.

Conformal prediction [2, 4] offers a powerful, distribution-free framework for this task. The core idea is to use a held-out calibration dataset to compute a scalar correction factor λ that adjusts the size of the predicted credible regions. For example, in the context of image-to-image regression, Angelopoulos et al. [3] introduce *Risk-Controlling Prediction Sets* (RCPS) [6]. This method allows a user to specify a tolerable error rate δ (e.g., ensuring that no more than 10% of pixels are incorrectly excluded from the credible intervals). The algorithm then uses the calibration data to find the smallest λ such that the risk is controlled at level α with high probability (e.g., $1-\delta$).

This approach transforms "heuristic" uncertainty maps into "rigorous" statistical bounds. A deep network might output a point estimate μ and a heuristic standard deviation σ that is uncalibrated. Conformal prediction allows us to wrap this output in a rigorous interval $[\mu - \lambda\sigma, \mu + \lambda\sigma]$ that is guaranteed to contain the ground truth with the specified probability, regardless of the distribution of the data or the architecture of the network. This provides a safety layer for scientific applications of AI, ensuring that even if the model is imperfect, its error bars are trustworthy. Crucially, while the *validity* of these guarantees holds regardless of the accuracy of the initial heuristic uncertainty, the *efficiency* of the resulting prediction sets does not. If the initial estimates are poor, conformal calibration will simply inflate the intervals to be large enough to satisfy the coverage requirement, resulting in uninformative error bars. Thus, accurate initial uncertainty estimation remains vital for producing tight, adaptive, and scientifically useful constraints. However, this guarantee comes with a caveat: it requires an exchangeable calibration dataset that is representative of the test distribution. If the distribution shifts (e.g., observing a new type of galaxy not seen during calibration), the guarantees may no longer hold, highlighting the continued need for

robust out-of-distribution detection.

5. SUMMARY & OUTLOOK

We have reviewed the landscape of high-dimensional uncertainty quantification, tracing the evolution from black-box deep learning to rigorous, physics-informed frameworks. The transition from point estimation to trustworthy UQ is essential for the scientific utility of AI, particularly in computationally demanding settings where MCMC sampling is infeasible due to the high-dimensional parameter spaces of large images, massive data volumes, and computationally costly measurement operators.

Contemporary methods generally face a trilemma between computational efficiency, reconstruction fidelity, and statistical rigour. While convex approaches like QuantifAI offer rigour and speed, they limit expressivity. Conversely, generative sampling, such as diffusion models, offers high fidelity but often at a prohibitive computational cost. We argue that the optimal strategy for highly computationally demanding problems lies in the intersection of these fields. By combining physics-informed unrolled architectures, with generative models and conformal prediction, we can satisfy the four key criteria identified in the introduction: (i) computational efficiency is achieved through unrolled optimization; (ii) the approach is physics-informed by design; (iii) expressive generative data-driven AI priors are leveraged to capture complex signal structure; and (iv) rigorous quantified uncertainties are provided via conformal calibration.

Looking forward, three key challenges remain. First, conformal guarantees rely on exchangeability, so developing methods that remain robust or fail gracefully when the test data drifts from the calibration distribution is critical, for example when discovering new physical phenomena. Second, the next generation of priors may move beyond specific datasets to large-scale foundation models. Adapting these universal priors to specific physical measurement operators while maintaining calibration will be a key frontier. Finally, calibrated images are often intermediate products. The field must move towards propagating these rigorous pixel-level uncertainties into reliable constraints on high-level scientific parameters.

By combining the expressivity of AI, the robustness of physics, and the rigour of conformal calibration, we can enable a new era of trustworthy scientific discovery for highly computationally demanding inverse problems.

ACKNOWLEDGEMENTS

This work is support by STFC (grant number ST/W001136/1). We acknowledge the use of large language models (LLMs) in the preparation of this manuscript; we take full responsibility for the accuracy of the content.

References

[1] Jonas Adler and Ozan Öktem. Deep bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[3] Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, pages 717–730, 2022.

[4] Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[6] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

[7] Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional gan for posterior sampling in image recovery problems. In *Advances in Neural Information Processing Systems*, volume 36, pages 68673–68684, 2023.

[8] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[9] Xiaohao Cai, Marcelo Pereyra, and Jason D. McEwen. Uncertainty quantification for radio interferometric imaging – i. proximal mcmc methods. *Monthly Notices of the Royal Astronomical Society*, 480(3):4154–4169, 2018.

[10] Xiaohao Cai, Marcelo Pereyra, and Jason D. McEwen. Uncertainty quantification for radio interferometric imaging: Ii. map estimation. *Monthly Notices of the Royal Astronomical Society*, 480(3):4170–4182, 2018.

[11] Mostafa Cherif, Tobías I Liaudat, Jonathan Kern, Christophe Kervazo, and Jérôme Bobin. Uncertainty quantification for fast reconstruction methods using augmented equivariant bootstrap: Application to radio interferometry. *arXiv preprint arXiv:2410.23178*, 2024.

[12] Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2023.

[13] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[14] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

[15] Noe Dia, M. J. Yantovski-Barth, Alexandre Adam, Micah Bowles, Laurence Perreault-Levasseur, Yashar Hezaveh, and Anna Scaife. Iris: A bayesian approach for image reconstruction in radio interferometry with expressive score-based priors. *arXiv preprint arXiv:2501.02473*, 2025.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[17] Alexis Goujon, Sebastian Neumayer, Pakshal Bohra, Stanislas Ducotterd, and Michael Unser. A neural-network-based convex regularizer for inverse problems. *IEEE Transactions on Computational Imaging*, 9: 781–795, 2023. doi: 10.1109/TCI.2023.3306100.

[18] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406, 2010.

[19] Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[21] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[22] Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

[23] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913643`.

[24] Tobías I. Liaudat, Matthijs Mars, Matthew A. Price, Marcelo Pereyra, Marta M. Betcke, and Jason D. McEwen. Scalable bayesian uncertainty quantification with data-driven priors for radio interferometric imaging. *RASTI*, 3:505–534, 2024.

[25] Matthijs Mars, Marta M Betcke, and Jason D McEwen. Learned interferometric imaging for the spider instrument. *RAS Techniques and Instruments*, 2(1):760–778, 2023.

[26] Matthijs Mars, Marta M. Betcke, and Jason D. McEwen. Learned radio interferometric imaging for varying visibility coverage. *RASTI*, 4:1–13, 2025.

[27] Matthijs Mars, Tobías I. Liaudat, Jessica J. Whitney, Marta M. Betcke, and Jason D. McEwen. Generative imaging for radio interferometry with fast uncertainty quantification. *arXiv preprint arXiv:2507.21270*, 2025.

[28] Jason D. McEwen, Tobías I. Liaudat, Matthew A. Price, Xiaohao Cai, and Marcelo Pereyra. Proximal nested sampling with data-driven priors for physical scientists. *Physical Sciences Forum*, 9(1), 2023. ISSN 2673-9984. doi: 10.3390/psf2023009013. URL https://www.mdpi.com/2673-9984/9/1/13.

[29] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. doi: 10.1109/ICNN.1994.374138.

[30] Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

[31] Marcelo Pereyra. Maximum-a-posteriori estimation with bayesian confidence regions. *SIAM Journal on Imaging Sciences*, 10(1):285–302, 2017.

[32] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021. doi: 10.1137/20M1387961. URL https://doi.org/10.1137/20M1387961.

[33] Luke Pratley and Jason D McEwen. Sparse image reconstruction for the spider optical interferometric telescope. *arXiv preprint arXiv:1903.05638*, 2019.

[34] Luke Pratley, Melanie Johnston-Hollitt, and Jason D McEwen. A fast and exact w-stacking and w-projection hybrid algorithm for wide-field interferometric imaging. *The Astrophysical Journal*, 874(2):174, 2019.

[35] Luke Pratley, Melanie Johnston-Hollitt, and Jason D McEwen. w-stacking w-projection hybrid algorithm for wide-field interferometric imaging: implementation details and improvements. *Publications of the Astronomical Society of Australia*, 37:e041, 2020.

[36] B. Remy, F. Lanusse, N. Jeffrey, J. Liu, J.-L. Starck, K. Osato, and T. Schrabback. Probabilistic mass-mapping with neural score estimation. *Astronomy & Astrophysics*, 672:A51, 2023.

[37] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[39] Ernest K. Ryu, Jialin Liu, Shuxiao Wang, Xili Chen, Zhuotao Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557, 2019.

[40] A. M. M. Scaife. Big telescope, big data: towards exascale with the square kilometre array. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2166): 20190060, 01 2020. ISSN 1364-503X. doi: 10.1098/rsta.2019.0060. URL https://doi.org/10.1098/rsta.2019.0060.

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[42] Julián Tachella and Marcelo Pereyra. Equivariant bootstrapping for uncertainty quantification in imaging inverse problems. *arXiv preprint arXiv:2310.11838*, 2023.

[43] Matthieu Terris, Chao Tang, Adrian Jackson, and Yves Wiaux. The airi plug-and-play algorithm for image reconstruction in radio-interferometry: variations and robustness. *Monthly Notices of the Royal Astronomical Society*, 537(2):1608–1619, 01 2025. ISSN 0035-8711. doi: 10.1093/mnras/staf022. URL https://doi.org/10.1093/mnras/staf022.

[44] David Y. W. Thong, Charlesquin Kemajou Mbakam, and Marcelo Pereyra. Do bayesian imaging methods report trustworthy probabilities? *arXiv preprint arXiv:2405.08179*, 2024.

[45] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948, 2013.

[46] Jessica J Whitney, Tobías I Liaudat, Matthew A Price, Matthijs Mars, and Jason D McEwen. Generative modelling for mass-mapping with fast uncertainty quantification. *Monthly Notices of the Royal Astronomical Society*, 542(3):2464–2479, 2025.