WHERE PRIORS LEAVE FINGERPRINTS: IDENTIFYING LEARNED REGULARIZERS FROM RECONSTRUCTION RESPONSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Many learned priors for inverse problems reconstruct equally well, so the reconstruction alone reveals little about which prior produced it. Model fingerprints read the outputs of a model, and the classical analysis of how a solver reacts to its data assumes a known regularizer. In this paper, we ask what the behavior of a solver reveals about its learned prior beyond the reconstruction, and when it suffices to distinguish two priors. We study the *response* of a solver, namely how its reconstruction changes when the measurement is randomly perturbed. For solvers that combine a least-squares data fit with a learned regularizer, we prove that, when the operator and the perturbation level are known, the covariance of random re-solves determines the Jacobian of the solver in closed form, which identifies the curvature of the regularizer on the subspace that the solver can reach. Guided by this result, we introduce the response fingerprint, built from the centered energies of the response in frequency bands, which applies to any differentiable solver. Across denoising, deblurring, inpainting, super-resolution, MRI and CT, it separates learned priors that are hard to tell apart from their outputs, and it transfers without retraining to external priors, which played no role in its design. It also provides a label-free reliability indicator that flags the priors it separates poorly.

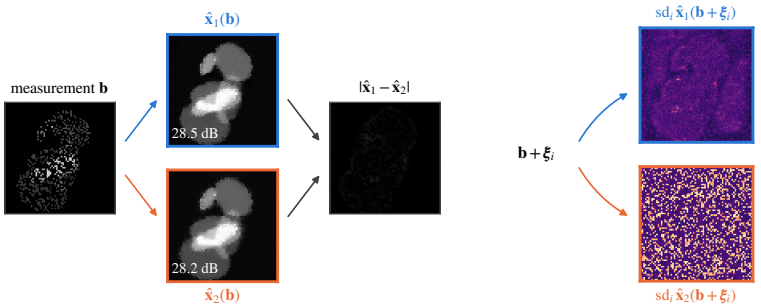
1 INTRODUCTION

In an inverse problem, an image is recovered from indirect and incomplete measurements, such as the projections acquired in computed tomography (CT) or the undersampled Fourier coefficients acquired in magnetic resonance imaging (MRI). Because the measurements do not determine the image uniquely, the solver relies on a prior that fills in what they leave undetermined (Engl et al., 1996; Arridge et al., 2019). Once a hand-crafted penalty, such as Tikhonov or total variation (Tikhonov & Arsenin, 1977; Rudin et al., 1992), this prior is increasingly learned from data (Arridge et al., 2019; Habring & Holler, 2024): it can be an explicit learned regularizer (Roth & Black, 2009; Lunz et al., 2018; Goujon et al., 2023; 2024; Fang et al., 2024; Eliasof et al., 2024), a *plug-and-play* denoiser within an iterative solver (Venkatakrishnan et al., 2013; Romano et al., 2017; Hurault et al., 2022), or an *unrolled network* that alternates learned steps with data-consistency steps (Adler & Öktem, 2018; Aggarwal et al., 2019; Monga et al., 2021). As these families mature, many of their members reconstruct equally well. For example, the two learned regularizers in Figure 1(a) reconstruct within a fraction of a decibel in peak signal-to-noise ratio (PSNR), and the learned CT regularizers of a recent comparative study (Hertrich et al., 2026) lie within two decibels, as detailed in Section E.5.

Similar reconstructions can come from different priors. The prior decides what the measurements leave open, but it leaves little trace in the reconstruction, which is what is usually inspected. Identifying the prior from the behavior of the solver, namely determining which of several known *candidate priors* a solver uses, is a task we call *attribution*. Attribution supports the provenance and comparison of learned solvers, for instance in medical and scientific imaging, where the same measurement can be reconstructed with many different priors. This leads to the question that we study:

What does the behavior of a solver reveal about its learned prior beyond the reconstruction, and when is it enough to distinguish two priors?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) the reconstruction hides the prior (b) the response reveals it

Figure 1: **The reconstruction hides the prior, while the response reveals it. (a)** An adversarial regularizer (blue) and Fields of Experts (orange) reconstruct $\hat{x}_1(b)$ and $\hat{x}_2(b)$ of one partially observed phantom within 0.3 dB PSNR, and their difference, on the same intensity scale, is small. **(b)** The per-pixel standard deviation of re-solves at $b + \xi_i$, for random perturbations ξ_i of the measurement, differs visibly. The protocol is in Section C.

Telling trained models apart from their behavior has been studied mainly outside inverse problems and for a different task, namely detecting whether a model was reused or copied. Model fingerprinting methods such as ModelDiff (Li et al., 2021), SAC (Guan et al., 2022) and ProbeSpec (Lou et al., 2026) compare the predictions of two classifiers on the same inputs. For image restoration, networks have been fingerprinted from artifacts in their outputs (Vonderfecht & Liu, 2022) or from a *critical image*, an input optimized through the network (Quan et al., 2023). Within inverse problems, the *response* of a solver, namely how its reconstruction changes when the measurement changes, is a classical object: for a known regularizer, it gives the variance and spatial resolution of the reconstruction (Fessler, 1996; Fessler & Rogers, 1996; Tenorio, 2001). These works offer valuable insight into distinguishing trained models and into how a known solver reacts to its data. Nonetheless, fingerprints read the outputs of a model, and priors that reconstruct alike produce similar outputs, as we show in Section 5.2, while the classical analysis of the response assumes that the regularizer is known. Neither identifies an unknown prior from the behavior of its solver, which is the gap that we address.

Our approach. We propose to identify a learned prior from the *response* of its solver rather than from the reconstruction itself. Intuitively, solvers with different priors are tied to the same measurements, while the prior determines how a change of the measurements spreads into the part of the image that they leave undetermined. Two priors that agree on the reconstruction can therefore respond differently, as illustrated in Figure 1(b). Measuring the response means running the solver on randomly perturbed measurements, with no access to the weights, gradients, or objective. Because the prior shapes the response at the frequencies that the measurement attenuates or leaves undetermined, we summarize it by its energy at different spatial frequencies, which yields a compact *response fingerprint*, estimated from a few re-solves and applicable to any differentiable solver, as illustrated in Figure 5. Attribution then compares it with reference fingerprints of the candidate priors and requires no training. We then show that, for *variational solvers*, which combine a least-squares data fit with a regularizer, the covariance of these re-solves identifies the curvature of the regularizer in every direction in which the reconstruction can move, given the operator and the perturbation level. In other words, *the reconstruction hides the prior, but the response reveals it on the subspace that the solver can reach*.

Main contributions. Our contributions are as follows:

- We introduce the response fingerprint, built from the response energy in frequency bands, which is estimated from a few re-solves and has a label-free reliability indicator, as developed in Section 3.
- We prove that, for variational solvers, the covariance of random re-solves determines in closed form the Jacobian, namely the derivative of the reconstruction with respect to the measurement, given the operator and the perturbation level, and with it the curvature of the regularizer on the subspace that the solver can reach, as shown in Section 4.
- We show on nine inverse problems that our fingerprint outperforms those reading the outputs, transfers without retraining to external priors that played no role in its design, and is robust to unknown acquisition conditions, as reported in Section 5.

2 Related Work

Response and sensitivity of variational solvers. Implicit differentiation gives the mean, covariance, and local impulse response of variational solvers (Fessler, 1996; Fessler & Rogers, 1996; Qi & Leahy, 2000). Statistical regularization theory (Tenorio, 2001; Kaipio & Somersalo, 2005), Stein identities (Stein, 1981; Ramani et al., 2008; Eldar, 2009; Tibshirani & Taylor, 2012; Tachella et al., 2025) and uncertainty analyses (Antorán et al., 2023; Dölz & Ebert, 2026) use them to estimate bias, risk and uncertainty, and robustness analyses bound how learned solvers react to perturbations (Antun et al., 2020; Genzel et al., 2023; Gottschling et al., 2025). These works evaluate a given solver, while we invert the response of an unknown solver to identify its prior.

Data and prior contributions to a reconstruction. Filter factors and source conditions describe how a penalty acts on the directions that the forward operator attenuates (Engl et al., 1996; Hansen, 1998; Benning & Burger, 2018), and null-space networks separate the part of a reconstruction fixed by the data from the part supplied by the prior (Schwab et al., 2019; Boink et al., 2023). We make this split for the prior itself: the response determines the filter factors of a stationary prior, given the operator and the perturbation level, and our fingerprint summarizes them in ring energies. We discuss work on identifying an objective from its solver and on model fingerprinting in Section H.

3 The Response Fingerprint

We identify which of K known candidate priors a solver uses, with access only to the ability to run the solver, in three steps, which we discuss in the subsections below. First, we perturb the measurement with small random noise and solve again, which yields a cloud of re-solves whose covariance is the response of the solver, as defined in Section 3.1. Second, we reduce the response to its energies in frequency bands, illustrated in Figure 2, which are estimable from a few re-solves, as we show in Section 3.2, and whose centered logarithms form our fingerprint, as described in Section 3.3. Third, we attribute the fingerprint to the nearest candidate prior and assess the reliability of this decision without labels, as described in Section 3.4. The entire process is summarized in Algorithm 1.

3.1 Random Re-solves Measure the Response

Notations. Standard notation, including norms, projectors and asymptotic symbols, is collected in Section A. We consider recovering an image $\boldsymbol{x}^\star \in \mathbb{R}^n$ with n pixels from a measurement

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\epsilon} \in \mathbb{R}^m, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the forward operator, of rank r, and $\boldsymbol{\epsilon}$ is measurement noise. When $r < n$, $\boldsymbol{A}$ has a nontrivial null space $\mathrm{Null}(\boldsymbol{A})$, and $\boldsymbol{b}$ does not determine $\boldsymbol{x}^\star$ uniquely. A classical way to resolve this ambiguity is the variational formulation (Tikhonov & Arsenin, 1977; Engl et al., 1996)

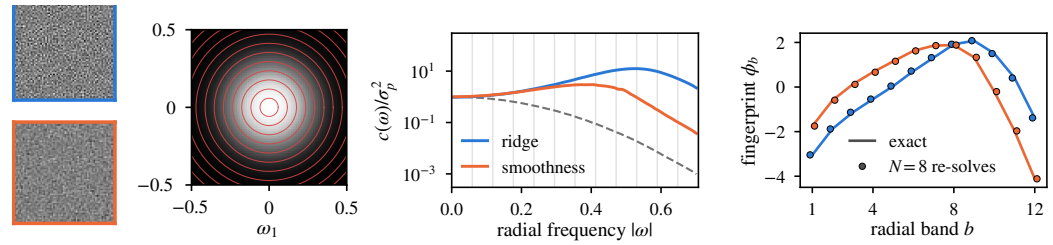
$$\hat{\boldsymbol{x}}(\boldsymbol{b}) \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \ \tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda R(\boldsymbol{x}), \tag{2}$$

where the first term measures the fit to the measurement, $R : \mathbb{R}^n \to \mathbb{R}$ is the *regularizer*, which encodes the prior, and $\lambda > 0$ is the *regularization parameter*, which sets its strength. More generally, a solver maps $\boldsymbol{b}$ to a reconstruction $\hat{\boldsymbol{x}}(\boldsymbol{b}) \in \mathbb{R}^n$, and its prior is the part of this map that is learned from data, which decides how the solver fills in what the measurement leaves undetermined. For a solver of the form of Equation (2), the prior is the learned regularizer R.

Random re-solves. At a fixed measurement $\boldsymbol{b}_0 \in \mathbb{R}^m$ with reconstruction $\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}(\boldsymbol{b}_0)$, we run the solver on randomly perturbed measurements $\boldsymbol{b}_0 + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_p^2 \boldsymbol{I}_m)$, where $\boldsymbol{I}_m$ is the $m \times m$ identity and $\sigma_p > 0$ is the noise level of the perturbations. We call these reconstructions *random re-solves*, and the N re-solves of one measurement a *cloud*. For a Lipschitz solver that is differentiable at $\boldsymbol{b}_0$, with Jacobian $\boldsymbol{J} = \partial\hat{\boldsymbol{x}}/\partial\boldsymbol{b} \in \mathbb{R}^{n\times m}$ there, a first-order expansion yields $\hat{\boldsymbol{x}}(\boldsymbol{b}_0 + \boldsymbol{\xi}) = \hat{\boldsymbol{x}}_0 + \boldsymbol{J}\boldsymbol{\xi} + o(\|\boldsymbol{\xi}\|_2)$, so the covariance $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ of the re-solves is

$$\boldsymbol{C} = \mathrm{Cov}[\hat{\boldsymbol{x}}(\boldsymbol{b}_0 + \boldsymbol{\xi})] = \sigma_p^2 \boldsymbol{J}\boldsymbol{J}^\top + o(\sigma_p^2), \tag{3}$$

where the leading term is the first-order covariance of Fessler (1996), and the remainder is $O(\sigma_p^4)$ for a C^3 solver with bounded derivatives, because Gaussian perturbations have zero third moments. We



(a) re-solves (b) radial bands (c) response spectrum (d) fingerprint

Figure 2: **Band energies read the prior where the operator attenuates the measurement.** Two stationary priors of Example 1, ridge ($h = 1$, blue) and smoothness (h the symbol of the negative Laplacian, orange), under a periodic Gaussian blur, in closed form. **(a)** One re-solve deviation of each prior. **(b)** The $B = 12$ radial bands of Equation (7), shaded by $|a(\omega)|^2$, bright where the blur passes a frequency. **(c)** Response spectra versus radial frequency, with $|a(\omega)|^2$ dashed. **(d)** Fingerprints, exact and from $N = 8$ re-solves, as the mean over 200 clouds; the error bars, the standard deviation of a single-cloud fingerprint over the clouds, are smaller than the markers. The protocol is in Section C.

call $\boldsymbol{C}$ the *response* of the solver, and in what follows we identify it with its first-order term $\sigma_p^2 \boldsymbol{J}\boldsymbol{J}^\top$. Each re-solve moves the reconstruction by approximately $\boldsymbol{J}\boldsymbol{\xi}$, a random combination of the columns of $\boldsymbol{J}$, so two solvers with the same reconstruction produce clouds of different shapes whenever their products $\boldsymbol{J}\boldsymbol{J}^\top$ differ, as illustrated in Figure 1(b). For variational solvers, this happens whenever their Jacobians differ, as we show in Section 4. Clouds arise from repeated acquisitions or Monte Carlo re-noising (Ramani et al., 2008) and require no access to the weights, gradients or objective.

Requirements for a fingerprint. In imaging, the number N of re-solves is much smaller than the number n of pixels, and the sample covariance of N re-solves has rank at most $N - 1$, so it does not estimate $\boldsymbol{C}$ reliably. A fingerprint is therefore a low-dimensional summary of the re-solves that *(R1)* is estimable from $N \ll n$ re-solves, *(R2)* depends on the solver only through its response $\boldsymbol{C}$, and not on its reconstruction, so that Section 4 characterizes what it can read, and *(R3)* is invariant to the scale of $\boldsymbol{C}$, which grows with σ_p^2, because σ_p may be unknown. The energy of the response at each spatial frequency, which is the noise power spectrum of the reconstruction when the perturbations are acquisition noise (Barrett & Myers, 2013), is a natural choice for such a summary.

3.2 Band Energies Are Estimable from a Small Number of Re-solves

To satisfy *(R1)*, we pool the response over subspaces with many response directions. We divide the image space into B *bands*, namely mutually orthogonal subspaces with orthogonal projectors $\boldsymbol{P}_b \in \mathbb{R}^{n \times n}$. The *band energy* of band b is the expected squared norm of the projected response,

$$e_b = \operatorname{tr}(\boldsymbol{P}_b \boldsymbol{C}), \qquad b = 1, \dots, B. \tag{4}$$

We estimate it from the deviations $\boldsymbol{z}_i = \hat{\boldsymbol{x}}(\boldsymbol{b}_0 + \boldsymbol{\xi}_i) - \hat{\boldsymbol{x}}_0$ of N re-solves from the reconstruction,

$$\hat{e}_b = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{P}_b \boldsymbol{z}_i\|_2^2. \tag{5}$$

To first order in σ_p, the deviations are draws from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$. Each $\|\boldsymbol{P}_b \boldsymbol{z}_i\|_2^2$ is then a Gaussian quadratic form (Mathai & Provost, 1992), and the estimate satisfies

$$\mathbb{E}\hat{e}_b = e_b, \qquad \frac{\operatorname{sd}(\hat{e}_b)}{e_b} = \sqrt{\frac{2}{N r_{\mathrm{eff},b}}}, \qquad r_{\mathrm{eff},b} = \frac{e_b^2}{\operatorname{tr}[(\boldsymbol{P}_b \boldsymbol{C})^2]}, \tag{6}$$

as detailed in Proposition 3. Here, sd is the standard deviation over the draws, and $r_{\mathrm{eff},b}$ is the *effective rank* of the band, namely the number of its response directions if these carry equal energy.

Pooling makes the response measurable. The effective rank sets the relative error of $\hat{e}_b$. For example, the energy along one response direction has $r_{\mathrm{eff},b} = 1$, and $N = 8$ re-solves estimate it with a relative error of 50%. In contrast, a band of 100 response directions of equal energy has a relative error of 5%. Pooling makes the response measurable from a few re-solves, which gives *(R1)*.

3.3 Radial Band Energies Form the Fingerprint

The estimates of Section 3.2 hold for any choice of bands. We now choose the bands, define the fingerprint, and interpret what it captures.

Radial bands. We take the bands to be rings of spatial frequency. Let $\boldsymbol{F} \in \mathbb{C}^{n\times n}$ be the unitary 2D discrete Fourier transform, whose frequencies ω, in cycles per pixel, have radial frequency $|\omega| \in [0, \sqrt{2}/2]$. Let $\mathcal{R}_b$ be the set of frequencies whose radial frequency lies in the b-th of B equal-width intervals of $[0, \sqrt{2}/2]$, and let $\mathbf{1}_{\mathcal{R}_b} \in \{0,1\}^n$ be its indicator. The *band projectors* are

$$\boldsymbol{P}_b = \boldsymbol{F}^{\mathsf{H}} \operatorname{diag}(\mathbf{1}_{\mathcal{R}_b})\, \boldsymbol{F}, \qquad b = 1, \dots, B, \tag{7}$$

so that $\boldsymbol{P}_b \boldsymbol{z}$ is the component of $\boldsymbol{z} \in \mathbb{R}^n$ whose spectrum lies in the b-th ring. The intervals are half-open, the last one closed, and each ring is symmetric under $\omega \mapsto -\omega$, so $\boldsymbol{P}_b$ is real. A ring groups frequencies of similar magnitude, which carry comparable energy when the response is approximately isotropic, so each band has a large effective rank.

The fingerprint. Our *response fingerprint* is the vector $\boldsymbol{\phi} \in \mathbb{R}^B$ of centered log band energies,

$$\phi_b = \log \hat{e}_b - \frac{1}{B}\sum_{q=1}^{B} \log \hat{e}_q, \qquad b = 1, \dots, B, \tag{8}$$

where subtracting the mean log energy removes any common scale of the band energies, including σ_p^2 of Equation (3), which gives *(R3)*. We assume every $e_b > 0$, so $\hat{e}_b > 0$ almost surely. Because $\hat{\boldsymbol{x}}_0$ is not observed under repeated acquisitions, we compute $\hat{e}_b$ from the re-solves centered at their mean and divide by $N-1$, for which Equation (6) holds with $N-1$ in place of N, as Proposition 3 shows.

Interpretation of the band energies. To understand what the band energies capture, we first consider a linear solver with a stationary quadratic prior, for which the response is available in closed form.

Example 1 (Band energies of a stationary prior; cf. Hansen, 1998; Engl et al., 1996)**.** Let $\boldsymbol{A}$ be a periodic convolution, and let the solver be Equation (2) with the quadratic regularizer $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}$, where $\boldsymbol{M} \succeq 0$ is a convolution and $\boldsymbol{A}^\top\boldsymbol{A} + \lambda \boldsymbol{M} \succ 0$, so that $\boldsymbol{J} = (\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{M})^{-1}\boldsymbol{A}^\top$. The unitary transform $\boldsymbol{F}$ diagonalizes $\boldsymbol{A}$ and $\boldsymbol{M}$, with symbols $a(\omega)$ and $h(\omega)$, namely their eigenvalues at frequency ω, and by Equation (3), $\boldsymbol{C} = \boldsymbol{F}^{\mathsf{H}}\operatorname{diag}(\boldsymbol{c})\boldsymbol{F}$ with

$$c(\omega) = \sigma_p^2 \frac{|a(\omega)|^2}{\big(|a(\omega)|^2 + \lambda h(\omega)\big)^2}, \qquad e_b = \sum_{\omega \in \mathcal{R}_b} c(\omega). \tag{9}$$

At frequencies that the operator passes, $c(\omega) \approx \sigma_p^2/|a(\omega)|^2$ is the same for every prior, whereas at those that it attenuates, $c(\omega) \approx \sigma_p^2 |a(\omega)|^2/(\lambda h(\omega))^2$ is set by the prior. The band energies therefore capture the prior where the measurement barely constrains the image. As we show in Figure 2 for two such priors under a Gaussian blur, their response spectra coincide at low frequencies and separate at high ones, and $N = 8$ re-solves estimate their fingerprints closely, as Equation (6) states.

3.4 Attribution Requires No Training and Indicates Its Reliability

We attribute a fingerprint to a candidate prior and quantify the reliability of the attribution.

Attribution. For each candidate prior $k = 1, \dots, K$, we run its solver on reference images, compute the fingerprint of each cloud, and average these fingerprints into the *reference centroid* $\boldsymbol{\mu}_k \in \mathbb{R}^B$. A test cloud with fingerprint $\boldsymbol{\phi}$ is then attributed to the candidate prior with the nearest centroid,

$$\hat{k} = \arg\min_{k \in \{1,\dots,K\}} \|\boldsymbol{\phi} - \boldsymbol{\mu}_k\|_2. \tag{10}$$

This rule needs no training, because centroids are fingerprint averages. This distance is also an attribution score for candidate prior k, where smaller values mean greater agreement with $\boldsymbol{\mu}_k$.

A label-free reliability indicator. A nearest-centroid rule has two failure modes. In a *collapse*, the centroids of two candidate priors nearly coincide. In a *dispersion*, the fingerprints of one candidate prior scatter across the gap between its centroid and that of another. To detect both, for candidate prior

k, we let s_k be the mean pairwise distance among its test fingerprints, which measures its scatter, and let k^* be the other candidate prior whose reference centroid is closest on average to these fingerprints. Our *reliability indicator* compares the scatter with the gap between the two centroids,

$$\rho_k = \frac{s_k}{g_k}, \qquad g_k = \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^*}\|_2. \tag{11}$$

A collapse makes g_k small and a dispersion makes s_k large, so both increase ρ_k, and large values, such as $\rho_k > 1$, indicate a candidate prior that the fingerprint separates poorly, with a cut-off that we calibrate per problem, as detailed in Section E.1. Both quantities use fingerprints that we compute by running candidate prior k on held-out images, and the indicator uses no information on whether any attribution is correct, which makes it label-free.

Computational complexity. Beyond the N solves, the fingerprint costs $O(Nn \log n)$ operations, one 2D fast Fourier transform per re-solve, as detailed in Section C, and attribution costs $O(KB)$ operations for the K distances in $\mathbb{R}^B$.

4 The Response Identifies the Curvature of the Regularizer

To understand what the fingerprint reads, we study variational solvers, namely solvers of the form of Equation (2) with a learned regularizer R. This form covers explicit learned regularizers (Roth & Black, 2009; Lunz et al., 2018; Mukherjee et al., 2020; Goujon et al., 2023; 2024), including 7 of the 9 learned priors of Section 5.1, and for a differentiable R, every minimizer satisfies

$$\boldsymbol{A}^\top(\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}) + \lambda \nabla R(\hat{\boldsymbol{x}}) = \boldsymbol{0}, \tag{12}$$

and evaluating Equation (12) for a known candidate prior yields the *optimality certificate*, which we use as a baseline (Keshavarz et al., 2011; Aswani et al., 2018). We write $\boldsymbol{H} := \nabla^2 R(\hat{\boldsymbol{x}}_0) \in \mathbb{R}^{n\times n}$ for the Hessian of R at $\hat{\boldsymbol{x}}_0$ and $\boldsymbol{G} := \boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{H} \in \mathbb{R}^{n\times n}$ for the *local system matrix*. Throughout this section, we make the following assumption.

Assumption A1 (Local regularity)**.** *The regularizer R is C^2 near $\hat{\boldsymbol{x}}_0$, $\boldsymbol{G}$ is positive definite, and $\hat{\boldsymbol{x}}$ is continuous at $\boldsymbol{b}_0$.*

Under Assumption A1, the implicit function theorem (Krantz & Parks, 2002) applied to Equation (12), as for implicitly defined estimators (Fessler, 1996), gives a unique continuously differentiable branch of critical points through $\hat{\boldsymbol{x}}_0$, and continuity makes the solver follow it, which yields

$$\boldsymbol{J} = \boldsymbol{G}^{-1}\boldsymbol{A}^\top, \qquad \boldsymbol{S} := \boldsymbol{A}\boldsymbol{J} = \boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A}^\top, \qquad \mathcal{T} := \mathrm{Range}(\boldsymbol{J}), \tag{13}$$

where $\boldsymbol{S} \in \mathbb{R}^{m\times m}$ is the Jacobian of the predicted measurement $\boldsymbol{A}\hat{\boldsymbol{x}}(\boldsymbol{b})$, and $\mathcal{T} \subseteq \mathbb{R}^n$ is the r-dimensional subspace of directions in which the reconstruction moves when the measurement changes, which we call the *subspace that the solver can reach*. The data fit contributes $\boldsymbol{A}^\top\boldsymbol{A}$ to $\boldsymbol{G}$ and the prior contributes $\lambda\boldsymbol{H}$, so the regularizer affects the response only through its curvature at $\hat{\boldsymbol{x}}_0$. We discuss the part of the response in $\mathrm{Null}(\boldsymbol{A})$ in Section A.1.

Random re-solves identify the curvature on $\mathcal{T}$. We note that the response $\boldsymbol{C} = \sigma_p^2\boldsymbol{J}\boldsymbol{J}^\top$ determines $\boldsymbol{J}$ only up to a rotation, because $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{U}$ yield the same $\boldsymbol{C}$ for every orthogonal $\boldsymbol{U} \in \mathbb{R}^{m\times m}$. For denoising, $\boldsymbol{J}$ is symmetric positive definite, which removes this ambiguity, as we discuss in Section A.1. For a general operator, the Jacobian $\boldsymbol{S}$ of the predicted measurement is symmetric positive semidefinite, and we now show in Theorem 1 that this suffices to recover $\boldsymbol{J}$ from $\boldsymbol{C}$, and with it the curvature of the regularizer on $\mathcal{T}$.

Theorem 1 (The response identifies the curvature on $\mathcal{T}$)**.** *Under Assumption A1, let $\boldsymbol{C} = \sigma_p^2\boldsymbol{J}\boldsymbol{J}^\top$ with σ_p known, and let $\boldsymbol{P}_\mathcal{T}$ be the orthogonal projector onto $\mathcal{T}$. Then*

$$\boldsymbol{S} = \sigma_p^{-1}\big(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top\big)^{1/2}, \qquad \boldsymbol{J} = \sigma_p^{-1}\,\boldsymbol{C}\boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top\big)^{+1/2}, \qquad \lambda\boldsymbol{H}\boldsymbol{J} = \boldsymbol{A}^\top(\boldsymbol{I}_m - \boldsymbol{S}), \tag{14}$$

where $(\cdot)^{1/2}$ is the positive semidefinite square root and $(\cdot)^{+1/2} := ((\cdot)^{1/2})^+$, so $\boldsymbol{C}$ fixes $\lambda\boldsymbol{H}\boldsymbol{P}_\mathcal{T}$, where $\mathcal{T} = \mathrm{Range}(\boldsymbol{C})$. Conversely, for any symmetric $\boldsymbol{\Delta} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{\Delta}\boldsymbol{P}_\mathcal{T} = \boldsymbol{0}$ and $\boldsymbol{G}+\lambda\boldsymbol{\Delta} \succ 0$, replacing R by $R + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_0)^\top\boldsymbol{\Delta}(\boldsymbol{x} - \hat{\boldsymbol{x}}_0)$ keeps $\hat{\boldsymbol{x}}_0$ a strict local minimizer, and a solver that follows the minimizers through $\hat{\boldsymbol{x}}_0$ has the same $\boldsymbol{J}$ and $\boldsymbol{C}$.

In other words, Theorem 1 shows that *the reconstruction hides the prior, but random re-solves reveal its curvature on the subspace that the solver can reach*, and by *(R2)*, our fingerprint, which needs neither $\boldsymbol{A}$ nor σ_p, depends on the regularizer only through this curvature. We prove Theorem 1 and discuss its consequences in Section A.1, and verify it to machine precision on quadratic regularizers in Section B. By Proposition 2 in Section A.2, generic plug-and-play and unrolled solvers whose data steps use $\boldsymbol{A}^\top \boldsymbol{b}$ also have the response of a variational solver.

5 Experiments

In this section, we conduct extensive experiments on 9 inverse problems. By Sections 3 and 4, band energies need neither retraining nor the forward operator, and depend on the prior only through its curvature on $\mathcal{T}$. We measure whether they separate priors that reconstruct alike. Our experiments seek to address the following questions: **(Q1)** Does the fingerprint identify learned priors better than fingerprints that read the outputs? **(Q2)** Does it transfer, without retraining, to *external priors*, namely priors that played no role in its design? **(Q3)** Is it robust to unknown acquisition conditions? We also show in Section G that the advantage of the response is smallest at denoising, the one operator without an invisible subspace, and that ρ_k tracks the accuracy of each candidate prior.

5.1 Experimental Setup

Candidate priors. Our *internal set* contains the $K = 9$ learned priors on which we develop and evaluate the fingerprint, listed in Table 4: seven explicit learned regularizers, namely Fields of Experts (Roth & Black, 2009), adversarial regularizers (Lunz et al., 2018), an input-convex regularizer (Amos et al., 2017; Mukherjee et al., 2020) and a learned quadratic (Kunisch & Pock, 2013), and two plug-and-play priors, a DnCNN (Zhang et al., 2017; Venkatakrishnan et al., 2013) and a gradient-step denoiser (Hurault et al., 2022), used in the solver as denoisers. Section 5.3 describes the external sets.

Table 1: **Reading the response outperforms reading the outputs at a matched budget.** Top-1 accuracy (%) among 9 learned priors on 9 inverse problems with 8 solver calls per test image (random guess baseline 11.1%), with 95% confidence intervals (CI). Blue marks our fingerprint and beige the baselines; the highest is in **bold**, the second underlined.

Method	Top-1 (%) ↑	95% CI
ModelDiff, Random Perturbations	16.5	[12.8, 20.2]
ModelDiff, Selected Perturbations	<u>17.3</u>	[13.6, 20.6]
Response Fingerprint (Ours)	**80.0**	[69.1, 87.2]

Forward operators and data. We consider 9 inverse problems, from denoising to problems where the measurement leaves much of the image open: denoising, Gaussian and motion deblurring, inpainting from 30% of the pixels, $4\times$ super-resolution, Cartesian and radial MRI, and CT with 18 projection angles and with a limited angular range. Images come from CIFAR-10 (Krizhevsky, 2009), MNIST (LeCun et al., 1998), OrganAMNIST (Yang et al., 2023), STL-10 (Coates et al., 2011) and synthetic phantoms, at 64^2 to 128^2 pixels, as detailed in Section C.

Baselines and controls. We compare with three families of methods, each testing a different explanation of what attribution needs. *(i) Fingerprints that read the outputs* test whether the behavior of a model on shared inputs is enough: ModelDiff (Li et al., 2021), in its decision distance vector (DDV) variant, compares pairwise output distances at shared perturbations, SAC (Guan et al., 2022) correlates outputs on common inputs, and ProbeSpec (Lou et al., 2026) pools the output spectrum at increasing perturbation levels. *(ii) Statistics of the same re-solves* test whether any summary of the cloud suffices: a Gaussian likelihood, a log-Euclidean distance (Arsigny et al., 2007), the maximum mean discrepancy (Gretton et al., 2012), a convolutional neural network (CNN) classifier, a low-rank covariance and the total response $\mathrm{tr}\,\boldsymbol{C}$. *(iii)* The *optimality certificate* of Equation (12) tests whether knowing the operator and λ suffices. Two *controls* isolate what our fingerprint uses: a random-basis covariance, which pools along random directions, not frequency bands, and the mean of the re-solves.

Protocol. Reference and test clouds use disjoint images and fixed pretrained weights, each with $N = 8$ re-solves unless stated otherwise. We report top-1 accuracy under Equation (10) with balanced test sets, so the random guess baseline is $100\%/K$. For fairness, every method reads the same clouds, or the same budget of solver calls if it selects its perturbations, and uses the distance rule of its authors. Confidence intervals resample operators or test images, as detailed in Section C.

5.2 The Fingerprint Identifies Learned Priors (Q1)

Reading the response outperforms reading the outputs. As we show in Table 1, with 8 solver calls per test image, our fingerprint attributes the 9 learned priors with an accuracy of 80.0%, more than four times the 17.3% of ModelDiff with selected perturbations and the 16.5% with random ones. This advantage is consistent across inverse problems: the mean paired improvement is +62.8%, and it is positive on all 9 operators, as shown in Section D.1. Moreover, the advantage over ModelDiff and SAC grows with the budget. On the same clouds, our fingerprint improves from 80.2% at $N = 8$ to 83.2% at $N = 64$, whereas ModelDiff and SAC stay between 14.5% and 20.2%, the strongest statistical distance reaches 62.5% at $N = 8$ and 68.1% at $N = 64$, and ProbeSpec, with 80 solver calls and shared noise draws, reaches 35.4%, as reported in Table 5. The separation therefore comes from how the reconstruction moves, rather than from the reconstruction itself.

Table 2: **The frequency content of the response separates the priors.** Summaries of the same re-solves (56 configurations, $N = 64$), marked as in Table 1; controls in rose.

Summary of the Re-solves	Top-1 (%) ↑
Random-Basis Covariance	13.0
Mean of the Re-solves	25.6
CNN Classifier on Stacked Re-solves	39.6
Optimality Certificate	44.0
Low-Rank Covariance	64.6
Total Response $\mathrm{tr}\,\boldsymbol{C}$	73.6
Response Fingerprint (Ours)	**83.5**

The fingerprint is stable across its design choices. As we show in Sections D.5 and D.6, the accuracy stays between 88.8% and 90.2% from 4 to 24 bands over 12 configurations, and a Cartesian partition reaches 86.2%, compared with 83.4% for the radial one over 108 configurations.

The frequency of the response separates the priors. To understand which part of the response the fingerprint relies on, we compare it with other summaries of the same re-solves. As we show in Table 2, the band energies, which record where in frequency the reconstruction moves, reach a top-1 accuracy of 83.5%, compared with 73.6% for the total response $\mathrm{tr}\,\boldsymbol{C}$, which records how much it moves. In contrast, the random-basis covariance reaches 13.0%, close to the random guess baseline of 11.1%, and the mean of the re-solves, which carries the reconstruction but not the response, reaches 25.6%. The priors are thus separated by the frequency structure of the response. Moreover, the fingerprint separates the convex ridge regularizer CRR (Goujon et al., 2023) from its weakly convex counterpart WCRR (Goujon et al., 2024): CRR is classified as WCRR on 10.2% of the test images, and WCRR as CRR on 8.9%, as detailed in Section D.4.

Priors that reconstruct alike respond differently. We illustrate this in Figure 3, with the protocol of Section D.7. On three inverse problems, four of the 9 learned priors of Table 1 reconstruct with mean PSNRs that span at most 4.0 dB. However, their responses to the same random perturbations of one measurement, in the bottom row, show visibly different spatial structure, and their fingerprints, on the right, form distinct curves across the radial bands. The exception is the adversarially trained pair, Fields of Experts and the learned quadratic, whose fingerprints nearly coincide, so attribution can confuse them. The fingerprint thus errs where the responses themselves agree, which collapses two centroids, as discussed in Section 3.4.

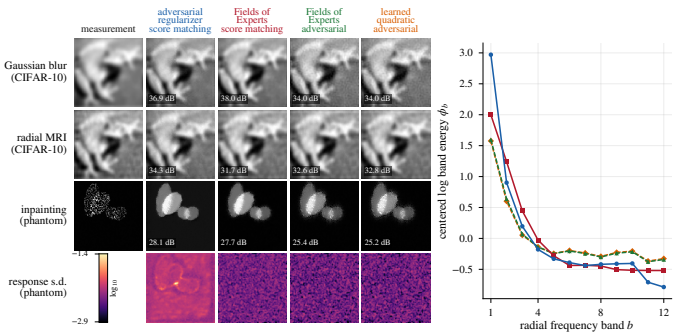


Figure 3: **Priors that reconstruct alike respond differently. Left:** reconstructions of four priors under three inverse problems with their mean PSNR, and the per-pixel standard deviation of the response on a shared $\log_{10}$ scale. **Right:** their fingerprints.

5.3 The Fingerprint Transfers to External Priors (Q2)

External sets. We use three *external sets* of learned priors, each with the publicly available trained weights of its authors and without any retraining. The first two consist of the learned regularizers of

Table 3: **The fingerprint transfers to external priors, without retraining.** Each row is a separate set of external priors: the learned regularizers of Hertrich et al. (2026) and the MRI networks of Yiasemis et al. (2022a); AUC is the area under the receiver operating characteristic curve for verifying a claimed candidate prior, averaged over candidate priors.

Set	Forward Operator	K	Top-1 (%) ↑	95% CI	Random Guess (%)	AUC ↑
Learned regularizers	denoising, $\boldsymbol{A} = \boldsymbol{I}$	7	58.9	[46.4, 72.3]	14.3	0.896
Learned regularizers	60-angle CT, m/n=0.234	7	66.1	[58.9, 72.3]	14.3	0.925
MRI networks	real measurements, 5× undersampled	10	68.0	[62.5, 72.5]	10.0	0.946

the comparative study of Hertrich et al. (2026), run with the solvers of its authors. On denoising, this set contains CRR (Goujon et al., 2023), WCRR (Goujon et al., 2024), ICNN (Mukherjee et al., 2020), TDV (Kobler et al., 2020), NETT (Li et al., 2020), PatchNR (Altekrüger et al., 2023) and LPN (Fang et al., 2024), and on 60-angle CT, EPLL (Zoran & Weiss, 2011) replaces LPN. The third consists of ten pretrained MRI reconstruction networks (Yiasemis et al., 2022a), which we apply to real MRI measurements from Calgary-Campinas (Souza et al., 2018), as detailed in Sections E.1 and E.4.

Attribution transfers without retraining. As we show in Table 3, the fingerprint attributes the three external sets with top-1 accuracies of 58.9%, 66.1% and 68.0%, above the random baselines of 14.3%, 14.3% and 10.0%. It verifies a claimed candidate prior with AUCs of 0.896, 0.925 and 0.946. On CT, the largest error is WCRR. With a perturbation of 4% instead of 1% of $\|\boldsymbol{b}\|$ and re-solves run to convergence, WCRR is correct on all 16 test images, as detailed in Section E.3. The four external priors that match the setting of Section 4 most closely are CRR, WCRR, ICNN and TDV, whose solvers deterministically reach a stationary point of Equation (2), and their regularizers are twice differentiable almost everywhere. The fingerprint attributes these priors at 70.3% on CT and 54.7% on denoising, compared with a random guess baseline of 25.0%, as detailed in Section E.2.

5.4 The Fingerprint Is Robust to Unknown Acquisition Conditions (Q3)

In practice, the forward operator $\boldsymbol{A}$, the law of the perturbations $\boldsymbol{\xi}$ and the regularization parameter λ are not always known exactly. To understand how a mismatch in each of them affects attribution, we vary them one at a time, as shown in Figure 8 in Section F. *(i) Forward operator.* As we show in Figure 8(a), our fingerprint stays at 83.5% under every tested error in the operator used for attribution, because $\boldsymbol{C}$ is computed from the re-solves, whereas the optimality certificate, which evaluates Equation (12) with this operator, decreases from 40.0% to 19.2%. *(ii) Perturbation law.* To first order, the law of the perturbations enters the response only through its covariance, because the re-solves move by $\boldsymbol{J}\boldsymbol{\xi}$. As we show in Figure 8(b), with references built from Gaussian perturbations, test clouds with Rademacher perturbations, namely random signs of matched energy, which have the same covariance, are attributed at 79.8%, close to the 84.2% with Rademacher references. Colored perturbations have a different covariance and are attributed at 79.4%, close to the 84.0% with colored references. *(iii) Regularization parameter.* Because λ weights each band differently, as Equation (9) shows, a mismatched λ moves the fingerprint gradually. As we show in Figure 8(c), over a 16-fold range of λ, our fingerprint retains 58.2% of its accuracy, compared with 43.0% for the certificate.

6 Conclusion

In this paper, we asked what the behavior of a solver reveals about its learned prior beyond the reconstruction, and when it is enough to distinguish two priors. We proved that, for variational solvers, the covariance of random re-solves determines the Jacobian in closed form, given the operator and the perturbation level, and with it the curvature of the regularizer on the subspace that the solver can reach. Built on this result, the response fingerprint applies to any differentiable solver, needs no training and comes with a label-free reliability indicator. It outperforms fingerprints that read the outputs, transfers without retraining to external priors and is robust to unknown acquisition conditions. A learned prior therefore leaves its fingerprint in the response of its solver, making response measurements a diagnostic for the provenance and comparison of learned imaging solvers.

AI USE STATEMENT

The authors are fully accountable for the content of this paper, including any text, claims or artifacts prepared with the help of generative AI. These tools supported the writing, namely language editing, the organization and presentation of the manuscript, and checks of the related literature. The authors reviewed all AI-assisted output for correctness, and they independently assessed and approved the scientific ideas, the methodological and experimental choices, the mathematical statements, and the interpretations and conclusions of the final manuscript.

ETHICS STATEMENT

Our response fingerprint identifies which known prior a solver uses, and applied to deployed imaging solvers, such attribution may support their provenance and comparison. We treat attribution from random re-solves as a diagnostic tool rather than a cryptographic guarantee, because candidate priors optimized to imitate one another are outside the scope of this work. The medical imaging data are public datasets used under the terms of their providers, as detailed in Section I.

REPRODUCIBILITY STATEMENT

Forward operators, candidate priors, data splits, numbers of re-solves, and perturbation and solver settings are specified in Section I, experimental protocols in Section C, and proofs with their assumptions in Section A. All methods are implemented in Python (Van Rossum & Drake, 2009) with PyTorch (Paszke et al., 2019). The external sets use the pretrained weights of their authors, without retraining, and all datasets are public. The code will be published upon acceptance.

REFERENCES

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. doi: 10.1109/TMI.2018.2799231.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019. doi: 10.1109/TMI.2018.2865356.

Fabian Altekrüger, Alexander Denker, Paul Hagemann, Johannes Hertrich, Peter Maass, and Gabriele Steidl. PatchNR: Learning from very few images by patch normalizing flow regularization. *Inverse Problems*, 39(6):064006, 2023. doi: 10.1088/1361-6420/acce5e.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017.

Javier Antorán, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. Uncertainty estimation for computed tomography with a linearised deep image prior. *Transactions on Machine Learning Research*, 2023.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. doi: 10.1073/pnas.1907377117.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996.

Anil Aswani, Zuo-Jun Max Shen, and Auyon Siddiq. Inverse optimization with noisy data. *Operations Research*, 66(3):870–892, 2018. doi: 10.1287/opre.2017.1705.

Harrison H Barrett and Kyle J Myers. *Foundations of image science*. John Wiley & Sons, 2013.

Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018. doi: 10.1017/S0962492918000016.

Yoeri E. Boink, Markus Haltmeier, Sean Holman, and Johannes Schwab. Data-consistent neural networks for solving nonlinear inverse problems. *Inverse Problems and Imaging*, 17(1):203–229, 2023. doi: 10.3934/ipi.2022037.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989. doi: 10.1214/aos/1176347115.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223. PMLR, 2011.

Jürgen Dölz and David Ebert. Local sensitivity analysis for Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 14(2):341–368, 2026. doi: 10.1137/25M1745350.

Yonina C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009. doi: 10.1109/TSP.2008.2008212.

Moshe Eliasof, Eldad Haber, and Eran Treister. DRIP: Deep regularizers for inverse problems. *Inverse Problems*, 40(1):015006, 2024. doi: 10.1088/1361-6420/ad0f3b.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. KIKI-net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic Resonance in Medicine*, 80(5):2188–2201, 2018. doi: 10.1002/mrm.27201.

Zhenghan Fang, Sam Buchanan, and Jeremias Sulam. What's in a prior? learned proximal networks for inverse problems. In *International Conference on Learning Representations*, 2024.

J.A. Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography. *IEEE Transactions on Image Processing*, 5(3):493–506, 1996. doi: 10.1109/83.491322.

Jeffrey A. Fessler and W. Leslie Rogers. Spatial resolution properties of penalized-likelihood image reconstruction: Space-invariant tomographs. *IEEE Transactions on Image Processing*, 5(9): 1346–1358, 1996.

Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks – robustness included? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1119–1134, 2023. doi: 10.1109/TPAMI.2022.3148324.

Nina M. Gottschling, Vegard Antun, Anders C. Hansen, and Ben Adcock. The troublesome kernel: On hallucinations, no free lunches, and the accuracy-stability tradeoff in inverse problems. *SIAM Review*, 67(1):73–104, 2025. doi: 10.1137/23M1568739.

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.

Alexis Goujon, Sebastian Neumayer, and Michael Unser. Learning weakly convex regularizers for convergent image-reconstruction algorithms. *SIAM Journal on Imaging Sciences*, 17(1):91–115, 2024. doi: 10.1137/23M1565243.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Rémi Gribonval and Mila Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62(6–7):773–789, 2020. doi: 10.1007/s10851-020-00951-y.

Jiyang Guan, Jian Liang, and Ran He. Are you stealing my model? sample correlation for fingerprinting deep neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36571–36584, 2022. doi: 10.52202/068431-2650.

Eldad Haber and Luis Tenorio. Learning regularization functionals—a supervised training approach. *Inverse Problems*, 19(3):611–626, 2003. doi: 10.1088/0266-5611/19/3/309.

Eldad Haber, Moshe Eliasof, and Luis Tenorio. Estimating a potential without the agony of the partition function. *SIAM Journal on Mathematics of Data Science*, 5(4):1005–1027, 2023. doi: 10.1137/22M1517135.

Andreas Habring and Martin Holler. Neural-network-based regularization methods for inverse problems in imaging. *GAMM-Mitteilungen*, 47(4):e202470004, 2024. doi: 10.1002/gamm.202470004.

Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, volume 4 of *SIAM Monographs on Mathematical Modeling and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, 1998. doi: 10.1137/1.9780898719697.

Johannes Hertrich, Hok Shing Wong, Alexander Denker, Stanislas Ducotterd, Zhenghan Fang, Markus Haltmeier, Željko Kereta, Erich Kobler, Oscar Leong, Mohammad Sadegh Salehi, Carola-Bibiane Schönlieb, Johannes Schwab, Zakhar Shumaylov, Jeremias Sulam, German Shâma Wache, Martin Zach, Yasi Zhang, Matthias J. Ehrhardt, and Sebastian Neumayer. Learning regularization functionals for inverse problems: a comparative study. In *Machine Learning Solutions for Inverse Problems: Part B*, volume 27 of *Handbook of Numerical Analysis*, pp. 85–135. Elsevier, 2026. doi: 10.1016/bs.hna.2026.04.001.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022.

Yohan Jun, Hyungseob Shin, Taejoon Eo, and Dosik Hwang. Joint deep model-based MR image and coil sensitivity reconstruction network (Joint-ICNet) for fast MRI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5266–5275, 2021. doi: 10.1109/CVPR46437.2021.00523.

Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer, New York, 2005. doi: 10.1007/b138659.

Arezou Keshavarz, Yang Wang, and Stephen Boyd. Imputing a convex objective function. In *2011 IEEE International Symposium on Intelligent Control*, pp. 613–619, 2011. doi: 10.1109/ISIC.2011.6045410.

Erich Kobler, Alexander Effland, Karl Kunisch, and Thomas Pock. Total deep variation for linear inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7546–7555, 2020. doi: 10.1109/CVPR42600.2020.00757.

Steven G. Krantz and Harold R. Parks. *The Implicit Function Theorem: History, Theory, and Applications*. Birkhäuser, Boston, 2002.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013. doi: 10.1137/120882706.

Edmund Landau. Einige Ungleichungen für zweimal differentiierbare Funktionen. *Proceedings of the London Mathematical Society*, s2-13(1):43–49, 1914. doi: 10.1112/plms/s2-13.1.43.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8:109, 2021. doi: 10.1038/s41597-021-00893-z.

Housen Li, Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. NETT: Solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020. doi: 10.1088/1361-6420/ab6d57.

Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. ModelDiff: Testing-based DNN similarity comparison for model reuse detection. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '21)*, pp. 139–151, 2021. doi: 10.1145/3460319.3464816.

Kai Lønning, Patrick Putzky, Jan-Jakob Sonke, Liesbeth Reneman, Matthan W. A. Caan, and Max Welling. Recurrent inference machines for reconstructing heterogeneous MRI data. *Medical Image Analysis*, 53:64–78, 2019. doi: 10.1016/j.media.2019.01.005.

Shanshan Lou, Hanzhe Yu, and Qi Xuan. ProbeSpec: Robust model fingerprinting via dynamic perturbation response spectrum. *Electronics*, 15(4):729, 2026. doi: 10.3390/electronics15040729.

Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8507–8516, 2018.

Arak Mathai and Serge Provost. *Quadratic Forms in Random Variables: Theory and Applications*, volume 87. 12 1992. doi: 10.2307/2290674.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965. doi: 10.24033/bsmf.1625.

Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Learned convex regularizers for inverse problems, 2020.

Frank Ong, Shahab Amin, Shreyas Vasanawala, and Michael Lustig. mridata.org: An open archive for sharing MRI raw data. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Patrick Putzky and Max Welling. Recurrent inference machines for solving inverse problems. arXiv preprint arXiv:1706.04008, 2017.

Jinyi Qi and Richard M. Leahy. Resolution and noise properties of MAP reconstruction for fully 3-D PET. *IEEE Transactions on Medical Imaging*, 19(5):493–506, 2000.

Yuhui Quan, Huan Teng, Ruotao Xu, Jun Huang, and Hui Ji. Fingerprinting deep image restoration models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13239–13249, October 2023. doi: 10.1109/ICCV51070.2023.01222.

Sathish Ramani, Thierry Blu, and Michael Unser. Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554, 2008. doi: 10.1109/TIP.2008.2001404.

Zaccharie Ramzi, Philippe Ciuciu, and Jean-Luc Starck. XPDNet for MRI reconstruction: An application to the 2020 fastMRI challenge. arXiv preprint arXiv:2010.07290, 2020.

Edward T. Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Transactions on Computational Imaging*, 5(1):52–67, 2019. doi: 10.1109/TCI.2018.2880326.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. doi: 10.1137/16M1102884.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Cham, 2015. Springer. doi: 10.1007/978-3-319-24574-4_28.

Stefan Roth and Michael J. Black. Fields of experts. *International Journal of Computer Vision*, 82 (2):205–229, 2009. doi: 10.1007/s11263-008-0197-6.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. doi: 10.1016/0167-2789(92)90242-F.

Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Deep null space learning for inverse problems: Convergence analysis and rates. *Inverse Problems*, 35(2):025008, 2019. doi: 10.1088/1361-6420/aaf14a.

Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical Report CMU-CS-94-125, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994.

Hae Jin Song and Laurent Itti. Riemannian-geometric fingerprints of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11425–11435, 2025. doi: 10.1109/ICCV51701.2025.01063.

Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018. doi: 10.1016/j.neuroimage.2017.08.021.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12262 of *Lecture Notes in Computer Science*, pp. 64–73, Cham, 2020. Springer. doi: 10.1007/978-3-030-59713-9_7.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Julián Tachella, Mike Davies, and Laurent Jacques. UNSURE: Self-supervised learning with unknown noise level and Stein's unbiased risk estimate. In *International Conference on Learning Representations*, 2025.

Luis Tenorio. Statistical regularization of inverse problems. *SIAM Review*, 43(2):347–366, 2001. doi: 10.1137/S0036144500358232.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012. doi: 10.1214/12-AOS1003.

Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of Ill-Posed Problems*. Scripta Series in Mathematics. V. H. Winston & Sons, Washington, D.C., 1977.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, 2009.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

Emilio Vilches. Proximal determination of convex functions. *Journal of Convex Analysis*, 28(4): 1187–1192, 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Jeremy Vonderfecht and Feng Liu. Fingerprints of super resolution networks. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=Jj0qSbtwdb.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

George Yiasemis, Nikita Moriakov, Dimitrios Karkalousos, Matthan Caan, and Jonas Teuwen. DIRECT: Deep image REConstruction toolkit. *Journal of Open Source Software*, 7(73):4278, 2022a. doi: 10.21105/joss.04278.

George Yiasemis, Jan-Jakob Sonke, Clarisa Sánchez, and Jonas Teuwen. Recurrent variational network: A deep learning inverse problem solver applied to the task of accelerated MRI reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2022b. doi: 10.1109/CVPR52688.2022.00081.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7555–7565, 2019. doi: 10.1109/ICCV.2019.00765.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017. doi: 10.1109/TIP.2017.2662206.

Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 479–486, 2011. doi: 10.1109/ICCV.2011.6126278.

APPENDIX OUTLINE

Main-text item	Where it is supported
Related work, Section 2	identifying an objective from its solver and model fingerprinting in Section H
Theorem 1, Equation (6) and the characterization for general differentiable solvers in Section 4	proof of Theorem 1, Proposition 2 and Proposition 3 in Section A; closed-form validation on quadratic regularizers, Figure 4, in Section B
Method of Section 3	Algorithm 1 in Section C; hyperparameters Section C; protocol of Figure 1 and fingerprints of its two solvers, Figure 5, in Section C; filter factors of Example 1 in Section A.3
Q1 attribution of learned priors, Section 5.2	experimental settings Section C; candidate priors Table 4; selected perturbations and per operator Section D.1; all baselines on the same clouds Section D.2; all summaries of the response Section D.3; CRR and WCRR Section D.4; partitions, bands and noise level Sections D.5 and D.6; protocol of Figure 3 Section D.7
Q2 external priors, Section 5.3	Section E, including solver noise and Assumption A1 Section E.2 and the CT pair Section E.3
Q3 robustness to acquisition conditions, Section 5.4	Figure 8 and its protocol Section F; perturbation law Section F.1; operator error Section F.2; λ Section F.3
Reliability of attribution, Section 3.4	Section G; white-box comparison and invisible fraction Section D.8; accuracy and identifiability Section E.5; indicator per set Section E.1

A PROOFS AND SUPPORTING RESULTS

This section proves Theorem 1 and the two supporting statements of the main text: which solvers have the response of a variational solver, and that band energies are estimable from a few re-solves.

Notation. Throughout, we write $\mathrm{tr}(\cdot)$ for the trace, $\|\cdot\|_2$ for the Euclidean norm of a vector and the spectral norm of a matrix, $\mathrm{Range}(\cdot)$ and $\mathrm{Null}(\cdot)$ for the range and the null space of a matrix, $(\cdot)^\top$ and $(\cdot)^\mathsf{H}$ for the transpose and the conjugate transpose, $\boldsymbol{X}^+$ for the Moore–Penrose pseudoinverse, $\boldsymbol{X} \succeq 0$ and $\boldsymbol{X} \succ 0$ for a positive semidefinite and a positive definite symmetric matrix, $\mathrm{diag}(\boldsymbol{v})$ for the diagonal matrix with diagonal $\boldsymbol{v}$, ∇ and ∇^2 for the gradient and the Hessian, C^k for k times continuously differentiable maps, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\mathbb{E}$ and Cov for the expectation and the covariance over random perturbations, and $O(\cdot)$ and $o(\cdot)$ for the usual asymptotic notation, namely $g = O(t)$ if $|g|/t$ stays bounded and $g = o(t)$ if $|g|/t \to 0$ as $t \to 0$. In the appendix, we also write $\boldsymbol{P}_\mathcal{V}$ for the orthogonal projector onto a subspace $\mathcal{V}$, $\mathcal{V}^\perp$ for its orthogonal complement, and $\|\cdot\|_F$ for the Frobenius norm.

A.1 PROOF OF THEOREM 1

Denoising. For denoising, where $\boldsymbol{A} = \boldsymbol{I}$, the rotation ambiguity of Section 4 disappears: the Jacobian $(\boldsymbol{I} + \lambda \boldsymbol{H})^{-1}$ is a symmetric positive definite smoother matrix (Buja et al., 1989), and therefore the unique positive definite square root of $\boldsymbol{C}/\sigma_p^2$ (Horn & Johnson, 2012). For a general operator, $\boldsymbol{J}$ is rectangular, and this argument does not apply directly.

Recovery of the Jacobian. Because $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A}^\top$ is symmetric, $\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top = \sigma_p^2(\boldsymbol{A}\boldsymbol{J})(\boldsymbol{A}\boldsymbol{J})^\top = \sigma_p^2\boldsymbol{S}^2$. As $\boldsymbol{S} \succeq 0$ and the positive semidefinite square root is unique (Horn & Johnson, 2012), $\boldsymbol{S} = \sigma_p^{-1}(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top)^{1/2}$. Next, $\boldsymbol{C}\boldsymbol{A}^\top = \sigma_p^2\boldsymbol{J}(\boldsymbol{A}\boldsymbol{J})^\top = \sigma_p^2\boldsymbol{J}\boldsymbol{S}$. Because $\boldsymbol{S}\boldsymbol{S}^+$ is the orthogonal projector onto $\mathrm{Range}(\boldsymbol{S}) = \mathrm{Range}(\boldsymbol{A})$ and $\boldsymbol{J} = \boldsymbol{G}^{-1}\boldsymbol{A}^\top$ vanishes on its complement $\mathrm{Null}(\boldsymbol{A}^\top)$, we obtain $\boldsymbol{C}\boldsymbol{A}^\top\boldsymbol{S}^+ = \sigma_p^2\boldsymbol{J}$, and substituting $\boldsymbol{S}^+ = \sigma_p(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top)^{+1/2}$ gives the formula for $\boldsymbol{J}$.

Curvature on $\mathcal{T}$. Substituting $\boldsymbol{G} = \boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{H}$ into $\boldsymbol{G}\boldsymbol{J} = \boldsymbol{A}^\top$ gives $\lambda\boldsymbol{H}\boldsymbol{J} = \boldsymbol{A}^\top - \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{J} = \boldsymbol{A}^\top(\boldsymbol{I}_m - \boldsymbol{S})$. Because $\mathrm{Range}(\boldsymbol{J}) = \mathcal{T}$, this identity fixes $\lambda\boldsymbol{H}\boldsymbol{P}_\mathcal{T}$. For the converse, replace R by $R' = R + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_0)^\top\boldsymbol{\Delta}(\boldsymbol{x} - \hat{\boldsymbol{x}}_0)$. The gradient of the added term vanishes at $\hat{\boldsymbol{x}}_0$, so $\hat{\boldsymbol{x}}_0$ still satisfies Equation (12) at $\boldsymbol{b}_0$, and the new system matrix $\boldsymbol{G}' = \boldsymbol{G} + \lambda\boldsymbol{\Delta} \succ 0$ keeps it a strict local minimizer. By the implicit function theorem, a solver that follows the minimizers through $\hat{\boldsymbol{x}}_0$ is

then differentiable at $\boldsymbol{b}_0$ with Jacobian $\boldsymbol{G}'^{-1}\boldsymbol{A}^\top$. Moreover, $\boldsymbol{G}'\boldsymbol{J} = \boldsymbol{G}\boldsymbol{J} + \lambda\boldsymbol{\Delta}\boldsymbol{J} = \boldsymbol{A}^\top$ because $\boldsymbol{\Delta}\boldsymbol{J} = \boldsymbol{0}$, so this solver has the Jacobian $\boldsymbol{J}$ and the response $\boldsymbol{C}$. □

The response in $\mathrm{Null}(\boldsymbol{A})$**.** The data fit contributes $\boldsymbol{A}^\top\boldsymbol{A}$ to $\boldsymbol{G}$ and the prior contributes $\lambda\boldsymbol{H}$. On $\mathrm{Null}(\boldsymbol{A})$, where $\boldsymbol{A}^\top\boldsymbol{A}$ vanishes, $\lambda\boldsymbol{H}$ is the only contribution to $\boldsymbol{G}$, so the part of the response in $\mathrm{Null}(\boldsymbol{A})$ is obtained from its part in $\mathrm{Range}(\boldsymbol{A}^\top)$ through a map that depends only on the curvature. For the quadratic regularizer of Example 1, whose Hessian is $\boldsymbol{H} = \boldsymbol{M}$, this part vanishes, and the prior is read at the frequencies that the operator attenuates, where $\lambda h(\omega)$ dominates $|a(\omega)|^2$.

Known operator and noise level. Theorem 1 describes the first-order response with known $\boldsymbol{A}$ and σ_p. Our fingerprint, computed without $\boldsymbol{A}$ and invariant to σ_p, needs neither, and Section 5 measures how well it separates priors.

Priors that agree on $\mathcal{T}$ **have the same response.** Two priors with the same reconstruction $\hat{\boldsymbol{x}}_0$, both satisfying Assumption A1, whose curvatures $\lambda\boldsymbol{H}$ agree on $\mathcal{T}$ have the same response, whereas priors whose curvatures differ on $\mathcal{T}$, such as CRR and WCRR in Section 5.2, respond differently. Finally, a C^1 solver has the response of a variational solver if and only if its response has rank r and $\boldsymbol{A}$ is injective on its range, as Proposition 2 in Section A.2 shows, and its fingerprint is then that of a variational solver.

Random and chosen perturbations. For a variational solver with known $\boldsymbol{A}$ and σ_p, $\boldsymbol{C}$ determines $\boldsymbol{J}$, so in the population limit random perturbations identify as much of the prior as perturbations along directions that we choose, which measure $\boldsymbol{J}$ one direction at a time.

Colored perturbations. For perturbations $\mathcal{N}(\boldsymbol{0}, \sigma_p^2\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ 0$, the response is $\sigma_p^2\bar{\boldsymbol{J}}\bar{\boldsymbol{J}}^\top$ with $\bar{\boldsymbol{J}} = \boldsymbol{J}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{G}^{-1}\bar{\boldsymbol{A}}^\top$ and $\bar{\boldsymbol{A}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}$. The recovery argument uses only that $\bar{\boldsymbol{A}}\bar{\boldsymbol{J}}$ is symmetric positive semidefinite with range $\mathrm{Range}(\bar{\boldsymbol{A}})$ and that $\bar{\boldsymbol{J}}$ vanishes on $\mathrm{Null}(\bar{\boldsymbol{A}}^\top)$, so it applies to $\bar{\boldsymbol{A}}$ and returns $\boldsymbol{J} = \bar{\boldsymbol{J}}\boldsymbol{\Sigma}^{-1/2}$.

A.2 When a differentiable solver has the response of a variational solver

Plug-and-play and unrolled solvers need not minimize an objective of the form of Equation (2). We now show in Proposition 2 exactly when such a solver has the response of some variational solver, in which case its fingerprint is the fingerprint of the regularizer with the same response. Unlike the characterization of proximity operators (Gribonval & Nikolova, 2020), which requires a symmetric Jacobian, Proposition 2 matches only the response and needs no symmetry.

Proposition 2 (Response of a differentiable solver; cf. Buja et al., 1989 for $\boldsymbol{A} = \boldsymbol{I}$)**.** *Let $\hat{\boldsymbol{x}}(\cdot)$ be a C^1 solver with response $\boldsymbol{C} = \sigma_p^2\boldsymbol{J}\boldsymbol{J}^\top$ at $\boldsymbol{b}_0$. Some regularizer that satisfies Assumption A1 yields a variational solver with the same reconstruction and the same response at $\boldsymbol{b}_0$ if and only if $\boldsymbol{C}$ has rank r and $\boldsymbol{A}$ is injective on* $\mathrm{Range}(\boldsymbol{C})$.

Proof. Let $\boldsymbol{J}^* = \sigma_p^{-1}\boldsymbol{C}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top)^{+1/2}$, the formula of Equation (14). Because $\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top$ has rank r, $\boldsymbol{S}^* = \boldsymbol{A}\boldsymbol{J}^* = \sigma_p^{-1}(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top)^{1/2}$ is symmetric and positive definite on $\mathrm{Range}(\boldsymbol{A})$, and $\mathrm{Null}(\boldsymbol{J}^*) = \mathrm{Null}(\boldsymbol{A}^\top)$. Because $\boldsymbol{A}$ is injective on $\mathrm{Range}(\boldsymbol{C})$, $\boldsymbol{C}^{1/2}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top)^{+}\boldsymbol{A}\boldsymbol{C}^{1/2}$ is the orthogonal projector onto $\mathrm{Range}(\boldsymbol{C}^{1/2})$, which gives $\sigma_p^2\boldsymbol{J}^*\boldsymbol{J}^{*\top} = \boldsymbol{C}$. We now build $\boldsymbol{G} \succ 0$ with $\boldsymbol{G}\boldsymbol{J}^* = \boldsymbol{A}^\top$. On $\mathcal{T}^* = \mathrm{Range}(\boldsymbol{J}^*)$, set $\boldsymbol{G}\boldsymbol{J}^*\boldsymbol{w} = \boldsymbol{A}^\top\boldsymbol{w}$. This is well defined because $\mathrm{Null}(\boldsymbol{J}^*) = \mathrm{Null}(\boldsymbol{A}^\top)$, symmetric because $\langle \boldsymbol{G}\boldsymbol{J}^*\boldsymbol{w}, \boldsymbol{J}^*\boldsymbol{w}'\rangle = \boldsymbol{w}^\top\boldsymbol{S}^*\boldsymbol{w}'$ for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^m$, and positive because $\boldsymbol{S}^*$ is positive definite on $\mathrm{Range}(\boldsymbol{A})$. This defines the columns of $\boldsymbol{G}$ on $\mathcal{T}^*$, symmetry fixes the block that maps $(\mathcal{T}^*)^\perp$ into $\mathcal{T}^*$, and setting the block on $(\mathcal{T}^*)^\perp$ to a large multiple of the identity makes $\boldsymbol{G} \succ 0$ by the Schur complement. The quadratic regularizer with Hessian $\lambda^{-1}(\boldsymbol{G} - \boldsymbol{A}^\top\boldsymbol{A})$ and a linear term that makes $\hat{\boldsymbol{x}}(\boldsymbol{b}_0)$ stationary at $\boldsymbol{b}_0$ then satisfies Assumption A1, because its solver is affine in $\boldsymbol{b}$, and this solver has the Jacobian $\boldsymbol{G}^{-1}\boldsymbol{A}^\top = \boldsymbol{J}^*$ and the response $\boldsymbol{C}$. Conversely, every variational solver satisfies both conditions, because $\mathrm{Range}(\boldsymbol{C}) = \mathrm{Range}(\boldsymbol{J}) = \boldsymbol{G}^{-1}\,\mathrm{Range}(\boldsymbol{A}^\top)$ has dimension r, and $\boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A}^\top\boldsymbol{w} = \boldsymbol{0}$ implies $\boldsymbol{A}^\top\boldsymbol{w} = \boldsymbol{0}$. □

Both conditions hold for every variational solver, for generic C^1 solvers whenever $\boldsymbol{A}$ has full row rank or is injective, and for generic plug-and-play and unrolled solvers whose data steps use $\boldsymbol{A}^\top\boldsymbol{b}$.

For a fixed-depth unrolled network, $\boldsymbol{J}$ exists and the response is estimated as before, but a truncated iteration need not make $\boldsymbol{J}$ the inverse of a curvature operator. Assumption A1 therefore governs the

reading of the fingerprint as curvature and not its existence, which is why we report the unrolled networks of Section E.4 as attribution results.

A.3 Band energies

We now show in Proposition 3 that the moments of the band-energy estimate in Equation (6) hold for any bands and responses, and that the centered log energies do not depend on the response scale.

Proposition 3 (Band-energy estimability; based on Mathai & Provost, 1992)**.** *Let $\boldsymbol{z}_1, \dots, \boldsymbol{z}_N \in \mathbb{R}^n$ be i.i.d. draws from $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ and let $\{\boldsymbol{P}_b\}_{b=1}^B \subset \mathbb{R}^{n\times n}$ be orthogonal projectors onto B mutually orthogonal bands. For $\hat{e}_b = N^{-1}\sum_i \|\boldsymbol{P}_b\boldsymbol{z}_i\|_2^2$ and $e_b = \operatorname{tr}(\boldsymbol{P}_b\boldsymbol{C})$, the moments of Equation (6) hold. For deviations from the sample mean $\bar{\boldsymbol{z}}$, the estimate $\hat{e}_b = (N-1)^{-1}\sum_i \|\boldsymbol{P}_b(\boldsymbol{z}_i - \bar{\boldsymbol{z}})\|_2^2$ satisfies them with $N-1$ in place of N. Centered log energies are invariant to $\boldsymbol{C} \mapsto \alpha\boldsymbol{C}$, $\alpha > 0$.*

Proof. For $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$, the quadratic form $\boldsymbol{z}^\top\boldsymbol{P}_b\boldsymbol{z}$ has mean $\operatorname{tr}(\boldsymbol{P}_b\boldsymbol{C}) = e_b$ and variance $2\operatorname{tr}[(\boldsymbol{P}_b\boldsymbol{C})^2]$ (Mathai & Provost, 1992). Averaging N independent draws keeps the mean and divides the variance by N, so $\operatorname{sd}(\hat{e}_b)^2 = 2\operatorname{tr}[(\boldsymbol{P}_b\boldsymbol{C})^2]/N = 2e_b^2/(Nr_{\mathrm{eff},b})$, which is Equation (6). For the centered estimate, an orthogonal (Helmert) transformation of the N draws maps $\sum_i(\boldsymbol{z}_i-\bar{\boldsymbol{z}})(\boldsymbol{z}_i-\bar{\boldsymbol{z}})^\top$ to $\sum_{j=1}^{N-1}\boldsymbol{w}_j\boldsymbol{w}_j^\top$ with $\boldsymbol{w}_j$ i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$, so the same argument applies to $N-1$ draws. Replacing $\boldsymbol{C}$ by $\alpha\boldsymbol{C}$ multiplies every band energy by α, and

$$\log(\alpha e_b) - \frac{1}{B}\sum_{q=1}^{B}\log(\alpha e_q) = \log e_b - \frac{1}{B}\sum_{q=1}^{B}\log e_q. \tag{15}$$

□

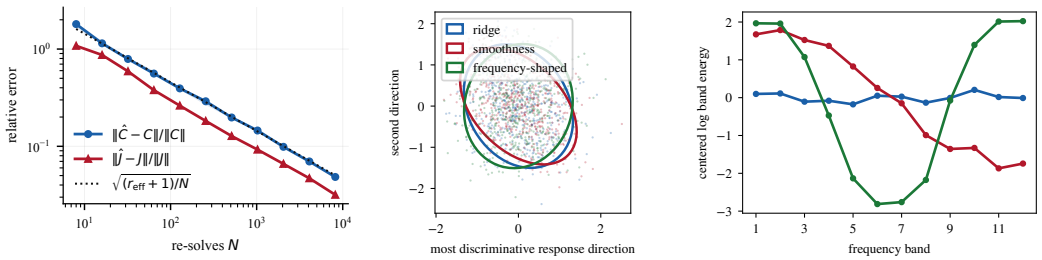
Filter factors. In Example 1, the response spectrum of Equation (9) can be written as $c(\omega) = \sigma_p^2 f(\omega)^2/|a(\omega)|^2$, where, for $a(\omega) \neq 0$, $f(\omega) = |a(\omega)|^2/(|a(\omega)|^2 + \lambda h(\omega))$ is the *filter factor* of regularization theory (Hansen, 1998; Engl et al., 1996).

The fingerprint satisfies the three requirements. *(R1)* By Proposition 3, each band energy is estimated with relative error $\sqrt{2/((N-1)r_{\mathrm{eff},b})}$ from centered deviations, and the radial bands of Equation (7) pool many response directions, so the fingerprint is estimable from $N \ll n$ re-solves. *(R2)* The population fingerprint, computed from the band energies e_b, is a function of $\boldsymbol{C}$ and not of $\hat{\boldsymbol{x}}_0$, and by Theorem 1 a change of the curvature outside $\mathcal{T}$ leaves $\boldsymbol{C}$ unchanged, so the fingerprint depends on the regularizer only through its curvature on $\mathcal{T}$. *(R3)* By Equation (15), the fingerprint is invariant to the scale of $\boldsymbol{C}$, and hence to the factor σ_p^2 of Equation (3). In contrast, the regularization strength λ changes the fingerprint, because it enters each band differently, as Equation (9) shows, and we measure this dependence in Section 5.4.

B Closed-form validation on quadratic regularizers

For a quadratic regularizer, the solver is linear and every quantity of Theorem 1 is available in closed form, so we can validate the algebra of each of its claims to machine precision. Across 5 operators, which include $r = m$ and $r < m$, graded spectra and a singular $\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top$, and under two perturbation laws, Equation (14) recovers the Jacobian and the response with a relative error of at most 2.5×10^{-14}. A change of the curvature outside $\mathcal{T}$ moves the response by at most 4.2×10^{-14}, whereas a change inside $\mathcal{T}$ moves it by more than 10^{-6}, as Theorem 1 states. With a sampled instead of an exact response, the error follows the rate $\sqrt{(r_{\mathrm{eff}}+1)/N}$ of the sample covariance, with $r_{\mathrm{eff}} = \operatorname{tr}(\boldsymbol{C})^2/\operatorname{tr}(\boldsymbol{C}^2) = 19.7$, and, for this well-conditioned operator, recovering $\boldsymbol{J}$ adds no error beyond it, as shown in Figure 4(a). As we show in Figure 4(b,c), band energies also separate three quadratic priors whose draws overlap.

Protocol of Figure 4(b,c). The three regularizers are quadratic, $R_{\boldsymbol{M}}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}$ on a one-dimensional signal with $n = 64$, with λ absorbed into $\boldsymbol{M}$: ridge $\boldsymbol{M} = 0.40\,\boldsymbol{I}$, smoothness $\boldsymbol{M} = 0.05\,\boldsymbol{L}_2^2 + 0.02\,\boldsymbol{I}$ with $\boldsymbol{L}_2$ the second-difference matrix, and a frequency-shaped prior $\boldsymbol{M} = \boldsymbol{F}^{\mathsf{H}}\operatorname{diag}\big(w(\nu)\big)\boldsymbol{F}$ in the unitary Fourier basis $\boldsymbol{F}$, with the mid-band weight $w(\nu) =$



(a) recovery error (b) draws of the response (c) band energies

Figure 4: **On quadratic regularizers, random re-solves recover the response at the rate of the sample covariance, and band energies separate priors whose draws overlap. (a)** Relative Frobenius error of the sampled covariance and of the Jacobian from Theorem 1, median of 7 runs; the dotted line is the rate $\sqrt{(r_{\text{eff}}+1)/N}$. **(b)** Exact draws from three quadratic regularizers in the two directions where their covariances differ most. **(c)** Their centered log band energies.

$0.02+2\exp\big(-((|\nu|-0.25)/0.08)^2\big)$ of the normalized frequency ν. The operator is a Gaussian random matrix $\boldsymbol{A}\in\mathbb{R}^{32\times 64}$ and $\sigma_p=0.35$. Each cloud holds $N=500$ exact draws from $\boldsymbol{C}=\sigma_p^2\boldsymbol{J}\boldsymbol{J}^\top$. The two-dimensional view is spanned by the two leading eigenvectors of $\sum_k(\boldsymbol{C}_k-\bar{\boldsymbol{C}})^2$, with $\bar{\boldsymbol{C}}$ the mean of the three covariances, after each cloud is scaled to unit root-mean-square norm. Panel (c) reads the same draws as centered log energies in 12 equal-width bands of $|\nu|$.

C Experimental details

The attribution algorithm. Algorithm 1 summarizes the attribution method of Section 3.

Experimental settings. Our imaging results come from six experimental settings: *(i)* the budget-matched comparison of Table 1 and Figure 9(a), in which every test perturbation is drawn independently of the references, as detailed in Section D.1; *(ii)* an independent-noise regeneration on the same 9 inverse problems, in which every method reads the same clouds, as detailed in Section D.2; *(iii)* a setting on the same problems in which matched reference and test clouds share their noise draws, which we use for ProbeSpec and the white-box comparison, as detailed in Sections D.2 and D.8; *(iv)* the estimator study over 56 configurations of Table 2, with disjoint reference and test images and separate perturbation seeds, as detailed in Section D.3; *(v)* the recent set of $K=13$ priors of Section D.4, which adds regularizers introduced after 2022 to the internal set; and *(vi)* the three external sets of Table 3, as detailed in Section E.

Fingerprint hyperparameters. The fingerprint has two hyperparameters, the number B of radial bands and the noise level σ_p of the perturbations, both fixed before evaluation. Unless stated otherwise, we use $B=12$, which we denote Radial-12, and $\sigma_p=10^{-2}$, and Section D.6 varies both.

Protocol of Figure 2. The illustration is in closed form and involves no trained model. The grid has 64×64 pixels, the operator is a periodic Gaussian blur of standard deviation 0.6 pixels, with symbol $a(\omega)=\exp(-2\pi^2\,0.6^2|\omega|^2)$, and both priors use $\lambda=0.02$, with $h(\omega)=1$ for ridge and $h(\omega)=4\sin^2(\pi\omega_1)+4\sin^2(\pi\omega_2)$ for smoothness. The response spectrum is Equation (9), and the exact band energies sum it over the $B=12$ bands of Equation (7). Re-solves are exact, $\boldsymbol{z}_i=\boldsymbol{J}\boldsymbol{\xi}_i$ with $\sigma_p=10^{-2}$, and each of the 200 clouds has $N=8$ re-solves. The measured relative standard deviation of $\hat{e}_b$ matches $\sqrt{2/((N-1)r_{\text{eff},b})}$ of Proposition 3 for the centered deviations, in all bands for both priors, and the mean estimated fingerprint differs from the exact one by at most 0.01.

Protocol of Figure 1. The measurement is a 128×128 phantom under inpainting with 30% of the pixels observed. The two solvers use an adversarial regularizer (Lunz et al., 2018) and Fields of Experts (Roth & Black, 2009), both trained by denoising score matching (Vincent, 2011), and their reconstructions are within 0.3 dB PSNR of each other on this image. The response maps are the per-pixel standard deviation over $N=64$ re-solves at $\boldsymbol{b}+\boldsymbol{\xi}_i$, each divided by its median because the fingerprint discards the overall response level; the response of the adversarial regularizer follows the image structure, while that of Fields of Experts is spatially white. The difference in Figure 1(a) is

Algorithm 1: Attribution by the response fingerprint.

Input: solver $\hat{\boldsymbol{x}}(\cdot)$, measurement $\boldsymbol{b}_0 \in \mathbb{R}^m$, noise level σ_p, number of re-solves N, band projectors $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_B$ of Equation (7), reference centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.

Output: attributed candidate prior $\hat{k}$.

1: For $i = 1, \ldots, N$, draw $\boldsymbol{\xi}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_p^2 \boldsymbol{I}_m)$ and solve $\boldsymbol{x}_i = \hat{\boldsymbol{x}}(\boldsymbol{b}_0 + \boldsymbol{\xi}_i)$. ▷ N calls to the solver

2: Compute the mean $\bar{\boldsymbol{x}} = N^{-1} \sum_i \boldsymbol{x}_i$ and the deviations $\boldsymbol{z}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

3: For $b = 1, \ldots, B$, compute $\hat{e}_b = (N-1)^{-1} \sum_i \|\boldsymbol{P}_b \boldsymbol{z}_i\|_2^2$. ▷ one 2D FFT per re-solve

4: Compute the fingerprint $\phi_b = \log \hat{e}_b - B^{-1} \sum_q \log \hat{e}_q$ of Equation (8).

5: Return $\hat{k} = \arg\min_k \|\boldsymbol{\phi} - \boldsymbol{\mu}_k\|_2$ of Equation (10).

The reference centroid $\boldsymbol{\mu}_k$ is the mean of the fingerprints computed by steps 1 to 4 from the solver of candidate prior k on its reference images.

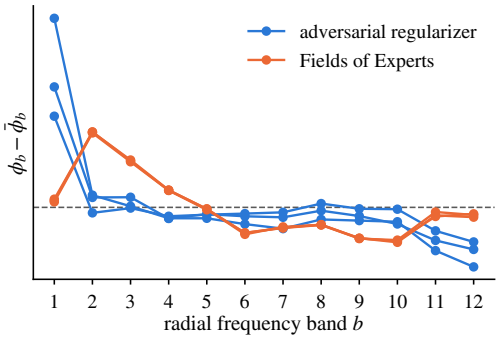


Figure 5: **The fingerprints of the two solvers of Figure 1 separate across test images.** Each curve is the fingerprint of Equation (8) for one test image, namely the centered log energies of the response in the twelve radial frequency bands, shown as its deviation from $\bar{\phi}_b$, the mean fingerprint of the four learned priors of the internal set that reconstruct this image above 20 dB. Blue curves are the adversarial regularizer and orange curves Fields of Experts, on three test images each, with one vertical scale for all curves; the dashed line marks zero deviation.

shown on the intensity scale of the reconstructions, and Figure 5 shows the fingerprints of the two solvers, which agree across test images of the same solver and differ most in the lowest bands.

Candidate priors. The internal set is listed in Table 4.

Baselines. *ModelDiff* (Li et al., 2021) is represented by its Decision Distance Vector (pairwise distances between reactions to a common set of inputs) under cosine similarity. Its adversarial input generator is defined through decision boundaries, so Section D.1 adds its continuous analogue, in which ModelDiff selects its own perturbations from a shared pool using references only. *SAC* (Guan et al., 2022) keeps its central object, the correlation among model outputs. Its input selection, through misclassified samples in SAC-w and CutMix in SAC-m, is defined only for classifiers, so we evaluate the correlations on our random perturbations. *ProbeSpec* (Lou et al., 2026), reimplemented from its description, preserves the original feature construction (corruption levels $\{0, 0.05, 0.10, 0.15, 0.20\}$, 2D DCT of the response, zig-zag ordering, band energies) and replaces only the class-dependent input selection with 16 seeded directions; it issues 80 solves per candidate prior and image, compared with our 64 at $N = 64$. Three statistical distances between clouds read a 64-dimensional Gaussian sketch of each cloud. The *Gaussian likelihood* fits a shrunk covariance (Ledoit & Wolf, 2004) to each reference cloud, averages it per candidate prior, and scores the mean log-likelihood of the centered test re-solves. The *log-Euclidean distance* (Arsigny et al., 2007) compares the test covariance with the log-Euclidean mean of the reference covariances of the candidate prior. *Kernel MMD* (Gretton et al., 2012) is the unbiased squared maximum mean discrepancy between the centered test re-solves and the pooled centered reference re-solves, under a Gaussian kernel whose bandwidth is the median pairwise distance among reference re-solves. All baselines read isotropic Gaussian perturbations at our default noise level, except in the budget-matched comparison of Section D.1.

Table 4: **The internal set spans convex, nonconvex and plug-and-play priors.** Nine learned priors of the internal set. Two entries share architecture and training objective and differ only by seed, which isolates the finest distinction in the set.

Prior	Reference	Class	Trained by
Learned quadratic	(Kunisch & Pock, 2013; Haber & Tenorio, 2003)	convex (quadratic)	adversarial
Fields of Experts	(Roth & Black, 2009)	nonconvex	adversarial
Input-convex network	(Amos et al., 2017; Mukherjee et al., 2020)	convex	adversarial
Adversarial regularizer	(Lunz et al., 2018)	nonconvex	adversarial
Fields of Experts	(Roth & Black, 2009)	nonconvex	denoising score matching
Adversarial regularizer	(Lunz et al., 2018)	nonconvex	denoising score matching
Adversarial regularizer†	(Lunz et al., 2018)	nonconvex	adversarial
DnCNN, plug-and-play	(Zhang et al., 2017; Venkatakrishnan et al., 2013)	plug-and-play	Gaussian denoising
Gradient-step denoiser	(Hurault et al., 2022)	plug-and-play (conservative)	denoising

†a second training seed of the same architecture and objective.

Confidence intervals. Configurations are resampled with replacement, then images within each sampled configuration, with every estimator recomputed on the same resampled units so that differences stay paired; 10000 draws, percentile intervals, fixed seed. The estimand is the performance of fixed trained checkpoints across the tested configurations, not the effect of retraining.

D Additional results

D.1 Budget-matched comparison with method-chosen perturbations

Protocol. This is the comparison of Table 1. As in the independent-noise regeneration of Section D.2, no test perturbation shares noise with a reference; unlike both settings there, the baseline may choose its own perturbations. For every (operator, image) there is one measurement and a pool of 256 seeded directions, *shared by all candidate priors*, solved at the default noise level. ModelDiff selects 8 perturbations greedily from the pool to maximize a reference-only separability of its DDVs,

$$\mathrm{sep}(\mathcal{S}) = \frac{\mathrm{mean}_k \ \min_{l \neq k} d_{\cos}(\bar{v}_k, \bar{v}_l)}{\mathrm{mean}_k \ \mathrm{mean}_i \ d_{\cos}(v_{ki}, \bar{v}_k)},$$

where $d_{\cos}$ is the cosine distance, v_{ki} is the DDV of candidate prior k on reference image i restricted to the selected set $\mathcal{S}$, and $\bar{v}_k$ its reference mean. Selection starts from the highest-scoring of 400 random pairs and adds, at each step, the highest-scoring of 64 random directions; no test image is touched. *ModelDiff, random perturbations* uses the first 8 pool perturbations. *Response Fingerprint (Ours)* reads our fingerprint from the response to the same first 8 perturbations (pool indices 0 to 7 on every operator). Each method keeps its own distance rule. We use the internal set ($K = 9$) on 9 operators, with 6 reference and 6 test images.

Result. As shown in Figure 6, the response fingerprint leads on each of the 9 operators. Operator-resampled intervals are $[0.691, 0.872]$ for ours and $[0.136, 0.206]$ and $[0.128, 0.202]$ for ModelDiff with chosen and pool perturbations; the paired advantage over ModelDiff on the pool perturbations is $+0.636$ ($[+0.514, +0.737]$). Its paired advantage over ModelDiff with chosen perturbations is $+0.628$ ($[+0.510, +0.704]$, operator bootstrap). Choosing the perturbations changes ModelDiff by $+0.008$ ($[-0.051, +0.064]$). The gap therefore comes from what each statistic reads from the same re-solves, and not from our choice of perturbations.

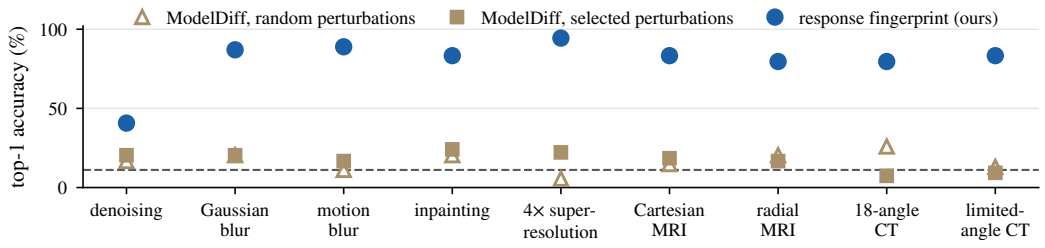


Figure 6: **The advantage of reading the response holds across a variety of inverse problems.** Top-1 accuracy (%) per operator at the matched budget of Table 1, for ours (blue) and ModelDiff with random or selected perturbations (beige); the dashed line marks the random guess baseline.

Table 5: **The fingerprint compares favorably with statistics of the same re-solves at each budget we test, and its advantage over ModelDiff and SAC increases with the number of re-solves.** Top-1 accuracy among 9 learned priors on 9 inverse problems when every method reads the same clouds, with test clouds seeded apart from the references (2,430 trials, random guess baseline 0.111). Blue marks our fingerprint and beige the baselines; in each column, the highest is in **bold** and the second highest underlined.

Method	Re-solves N			
	8	16	32	64
ModelDiff	0.202	0.193	0.167	0.146
SAC	0.181	0.166	0.155	0.145
Gaussian Likelihood	0.574	0.622	0.644	0.656
Log-Euclidean Distance	<u>0.625</u>	<u>0.662</u>	<u>0.665</u>	<u>0.681</u>
Kernel MMD	0.492	0.562	0.604	0.650
Response Fingerprint (Ours)	**0.802**	**0.819**	**0.819**	**0.832**

D.2 Behavioral and statistical fingerprints on the same response clouds

Protocol. To test whether the advantage comes only from calling the solver many times, one cloud is generated per (operator, candidate prior, image), and every method's fingerprint is computed from that cloud. The N sweep uses nested prefixes of one cloud, so a smaller budget is a strict subset of the same re-solves. Each method is scored by the distance rule its authors prescribe, fixed before any test number was read. We use the internal set on the 9 inverse problems, with six reference images per candidate prior and 30 test images per candidate prior and operator, seeded apart from the references, so no test cloud shares measurement noise or perturbation draws with a reference.

Independent-noise regeneration. Table 5 reports this setting over 2,430 configuration–image trials on 9 operators. The response fingerprint is more accurate than the other methods at each budget we test. Its paired advantage over ModelDiff is $+0.599$ at $N = 8$ ($[+0.505, +0.658]$) and $+0.686$ at $N = 64$ ($[+0.590, +0.755]$), and over SAC $+0.620$ ($[+0.548, +0.678]$) and $+0.686$ ($[+0.605, +0.748]$). The advantage over ModelDiff and SAC grows with the budget, because the fingerprint improves from 0.802 to 0.832 while ModelDiff and SAC change from 0.202 to 0.146 and from 0.181 to 0.145. The three statistical distances read the same re-solves through a likelihood, a covariance distance and a kernel two-sample statistic, and stay below the fingerprint by $+0.228$ ($[+0.163, +0.286]$), $+0.177$ ($[+0.115, +0.232]$) and $+0.310$ ($[+0.253, +0.370]$) at $N = 8$, and by at least $+0.151$ at $N = 64$.

Shared noise draws. In a second setting on the same 9 inverse problems, with six test images per candidate prior, the noise seed of a cloud is a function of the operator, the candidate prior and the image index, so the reference and test clouds of a candidate prior that share an image index also share their noise draws. This favors statistics that compare responses perturbation by perturbation, and ModelDiff rises to 0.490 at $N = 8$ and 0.586 at $N = 64$, while our fingerprint reaches 0.831 and 0.856, a paired advantage of $+0.342$ ($[+0.212, +0.455]$) at $N = 8$. ProbeSpec, which issues 80

Table 6: **Band energies compare favorably with other summaries of the same re-solves.** Top-1 accuracy (56 configurations, $N = 64$), mean over configurations $\pm$ standard error, with a random guess baseline of 0.111. The highest value is in **bold** and the second highest is underlined. Blue marks our fingerprint, beige the baselines, rose the controls, and unshaded rows are its ablations.

Summary of the Re-solves	Top-1 ↑
Optimality Certificate	0.440±0.033
8 First-Order Scalars	0.640±0.017
5 Second-Order Scalars	0.696±0.021
CNN Classifier on Stacked Re-solves	0.396±0.021
PRNU Residual NCC	0.496±0.026
Total Response $\mathrm{tr}\,\boldsymbol{C}$	0.736±0.012
Low-Rank Covariance	0.646±0.016
Low-Rank Covariance, Learned Basis	0.353±0.019
Eigenvalues Only	0.543±0.017
Eigenvectors Only	0.370±0.018
Random-Basis Covariance	0.130±0.004
Mean of the Re-solves	0.256±0.018
Uncentered Band Energies	0.815±0.014
Centered, 2× Bands	0.810±0.015
Response Fingerprint (Ours)	**0.835**±0.017

solves per candidate prior and image, reaches 0.354 (AUC 0.704). Comparing each test cloud only with references from other images gives our fingerprint 0.780 at $N = 8$ and 0.842 at $N = 64$.

In both settings, the dependence unit is the forward operator, so the intervals resample 9 clusters, and the paired advantage of our fingerprint has the same sign at every budget and in both settings.

D.3 All summaries of the response covariance

Table 6 extends Table 2 with further summaries of the response and with two ablations of our fingerprint. All methods take the same re-solves as input, and reference fingerprints come from images disjoint from the test images, so every method has the same data budget. Among the additional baselines, PRNU Residual NCC adapts the photo-response non-uniformity (PRNU) fingerprint of cameras (Lukas et al., 2006) and compares residuals by normalized cross-correlation (NCC). A *random* basis of the same covariance remains at the level of a random guess, while our response fingerprint of Equation (8) compares favorably with the other summaries, and its two ablations reach 0.815 without centering and 0.810 with twice as many bands.

Seeding. This setting and its 108-configuration extension draw reference and test clouds independently. For image i and perturbation k, reference clouds use perturbation seed $1000 + 31i + k$ and test clouds use $7000 + 53i + k$, on disjoint reference and test images and with separate measurement-noise seeds, so no test perturbation repeats a reference draw.

D.4 The fingerprint separates a convex regularizer from its weakly convex counterpart

The convex ridge regularizer CRR (Goujon et al., 2023) and its weakly convex counterpart WCRR (Goujon et al., 2024) belong to one family and differ in their convexity constraint. By Theorem 1, any difference between them reaches the response through the curvature on $\mathcal{T}$. In the set of $K = 13$ priors that adds regularizers introduced after 2022 to the internal set (random guess baseline 0.077), the fingerprint attributes CRR and WCRR correctly at 0.778 and 0.770, and attributes one as the other in 10.2% and 8.9% of the test images, as shown in Figure 7. The fingerprint is developed on the 9 priors of the internal set only, and the recent set, like the external sets, is used only for evaluation.

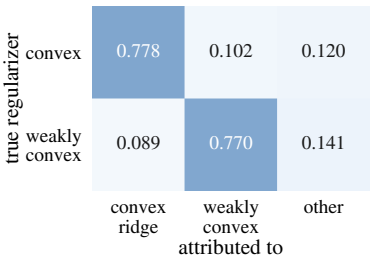


Figure 7: **The fingerprint separates the convex ridge regularizer from its weakly convex counterpart.** Fraction of the test images of CRR (convex ridge) and WCRR (weakly convex) attributed to each of the two or to another candidate prior, in the set of $K = 13$ priors; rows are the true regularizer.

Table 7: **Band-count sensitivity around the fixed $B = 12$ operating point.** The default value was fixed before this auxiliary sweep and is not reselected from it. Mean over the 12 configurations shared by all values; blue marks our fingerprint.

Radial bands B	4	8	16	24
Response Fingerprint (Ours)	$0.888_{\pm 0.010}$	$0.902_{\pm 0.012}$	$0.897_{\pm 0.012}$	$0.891_{\pm 0.010}$

D.5 Partition geometry

Radial pooling is natural for local, convolutional priors but discards orientation, so we compare it with matched 12-dimensional partitions fixed before evaluation, on identical clouds with the paired bootstrap of Section C. A Directional-12 partition ties with Radial-12 ($+0.006$, 95% CI $[-0.005, +0.017]$), including on anisotropic operators, and a Cartesian 12-cell partition is stronger, $0.834 \to 0.862$. Every reported number uses Radial-12, fixed before this comparison, so the reported accuracy is conservative. The Cartesian advantage is not explained by the alignment of the operators with the axes: a $45°$ rotation of the partition retains 67% of the Cartesian improvement, whereas equal-occupancy radial bands recover only 16% of that improvement, and deleting the axis-aligned or all anisotropic operators leaves a positive gap.

D.6 Band count and noise level

These sweeps, reported in Tables 7 and 8, test sensitivity rather than retune the reported method. The band-count experiment evaluates values on both sides of the fixed $B = 12$ choice: $B \in \{4, 8, 16, 24\}$ changes accuracy by 0.013 over 12 configurations. The default value therefore lies inside a broad plateau rather than at a sweep-selected maximum. The noise-level experiment similarly brackets the fixed $\sigma_p = 10^{-2}$ with neighboring noise levels. In the reported experiments, σ_p is 1.07 to 4.5 times the per-entry standard deviation of the acquisition noise across the 9 inverse problems, so a cloud resembles a repeated acquisition. Accuracy increases across the displayed range, from 0.886 at 3×10^{-3} to 0.926 at 10^{-1}. We keep the smaller default because the analysis is local, and larger noise levels trade a lower variance for a larger higher-order bias.

D.7 Responses of four learned priors

Protocol of Figure 3. The four columns of Figure 3 are learned priors of the internal set, colored consistently across panels. The top three rows are CIFAR-10 under Gaussian blur and radial MRI, and a 128×128 phantom under 30% inpainting; each PSNR is the mean over three test images, one of which is shown. The bottom row is the per-pixel standard deviation over $N = 64$ re-solves of that phantom with $\sigma_p = 0.01$, and the fingerprints on the right are computed from the same re-solves.

Table 8: **Noise-level sensitivity around the fixed $\sigma_p = 10^{-2}$ operating point.** The displayed values bracket the default noise level, which is not retuned from this trend, because the analysis holds for small perturbations; blue marks our fingerprint.

Noise level σ_p	3×10^{-3}	3×10^{-2}	10^{-1}
Response Fingerprint (Ours)	$0.886_{\pm 0.023}$	$0.915_{\pm 0.018}$	$0.926_{\pm 0.015}$

Table 9: **Our fingerprint, which needs no gradients, is competitive with a white-box method that optimizes an input through the solver.** Top-1 accuracy of the white-box method of Quan et al. (2023) and of our fingerprint on the operators on which that method is defined, under the original rule of each method and as the maximum over the same 16 decision rules. Blue marks our fingerprint and beige the baseline; in each pair of columns, the higher accuracy is in **bold**.

	Original rule		Maximum over the same 16-rule sweep	
Operator	Quan et al.	Ours	Quan et al.	Ours
denoising, $\boldsymbol{A} = \boldsymbol{I}$	**0.889**	0.648	0.963	**0.981**
$4\times$ super-resolution	**0.889**	0.870	0.981	**1.000**
Gaussian blur	0.796	**0.944**	0.833	**0.963**

D.8 The white-box comparison and the role of the operator

Quan et al. (2023) fingerprint a restoration network by optimizing a critical image through it. The method is white-box: it needs an end-to-end differentiable copy of the reconstruction pipeline and 135,000 backward passes through the solver per critical image, and it is defined only on the 3 of our operators whose forward model is an image-domain degradation. We ran it there with their original defaults, loss, seeding and post-processing, and read our fingerprint at $N = 8$ on the clouds of Section D.2 with shared noise draws, as reported in Table 9.

Their method is reported both under its original rule and as the maximum over sixteen decision-rule variants, and ours under its nearest-centroid Euclidean rule and under the same sweep. Under the original rules, their method is more accurate at $\boldsymbol{A} = \boldsymbol{I}$ and on super-resolution, and ours on Gaussian blur. Under the same sixteen-rule sweep, our unchanged fingerprint is more accurate on all three. Because both sweep maxima are taken over the same test images, they are optimistic for both methods, so we report them beside the original-rule results, and every other number in the paper uses the nearest-centroid rule of Section 3.4. The two methods also differ in access: ours needs N forward solves and no gradient, weights or computational graph.

The invisible fraction of each operator. For each operator we take the singular spectrum of $\boldsymbol{A}$ on a 64×64 grid and compute its *invisible fraction*, the fraction of image-space directions whose singular values are at most 10^{-3} times the largest, including the directions it annihilates, as reported in Table 10. Compared with ModelDiff at the matched budget of Table 1, with ModelDiff choosing its own perturbations at $N = 8$ with independent noise, the ordering at denoising is exact: the smallest margin, $+0.204$, is on denoising, whose invisible fraction is zero, and every other operator's margin is at least 2.9 times larger. Beyond that the relation is loose, with a Spearman correlation of $+0.54$ between margin and invisible fraction over 9 operators, and super-resolution, at the largest invisible fraction, has the same margin as motion blur at 0.0010, so we claim the ordering only. Section F.1 finds the same dependence on the operator from an independent contrast.

Table 10: **The margin is smallest on the one operator with no invisible subspace.** Invisible fraction of each $\boldsymbol{A}$ on a 64×64 grid, namely the fraction of image-space directions with singular values at most 10^{-3} times the largest, and margin of ours over ModelDiff at the matched budget of Table 1.

Operator	Invisible Fraction	Margin
$4\times$ super-resolution	0.938	+0.722
limited-angle Radon	0.812	+0.741
18-angle Radon	0.721	+0.722
inpainting	0.700	+0.593
radial MRI	0.539	+0.630
Cartesian MRI	0.438	+0.648
Gaussian blur	0.046	+0.667
motion blur	0.001	+0.722
denoising	0.000	+0.204

E External priors

E.1 Learned regularizers of a comparative study

Setup. We fingerprint the learned regularizers of the comparative study of Hertrich et al. (2026), with the weights provided by its authors: CRR (Goujon et al., 2023), WCRR (Goujon et al., 2024), ICNN (Amos et al., 2017; Mukherjee et al., 2020), TDV (Kobler et al., 2020), NETT (Li et al., 2020), PatchNR (Altekrüger et al., 2023) and LPN (Fang et al., 2024). No regularizer is retrained. We use their weights, forward operators, data-fidelity terms and solvers as provided, and only wrap the solver so that the perturbation acts on $\boldsymbol{b}$. Before attribution, we verify that their evaluation yields the reconstruction quality reported by the authors, to rule out a gross implementation or setup mismatch as an explanation of attribution errors. The realized relative perturbation is 1.00% of $\|\boldsymbol{b}\|$ and every denoising solve converged (minimum convergence fraction 1.00). References use 8 images and tests a disjoint 16, giving 112 trials; intervals resample test images, the dependence unit.

Denoising. Macro attribution is 0.589 (95% CI $[0.464, 0.723]$) compared with a random guess baseline of 0.143, AUC 0.896. Adding EPLL (Zoran & Weiss, 2011) gives the full set of 8 candidate priors, which attributes at 0.578 compared with a random guess baseline of 0.125 with $N = 8$ and at 0.656 with $N = 64$ (AUC 0.910 and 0.935). CRR and WCRR, the closest architectural pair, are confused 4 times in total over 16 test images per direction, so relaxing convexity to weak convexity moves the fingerprint by less than the distance to an unrelated architecture, but not by zero.

CT. The same set on the authors' 60-angle tomography at 362^2 has $30{,}720$ measurements for $131{,}044$ unknowns ($m/n = 0.234$), so $\mathrm{Null}(\boldsymbol{A})$ has dimension at least $100{,}324$. All 7 candidate priors reproduce the authors' CT evaluation first; LPN is not in the CT set because its CT pipeline is a plug-and-play ADMM solver without a one-step checkpoint. Attribution is 0.661 compared with a random guess baseline of 0.143 (AUC 0.925; 95% CI $[0.589, 0.723]$ from $20{,}000$ image resamples) and 0.670 at $N = 64$. Table 11 gives the breakdown; the WCRR confusion is analyzed in Section E.3.

The reliability indicator on each set. The indicator ρ_k of Equation (11), which requires no ground truth, tracks the accuracy of each candidate prior on every set: Pearson -0.891 on CT ($K = 7$), -0.594 on denoising ($K = 7$), -0.776 on real MRI ($K = 10$), Spearman -0.786, -0.821 and -0.782; all 6 correlations over 24 candidate priors are negative. On CT, every candidate prior with $\rho_k > 1$ stays at or below 0.688 accuracy, and every candidate prior with $\rho_k < 0.15$ reaches 1.000, while on denoising the indicator ranks the candidate priors without a common threshold, so a threshold is calibrated per problem. It therefore indicates, before an attribution is trusted, which candidate priors the response separates poorly.

E.2 Solver noise and local regularity

Attribution does not require Assumption A1, whereas reading the fingerprint as the curvature of a regularizer does. The 4 external priors CRR, WCRR, ICNN and TDV come closest to Assumption A1: their solvers are deterministic and reach a stationary point of Equation (2), and their regularizers

Table 11: **Under a rank-deficient CT operator, the candidate priors that reconstruct best are the hardest to attribute.** External priors under a rank-deficient operator (CT, 60 angles). PSNR follows the evaluation of the authors; accuracy is the top-1 accuracy of each candidate prior over 16 held-out images, compared with a random guess baseline of 0.143. The most accurately attributed candidate prior is not the most accurate reconstruction.

Candidate Prior	PSNR (dB)	Attribution Accuracy
CRR	32.30	1.000
ICNN	32.16	1.000
NETT	32.01	1.000
EPLL	32.55	0.688
PatchNR	32.64	0.562
TDV	33.67	0.312
WCRR	32.85	0.062
Macro mean	32.6	0.661

are twice differentiable almost everywhere, since CRR, WCRR and ICNN have piecewise-constant second derivatives and TDV is smooth. We do not certify $\boldsymbol{G} \succ 0$. Within these 4 candidate priors, accuracy spans 0.312 to 0.625, wider than any gap between this group and the rest, and, restricted to these 4 candidate priors, the fingerprint attributes at 0.703 on CT (95% CI $[0.609, 0.797]$) and 0.547 on denoising ($[0.406, 0.703]$), compared with a random guess baseline of 0.25. The fingerprint therefore remains discriminative on the subset that most closely satisfies the assumptions underlying the curvature interpretation.

Removing solver noise. Because $\mathrm{Cov}[\hat{\boldsymbol{x}}] = \sigma_p^2 \boldsymbol{J}\boldsymbol{J}^\top + \mathrm{Cov}[\boldsymbol{\eta}]$ to first order, for solver noise $\boldsymbol{\eta}$ independent of $\boldsymbol{\xi}$, and band energies are linear in the covariance, a *null cloud* of re-solves of the unperturbed measurement ($\sigma_p = 0$) estimates $\mathrm{Cov}[\boldsymbol{\eta}]$ directly, and subtracting its band energies gives an unbiased estimate of the response alone. Each null cloud holds 8 re-solves of the same unperturbed measurement and is subtracted band by band from the perturbed cloud of that image, and every corrected band energy is positive in every cloud of the three sets, so the fingerprint is defined throughout. Additivity holds per band for every candidate prior (0 violations over 7 candidate priors). The objective of PatchNR is stochastic by design, because it resamples 10^4 random patches per gradient call, and 0.544 of its measured band energy is not response; removing it takes PatchNR from 0.938 to 0.750. NETT, whose non-determinism is solver-level at 0.035, is unchanged ($1.000 \to 1.000$), as are the remaining denoising candidate priors up to a single test image, all 6 candidate priors on the CT set ($+0.0000$), and all 10 deterministic networks on the MRI set ($+0.0000$, with a ratio of null to perturbed band energy of at most 0.000000). On the denoising set, the correction lowers the macro accuracy from 0.643 to 0.607, mostly through PatchNR.

E.3 The convex and weakly convex pair on CT

The CT macro number averages over candidate priors that behave very differently, as reported in Table 11: CRR, ICNN and NETT are attributed at 1.000, while WCRR is attributed at 0.062. The loss in accuracy is dominated by a one-directional confusion: WCRR is attributed as CRR on 10 of 16 images, whereas CRR is never attributed as WCRR. The two failure modes of Section 3.4 identify its form. The WCRR and CRR centroids are 5.781 apart, compared with a median pairwise separation of 4.245, further apart than a typical pair, while the test fingerprints of WCRR scatter by 3.474 across a gap of 2.120 to the centroid of EPLL, the candidate prior nearest to them on average, so that $\rho_k = 1.64 > 1$ indicates that WCRR is poorly separated, and the confusion is a dispersion rather than a collapse of weak convexity onto convexity. It is not estimator variance, because WCRR attributes identically at $N = 8$ and $N = 64$.

A larger perturbation with converged re-solves removes the confusion. We raised the relative perturbation from 1% to 2% and 4% of $\|\boldsymbol{b}\|$ on the same 8 reference and 16 test images, and raised the iteration budget of each (candidate prior, perturbation) cell until at least 90% of its re-solves converged; every cell of the six deterministic candidate priors met this criterion, with a smallest converged fraction of 0.958 and up to $48{,}000$ iterations for WCRR at 4%. The set of the six

deterministic candidate priors then attributes at 0.677, 0.938 and 1.000 compared with a random guess baseline of 0.167, and WCRR at 0.188, 1.000 and 1.000, correct on all 16 test images at 4%. PatchNR is excluded from this set because its stochastic objective does not meet the convergence criterion at any budget we tried (at most 0.50 of its re-solves converge, at up to 48,000 iterations), so we treat this analysis of six candidate priors as secondary. The confusion is therefore a property of this pair on this operator at the default noise level, and it disappears once the perturbation is large enough and every re-solve converges.

E.4 MRI NETWORKS ON REAL MEASUREMENTS

This set tests attribution among accurate solvers on real measurements.

Set and protocol. We use 10 pretrained multi-coil brain-MRI reconstruction networks that are available as one family in a single framework (DIRECT (Yiasemis et al., 2022a), Calgary-Campinas (Souza et al., 2018)): RecurrentVarNet (Yiasemis et al., 2022b), LPDNet (Adler & Öktem, 2018), IterDualNet, which unrolls gradient descent with image- and Fourier-domain networks, ConjGradNet, which alternates a learned denoiser with conjugate-gradient data consistency (Shewchuk, 1994), RIM (Putzky & Welling, 2017; Lønning et al., 2019), VarNet (Sriram et al., 2020), Joint-ICNet (Jun et al., 2021), XPDNet (Ramzi et al., 2020), KIKI-net (Eo et al., 2018), and U-Net (Ronneberger et al., 2015). All ten run at the 5× rate with the challenge's own $R = 5$ undersampling masks, the setting in which they were trained, on slices at their native size (218×170 and 218×180), with no resizing or cropping. Acceleration, mask set, coil-sensitivity estimation, normalization and FFT convention are shared by all networks. We verified before evaluation that all configurations use the same physics and inference settings. Splits are at the volume level and grouped by exam identifier, so no subject appears on both sides (14 volumes, 13 exam groups). The perturbations act on the measured Fourier coefficients at the sampled locations, and no network is retrained. We compute fingerprints on magnitude images, the modulus of the coil-combined output of each network, rescaled per scene. The magnitude image lies outside the domain of $\boldsymbol{A}$, so these are attribution results, outside the reading of Section 4.

Reconstruction quality. To verify that every network reconstructs these measurements well, we compare it with the fully sampled reference and with *zero-filling*, namely the inverse Fourier transform of the measurement with the unsampled coefficients set to zero. All 10 networks are more accurate than zero-filling in PSNR and SSIM (Wang et al., 2004), with 29.6 to 36.1 dB, compared with 25.4 dB, so attribution on this set is among accurate solvers.

Result. At $N = 8$, verification of the claimed candidate prior reaches 0.946 AUC and closed-set attribution 0.680 (95% CI $[0.625, 0.725]$ over 8 volume clusters) compared with a random guess baseline of 0.100; $N = 16$ is unchanged, so the budget of re-solves does not limit the accuracy.

Accuracy of the networks and attribution. Attribution accuracy decreases as reconstruction quality increases. Networks with lower reconstruction accuracy are attributed at up to 1.000, while the most accurate family, the unrolled data-consistency networks, sits near 0.531 and exchanges errors almost exclusively within itself. This is in line with Proposition 4, which states that solvers close to a common reference have close responses, although the networks of that family also share an architecture, which may bring their responses closer as well, as detailed in Section E.5.

E.5 ACCURACY AND IDENTIFIABILITY

On all three external sets, more accurate candidate priors are harder to attribute. We now show in Proposition 4 why this is expected for any method that reads the behavior of the solver: two solvers uniformly close to a common reference solver have close Jacobians, and hence close fingerprints.

Proposition 4 (Solvers close to a common reference have close fingerprints; based on Landau, 1914)**.** *Let $\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2$ be C^2 solvers within δ of a common reference solver $\boldsymbol{x}^{\mathrm{ref}}$, such as the Bayes estimator, on the ball of radius τ around $\mathbf{b}_0$, with second derivatives bounded by β there. If $\tau \geq 2\sqrt{\delta/\beta}$, their Jacobians at $\mathbf{b}_0$ satisfy $\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|_2 \leq 4\sqrt{\delta\beta}$. If, moreover, $\|\boldsymbol{J}_i\|_2 \leq L$ and every band energy of both solvers is at least $e_{\min} > 0$, their fingerprints computed from $\boldsymbol{C}_1 = \sigma_p^2 \boldsymbol{J}_1 \boldsymbol{J}_1^\top$ and $\boldsymbol{C}_2 = \sigma_p^2 \boldsymbol{J}_2 \boldsymbol{J}_2^\top$ satisfy $\|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|_\infty \leq 16\, \sigma_p^2 L\, d_{\max} \sqrt{\delta\beta}/e_{\min}$, where $d_{\max}$ is the largest band dimension.*

Table 12: **The inversion between quality and identifiability follows a two-link chain on three external sets.** Link 1 relates reconstruction error to fingerprint isolation, the mean fingerprint distance of a candidate prior to the rest of its set; link 2 relates isolation to attribution accuracy. Proposition 4 states the bound behind link 1.

		Link 1: error → isolation		Link 2: isolation → accuracy	
Set	K	Pearson	Spearman	Pearson	Spearman
real MRI	10	+0.729	+0.588	+0.770	+0.745
external denoising	7	+0.766	+0.500	+0.906	+0.821
external CT	6	+0.586	+0.771	+0.580	+0.600

Proof. Let $\boldsymbol{d} = \hat{\boldsymbol{x}}_1 - \hat{\boldsymbol{x}}_2$, so $\|\boldsymbol{d}\|_2 \le 2\delta$ and $\|\nabla^2 \boldsymbol{d}\| \le 2\beta$ on the ball. For unit vectors $\boldsymbol{u}, \boldsymbol{v}$, the scalar function $q(t) = \langle \boldsymbol{u}, \boldsymbol{d}(\boldsymbol{b}_0 + t\boldsymbol{v})\rangle$ satisfies $|q| \le 2\delta$ and $|q''| \le 2\beta$, so Taylor's theorem gives $|\kappa|\,|q'(0)| \le |q(\kappa)| + |q(0)| + \frac{\kappa^2}{2}\sup|q''| \le 4\delta + \kappa^2\beta$ for every step $|\kappa| \le \tau$. Choosing $|\kappa| = 2\sqrt{\delta/\beta}$, which the radius condition allows, gives $|q'(0)| \le 4\sqrt{\delta\beta}$, and the supremum over $\boldsymbol{u}, \boldsymbol{v}$ bounds $\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|_2$ (Landau, 1914). Then $\|\boldsymbol{C}_1 - \boldsymbol{C}_2\|_2 \le \sigma_p^2(\|\boldsymbol{J}_1\|_2 + \|\boldsymbol{J}_2\|_2)\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|_2$, each band energy changes by at most $d_b\|\boldsymbol{C}_1 - \boldsymbol{C}_2\|_2$, each log energy by at most this amount divided by $e_{\min}$, and centering at most doubles the sup-norm. □

Outputs of two solvers within δ of a common reference differ by at most 2δ, whereas the bound on their Jacobians is of order $\sqrt{\delta}$, so for small δ the bound permits a larger separation of the responses than of the outputs. The bound concerns reconstruction error, and not data fit: two solvers that fit the measurement equally well can still fill in $\mathrm{Null}(\boldsymbol{A})$ differently, which is the favorable regime for attribution. It needs no assumption on $\boldsymbol{A}$, so it also applies to denoising.

Measurement. The learned regularizers of the comparative study span a narrow quality range, 32.01 to 33.67 dB mean PSNR on CT and 27.53 to 29.67 dB on denoising, and the MRI reconstruction networks span 29.6 to 36.1 dB. Mean PSNR measures mean-square closeness to the reference, whereas Proposition 4 assumes uniform closeness on a ball, so PSNR serves as an empirical proxy for the closeness condition in Proposition 4. The chain has two separately measurable links: reconstruction error → *fingerprint isolation* (the mean fingerprint distance of a candidate prior to the rest of its set) → attribution accuracy. Over 3 sets and 23 candidate priors, both links are positive on every set under both Pearson and Spearman correlation (12 of 12 coefficients), with link 1 Pearson 0.586 to 0.766 and link 2 Pearson 0.580 to 0.906, as reported in Table 12. The sets share no data, framework, architecture family or operator, and the operator of the denoising set has no null space. Dropping the lowest-PSNR candidate prior of each set leaves both links positive on all three sets (Spearman at least 0.20). With six to ten candidate priors over a narrow error range, the rate of the bound is not estimable, so we report the sign. In Figure 9(b), the CT set has $K = 6$ candidate priors and per-image PSNR.

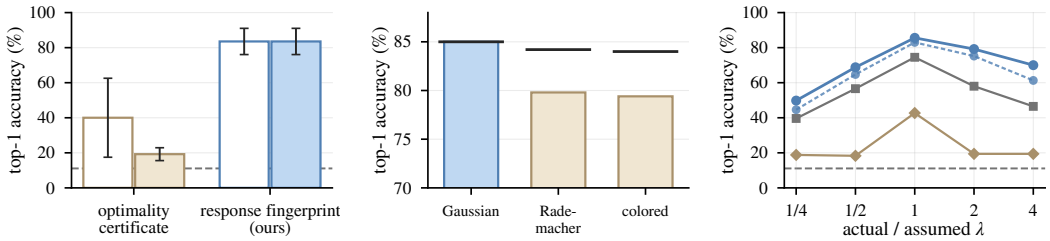
F Robustness to acquisition conditions

This section supports Section 5.4 and Figure 8.

Protocol of Figure 8. Each panel of Figure 8 changes one factor and keeps the rest of the protocol fixed. **(a)** The exact operator compared with the worst tested operator error, mean over the 41 configurations of Section F.2 with standard-deviation error bars. **(b)** Test-time perturbation laws with matched energy, with references built under Gaussian perturbations ($K = 9$, $N = 64$, 9 inverse problems, 8 for the colored laws); black ticks mark the ceiling of each law, as detailed in Section F.1. **(c)** Actual over assumed λ over 29 configurations, as detailed in Section F.3.

F.1 Transfer across the perturbation law

If attribution succeeded only under the perturbation law used for the references, the fingerprint would describe our perturbations rather than the solver. To first order, the re-solves move by $\boldsymbol{J}\boldsymbol{\xi}$ for any mean-zero $\boldsymbol{\xi}$, so the law enters the response only through its covariance. A law with the same



(a) error in $\boldsymbol{A}$ (b) perturbation law (c) regularization strength

Figure 8: **The fingerprint is robust to acquisition conditions that are unknown in practice. (a)** Accuracy with the exact operator (open bars) and under the worst tested error in the operator used for attribution (filled bars). **(b)** Accuracy under test-time perturbation laws of matched energy, with references from Gaussian perturbations; black ticks mark the accuracy with references of the same law. **(c)** Accuracy versus the ratio of the actual to the assumed λ for our fingerprint (solid blue), its uncentered ablation (dotted blue), $\mathrm{tr}\,\boldsymbol{C}$ (gray) and the optimality certificate (beige). Blue marks ours, and beige the certificate or a changed perturbation law; dashed lines mark the random guess baseline.

Table 13: **The fingerprint transfers across perturbation laws.** Accuracy with references from Gaussian perturbations, the accuracy with references under the test law, and their difference. The whitened row is a consistency check, whose perturbations equal those of the reference law.

Test Law	Cross-law ↑	Same-law ceiling	Transfer cost ↓
Gaussian (Reference Law)	0.850	0.850	+0.000
Rademacher	0.798	0.842	+0.043
Colored	0.794	0.840	+0.046
Colored, Whitened (Check)	0.838	0.838	+0.000

covariance but a different shape (Rademacher) therefore has the same first-order response, whereas a law with a different covariance (colored Gaussian $\mathcal{N}(0, \sigma_p^2\boldsymbol{\Sigma})$) changes the response, so its transfer is not implied by the analysis. As a consistency check of the implementation, we also whiten the colored perturbations with the known $\boldsymbol{\Sigma}^{-1/2}$ before each re-solve, which returns the white Gaussian perturbations of the reference law, drawn with the same seed.

Protocol. We build the references of every candidate prior under isotropic Gaussian perturbations and attribute under each law, over 9 operators, $K = 9$, $N = 64$, and a random guess baseline of 0.111. All laws are RMS-matched. Cross-law accuracy is compared with the same-law ceiling of that law, which isolates the cost of the law *change*. $\boldsymbol{\Sigma}$ is Fourier-diagonal (condition number 10, geometric spectrum), which gives exact $\boldsymbol{\Sigma}^{\pm 1/2}$ at measurement dimension $O(10^4)$; the colored laws exclude inpainting, whose measurements are not an image, leaving 8 operators.

As we show in Table 13, the fingerprint transfers to both laws: Rademacher and colored perturbations are attributed within $+0.043$ and $+0.046$ of the accuracy obtained when the references use the same law, and the whitened check reproduces its own same-law accuracy. The Gaussian reference law covers 9 operators, and the colored laws cover the 8 operators without inpainting. At the fingerprint level, a colored law shifts the fingerprint of a candidate prior by 0.019 of the separation between candidate priors. Under the reference law, attribution is 0.882 averaged over the operators with a non-trivial null space and 0.593 on denoising, providing an independent replication of the operator dependence in Figure 9(a).

F.2 Operator misspecification, per operator

Table 14 gives the full grid of operator errors; Figure 8(a) shows the mean over its non-vacuous rows.

Table 14: **An error in the operator used for attribution leaves our fingerprint unchanged on each tested configuration, while the accuracy of the certificate decreases toward the random guess baseline.** Only the copy of $\boldsymbol{A}$ used for attribution is perturbed; $\epsilon_{\boldsymbol{A}}$ is its relative error, and "worst" is the lowest accuracy over the tested errors. STL-10 appears at 64^2 and at 96^2, with one row per operator at each resolution. In each row, the higher value of each pair of columns is in **bold**. Blue marks our fingerprint and beige the certificate.

Data	Operator	Random Guess	Certificate		Response Fingerprint (Ours)	
			$\epsilon_{\boldsymbol{A}}=0$	Worst	$\epsilon_{\boldsymbol{A}}=0$	Worst
CIFAR-10	Gaussian blur	0.111	0.620	0.213	**0.824**	**0.824**
CIFAR-10	motion blur	0.111	0.472	0.204	**0.880**	**0.880**
CIFAR-10	inpainting	0.111	0.741	0.222	**0.815**	**0.815**
CIFAR-10	Cartesian MRI	0.111	0.213	0.213	**0.759**	**0.759**
CIFAR-10	radial MRI	0.111	0.213	0.213	**0.769**	**0.769**
CIFAR-10	limited-angle Radon	0.111	0.222	0.222	**0.843**	**0.843**
CIFAR-10	18-angle Radon	0.111	0.231	0.222	**0.824**	**0.824**
CIFAR-10	4× super-resolution	0.111	0.556	0.213	**0.880**	**0.880**
MNIST	Gaussian blur	0.111	0.241	0.222	**0.889**	**0.889**
MNIST	motion blur	0.111	0.546	0.222	**0.963**	**0.963**
MNIST	inpainting	0.111	0.796	0.222	**0.981**	**0.981**
MNIST	Cartesian MRI	0.111	0.222	0.222	**0.852**	**0.852**
MNIST	radial MRI	0.111	0.222	0.222	**0.824**	**0.824**
MNIST	18-angle Radon	0.111	0.296	0.222	**0.833**	**0.833**
MNIST	4× super-resolution	0.111	0.444	0.222	**0.991**	**0.991**
OrganAMNIST	Gaussian blur	0.111	0.602	0.204	**0.889**	**0.889**
OrganAMNIST	motion blur	0.111	0.491	0.222	**0.907**	**0.907**
OrganAMNIST	inpainting	0.111	**0.833**	0.222	0.713	**0.713**
OrganAMNIST	Cartesian MRI	0.111	0.204	0.204	**0.750**	**0.750**
OrganAMNIST	radial MRI	0.111	0.167	0.167	**0.796**	**0.796**
OrganAMNIST	18-angle Radon	0.111	0.222	0.222	**0.824**	**0.824**
OrganAMNIST	4× super-resolution	0.111	0.546	0.222	**0.926**	**0.926**
phantom	Gaussian blur	0.111	0.176	0.120	**0.926**	**0.926**
phantom	inpainting	0.111	**0.889**	0.222	**0.889**	**0.889**
phantom	Cartesian MRI	0.111	0.120	0.120	**0.806**	**0.806**
phantom	radial MRI	0.111	0.120	0.120	**0.704**	**0.704**
phantom	4× super-resolution	0.111	0.380	0.120	**0.981**	**0.981**
STL-10	Gaussian blur	0.111	0.583	0.148	**0.759**	**0.759**
STL-10	Gaussian blur	0.111	0.602	0.185	**0.880**	**0.880**
STL-10	motion blur	0.111	0.546	0.139	**0.870**	**0.870**
STL-10	motion blur	0.111	0.417	0.194	**0.833**	**0.833**
STL-10	inpainting	0.111	**0.731**	0.176	0.685	**0.685**
STL-10	inpainting	0.111	0.704	0.204	**0.787**	**0.787**
STL-10	Cartesian MRI	0.111	0.139	0.139	**0.750**	**0.750**
STL-10	Cartesian MRI	0.111	0.185	0.185	**0.778**	**0.778**
STL-10	radial MRI	0.111	0.139	0.139	**0.787**	**0.787**
STL-10	radial MRI	0.111	0.176	0.176	**0.806**	**0.806**
STL-10	18-angle Radon	0.111	0.241	0.222	**0.806**	**0.806**
STL-10	18-angle Radon	0.111	0.222	0.222	**0.824**	**0.824**
STL-10	4× super-resolution	0.111	0.444	0.130	**0.778**	**0.778**
STL-10	4× super-resolution	0.111	0.491	0.185	**0.870**	**0.870**

F.3 Regularization strength: actual vs assumed

Table 15 reports the sweep of Figure 8(c). References are built once at the assumed λ; test reconstructions come from the same candidate prior at λ scaled by the column factor. The certificate is scored with the assumed λ, which is the information available at attribution. "Retention" is the worst accuracy on the grid divided by the matched-λ accuracy.

Table 15: **Centered band energies retain more accuracy than the compared methods under a mismatched regularization strength.** Accuracy under a mismatch between the actual and assumed regularization strength, mean over 29 configurations, with a random guess baseline of 0.111. In each column, the highest value is in **bold** and the second highest is underlined. Blue marks our fingerprint, beige the baselines, and unshaded rows are its ablations.

Method	Actual λ / assumed λ					Retention
	0.25	0.5	1	2	4	
Optimality Certificate	0.188	0.183	0.427	0.194	0.194	0.430
Low-Rank Covariance, Learned Basis	0.151	0.181	0.336	0.450	0.428	0.450
Total Response $\mathrm{tr}\,\boldsymbol{C}$	0.396	0.566	0.745	0.580	0.465	0.531
Uncentered Band Energies	0.446	0.647	0.830	0.752	0.613	0.537
Response Fingerprint (Ours)	**0.498**	**0.688**	**0.856**	**0.792**	**0.700**	**0.582**

G WHEN ATTRIBUTION IS RELIABLE

This section supports the reliability indicator of Section 3.4 and complements Section 5. By Theorem 1, the response carries the curvature on $\mathcal{T}$, and Equation (10) separates two candidate priors when their fingerprints differ by more than their scatter. Two factors set this difference.

The forward operator. When $\boldsymbol{A}$ has a nontrivial null space, as in inpainting and undersampled MRI, the part of the response in $\mathrm{Null}(\boldsymbol{A})$ is set by the curvature of the regularizer, as Section 4 shows, whereas at $\boldsymbol{A} = \boldsymbol{I}$ the Jacobian $(\boldsymbol{I} + \lambda \boldsymbol{H})^{-1}$ is close to the identity whenever $\lambda \boldsymbol{H}$ is small. The advantage of the response is therefore expected to be smallest at $\boldsymbol{A} = \boldsymbol{I}$.

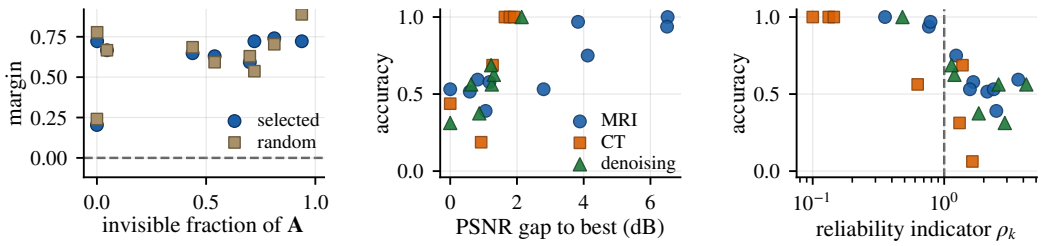
The accuracy of the candidate priors. Solvers that are uniformly close to a common reference solver, such as the Bayes estimator, have close Jacobians and hence close fingerprints, at rate $O(\sqrt{\delta})$ in their reconstruction error δ, as Proposition 4 shows (Landau, 1914).

Both factors set the gap g_k of the reliability indicator, so ρ_k indicates the regimes in which it is small. As we show in Figure 9, the advantage of the response is smallest at denoising, accurate candidate priors are harder to separate, and ρ_k tracks the accuracy of each candidate prior.

The advantage of the response is smallest at denoising. As we show in Figure 9(a), our fingerprint outperforms ModelDiff on each of the 9 operators, and its margin, $+0.204$ at denoising, is at least 2.9 times larger once the invisible fraction of $\boldsymbol{A}$, namely the fraction of image-space directions with singular values at most 10^{-3} times the largest, is nonzero. Consistent with the first factor, the white-box method of Quan et al. (2023), with $135{,}000$ backward passes per image, is more accurate at $\boldsymbol{A} = \boldsymbol{I}$, as detailed in Section D.8.

Accurate candidate priors are harder to separate. As we show in Figure 9(b) and Table 12, accuracy increases with the PSNR gap to the most accurate candidate prior, in line with Proposition 4.

The reliability indicator correlates with candidate-prior attribution accuracy. As we show in Figure 9(c), ρ_k tracks this accuracy, with correlations of -0.891, -0.776 and -0.594 on CT, MRI and denoising; on CT, candidate priors with $\rho_k > 1$ reach at most 0.688, while those with $\rho_k < 0.15$ reach 1.000.



(a) operator (b) reconstruction quality (c) reliability

Figure 9: **Attribution is harder at $A = I$ and for accurate candidate priors, and the indicator tracks both. (a)** Margin over ModelDiff, with selected (blue) or random (beige) perturbations, as a function of the invisible fraction of A. **(b, c)** Accuracy of each candidate prior as a function of the PSNR gap to the most accurate candidate prior and of ρ_k for the MRI (circles), CT (squares) and denoising (triangles) sets. Dashed lines mark zero margin and $\rho_k = 1$.

H Further related work

This section complements Section 2.

Identifying an objective from its solver. In denoising, a proximal map determines a convex potential up to a constant (Moreau, 1965), and so does its norm (Vilches, 2021), and tests of Jacobian symmetry decide whether a denoiser corresponds to an explicit objective (Reehorst & Schniter, 2019; Gribonval & Nikolova, 2020; Hurault et al., 2022). Inverse optimization (Keshavarz et al., 2011; Aswani et al., 2018), bilevel learning (Kunisch & Pock, 2013), and potential estimation (Haber et al., 2023) infer objectives within a specified family. These approaches rely on global or chosen access to the solver or on a parametric family, while we use random re-solves at a single measurement.

Model fingerprinting and attribution. Beyond the fingerprints of classifiers and restoration networks discussed in Section 1, generated images are traced back to their generator (Yu et al., 2019; Song & Itti, 2025). These fingerprints address a different task and read the outputs of a model.

I Reproducibility

All methods are implemented in Python (Van Rossum & Drake, 2009) with PyTorch (Paszke et al., 2019). Each experiment is specified by its forward operators, set of candidate priors, image distributions, disjoint reference and test splits, budget of re-solves, perturbation law and noise level, solver and iteration budget, and decision rule; these are stated in the corresponding appendix sections, together with their protocols and decision rules. The external regularizers (Hertrich et al., 2026) and MRI reconstruction networks (Yiasemis et al., 2022a) are used with the weights provided by their authors, without retraining. The datasets are public: Calgary-Campinas (Souza et al., 2018) and LoDoPaB-CT (Leuschner et al., 2021) require registration with their providers, and the knee volumes are openly licensed (Ong et al., 2018). The code will be published upon acceptance.