LEARNED INTENSITY SENSORS TRADE SEPARABILITY ACROSS ORIENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned sensor is frozen in hardware, so whatever it fails to record is lost to every decoder. We ask what such a sensor trades away, for intensity sensors, which record squared magnitudes of a coherent field under additive Gaussian noise. How well one separates two draws from a signal distribution, measured by the average Kullback-Leibler divergence between their measurement laws, is an inner product of two fourth moments, one of the sensor and one of the data, and equals the variance the data induce in its noiseless measurements. Averaged over rotations of the distribution, this separability is the same for every sensor with m unit-norm measurements: learning redistributes it and cannot create it, although on one distribution a sensor may hold far more than isotropy. The excess is an exact overlap with the sensor's fourth-order defect, which the distance to a complex projective 2-design bounds only uniformly over a distribution class, not on the task at hand. Under a controlled rotation, a learned sensor that beats a quadratic-mask design by 2.6 dB in distribution loses to it by 0.8 dB at 90° with retrained decoders, and a decoder-free overlap ranks this loss across 12 configurations. In a matched family where only the fourth-order direction varies, the effect is graded with the alignment but small, and its sign follows the local Cramér-Rao bound rather than the separation. Two boundaries are as informative: the ordering reverses at low noise, and in coded diffraction imaging of phase objects no learned mask set loses more than 0.11 dB of separability.

1 INTRODUCTION

End-to-end learning now designs the sensor together with its decoder: phase masks for coded diffraction imaging (Jerez et al., 2024), illumination for phase microscopy (Kellman et al., 2019), MRI sampling masks (Bahadir et al., 2020; Huijben et al., 2020) and diffractive optics (Sitzmann et al., 2018; Ho et al., 2025) are trained jointly with the reconstruction network, on the premise that a sensor adapted to the data records more of what the decoder needs. That premise hides a risk, because deployed systems routinely meet data that differ from their training set (Koh et al., 2021). A decoder can then be retrained or adapted at test time (Darestani et al., 2022), but a sensor is frozen in hardware: whatever it failed to record is lost to every decoder. Robustness studies of learned reconstruction have examined the decoder (Antun et al., 2020; Darestani et al., 2021) and priors for fixed forward models (Zheng et al., 2025). *What, then, does a learned sensor itself trade away, and can the trade be measured before deployment, without a decoder?*

We answer this for *intensity sensors*, which record the squared magnitude $|\langle a_i, x\rangle|^2$ of a coherent field under additive Gaussian noise with unit-norm rows: the idealized model of coded diffraction imaging and ptychography. The answer rests on one identity. How well such a sensor tells apart two draws from a signal distribution $\mathcal{D}$, measured by the average Kullback-Leibler divergence between their measurement laws, is an inner product $\langle \Phi_4(A), M(\mathcal{D})\rangle$ of the fourth moment of the measurement vectors with that of the data, and equals the variance the data induce in the noiseless measurements. We call it the *separation*. Averaged over rotations of $\mathcal{D}$ it is the same for every sensor with m such measurements (Theorem 1): learning *redistributes* separability across orientations and cannot create it. On any one distribution a sensor may hold far more than isotropy, and the excess is the exact overlap $\langle T(A), M(\mathcal{D})\rangle$ with its fourth-order defect $T(A)$, which the distance to a complex projective 2-design bounds uniformly over a class of distributions but not on the task at hand (Corollary 1). The mathematics is a Haar twirl and Cauchy-Schwarz; what is new is the object it applies to, a physical

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.

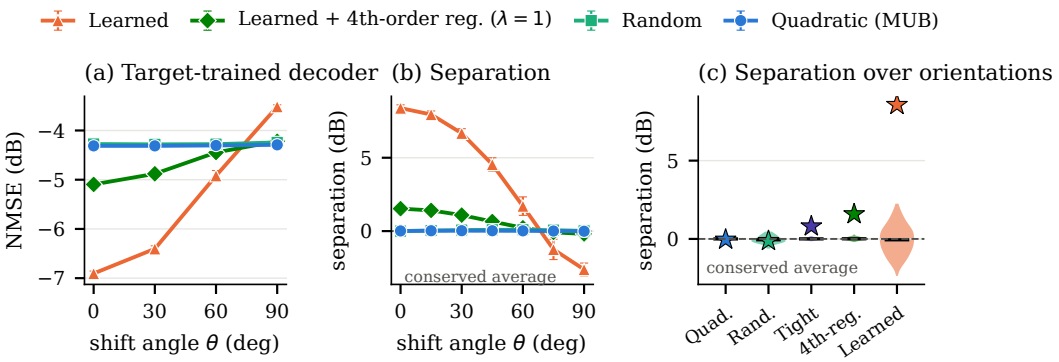


Figure 1: **A learned intensity sensor bets on the orientation of its training data.** (a) With decoders retrained on each shifted distribution (target-trained), the learned sensor wins in distribution ($\theta = 0$) but loses to the quadratic design (four mutually unbiased chirp bases, MUB) at $90°$, while the fixed designs are flat. (b) The decoder-free separation relative to its orientation average, the same number for all sensors by Theorem 1. (c) Separation over 150 Haar orientations (seed 0; "Tight": the tight-frame-penalty sensor of Table 1; 4th-reg.: $\lambda = 1$); stars: training orientation, bars: median. Error bars: 95% t-intervals over 5 seeds.

operator frozen at deployment while its data may rotate, whose adaptation is therefore zero-sum, measurable in $O(mnd)$ time before any decoder exists, and governed by the *direction* of its defect rather than by the distance to isotropy.

We test this with two sensor-level measurements independent of the deployed decoder: a *target-trained* decoder retrained on the shifted data with the sensor frozen, and the decoder-free statistics themselves. Under a controlled rotation of the dominant signal subspace at moderate noise, a learned sensor that beats a quadratic-mask design by 2.6 dB in distribution loses to it by 0.8 dB at $90°$ even with the target-trained decoder, while fixed designs stay flat (Figure 1); the separation bet, computed before any decoder is trained, ranks that loss across 12 configurations (Spearman 0.99), and the design-defect penalty moves it monotonically from 8.58 to 0.83 dB. A matched family isolates the mechanism: in exact tight frames whose second-order geometry and design defect are frozen by construction, so that only the fourth-order direction varies, reconstruction moves with the alignment and only with it (-0.00 to -0.30 dB as three of four bases are adapted to the training subspace), but in the direction the local bound predicts, opposite to the separation. We are equally explicit about where the account stops. At low noise the ordering reverses and the learned sensor wins at $90°$ by up to 2.50 dB, so the conservation law constrains sensing geometry and not the risk of a prior-driven decoder; and in coded diffraction imaging of natural-image phase objects no learned mask set loses more than 0.11 dB of separability under an exact rotation, while a penalty that lowers the mask-dependent flatness excess 7.5-fold, moving the design defect by only 0.11%, raises the alignment with the data and worsens the phase-object bound.

Contributions.

- **An exact principle.** The orientation average of the local Fisher form and of the pairwise KL separation is the same for every unit-norm intensity sensor, and 2-designs are exactly the sensors with no orientation dependence; they also minimize the orientation-averaged Cramér-Rao criterion, the intensity-sensing counterpart of Scott (2006) (Theorems 1, 2).
- **An exact mechanism.** The deviation from isotropy on a given distribution is the alignment $\langle T(A), M(\mathcal{D})\rangle$, while the design defect controls only a distribution-uniform envelope; we measure how loose that envelope is (0.00 to 0.25 of its ceiling here, 7×10^{-4} of it in coded diffraction) and show a case where lowering the defect raises the alignment (Corollary 1).
- **Controlled tests.** A matched family with everything but the fourth-order direction frozen makes the alignment the only free variable: its effect on reconstruction is graded, null without alignment, but signed by the local bound; and target-trained decoders follow the sensor ordering at moderate noise and reverse it at low noise (Sections 5.2 to 5.5).

2 Related work

Learned sensing, and robustness of learned reconstruction. Jointly optimizing a physical encoder and a digital decoder is standard in computational imaging and is surveyed for optical coding by Arguello et al. (2023): diffractive optics (Sitzmann et al., 2018; Chang & Wetzstein, 2019), wave-optics camera design (Ho et al., 2025), coded illumination for phase microscopy (Kellman et al., 2019), learned sampling for compressive and MRI acquisition (Baldassarre et al., 2016; Bahadir et al., 2020; Huijben et al., 2020), learned measurement matrices (Wu et al., 2019), once designed by minimizing mutual coherence (Elad, 2007), coded diffraction masks (Jerez et al., 2024), and optical convolutional front-ends (Xiang et al., 2025). These optimize accuracy on the training distribution, seeking robustness to fabrication and alignment error by perturbation training (Mengu et al., 2020) and transfer by retraining the back-end (Choi et al., 2025). On the reconstruction side, deep decoders can be unstable to perturbations (Antun et al., 2020) and degrade under natural shifts like classical methods tuned to a dataset (Darestani et al., 2021); test-time training closes much of the gap (Darestani et al., 2022), and diffusion priors or learned denoisers act as plug-and-play decoders (Chung et al., 2023; Wu et al., 2024; Zheng et al., 2025; Metzler et al., 2018). All of this adapts or evaluates the *decoder* for a fixed forward model; we ask what a learned *sensor* loses off-distribution and measure it before deployment.

Designs in phase retrieval, tomography and coded diffraction. Complete sets of mutually unbiased bases and symmetric informationally complete measurements are complex projective 2-designs (Wootters & Fields, 1989; Renes et al., 2004; Klappenecker & Rötteler, 2005); they make the lifted intensity map a tight frame, giving linear ("painless") reconstruction (Balan et al., 2009) and optimal linear state tomography over unitarily invariant ensembles (Scott, 2006; Roy & Scott, 2007). PhaseLift recovers signals from $O(d \log d)$ Gaussian measurements (Candès et al., 2013), higher-order designs partially derandomize that guarantee (Gross et al., 2015), and Cramér-Rao bounds for complex phase retrieval are derived by Bandeira et al. (2014) and Balan (2016). That a measurement trades Fisher information between parameters at a fixed total is the content of the Gill-Massar bound (Gill & Massar, 2000); measurements with uniform Fisher information are called Fisher-symmetric (Li et al., 2016; Zhu & Hayashi, 2018), under multinomial rather than additive noise; and in experimental design the trace of the inverse information is A-optimality, with invariant designs obtained by symmetrizing a convex criterion (Pukelsheim, 2006; Box & Hunter, 1957). For coded diffraction, random masks give recovery guarantees (Candès et al., 2015) and discrete chirps are a classical source of mutually unbiased bases (Alltop, 1980; Howard et al., 2006). All of this concerns optimal designs, which a compressive learned sensor cannot be, and we claim none of it. Our delta is the reading for a sensor that is *not* a design: the excess separability it collects on *any* distribution is the exact ledger (5), computable before a decoder exists, and the distance to a design bounds that excess only uniformly over a distribution class, not on the task at hand.

3 The fourth-order geometry of intensity sensing

Setting. A signal $x \in \mathbb{C}^d$ is observed by an *intensity sensor* with m unit-norm measurement vectors $a_1, \dots, a_m \in \mathbb{C}^d$,

$$y_i = |\langle a_i, x\rangle|^2 + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \dots, m, \tag{1}$$

with $\langle a, x\rangle = a^H x$, A the sensor ($m \times d$, rows a_i^H) and $P_i = a_i a_i^H$. Since x and $e^{\mathrm{i}\varphi}x$ give the same y, errors are measured after optimal global-phase alignment. *Every statement below assumes* (1): additive Gaussian noise of constant variance, unit-norm rows, and $\mathbb{E}\|x\|^4 < \infty$. This covers coded diffraction imaging and ptychography with a read-noise-limited detector, but not holography, incoherent imaging, or photon-limited detection, where the Fisher weights become $1/\mathbb{E}\, y_i$. Lifting $X = xx^H$ makes each measurement linear, $|\langle a_i, x\rangle|^2 = \langle P_i, X\rangle$ with $\langle M, N\rangle = \mathrm{tr}(M^H N)$ (Table 2). Two operators summarize the sensor, $\Phi_2(A) = \frac{1}{m}\sum_i P_i$ and $\Phi_4(A) = \frac{1}{m}\sum_i P_i \otimes P_i$, with Haar values $\Phi_2^\star = I/d$ and $\Phi_4^\star = (I + F)/(d(d+1))$, F the swap of the tensor factors. A sensor is *second-order isotropic* (a tight frame) if $\Phi_2(A) = \Phi_2^\star$ and *fourth-order isotropic* (a complex projective 2-design) if $\Phi_4(A) = \Phi_4^\star$, the first following from the second but not conversely. We write $T(A) = \Phi_4(A) - \Phi_4^\star$ for the *defect operator* and $\Delta(A) = \|T(A)\|_F$ for the *design defect*, so that $\Delta(A)^2 = \mathrm{FP}(A) - \mathrm{FP}^\star$ with $\mathrm{FP}^\star = 2/(d(d+1))$ and $\mathrm{FP}(A) = \frac{1}{m^2}\sum_{i,j}|\langle a_i, a_j\rangle|^4$ the

fourth-order frame potential (Renes et al., 2004; Scott, 2006). Exact 2-designs need $m \geq d^2$, and for $d \leq m < d^2$ a sharper bound (Appendix B) forces $\mathrm{FP}(A) \geq 4.29\,\mathrm{FP}^\star$ at the size we study, so a compressive sensor always has $\Delta(A) > 0$ and what matters is *where* $T(A)$ points (Remark 1).

Two fourth-order functionals. Perturbing the signal along v changes the lifted signal by $H_v = xv^H + vx^H$, so the Fisher information of y along v is

$$\mathcal{I}_A(x; v) = \frac{1}{\sigma^2} \sum_{i=1}^{m} \langle P_i, H_v \rangle^2 = \frac{m}{\sigma^2} \left\langle \Phi_4(A),\, H_v \otimes H_v \right\rangle, \tag{2}$$

a linear functional of $\Phi_4(A)$, whereas $\Phi_2(A)$ only fixes the expected total intensity. Second-order isotropy therefore says little: one orthonormal basis is a tight frame, the error-minimizing structure for *linear* sensing (Goyal et al., 2001), yet measures only the moduli $|x_k|^2$ and is blind to all $d-1$ relative phases. We read (2) through two scalar summaries, claiming nothing about mutual information or Bayes risk, neither of which is conserved below.

(i) A local benchmark. The real Fisher matrix $J_A(x) \in \mathbb{R}^{2d \times 2d}$ of (2) annihilates the global-phase direction $u_x = (-\operatorname{Im} x, \operatorname{Re} x)/\|x\|$, since $H_{\mathrm{i}x} = 0$. Let $K_A(x)$ be its restriction to the quotient tangent space $W_x = u_x^{\perp}$; when $K_A(x) \succ 0$, which needs $m \geq 2d-1$, we set $\mathrm{CRB}_A(x) = \operatorname{tr} K_A(x)^{-1}$, and ∞ otherwise. It bounds locally unbiased estimators in this chart (Bandeira et al., 2014; Balan, 2016), but not the phase-aligned risk of biased prior-driven decoders, which can fall below it (Table 1); being a trace of an inverse, it is dominated by the directions the sensor sees worst.

(ii) A global functional. The Kullback-Leibler divergence between the measurement laws of two signals is $\mathrm{KL}(p_x \| p_{x'}) = \frac{m}{2\sigma^2} \langle \Phi_4(A), (X - X')^{\otimes 2} \rangle$, so how well a sensor tells apart two draws from a distribution $\mathcal{D}$ is again linear in $\Phi_4(A)$. We call it the *separation*

$$\mathcal{S}_A(\mathcal{D}) \;=\; \mathbb{E}_{x,x' \sim \mathcal{D}}\, \mathrm{KL}(p_x \| p_{x'}) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{m} \mathrm{Var}_{\mathcal{D}} \left(|\langle a_i, x \rangle|^2 \right) \;=\; \frac{m}{\sigma^2} \left\langle \Phi_4(A),\, M(\mathcal{D}) \right\rangle, \tag{3}$$

with $M(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[X^{\otimes 2}] - (\mathbb{E}_{\mathcal{D}} X)^{\otimes 2}$ the lifted covariance of the data. The middle expression makes $\mathcal{S}$ practical: the variance the data induce in the noiseless measurements, needing no decoder, estimator or matrix inversion, at $O(mnd)$ on n signals. It is *average-case and pairwise*, not a mutual information: a sensor with large $\mathcal{S}$ separates typical pairs well but may still be blind in some direction, which is what the local benchmark reports.

4 A CONSERVATION LAW FOR LEARNED INTENSITY SENSORS

Learning adapts a sensor to the orientation of its training distribution; averaged over orientations that adaptation is exactly zero-sum. Proofs are in Appendix B, and $U \cdot \mathcal{D}$ is the law of Ux for $x \sim \mathcal{D}$.

Theorem 1 (Conservation of separability under rotations)**.** *Assume model* (1). (i) *For all* $x, v \in \mathbb{C}^d$, $\mathbb{E}_{U \sim \mathrm{Haar}(\mathrm{U}(d))}\, \mathcal{I}_A(Ux; Uv) = \mathcal{I}^\star(x; v) := \frac{m}{\sigma^2}(\operatorname{tr}(H_v)^2 + \|H_v\|_F^2)/(d(d+1))$. (ii) *For every distribution* $\mathcal{D}$*,*

$$\mathbb{E}_U\, \mathcal{S}_A(U \cdot \mathcal{D}) = \mathcal{S}^\star(\mathcal{D}) := \frac{m}{\sigma^2} \frac{2\,\mathbb{E}\|x\|^4 - (\mathbb{E}\|x\|^2)^2 - \|\mathbb{E}[xx^H]\|_F^2}{d(d+1)}. \tag{4}$$

Neither right-hand side depends on the sensor. (iii) *The pointwise quantities equal these Haar averages, that is* $\mathcal{I}_A(Ux; Uv) = \mathcal{I}^\star(x; v)$ *and* $\mathcal{S}_A(U \cdot \mathcal{D}) = \mathcal{S}^\star(\mathcal{D})$ *for every* U*,* x*,* v *and* $\mathcal{D}$*, if and only if* A *is a complex projective 2-design. Part (iii) is the standard characterization of complex projective 2-designs (Renes et al., 2004; Scott, 2006), restated for these two functionals.*

The conserved object is *co-rotated*: signal and tangent direction in (i), the whole distribution in (ii). Only the *orientation average* is sensor-independent, so on a fixed distribution a sensor may separate far better than isotropy, and does. Part (ii) is the form we test, being measurable through (3).

Corollary 1 (Exact ledger, and what the design defect does and does not control)**.** *Let* $\widetilde{M}(\mathcal{D})$ *be the orthogonal projection of* $M(\mathcal{D})$ *onto* $\mathrm{span}\{I, F\}^{\perp}$*. For every* $\mathcal{D}$*,*

$$\mathcal{S}_A(\mathcal{D}) - \mathcal{S}^\star(\mathcal{D}) = \frac{m}{\sigma^2} \left\langle T(A),\, M(\mathcal{D}) \right\rangle \quad \textit{and} \quad \left| \mathcal{S}_A(\mathcal{D}) - \mathcal{S}^\star(\mathcal{D}) \right| \leq \frac{m}{\sigma^2}\, \Delta(A)\, \|\widetilde{M}(\mathcal{D})\|_F, \tag{5}$$

with equality exactly when $T(A)$ *and* $\widetilde{M}(\mathcal{D})$ *are linearly dependent. Consequently* $\sup_{\mathcal{D} \in \mathcal{C}_C} |\mathcal{S}_A(\mathcal{D}) - \mathcal{S}^\star(\mathcal{D})| \leq (2mC/\sigma^2)\, \Delta(A)$ *on the class* $\mathcal{C}_C = \{\mathcal{D} : \mathbb{E}\|x\|^4 \leq C\}$.

Corollary 1 separates two questions that the design literature merges. The excess on the data at hand is one inner product, the *alignment* $\langle T(A), M(\mathcal{D})\rangle$, which averages to zero over orientations. The *magnitude* $\Delta(A)$ caps that excess uniformly over any class of distributions with a bounded fourth moment, which is the precise sense in which we call it a *budget*, and it is tight: for a tight frame A and $\mathcal{D}$ uniform on its own rows, $\widetilde{M}(\mathcal{D}) = T(A)$ and the inequality is an equality, so $\Delta(A)$ really is the most such a sensor could bet on any distribution in the class. It is nonetheless worst-case only: a sensor can shrink $\Delta(A)$ and still raise its excess on a particular $\mathcal{D}$ by rotating $T(A)$ towards $\widetilde{M}(\mathcal{D})$, and the inequality is loose whenever the two are far from parallel, which we measure (Sections 5.2 and 5.5). Because the trace inverse is convex in the Fisher matrix, the local benchmark obeys a matching minimax statement.

Theorem 2 (2-designs minimize the orientation-averaged local bound)**.** *Assume model (1), and let $x \neq 0$. Then $\mathbb{E}_U \operatorname{CRB}_A(Ux) \geq \operatorname{CRB}^\star(x) := \sigma^2 d(d+1)(d-\frac{7}{8})/(m\|x\|^2)$, with equality if and only if A is a 2-design; if $K_A(Ux)$ is singular on a set of positive measure the left side is $+\infty$. Consequently $\sup_U \operatorname{CRB}_A(Ux) \geq \operatorname{CRB}^\star(x)$, and a 2-design attains $\operatorname{CRB}^\star(x)$ at every orientation. For $m < d^2$ the inequality is strict for every sensor with a finite left side, and no result here ties $\Delta(A)$ to the size of that gap.*

The identities behind Theorems 1 and 2 are classical and we claim none of them (Section 2); what is new is the reading for a *learned, compressive* sensor, for which an exact design is impossible. A Bayesian van Trees version of part (i) holds for any prior rotated with the data (Proposition 1), but bounds the unaligned error and is uninformative for phase-invariant priors, so we do not rely on it.

Coded diffraction patterns: three quantities that must not be confused. In coded diffraction imaging (CDP) (Candès et al., 2015), a field x on a $p \times p$ grid ($d = p^2$) passes through L unimodular phase masks q_ℓ and the detector records $y_{\ell,k} = |(\mathcal{F}(q_\ell \odot x))_k|^2$, $\mathcal{F}$ the unitary 2D DFT. Each mask contributes an orthonormal basis, so *every* CDP sensor is a tight frame: second-order isotropy is free and cannot distinguish mask designs.

Theorem 3 (Fourth-order geometry of coded diffraction)**.** *Let $\hat{r}_{\ell\ell'} = \mathcal{F}(q_\ell \odot \overline{q_{\ell'}})$ be the spectrum of the difference mask of masks $\ell \neq \ell'$, normalized so that $\sum_\xi |\hat{r}_{\ell\ell'}(\xi)|^2 = d$. Then*

$$\operatorname{FP}(A) = \underbrace{\frac{Ld + L(L-1)}{(Ld)^2}}_{\operatorname{FP}_{\min}(L)} + \frac{1}{(Ld)^2} \sum_{\ell \neq \ell'} \underbrace{\frac{1}{d} \sum_\xi \left(|\hat{r}_{\ell\ell'}(\xi)|^2 - 1\right)^2}_{\text{flatness excess } E_{\ell\ell'}}, \tag{6}$$

so $\operatorname{FP}(A) \geq \operatorname{FP}_{\min}(L)$, with equality if and only if every difference mask has flat Fourier magnitude, that is, the L bases are mutually unbiased. For an odd prime p, $L \leq p$ and distinct $t \in \mathbb{F}_p$, the finite-field quadratic masks $q_t[n] = \omega^{t(n_1^2+n_2^2)}$, $\omega = e^{2\pi \mathrm{i}/p}$, attain $\operatorname{FP}_{\min}(L)$; i.i.d. uniform random-phase masks have expected flatness excess $1 - 1/d$ per ordered pair.

Three quantities must be kept apart. The *total design defect* $\Delta(A)$ is dominated by the constant $\operatorname{FP}_{\min}(L) - \operatorname{FP}^\star \approx 1/(Ld)$, which no choice of masks can change. The *mask-dependent flatness excess* $E_{\ell\ell'}$ is the only part of $\Delta(A)$ the masks control, and at $L = 3$, $d = 961$ it is at most 0.25% of $\Delta(A)^2$: driving the mean flatness excess from 1.18 to 0.16 changes $\Delta(A)$ by only 0.11%. The *alignment* $\langle T(A), M(\mathcal{D})\rangle$ is what a given object class sees, and (5) leaves it free to move either way while $\Delta(A)$ stays put. Section 5.5 asks what a penalty on the part the masks do control does to the rest.

Two interventions. Corollary 1 gives one diagnostic and one penalty. The *separation bet* $\beta_A(\mathcal{D}) = 10 \log_{10}(\mathcal{S}_A(\mathcal{D}) / \operatorname{median}_U \mathcal{S}_A(U \cdot \mathcal{D}))$ measures in dB, without a decoder, how much a sensor favours its training orientation over a typical one; it is zero for a 2-design and its arithmetic counterpart is exactly zero-sum by (4). The *design-defect regularizer* presses on the budget: we add $\lambda\, \Delta(A)^2/\operatorname{FP}^\star$ to the end-to-end reconstruction loss, at a cost of $O(m^2 d)$ or $L(L-1)$ FFTs through (6). Since (5) bounds the excess by $\Delta(A)\|\widetilde{M}\|_F$ and not by $\Delta(A)$ alone, this is a worst-case control that *permits*, but does not predict, an increase of the alignment on a given task; Section 5.5 reports a case where that is what happens.

Table 1: **Controlled rotations** ($d = 31$, $m = 124$, $\sigma = 0.03$; mean over 5 seeds $\pm$ 95% t-interval, except the rotation-augmented row, which uses 3). Geometry: second-order gap $d\|\Phi_2 - \Phi_2^\star\|_F$, $\mathrm{FP}/\mathrm{FP}^\star = 1 + \Delta^2/\mathrm{FP}^\star$ and the separation bet β_A; then target-trained NMSE and local bound at $\theta = 0°$ and $90°$. Differences in the text use unrounded means, paired over seeds.

	sensor geometry			target-trained NMSE (dB)		CRB (dB)	
Sensor	$d\|\Phi_2-\Phi_2^\star\|_F$	$\mathrm{FP}/\mathrm{FP}^\star$	bet β_A (dB)	θ=0°	θ=90°	θ=0°	θ=90°
Random phases	2.73	5.0	+0.04	$-4.28_{\pm 0.05}$	$-4.24_{\pm 0.07}$	$-1.29_{\pm 0.24}$	$-1.31_{\pm 0.21}$
Quadratic (4 MUBs)	0.00	4.4	+0.01	$-4.31_{\pm 0.03}$	$-4.29_{\pm 0.01}$	$-2.56_{\pm 0.19}$	$-2.58_{\pm 0.22}$
Learned	13.01	26.2	+8.58	$-6.90_{\pm 0.05}$	$-3.52_{\pm 0.04}$	$-3.46_{\pm 0.44}$	$1.35_{\pm 0.20}$
Learned + tight-frame penalty	0.03	4.6	+0.75	$-4.72_{\pm 0.02}$	$-4.27_{\pm 0.02}$	$-2.31_{\pm 0.22}$	$-2.54_{\pm 0.22}$
Learned + phase noise	11.68	19.5	+8.18	$-6.87_{\pm 0.04}$	$-3.56_{\pm 0.03}$	$-3.43_{\pm 0.52}$	$1.00_{\pm 0.21}$
Learned + proximity	3.46	5.2	+3.65	$-5.69_{\pm 0.02}$	$-3.94_{\pm 0.05}$	$-2.99_{\pm 0.19}$	$-0.93_{\pm 0.23}$
Learned, quadratic init	12.94	25.8	+8.56	$-6.91_{\pm 0.02}$	$-3.48_{\pm 0.04}$	$-3.29_{\pm 0.23}$	$1.32_{\pm 0.33}$
Learned, rotation-augmented	13.15	26.9	+5.05	$-5.93_{\pm 0.17}$	$-5.90_{\pm 0.13}$	$-1.62_{\pm 0.35}$	$-1.52_{\pm 0.78}$
Learned + 4th-order, λ=0.03	3.98	5.0	+5.08	$-6.45_{\pm 0.03}$	$-3.77_{\pm 0.03}$	$-4.53_{\pm 0.09}$	$-1.22_{\pm 0.22}$
Learned + 4th-order, λ=0.1	2.65	4.6	+3.84	$-6.12_{\pm 0.02}$	$-3.92_{\pm 0.03}$	$-4.35_{\pm 0.15}$	$-1.77_{\pm 0.20}$
Learned + 4th-order, λ=0.3	1.63	4.4	+2.65	$-5.66_{\pm 0.02}$	$-4.08_{\pm 0.03}$	$-3.90_{\pm 0.21}$	$-2.19_{\pm 0.22}$
Learned + 4th-order, λ=1	0.82	4.3	+1.53	$-5.10_{\pm 0.02}$	$-4.22_{\pm 0.03}$	$-3.28_{\pm 0.21}$	$-2.51_{\pm 0.20}$
Learned + 4th-order, λ=3	0.44	4.3	+0.83	$-4.74_{\pm 0.02}$	$-4.27_{\pm 0.01}$	$-2.89_{\pm 0.21}$	$-2.60_{\pm 0.24}$

5 Experiments

We ask what a learned intensity sensor trades away when the data rotate (Q1), which decoder-free statistic ranks the trade and whether it can be controlled without giving up the gain (Q2), whether the fourth-order *direction* is what matters once every other geometric property of the sensor is frozen (Q3), when the sensor-level ordering is reflected in reconstruction (Q4), and what the picture becomes in coded diffraction imaging (Q5). A decoder degrades under shift even with a perfect sensor, so we never read sensor quality off the deployed decoder: the *target-trained decoder* has the architecture and training budget of the deployed one but is trained afresh, with the sensor frozen, on the shifted distribution. It removes decoder-level overfitting, but is not an oracle for the sensor, since it still depends on how easily this architecture inverts it, and Section 5.4 shows it far from saturation at low noise. Errors are phase-aligned normalized mean squared errors (NMSE) in dB.

5.1 Learned sensors specialize at the sensor level (Q1)

Setup. Signals $x \in \mathbb{C}^{31}$ are complex Gaussian with four dominant directions carrying 60% of the energy; at test time that subspace is rotated by an angle θ toward an orthogonal one, leaving the norm distribution unchanged, so every θ is a rotation of the training distribution. Sensors have $m = 124$ phase-only rows, just above the generic injectivity threshold $4d - 4$ for complex phase retrieval (Bandeira et al., 2014; Conca et al., 2015). We compare two fixed designs, random phases and the quadratic design (four mutually unbiased chirp bases), with a learned sensor and with learned sensors under the controls below; none is a 2-design. Decoders are MLPs and $\sigma = 0.03$ was chosen after an exploratory sweep in which the effect reverses at low noise. Numbers are means over 5 seeds with 95% t-intervals unless stated otherwise (Table 3).

Results. With the deployed decoder every sensor trained at $\theta = 0$ degrades under rotation (random: from -4.28 to 0.20 dB), which is decoder-level overfitting. With target-trained decoders the fixed designs are flat (random loses 0.03 ± 0.05 dB between $0°$ and $90°$), whereas the learned sensor loses 3.39 ± 0.07 dB: it is 2.59 ± 0.07 dB better than the quadratic design at $\theta = 0$ and 0.78 ± 0.04 dB worse at $90°$ (paired over seeds; a larger decoder gives the same picture, Appendix E). Both decoder-free statistics show the same crossover (Figure 1, Table 1): the learned sensor's separation at its training orientation is 8.4 dB above the orientation average Theorem 1 fixes for every sensor and 8.58 ± 0.13 dB above its own median over 150 orientations, against at most 0.04 dB for the fixed designs, and its local bound degrades by 4.8 dB against at most 0.02 dB for them. Conservation is not an approximation: over all 60 trained sensors the Haar mean of the separation matches (4) to within 0.4%, the Monte-Carlo error of 150 orientations. What the learned sensor gained at its training orientation was taken from other orientations, exactly.

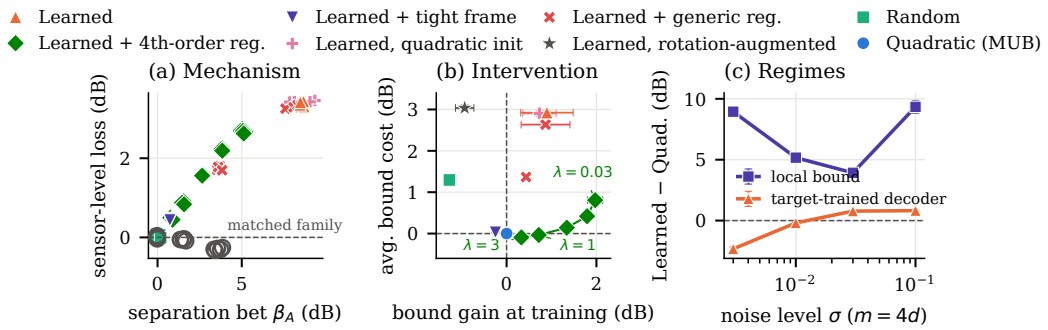


Figure 2: **Ranking, intervention and regimes.** (a) Sensor-level loss (target-trained NMSE at 90° minus 0°) against the separation bet, all sensors trained at $\theta = 0$, all seeds (generic reg.: phase noise or proximity); open markers are the matched family of Section 5.3, where the bet over-predicts. (b) Local-bound gain at the training orientation against its cost averaged over 150 orientations, relative to the quadratic design (95% intervals over seeds); the line joins $\lambda \in \{0.03, 0.1, 0.3, 1, 3\}$. (c) Learned minus quadratic at 90° against the noise level, $m = 4d$.

5.2 What ranks the trade, and what controls it (Q2)

The separation bet ranks the trade in this regime. The bet $\beta_A(\mathcal{D})$ needs the sensor and the training signals, no decoder and no matrix inverse. Over the 60 runs trained at $\theta = 0$ it ranks the sensor-level loss with Spearman 0.98 and over the 12 configuration means with 0.99 (bootstrap $[0.89, 1.00]$), against 0.97 for the bet computed from the local bound (Figure 2a, Table 4). This says something about the specialization mode these sensors use, not about reconstruction risk in general: they specialize by concentrating row energy on the training subspace (4.2 times isotropic for the learned sensor), and an aligned second-order statistic ranks the loss just as well (0.99), as does a purely data-driven second-order bet (0.99; Table 4).

Capping the budget controls the bet. Corollary 1 bounds every excess by $\Delta(A)\|\widetilde{M}\|_F$, so penalizing $\Delta(A)^2$ should shrink the bet monotonically, and it does: $\lambda \in \{0.03, 0.1, 0.3, 1, 3\}$ gives 5.08, 3.84, 2.65, 1.53 and 0.83 dB against 8.58 dB unconstrained. The bound is loose, as it must be: the Cauchy-Schwarz ratio $|\langle T, \widetilde{M}\rangle|/(\Delta\|\widetilde{M}\|_F)$ is 0.24 for the learned sensor and 0.00 to 0.25 over all sensors, against a ceiling of 0.78 for this $\mathcal{D}$. The intervention keeps separability where it is wanted (Figure 2b): with $\lambda = 1$ the local bound at the training orientation is 0.72 ± 0.05 dB below the quadratic design's while its Haar average differs by only -0.03 ± 0.00 dB, and every regularized sensor with $\lambda \leq 0.3$ dominates the unconstrained one, which gains 0.90 ± 0.58 dB at an average cost of 2.92 dB. Three controls, each run at a single weight and therefore not tuned against the λ sweep, fail to reproduce the trade-off (Table 1): phase-noise injection, a variant of the perturbation training Mengu et al. (2020) use against physical misalignment; proximity to the random initialization; and quadratic initialization. Augmenting over the whole family $\theta \in [0°, 90°]$, possible only when it is known, removes the loss inside it (0.03 ± 0.30 dB) but bets on the family (2.15 ± 0.28 dB).

5.3 A matched family: only the fourth-order direction varies (Q3)

Row-energy concentration is a second-order property, so the ranking above does not yet show that the fourth-order geometry is what matters; we therefore freeze everything else. A union of 4 orthonormal bases of $\mathbb{C}^{31}$ is an exact tight frame, so $\Phi_2(A) = \Phi_2^\star$ to machine precision and, by Proposition 2 (Appendix B), its mean row energy on *every* subspace equals $\mathrm{tr}(\Sigma)/d$: the aligned second-order statistic is identically 0 dB at every orientation. Holding one such frame fixed and rotating the data leaves $\Phi_2(A)$, $\mathrm{FP}(A)$ and $\Delta(A)$ unchanged, so only the direction of $T(A)$ relative to $M(\mathcal{D}_\theta)$ varies. We build three members: TF_0 with four Haar-random bases, TF_1 with one basis containing an orthonormal basis of the training subspace, and TF_3 with three such bases, the largest alignment that keeps the sensor locally identifiable (with four, no measurement mixes that subspace with its complement and the bound is infinite). Across 5 seeds the second-order gap stays below 3×10^{-15} and the aligned statistic below 2×10^{-15} dB at every angle, and $\mathrm{FP}/\mathrm{FP}^\star$ is 4.75, 4.75 and 5.08.

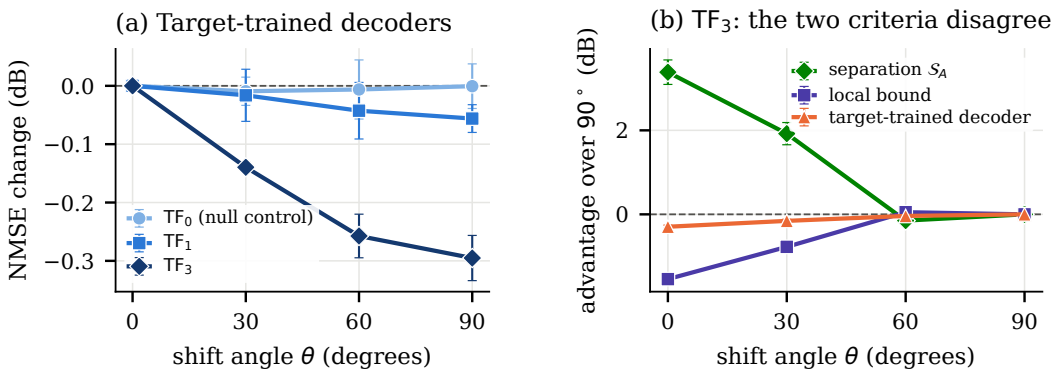


Figure 3: **A matched family in which only the fourth-order direction varies.** TF_j: an exact tight frame of four orthonormal bases of $\mathbb{C}^{31}$, j of them adapted to the training subspace. (a) Target-trained decoder error relative to the training orientation (5 seeds, paired, 95% t-intervals). (b) For TF_3, what each quantity says about orientation θ relative to 90°, positive meaning θ is the better one.

The two criteria make opposite predictions here: TF_3 separates the training orientation 3.39 dB better than the shifted one (3.58 against 0.20 dB above isotropic), while its local bound at $\theta = 0$ is 1.54 dB *worse* than at 90°. Target-trained decoders (Figure 3) settle three things. The fourth-order direction does change reconstruction: the change from 0° to 90° is -0.00 ± 0.04 dB for the null control TF_0, -0.06 ± 0.02 dB for TF_1 and -0.30 ± 0.04 dB for TF_3 (5 seeds, paired), monotone in the alignment and zero without it. The effect is small, 0.30 dB of reconstruction for 3.39 dB of separation. And its sign follows the local bound rather than the separation: reconstruction is *better* at 90°, where the sensor separates typical pairs less well but its hardest direction is 1.54 dB better. The bet's ranking in Section 5.2 therefore rides on the second-order concentration that accompanies it in learned sensors; with that frozen, the two statistics come apart and the decoder follows the worst direction.

5.4 When is the sensor-level ordering reflected in reconstruction? (Q4)

The local benchmark does not bound the risk of a biased prior-driven decoder: the target-trained errors in Table 1 lie below it because it applies to locally unbiased estimators while the decoders exploit the prior. We therefore ask only when the decoder-free ordering agrees with target-trained reconstruction, calling the regimes *bound-aligned* and *bound-misaligned* without claiming the decoder has exhausted the measurements in either. Across $\sigma \in [0.003, 0.1]$ at $m = 4d$ with 5 seeds (Figure 2c) and $m/d \in \{2, 8\}$ at seed 0 (Table 5), the learned sensor's bound at 90° is 3.9 to 9.3 dB worse than the quadratic design's wherever it is well conditioned, yet the target-trained decoders reproduce that ordering only when $\sigma \geq 0.03$ (0.81 ± 0.04 dB at $\sigma = 0.1$, $m = 4d$); at $\sigma = 0.003$ the learned sensor instead *wins*, by 2.32 ± 0.16 dB at $m = 4d$ and by up to 2.50 dB elsewhere. There the decoders sit 15.7 dB (quadratic) and 4.5 dB (learned) above their bounds and are still improving with more training, so neither sensor is information-limited and the ordering is set by how easily each is inverted.

5.5 Coded diffraction imaging: a stress test for the global criterion (Q5)

We simulate coded diffraction imaging of 31×31 phase objects $x = \exp(\mathrm{i}\pi\,\mathrm{img})$ with $L = 3$ masks and Gaussian noise $\sigma = 0.3$ on intensities of mean 1, trained on CIFAR-10 (Krizhevsky, 2009) and tested on 13 sets: CIFAR-10, STL-10 (Coates et al., 2011), SVHN (Netzer et al., 2011), Fashion-MNIST (Xiao et al., 2017), CIFAR-10-C (Hendrycks & Dietterich, 2019) and CIFAR-10 with a fraction of pixels scrambled, an exact rotation of the object distribution. Masks are random, quadratic, or learned from random, quadratic or flat initializations, the last also with the design-defect penalty at $\lambda = 1$ (Appendix D). Phase objects lie on a torus, outside the Haar setting, so this is a stress test of the global criterion rather than a benchmark.

The flatness excess and the alignment move in opposite directions. Every mask set is a tight frame and, with $m = 3d \ll d^2$, all have essentially the same budget: their defects differ by 0.12%, and the largest alignment measured on any test set is 7×10^{-4} of the Cauchy-Schwarz bound of

Corollary 1 (Table 6), about 1400 times too large to order these masks. They differ on phase objects, and the two criteria then disagree (Figure 4). The clearest case is the penalty: on flat-initialized masks it lowers the mean flatness excess from 1.18 to 0.16, a factor of 7.5, which moves $\Delta(A)$ by only 0.11%, while the alignment $10\log_{10}(\mathcal{S}_A/\mathcal{S}^\star)$ rises from 3.7 to 7.0 dB and the phase-object bound worsens by 0.08 to 0.78 dB: pressing on the only part of the budget the masks control turned the defect towards the data, an outcome Corollary 1 permits and does not predict. Across the 13 test sets the alignment orders the measured bounds better than the flatness excess (mean Spearman $+0.66$ against $+0.35$; 6 mask sets on correlated test sets, so we read only the sign), with the opposite sign to the synthetic study: here the mask sets that separate typical phase objects best are those whose hardest direction is worst.

The global criterion is not refuted: its exact minimizer, the quadratic masks, has the lowest bound on 12 of the 13 sets; what fails is a *local* penalty applied from a degenerate starting point. Neither scalar resolves the four mask sets close to isotropic here, whose alignments span 0.09 to 0.57 dB and whose bounds span 0.15 to 1.61 dB, and no learned mask set loses more than 0.11 dB of separability under an exact rotation, so the sensor-level effect of Section 5.1 does not appear here at all. Reconstruction is bound-misaligned (3.4 to 5.2 dB above the bound), and learned masks beat the quadratic ones on natural images by 0.81 to 1.67 dB while untrained random masks do too, so most of that advantage is not attributable to learning (Table 6).

6 Discussion, limitations and conclusion

A budget is not a prediction, and the fourth order is not always what does the work. Corollary 1 splits a design question usually treated as one. The design defect $\Delta(A)$ is a *budget* on what a sensor could bet on *any* distribution in a class, so minimizing it is right when nothing is known about the data (Theorem 2); what a task sees is an *alignment*, and the budget can be far from spent. In coded diffraction all six mask sets have the same defect to within 0.12% and realize at most 7×10^{-4} of the Cauchy-Schwarz bound, so it orders nothing there while the alignment orders the measured bounds. Penalize the defect when the deployment distribution is unknown; minimize the alignment over a plausible family of $\mathcal{D}$ when it is known. As for the fourth order, on the main grid a purely second-order, data-driven statistic ranks the sensor-level loss as well as the separation bet (0.99 against 0.99), because those sensors specialize by concentrating row energy and the two orders move together. The matched family separates them: the fourth-order direction is causal, graded with the alignment and null without it, but worth only 0.30 dB of reconstruction for 3.39 dB of separation, and its sign is set by the worst-direction bound. We therefore read the separation as a cheap detector that a sensor has specialized, and the local bound as the better guide to which orientation reconstructs well.

Limitations. The theory assumes model (1); photon-limited detection weights each measurement by $1/\mathbb{E}\, y_i$ and breaks the exact conservation. The separation is average-case and the Cramér-Rao expression local and worst-direction, bounding no biased decoder's risk; Section 5.3 is a case where the separation predicts the wrong sign, and Proposition 1 does not close the gap. Natural shifts are not Haar rotations, and phase objects lie outside the Haar setting, where neither scalar resolved the near-isotropic mask sets. The matched family is built from orthonormal bases and is therefore not phase-only, unlike the sensors of Section 5.1. The synthetic loss was measured at $d = 31$ at a noise level chosen after an exploratory sweep, and reverses at low noise; coded diffraction used 31×31 objects, three seeds per mask set and MLP decoders, so hardware remains untested.

Conclusion. How well an intensity sensor separates a signal distribution is an inner product of two fourth moments, one of the sensor and one of the data, whose orientation average is sensor-independent: learning redistributes separability rather than creating it, and the orientation that separates best need not be the one that reconstructs best.

AI use statement

We used generative AI tools (a large-language-model assistant, including its coding-agent mode) in this work. Their uses were: aiding and polishing the writing; reviewing drafts of the manuscript and raising objections we then answered; writing, refactoring and debugging the simulation, training, evaluation and analysis code; discussion, that is, proposing and criticizing hypotheses, framings and

experiment designs that we then decided on; and generating the plots. We also used them to create slides for internal meetings, which are not part of this submission. The remaining categories of the venue's policy do not apply: no translation, no data collection or cleaning beyond the public datasets' own preprocessing, and no qualitative data analysis.

We reviewed all AI-assisted work and are responsible for it. Concretely: the identity behind every theorem, corollary and proposition is checked numerically, and those checks are listed in Appendix C; every number in Section 5 and in Appendices D and E is produced by analysis scripts that read the saved experiment outputs, so no number in the text was transcribed by hand; and every reference was opened and checked against its publisher, DOI or proceedings record. We take responsibility for the final content of this work, including its text, claims, proofs, code and figures.

Ethics statement

This work uses only public image datasets (CIFAR-10, CIFAR-10-C, STL-10, SVHN and Fashion-MNIST) under their original terms, involves no human subjects and no personal data, and all experiments are simulations. Learned sensors are used in medical and scientific imaging, where a sensor that silently loses separability under a distribution shift can cause errors that a downstream model cannot repair; our results are intended to help anticipate and limit such failures.

Reproducibility statement

Section 3 defines the measurement model and every quantity we report; the assumptions of each result are stated in its statement, complete proofs are in Appendix B, and the numerical checks of those proofs are in Appendix C. Section 5 and Appendix D specify the signal models, the sensors, the decoder architectures, optimizers and training budgets, the noise levels, the dataset preprocessing and splits, the target-trained decoder protocol, the Cramér-Rao computation, and the number of seeds per configuration (Table 3). Decoder architectures, optimizers and training budgets are shared by all sensors and were fixed in advance; the noise level of the main synthetic grid was chosen after an exploratory sweep, and Section 5.4 and Table 5 report every noise level and measurement budget we ran.

References

Andrés Jerez, Juan Estupiñán, Jorge Bacca, and Henry Arguello. Deep unrolled single snapshot phase retrieval via non-convex formulation and phase mask design. *IEEE Journal of Selected Topics in Signal Processing*, 18(4):694–703, 2024.

Michael R. Kellman, Emrah Bostan, Nicole A. Repina, and Laura Waller. Physics-based learned design: Optimized coded-illumination for quantitative phase imaging. *IEEE Transactions on Computational Imaging*, 5(3):344–353, 2019.

Cagla D. Bahadir, Alan Q. Wang, Adrian V. Dalca, and Mert R. Sabuncu. Deep-learning-based optimization of the under-sampling pattern in MRI. *IEEE Transactions on Computational Imaging*, 6:1139–1152, 2020.

Iris A. M. Huijben, Bastiaan S. Veeling, and Ruud J. G. van Sloun. Deep probabilistic subsampling for task-adaptive compressed sensing. In *International Conference on Learning Representations (ICLR)*, 2020.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics*, 37(4):114:1–114:13, 2018.

Chi-Jui Ho, Yash Belhe, Steve Rotenberg, Ravi Ramamoorthi, Tzu-Mao Li, and Nicholas Antipa. A differentiable wave optics model for end-to-end computational imaging system optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28042–28051, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021.

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4754–4776, 2022.

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.

Mohammad Zalbagi Darestani, Akshay S. Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2433–2444, 2021.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T. Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. InverseBench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations (ICLR)*, 2025.

A. J. Scott. Tight informationally complete quantum measurements. *Journal of Physics A: Mathematical and General*, 39(43):13507–13530, 2006.

Henry Arguello, Jorge Bacca, Hasindu Kariyawasam, Edwin Vargas, Miguel Marquez, Ramith Hettiarachchi, Hans Garcia, Kithmini Herath, Udith Haputhanthri, Balpreet Singh Ahluwalia, Peter So, Dushan N. Wadduwage, and Chamira U. S. Edussooriya. Deep optical coding design in computational imaging: A data-driven framework. *IEEE Signal Processing Magazine*, 40(2): 75–88, 2023.

Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10192–10201, 2019.

Luca Baldassarre, Yen-Huan Li, Jonathan Scarlett, Baran Gözcü, Ilija Bogunovic, and Volkan Cevher. Learning-based compressive subsampling. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):809–822, 2016.

Shanshan Wu, Alex Dimakis, Sujay Sanghavi, Felix X. Yu, Daniel Holtmann-Rice, Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar. Learning a compressed sensing measurement matrix via gradient unrolling. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6828–6839, 2019.

Michael Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55(12):5695–5702, 2007.

Jinlin Xiang, Minho Choi, Yubo Zhang, Zhihao Zhou, Arka Majumdar, and Eli Shlizerman. Neural tangent knowledge distillation for optical convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 161177–161197, 2025.

Deniz Mengu, Yifan Zhao, Nezih T. Yardimci, Yair Rivenson, Mona Jarrahi, and Aydogan Ozcan. Misalignment resilient diffractive optical networks. *Nanophotonics*, 9(13):4207–4219, 2020.

Minho Choi, Jinlin Xiang, Anna Wirth-Singh, Seung-Hwan Baek, Eli Shlizerman, and Arka Majumdar. Transferable polychromatic optical encoder for neural networks. *Nature Communications*, 16: 5623, 2025.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 118389–118427, 2024.

Christopher A. Metzler, Philip Schniter, Ashok Veeraraghavan, and Richard G. Baraniuk. prDeep: Robust phase retrieval with a flexible deep network. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3501–3510, 2018.

William K. Wootters and Brian D. Fields. Optimal state-determination by mutually unbiased measurements. *Annals of Physics*, 191(2):363–381, 1989.

Joseph M. Renes, Robin Blume-Kohout, A. J. Scott, and Carlton M. Caves. Symmetric informationally complete quantum measurements. *Journal of Mathematical Physics*, 45(6):2171–2180, 2004.

Andreas Klappenecker and Martin Rötteler. Mutually unbiased bases are complex projective 2-designs. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pp. 1740–1744, 2005.

Radu Balan, Bernhard G. Bodmann, Peter G. Casazza, and Dan Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

Aidan Roy and A. J. Scott. Weighted complex projective 2-designs from bases: Optimal state determination by orthogonal measurements. *Journal of Mathematical Physics*, 48(7):072110, 2007.

Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

David Gross, Felix Krahmer, and Richard Kueng. A partial derandomization of PhaseLift using spherical designs. *Journal of Fourier Analysis and Applications*, 21(2):229–266, 2015.

Afonso S. Bandeira, Jameson Cahill, Dustin G. Mixon, and Aaron A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1): 106–125, 2014.

Radu Balan. Reconstruction of signals from magnitudes of redundant representations: The complex case. *Foundations of Computational Mathematics*, 16(3):677–721, 2016.

Richard D. Gill and Serge Massar. State estimation for large ensembles. *Physical Review A*, 61(4): 042312, 2000.

Nan Li, Christopher Ferrie, Jonathan A. Gross, Amir Kalev, and Carlton M. Caves. Fisher-symmetric informationally complete measurements for pure states. *Physical Review Letters*, 116(18):180402, 2016.

Huangjun Zhu and Masahito Hayashi. Universally Fisher-symmetric informationally complete measurements. *Physical Review Letters*, 120(3):030404, 2018.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

G. E. P. Box and J. S. Hunter. Multi-factor experimental designs for exploring response surfaces. *The Annals of Mathematical Statistics*, 28(1):195–241, 1957.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

W. O. Alltop. Complex sequences with low periodic correlations. *IEEE Transactions on Information Theory*, 26(3):350–354, 1980.

Stephen D. Howard, A. Robert Calderbank, and William Moran. The finite Heisenberg-Weyl groups in radar and communications. *EURASIP Journal on Advances in Signal Processing*, 2006:085685, 2006.

Vivek K. Goyal, Jelena Kovačević, and Jonathan A. Kelner. Quantized frame expansions with erasures. *Applied and Computational Harmonic Analysis*, 10(3):203–233, 2001.

Aldo Conca, Dan Edidin, Milena Hering, and Cynthia Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356, 2015.

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

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 (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, 2011.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Aram W. Harrow. The church of the symmetric subspace. *arXiv preprint arXiv:1308.6595*, 2013.

Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

A Notation

Table 2 lists every symbol; each is also defined at its first use in the main text.

Table 2: **Notation.** Every symbol used in the paper, in order of first use; the four confusable scalars are in the last block.

Symbol	Meaning
$x \in \mathbb{C}^d$, d	signal (coherent field) and its dimension
a_i, m, A	unit-norm measurement vectors, their number, the sensor ($m \times d$, rows a_i^H)
y_i, ε_i, σ	intensity measurement $\lvert\langle a_i, x\rangle\rvert^2 + \varepsilon_i$, Gaussian noise, noise level
$P_i = a_i a_i^H$, $X = xx^H$	rank-one projector of row i, lifted signal
$\Phi_2(A)$, $\Phi_4(A)$	second- and fourth-moment operators of the rows (Section 3)
$\Phi_2^\star$, $\Phi_4^\star$, F	Haar values I/d and $(I+F)/(d(d+1))$; F swaps the factors of $\mathbb{C}^d \otimes \mathbb{C}^d$
$T(A)$, $\Delta(A)$	defect operator $\Phi_4(A) - \Phi_4^\star$ and design defect $\lVert T(A)\rVert_F$ (Section 3)
$\mathrm{FP}(A)$, $\mathrm{FP}^\star$	fourth-order frame potential $\lVert\Phi_4(A)\rVert_F^2$ and its 2-design value $2/(d(d+1))$

Table 2: **Notation** (continued).

Symbol	Meaning
$\mathrm{FP}_{\min}(L)$	smallest frame potential of an L-mask coded-diffraction sensor, attained for $L \le p$, Eq. (6)
$H_v = xv^H + vx^H$	change of the lifted signal along a perturbation v
$\mathcal{I}_A(x; v)$, $J_A(x)$	Fisher information along v and real $2d \times 2d$ Fisher matrix, Eq. (2)
$\mathcal{I}^\star(x; v)$, $J^\star(x)$	orientation-averaged (isotropic) Fisher information and its matrix (Theorem 1)
W_x, u_x, $K_A(x)$	quotient tangent space $u_x^\perp \subset \mathbb{R}^{2d}$, the global-phase direction $(-\operatorname{Im} x, \operatorname{Re} x)/\|x\|$ it removes, and the restriction of $J_A(x)$ to it
$\mathrm{CRB}_A(x)$, $\mathrm{CRB}^\star(x)$	local Cramér-Rao benchmark $\operatorname{tr} K_A(x)^{-1}$ (∞ if x is not locally identifiable) and its isotropic value (Section 3)
$\mathrm{KL}(p_x \Vert p_{x'})$	Kullback-Leibler divergence between the measurement laws of two signals
$\mathcal{D}$, $U \cdot \mathcal{D}$	signal distribution; the law of Ux for $x \sim \mathcal{D}$
$M(\mathcal{D})$, $\widetilde{M}(\mathcal{D})$	lifted covariance $\mathbb{E}[X^{\otimes 2}] - (\mathbb{E}X)^{\otimes 2}$ of the data, and its orthogonal projection onto $\operatorname{span}\{I, F\}^\perp$
$\mathcal{S}_A(\mathcal{D})$, $\mathcal{S}^\star(\mathcal{D})$	separation of A on $\mathcal{D}$ and its conserved orientation average, Eqs. (3) and (4)
$\mathcal{C}_C$	the class $\{\mathcal{D} : \mathbb{E}\|x\|^4 \le C\}$ over which $\Delta(A)$ bounds the excess (Corollary 1)
$e_i = a_i^H \Sigma a_i$, $\bar{e}$	energy of row i under covariance Σ, and its mean over rows (Proposition 2)
$U \sim \mathrm{Haar}(\mathrm{U}(d))$, O_U	uniformly random unitary (a random orientation) and its real $2d \times 2d$ representation
J_π, $B_A(U)$, $B^\star$	Fisher information of a prior, van Trees bound under the rotated prior, isotropic value (Proposition 1)
NMSE	normalized mean squared error: phase-aligned squared error over $\mathbb{E}\|x\|^2$, in dB
MUB	mutually unbiased bases; the quadratic design uses four (synthetic) or three (coded diffraction)
L, q_ℓ, $\mathcal{F}$	number of masks, unimodular phase masks, unitary 2D DFT (coded diffraction)
$\hat{r}_{\ell\ell'}$, $E_{\ell\ell'}$, p, ω	spectrum of the difference mask $q_\ell \overline{q_{\ell'}}$, its flatness excess, prime grid side, $e^{2\pi \mathrm{i}/p}$
λ, $\mathcal{R}(A)$	weight and value of the design-defect penalty (Section 4)
θ, V_0, V_θ, r, $\mathcal{D}_\theta$	shift angle, bases of the training and rotated dominant subspaces, their dimension $r = 4$, rotated signal distribution (Section 5.1, Appendix D)
TF_j	matched tight frame with j of four bases adapted to V_0 (Section 5.3)
ρ	fraction of permuted pixels in scrambled images (Section 5.5)
separation bet $\beta_A(\mathcal{D})$	$10 \log_{10}$ of $\mathcal{S}_A(\mathcal{D})$ over its median across 150 orientations of $\mathcal{D}$, in dB; zero for a 2-design
local bet	the same construction applied to CRB_A in place of $\mathcal{S}_A$, in dB
alignment	the overlap $\langle T(A), M(\mathcal{D})\rangle$ of Corollary 1, reported in dB as $10 \log_{10}(\mathcal{S}_A(\mathcal{D})/\mathcal{S}^\star(\mathcal{D}))$
flatness excess	the dimensionless mask-dependent term $E_{\ell\ell'}$ of Eq. (6); at most 0.25% of $\Delta(A)^2$ at $L = 3$, $d = 961$
aligned second-order statistic	mean row energy on V_0 over its isotropic value, $\frac{d}{r} \cdot \frac{1}{m} \sum_i \|P_{V_0} a_i\|^2$; identically 1 for every tight frame

B PROOFS

Throughout, $\langle M, N\rangle = \operatorname{tr}(M^H N)$, F is the swap operator on $\mathbb{C}^d \otimes \mathbb{C}^d$, $\Pi_{\mathrm{sym}} = (I + F)/2$ is the projector onto the symmetric subspace $\mathrm{Sym}^2 \subset \mathbb{C}^d \otimes \mathbb{C}^d$, and u denotes a Haar-random unit vector

in $\mathbb{C}^d$. We use two standard facts (Harrow, 2013): (F1) $\mathbb{E}_u(uu^H)^{\otimes 2} = 2\Pi_{\mathrm{sym}}/(d(d+1)) = \Phi_4^\star$; (F2) for Hermitian H, $\langle I, H\otimes H\rangle = \mathrm{tr}(H)^2$ and $\langle F, H\otimes H\rangle = \mathrm{tr}(H^2) = \|H\|_F^2$.

Design defect identity $\Delta(A)^2 = \mathrm{FP}(A) - \mathrm{FP}^\star$. $\|\Phi_4(A)\|_F^2 = \frac{1}{m^2}\sum_{i,j}\langle P_i, P_j\rangle^2 = \frac{1}{m^2}\sum_{i,j}|\langle a_i, a_j\rangle|^4 = \mathrm{FP}(A)$. By (F2) with $H = P_i$, $\langle \Phi_4(A), \Phi_4^\star\rangle = \frac{1}{m}\sum_i \frac{\mathrm{tr}(P_i)^2+\mathrm{tr}(P_i^2)}{d(d+1)} = \frac{2}{d(d+1)}$, and $\|\Phi_4^\star\|_F^2 = \frac{\mathrm{tr}((I+F)^2)}{d^2(d+1)^2} = \frac{2d^2+2d}{d^2(d+1)^2} = \frac{2}{d(d+1)}$. Expanding $\|\Phi_4(A) - \Phi_4^\star\|_F^2$ gives $\mathrm{FP}(A) - \frac{2}{d(d+1)}$. In particular $\mathrm{FP}(A) \geq 2/(d(d+1))$ (the Welch bound (Welch, 1974)) with equality if and only if A is a 2-design. For $m \geq d$ a sharper bound holds. By Cauchy-Schwarz, $\sum_{i\neq j}|\langle a_i, a_j\rangle|^4 \geq \big(\sum_{i\neq j}|\langle a_i, a_j\rangle|^2\big)^2/(m(m-1))$, and $\sum_{i,j}|\langle a_i, a_j\rangle|^2 = m^2\|\Phi_2(A)\|_F^2 \geq m^2/d$ with equality if and only if A is a tight frame. Hence $\mathrm{FP}(A) \geq \frac{1}{m}\big(1 + \frac{(m/d-1)^2}{m-1}\big)$, with equality if and only if A is a tight frame with constant $|\langle a_i, a_j\rangle|$ for $i \neq j$ (an equiangular tight frame); at $m = d^2$ the bound equals $\mathrm{FP}^\star$.

Lemma 1. *$\Phi_4(A)$ and $\Phi_4^\star$ are supported on Sym^2. If a Hermitian operator T supported on Sym^2 satisfies $\langle x^{\otimes 2}, Tx^{\otimes 2}\rangle = 0$ for all $x \in \mathbb{C}^d$, then $T = 0$.*

Proof. $P_i \otimes P_i = (a_i\otimes a_i)(a_i\otimes a_i)^H$ and $a_i \otimes a_i \in \mathrm{Sym}^2$; $\Phi_4^\star \propto \Pi_{\mathrm{sym}}$. For the second claim, $f(x) = \langle x^{\otimes 2}, Tx^{\otimes 2}\rangle = \sum_{ijkl}\bar{x}_i\bar{x}_j T_{ij,kl}x_k x_l$ is a polynomial in $(x, \bar{x})$ that vanishes identically, so the coefficient of each monomial $\bar{x}_i\bar{x}_j x_k x_l$ vanishes. That coefficient is a positive multiple of the entry of T symmetrized over (i, j) and over (k, l), which equals $T_{ij,kl}$ because $T = \Pi_{\mathrm{sym}} T \Pi_{\mathrm{sym}}$. Hence $T = 0$. □

Proof of Theorem 1(i). Since $H_{Uv}(Ux) = Ux(Uv)^H + Uv(Ux)^H = UH_vU^H$ and $\langle P_i, UHU^H\rangle = \langle U^H P_i U, H\rangle$,

$$\mathcal{I}_A(Ux; Uv) = \frac{1}{\sigma^2}\sum_{i=1}^m \big\langle (U^H a_i a_i^H U)^{\otimes 2},\, H_v^{\otimes 2}\big\rangle.$$

For fixed unit a_i and Haar U, $U^H a_i$ is a Haar-random unit vector, so by (F1) each summand has expectation $\langle \Phi_4^\star, H_v^{\otimes 2}\rangle$, and by (F2) $\langle \Phi_4^\star, H_v^{\otimes 2}\rangle = (\mathrm{tr}(H_v)^2 + \|H_v\|_F^2)/(d(d+1))$, which proves Theorem 1(i). For the explicit form, write $v = \alpha\hat{x} + \beta\mathrm{i}\hat{x} + w$ with $w \perp x$. Then $x^H v = \|x\|(\alpha + \mathrm{i}\beta)$, $\mathrm{tr}\, H_v = 2\,\mathrm{Re}(x^H v) = 2\alpha\|x\|$ and $\|H_v\|_F^2 = 2\|x\|^2\|v\|^2 + 2\,\mathrm{Re}((x^H v)^2) = \|x\|^2(4\alpha^2 + 2\|w\|^2)$, so $\mathrm{tr}(H_v)^2 + \|H_v\|_F^2 = \|x\|^2(8\alpha^2 + 2\|w\|^2)$. If A is a 2-design then $\Phi_4(A) = \Phi_4^\star$ and $\mathcal{I}_A(y; w) = \mathcal{I}^\star(y; w)$ for every y, w by (2). Conversely, if $\mathcal{I}_A(x; v) = \mathcal{I}^\star(x; v)$ for all x, v, take $v = x$, so $H_x = 2xx^H$, and obtain $\langle \Phi_4(A) - \Phi_4^\star, (xx^H)^{\otimes 2}\rangle = 0$ for all x; Lemma 1 gives $\Phi_4(A) = \Phi_4^\star$. □

Proof of Theorem 1(ii) and of the identity (3). With Gaussian noise the measurements of two signals x, x' are distributed as $\mathcal{N}(\mu(x), \sigma^2 I)$ and $\mathcal{N}(\mu(x'), \sigma^2 I)$ with $\mu(x)_i = \langle P_i, X\rangle$, so $\mathrm{KL}(p_x\|p_{x'}) = \frac{1}{2\sigma^2}\sum_i\langle P_i, X - X'\rangle^2 = \frac{m}{2\sigma^2}\langle \Phi_4(A), (X - X')^{\otimes 2}\rangle$. Averaging over two independent draws from $\mathcal{D}$ and using $\mathbb{E}(\zeta - \zeta')^2 = 2\,\mathrm{Var}\,\zeta$ for i.i.d. ζ, ζ' with $\zeta = \langle P_i, X\rangle = |\langle a_i, x\rangle|^2$ gives the second expression in (3); and $\mathrm{Var}\langle P_i, X\rangle = \langle P_i^{\otimes 2}, M(\mathcal{D})\rangle$ with $M(\mathcal{D}) = \mathbb{E}[X^{\otimes 2}] - (\mathbb{E}X)^{\otimes 2}$ gives the third. For the conservation law, $M(U\cdot\mathcal{D}) = U^{\otimes 2}M(\mathcal{D})(U^H)^{\otimes 2}$, so by Schur's lemma the Haar average $\mathbb{E}_U M(U\cdot\mathcal{D}) = \alpha\,\Pi_{\mathrm{sym}} + \beta\,\Pi_{\mathrm{anti}}$ with $\alpha = \mathrm{tr}(\Pi_{\mathrm{sym}}M)/\,\mathrm{tr}\,\Pi_{\mathrm{sym}}$ and $\beta = \mathrm{tr}(\Pi_{\mathrm{anti}}M)/\,\mathrm{tr}\,\Pi_{\mathrm{anti}}$. By (F2), $\langle \Phi_4(A), \Pi_{\mathrm{sym}}\rangle = \frac{1}{2m}\sum_i(\mathrm{tr}(P_i)^2 + \mathrm{tr}(P_i^2)) = 1$ and $\langle \Phi_4(A), \Pi_{\mathrm{anti}}\rangle = 0$ for unit-norm rows, so $\mathbb{E}_U\mathcal{S}_A(U\cdot\mathcal{D}) = \frac{m}{\sigma^2}\alpha$, which depends on $\mathcal{D}$ only. Finally $\mathrm{tr}\,\Pi_{\mathrm{sym}} = d(d+1)/2$ and $2\,\mathrm{tr}(\Pi_{\mathrm{sym}}M) = \mathrm{tr}\,M + \mathrm{tr}(FM) = \big(\mathbb{E}\|x\|^4 - (\mathbb{E}\|x\|^2)^2\big) + \big(\mathbb{E}\|x\|^4 - \|\mathbb{E}\,xx^H\|_F^2\big)$, using $\mathrm{tr}(X^{\otimes 2}) = (\mathrm{tr}\,X)^2$ and $\mathrm{tr}(F\,X\otimes X') = \mathrm{tr}(XX')$, which gives (4). The same expression equals $\frac{m}{\sigma^2}\langle \Phi_4^\star, M(\mathcal{D})\rangle$, so $\mathcal{S}^\star$ is also the value taken by any 2-design at every orientation. If A is a 2-design, $\mathcal{S}_A = \mathcal{S}^\star$ for every $\mathcal{D}$ by (3). Conversely, taking $\mathcal{D}$ uniform on $\{x, x'\}$ makes $M(\mathcal{D}) = \frac{1}{4}(X - X')^{\otimes 2}$, so equality for all such $\mathcal{D}$ forces $\langle T, (X - X')^{\otimes 2}\rangle = 0$ for all x, x'; setting $x' = 0$ and applying Lemma 1 gives $T = 0$. □

Proof of Corollary 1. The identity is (3) minus $\mathcal{S}^\star(\mathcal{D}) = \frac{m}{\sigma^2}\langle \Phi_4^\star, M(\mathcal{D})\rangle$. For the bound, $\langle T, I\rangle = \frac{1}{m}\sum_i \mathrm{tr}(P_i)^2 - 1 = 0$ and $\langle T, F\rangle = \frac{1}{m}\sum_i \mathrm{tr}(P_i^2) - 1 = 0$ by (F2) and unit-norm rows, so

$T \perp \operatorname{span}\{I, F\}$ and $\langle T, M\rangle = \langle T, \widetilde{M}\rangle$; Cauchy-Schwarz gives $|\langle T, \widetilde{M}\rangle| \le \|T\|_F \|\widetilde{M}\|_F = \Delta(A)\|\widetilde{M}\|_F$, with equality if and only if $T \propto \widetilde{M}$. □

Proof of Theorem 2. Let $J^\star(x)$ be the real $2d \times 2d$ matrix of the quadratic form $\mathcal{I}^\star(x;\cdot)$ of Theorem 1, and $O_U \in \mathrm{O}(2d)$ the real representation of U and $U^H A$ the sensor with rows $U^H a_i$. Since $\langle a_i, Ux\rangle = \langle U^H a_i, x\rangle$, we have $\mathcal{I}_A(Ux; Uv) = \mathcal{I}_{U^H A}(x; v)$, i.e. $J_A(Ux) = O_U J_{U^H A}(x) O_U^\top$, and therefore $\mathrm{CRB}_A(Ux) = \operatorname{tr} J_{U^H A}(x)^+$. Every $J_{U^H A}(x)$ annihilates $(\mathrm{i}x)_\mathbb{R}$; let W be its orthogonal complement ($\dim W = 2d-1$) and K_U the restriction of $J_{U^H A}(x)$ to W. If K_U is singular on a set of positive measure, the left side of the bound is infinite and there is nothing to prove. Otherwise $K_U \succ 0$ almost surely, and by Theorem 1 (applied to all v and polarized) $\mathbb{E}_U K_U = K^\star$, the restriction of $J^\star(x)$ to W. The map $K \mapsto \operatorname{tr} K^{-1}$ is strictly convex on positive definite matrices, so Jensen's inequality gives $\mathbb{E}_U \operatorname{tr} K_U^{-1} \ge \operatorname{tr}(K^\star)^{-1}$ with equality if and only if $K_U = K^\star$ almost surely. From the explicit form of $\mathcal{I}^\star$, $J^\star(x)$ has eigenvalue $8c$ on $\hat{x}_\mathbb{R}$, $2c$ on the $(2d-2)$-dimensional real subspace $\{w_\mathbb{R} : w \perp x\}$ and 0 on $(\mathrm{i}x)_\mathbb{R}$, where $c = m\|x\|^2/(\sigma^2 d(d+1))$; hence $\operatorname{tr}(K^\star)^{-1} = \frac{1}{c}\left(\frac{1}{8} + d - 1\right) = \mathrm{CRB}^\star(x)$. If equality holds, continuity in U gives $\mathcal{I}_A(Ux; Uv) = \mathcal{I}^\star(x; v)$ for all U and v; as $\{Ux\}$ is the whole sphere of radius $\|x\|$ and both sides are homogeneous of degree two in x, $\mathcal{I}_A = \mathcal{I}^\star$ everywhere and A is a 2-design by Theorem 1. The worst-case statement follows from $\sup \ge$ mean. □

A Bayesian version of Theorem 2. Learned decoders exploit a prior and are biased, so the relevant benchmark for them is the van Trees inequality (Gill & Levit, 1995): for a prior density π on $\mathbb{C}^d \cong \mathbb{R}^{2d}$ satisfying its regularity conditions, with Fisher information matrix J_π, every estimator $\hat{x}(y)$ of $x \sim \pi$ satisfies $\mathbb{E}\|\hat{x} - x\|^2 \ge \operatorname{tr}\left(\mathbb{E}_{x\sim\pi} J_A(x) + J_\pi\right)^{-1}$, without phase alignment.

Proposition 1. *Let $J_\pi \succ 0$ and let π_U be the law of Ux for $x \sim \pi$ (the prior rotated along with the data). Write $B_A(U) = \operatorname{tr}\left(\mathbb{E}_{x\sim\pi_U} J_A(x) + J_{\pi_U}\right)^{-1}$ for the van Trees bound of sensor A under π_U. Then*

$$\mathbb{E}_{U\sim\mathrm{Haar(U}(d))}\, B_A(U) \;\ge\; B^\star := \operatorname{tr}\left(\mathbb{E}_{x\sim\pi} J^\star(x) + J_\pi\right)^{-1},$$

and every complex projective 2-design has $B_A(U) = B^\star$ for all U.

Proof. The density of π_U is $\pi(U^H\cdot)$, so $J_{\pi_U} = O_U J_\pi O_U^\top$, and as in the proof of Theorem 2, $\mathbb{E}_{x\sim\pi_U} J_A(x) = \mathbb{E}_{x\sim\pi} J_A(Ux) = O_U\, \mathbb{E}_{x\sim\pi} J_{U^H A}(x)\, O_U^\top$. Hence $B_A(U) = \operatorname{tr} K_U^{-1}$ with $K_U = \mathbb{E}_{x\sim\pi} J_{U^H A}(x) + J_\pi \succ 0$. Theorem 1, polarized and averaged over x, gives $\mathbb{E}_U K_U = \mathbb{E}_{x\sim\pi} J^\star(x) + J_\pi$, and Jensen's inequality for the convex map $K \mapsto \operatorname{tr} K^{-1}$ gives the claim. For a 2-design, $J_{U^H A}(x) = J^\star(x)$ for every U and x, so K_U does not depend on U. □

Scope. If π is invariant under a global phase, as circular Gaussian priors are, so is the posterior; the unaligned Bayes estimator is then 0 and every sensor has unaligned risk $\mathbb{E}\|x\|^2$, so Proposition 1 is correct but uninformative. It is informative for densities that break the global-phase symmetry, for example Gaussian priors with a nonzero mean. A bound on the phase-aligned error of prior-driven decoders remains open.

Proposition 2 (Separation of a Gaussian distribution, and what a tight frame can bet)**.** *Let $\mathcal{D} = \mathcal{CN}(0, \Sigma)$ and let $e_i = a_i^H \Sigma a_i$ be the row energies. Then*

$$\mathcal{S}_A(\mathcal{D}) = \frac{1}{\sigma^2}\sum_{i=1}^m e_i^2 = \frac{m}{\sigma^2}\Big(\bar{e}^{\,2} + \frac{1}{m}\sum_i (e_i - \bar{e})^2\Big), \qquad \bar{e} = \frac{1}{m}\sum_i e_i = \big\langle \Phi_2(A), \Sigma\big\rangle. \tag{7}$$

If A is a tight frame, then $\bar{e} = \operatorname{tr}(\Sigma)/d$ is the same for every tight frame and every orientation of Σ, so a tight frame can bet on an orientation only through the spread *of its row energies, a fourth-order quantity that no second-order statistic of the sensor can see. If in addition A is a 2-design, the spread is fixed as well and the bet is zero.*

Proof of Proposition 2. $\langle a_i, x\rangle \sim \mathcal{CN}(0, e_i)$, so $|\langle a_i, x\rangle|^2$ is exponential with mean e_i and variance e_i^2; (3) gives (7), and the decomposition into mean and spread is elementary. For a tight frame $\Phi_2(A) = I/d$, so $\bar{e} = \operatorname{tr}(\Sigma)/d$, which is invariant under $\Sigma \mapsto U\Sigma U^H$. For a 2-design, $\frac{1}{m}\sum_i e_i^2 = \langle \Phi_4(A), \Sigma^{\otimes 2}\rangle = \langle \Phi_4^\star, \Sigma^{\otimes 2}\rangle = \frac{(\operatorname{tr}\Sigma)^2 + \|\Sigma\|_F^2}{d(d+1)}$ by (F2), which is also orientation invariant. □

Remark. Equation (7) is why the aligned second-order statistic used as a foil in Section 5.2 (the mean row energy on the training subspace) is identically zero for every tight frame, while the separation bet need not be: Section 5.3 exhibits exact tight frames with bets up to 3.59 dB.

Proof of Theorem 3. Let $\mathcal{F}_{kn} = \omega^{-k\cdot n}/\sqrt{d}$ on $\mathbb{Z}_p^2$. The row indexed by (ℓ, k) satisfies $\langle a_{\ell k}, x\rangle = \sum_n \mathcal{F}_{kn} q_\ell[n] x[n]$, so $a_{\ell k}[n] = \overline{\mathcal{F}_{kn} q_\ell[n]}$. For fixed ℓ the rows $\{a_{\ell k}\}_k$ form an orthonormal basis, hence $\sum_k a_{\ell k} a_{\ell k}^H = I$ and $\Phi_2(A) = I/d$ for any masks. For $\ell \neq \ell'$, $\langle a_{\ell k}, a_{\ell' k'}\rangle = \frac{1}{d}\sum_n \omega^{-(k-k')\cdot n} q_\ell[n]\overline{q_{\ell'}[n]} = \frac{1}{\sqrt{d}}\hat{r}_{\ell\ell'}(k-k')$, so $\sum_{k,k'} |\langle a_{\ell k}, a_{\ell' k'}\rangle|^4 = \frac{1}{d}\sum_\xi |\hat{r}_{\ell\ell'}(\xi)|^4$. Same-basis pairs contribute Ld in total. Since $\frac{1}{d}\sum_\xi |\hat{r}|^2 = 1$ (Parseval), $\frac{1}{d}\sum_\xi |\hat{r}|^4 = 1 + \frac{1}{d}\sum_\xi (|\hat{r}|^2 - 1)^2$, which yields (6); the excess vanishes if and only if $|\hat{r}_{\ell\ell'}| \equiv 1$, i.e. the two bases are mutually unbiased. For the quadratic masks with $t \neq t'$ and odd prime p, the unnormalized transform of $q_t\overline{q_{t'}}$ at ξ factors as $G(c,\xi_1)G(c,\xi_2)$ with $c = t - t' \neq 0$ and $G(c,b) = \sum_{n\in\mathbb{F}_p} \omega^{cn^2 - bn}$; completing the square, $|G(c,b)| = |\sum_n \omega^{cn^2}| = \sqrt{p}$ (quadratic Gauss sum), so $|\hat{r}(\xi)| = p/\sqrt{d} = 1$. For i.i.d. uniform phases, $r_n = q_\ell[n]\overline{q_{\ell'}[n]}$ are i.i.d. uniform on the unit circle, and $\mathbb{E}|\sum_n r_n \omega^{-\xi\cdot n}|^4 = d + 2d(d-1)$, so $\mathbb{E}\frac{1}{d}\sum_\xi |\hat{r}(\xi)|^4 = 2 - \frac{1}{d}$ and the expected excess is $1 - \frac{1}{d}$. □

Remark 1 (A certificate for near-designs). For any sensor, with $T = \Phi_4(A) - \Phi_4^\star$, $|\mathcal{I}_A(x;v) - \mathcal{I}^\star(x;v)| = \frac{m}{\sigma^2}|\langle T, H_v^{\otimes 2}\rangle| \leq \frac{m}{\sigma^2}\Delta(A)\|H_v\|_F^2$, while $\mathcal{I}^\star(x;v) \geq \frac{m}{\sigma^2}\|H_v\|_F^2/(d(d+1))$. Hence $|\mathcal{I}_A(x;v)/\mathcal{I}^\star(x;v) - 1| \leq \epsilon := d(d+1)\Delta(A)$ for every signal, direction and orientation, and $\mathrm{CRB}_A(x) \leq \mathrm{CRB}^\star(x)/(1-\epsilon)$ whenever $\epsilon < 1$. This distribution-free certificate is informative only for near-designs ($m \gtrsim d^2$); for the compressive sensors studied in Section 5, $\epsilon > 1$, which is why we measure orientation dependence directly.

C Numerical checks of the theory

The checks below test our implementations and the algebra of the identities in Sections 3 and 4, in double precision; they are not evidence for the proofs, which are in Appendix B.

- *2-designs.* For the complete set of $d+1 = 8$ mutually unbiased bases in $d = 7$ (chirp bases plus the standard basis, $m = 56$), $g_A(x) = \frac{1}{m}\sum_i |\langle a_i, x\rangle|^4$ is constant over 5,000 random directions to relative precision 10^{-15}, the lifted tightness ratio $\frac{1}{m}\sum_i \langle P_i, H\rangle^2/(\|H\|_F^2/(d(d+1)))$ equals 1 for random traceless Hermitian H, the Fisher matrix equals $J^\star(x)$ (relative error 6×10^{-16}), and $\mathrm{CRB}_A(x) = \mathrm{CRB}^\star(x)$.
- *Conservation (Theorem 1).* For a random phase-only sensor with $m = 4d$ and for a pair of mutually unbiased bases ($m = 2d$) in $d = 7$, the Monte-Carlo average of $O_U^\top J_A(Ux) O_U$ over 20,000 Haar unitaries matches $J^\star(x)$ to 0.5% and 0.6% (Monte-Carlo error).
- *Distribution-level conservation (Theorem 1(ii)).* The identity $\mathcal{S}_A(\mathcal{D}) = \frac{m}{\sigma^2}\langle \Phi_4(A), M(\mathcal{D})\rangle$ holds to 7×10^{-16} relative error for a random phase-only sensor in $d = 5$ against the operators formed explicitly, and $\mathcal{S}_A(\mathcal{D}) - \mathcal{S}^\star(\mathcal{D})$ equals $\frac{m}{\sigma^2}\langle T(A), M(\mathcal{D})\rangle$ to the same precision. Averaging $\mathcal{S}_A(U\cdot\mathcal{D})$ over 4,000 Haar unitaries reproduces $\mathcal{S}^\star(\mathcal{D})$ within the Monte-Carlo error for two sensors with very different geometry. In the experimental setting ($d = 31$, $m = 124$, 150 orientations), the Haar mean of the separation agrees with (4) to 0.4% for all 60 trained sensors and to 0.2% for the matched family of Section 5.3.
- *Jensen gap (Theorem 2).* For the random sensor, $\mathbb{E}_U \mathrm{CRB}_A(Ux) = 2.39\,\mathrm{CRB}^\star(x)$ and the minimum over the 20,000 orientations is $1.26\,\mathrm{CRB}^\star(x)$, never below the bound.
- *Bayesian version (Proposition 1).* For a Gaussian prior $\mathcal{CN}(\mu, \Sigma)$ with $\mu \neq 0$ (which breaks the global-phase symmetry) and anisotropic Σ in $d = 7$, $\sigma = 0.1$, the expected Fisher matrix has the closed form $\frac{4}{\sigma^2}\sum_i R(P_i)\big(\frac{1}{2}R(\Sigma) + \mu_{\mathbb{R}}\mu_{\mathbb{R}}^\top\big)R(P_i)$, with $R(\cdot)$ the real $2d\times 2d$ representation, and matches a Monte-Carlo average to 0.5%. For a random phase-only sensor with $m = 4d$, the van Trees bound averaged over 4,000 Haar orientations is $1.27\,B^\star$ (single orientations reach $0.97\,B^\star$); for the complete set of mutually unbiased bases it is the same at every orientation (relative deviation 7×10^{-16}).
- *Coded diffraction (Theorem 3).* For $p = 7$ and $L \in \{2,3,4\}$, the frame potential computed from the explicit $Ld \times d$ sensor matches (6) to 2×10^{-17}; quadratic masks have zero excess

($< 4 \times 10^{-16}$); the mean excess of i.i.d. random masks over 200 draws is 0.97 to 1.00 against the predicted $1 - 1/d = 0.98$.

- *Equality in Corollary 1.* For unions of orthonormal bases in $d \in \{4, 5, 7\}$ and $\mathcal{D}$ uniform on the rows, $\|\widetilde{M}(\mathcal{D}) - T(A)\|_F / \|T(A)\|_F \leq 6 \times 10^{-16}$, so the Cauchy-Schwarz bound is attained.
- *Gaussian separation (Proposition 2).* For a random phase-only sensor ($d = 7$, $m = 23$) and an anisotropic Σ, $\sum_i e_i^2/\sigma^2$ matches $\mathcal{S}_A(\mathcal{D})$ estimated from 4×10^5 samples to 1.9×10^{-3} (Monte-Carlo error); for the complete set of mutually unbiased bases in $d = 7$ the mean row energy equals $\mathrm{tr}(\Sigma)/d$ and $\frac{1}{m}\sum_i e_i^2$ equals $((\mathrm{tr}\,\Sigma)^2 + \|\Sigma\|_F^2)/(d(d+1))$, both to machine precision.
- *Chirp bases.* For prime $d \in \{7, 31\}$ the chirp bases $\{\omega^{\alpha n^2 + kn}/\sqrt{d}\}_k$ with distinct α satisfy $d\,|\langle a, b\rangle|^2 = 1$ for all cross-basis pairs to 10^{-14}, are tight frames, and their frame potential equals $(Ld + L(L-1))/(Ld)^2$.

D Experimental details

Controlled rotations (Sections 5.1 to 5.4). Signals $x \in \mathbb{C}^{31}$ are drawn from $\mathcal{D}_\theta$: $x = \sigma_{\mathrm{lo}} z + \sqrt{\sigma_{\mathrm{hi}}^2 - \sigma_{\mathrm{lo}}^2}\, V_\theta w$ with $z \sim \mathcal{CN}(0, I_{31})$, $w \sim \mathcal{CN}(0, I_r)$, $r = 4$, $\sigma_{\mathrm{hi}}^2 = 10\,\sigma_{\mathrm{lo}}^2$ and $\mathbb{E}\|x\|^2 = 1$, so the four dominant directions hold 60% of the energy. $V_\theta = \cos\theta\, V_0 + \sin\theta\, V_1$, where $[V_0, V_1]$ is a random 31×8 matrix with orthonormal columns drawn once per seed; every θ is a rotation of the training distribution $\mathcal{D}_0$ with an identical norm distribution. Sensors have $m = 124$ phase-only rows $a_{ij} = e^{\mathrm{i}\vartheta_{ij}}/\sqrt{31}$. *Random*: i.i.d. uniform phases (fixed). *Quadratic*: the chirp bases with $\alpha \in \{0, 1, 2, 3\}$. *Learned*: phases initialized at random (or at the quadratic design for "quadratic init") and trained jointly with the decoder. *Tight frame*: penalty $(d\|\Phi_2(A) - I/d\|_F)^2$ with weight 1 (the achieved gap is below 0.1). *Fourth-order*: $\lambda\,(\mathrm{FP}(A) - \mathrm{FP}^\star)/\mathrm{FP}^\star$ with $\lambda \in \{0.03, 0.1, 0.3, 1, 3\}$. *Generic*: phase-noise injection (i.i.d. $\mathcal{N}(0, 0.3^2)$ perturbation of all phases at every step) or a proximity penalty $\mathrm{mean}(1 - \cos(\vartheta - \vartheta_0))$ to the random initialization with weight 1. *Rotation-augmented*: every training step draws θ uniformly from $[0°, 90°]$. Only the fourth-order weight was swept; the tight-frame, proximity and phase-noise weights were each run at a single value, so comparisons *between* penalties are read from one point per competitor and not from a Pareto front, whereas the λ sweep is compared against the unconstrained sensor throughout. The decoder is an MLP with two hidden layers of 512 GELU units that maps $d \cdot y$ to $(\mathrm{Re}\,\hat{x}, \mathrm{Im}\,\hat{x})$; the loss is the phase-invariant error $\|x\|^2 + \|\hat{x}\|^2 - 2|\langle \hat{x}, x\rangle|$ normalized by $\mathbb{E}\|x\|^2$. Training uses Adam (learning rate 2×10^{-3} for the decoder and 2×10^{-2} for the phases, cosine decay) for 6,000 steps with 512 fresh samples per step. Target-trained decoders are freshly initialized and trained with the same budget on $\mathcal{D}_\theta$ with the sensor frozen. NMSE is evaluated on 20,000 fresh samples with a fixed evaluation seed. The CRB is averaged over 200 signals from $\mathcal{D}_\theta$ (fixed per seed) and normalized by their mean energy. Both decoder-free statistics rotate the same 200 signals at $\theta = 0$ by the same 150 Haar-random unitaries, shared across all sensors, so Table 1, Figure 1c and Figure 2 use one sample and every comparison is paired. The separation of one sensor over all 150 orientations costs about 0.03 s on one CPU core and is linear in d at fixed m/d; the local bound over the same orientations costs about 40 s and is cubic in d. Seeds control the sensor initialization, the subspace and the data stream. The main grid of Table 1, the matched family of Section 5.3 and the noise sweep of Figure 2c use 5 seeds; the rotation-augmented row and the coded-diffraction study use 3; the $m/d \in \{2, 8\}$ rows of Table 5 and the decoder-capacity and decoder-budget checks use seed 0 only. Table 3 lists this per configuration. Unless a number is marked as a standard deviation it is a mean with a 95% t-interval, and differences between sensors are paired over seeds and computed before rounding. The synthetic experiments ran on a laptop CPU (Intel i9-9880H) and the coded diffraction experiments on its GPU (AMD Radeon Pro 5500M).

Coded diffraction imaging (Section 5.5). Images are converted to grayscale in $[0, 1]$ and brought to 31×31: CIFAR-10 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011) are cropped from 32×32, STL-10 (Coates et al., 2011) is area-downsampled from 96×96, and Fashion-MNIST (Xiao et al., 2017) is bilinearly upsampled from 28×28. CIFAR-10-C (Hendrycks & Dietterich, 2019) uses the same preprocessing as CIFAR-10 for its 15 corruptions at severities 3 and 5. The phase object is $x = \exp(\mathrm{i}\pi\,\mathrm{img})$ (unit modulus, $\|x\|^2 = d$ for every image, so all shifts preserve the norm). Pixel scrambling applies one fixed permutation to a fixed random subset containing a fraction ρ of the pixel positions (a cyclic shift within the subset), which is an exact unitary rotation of the object distribution.

The sensor has $L = 3$ masks; the quadratic design uses $t \in \{1, 2, 3\}$. Gaussian noise with standard deviation $\sigma = 0.3$ is added to intensities whose mean is 1. The decoder is an MLP with two hidden layers of 2048 GELU units acting on $\sqrt{y + 10^{-3}}$ (clamped at 0), trained with SGD (Nesterov momentum 0.9, learning rate 0.1 with cosine decay, gradient clipping at norm 5) for 6,000 steps of 128 CIFAR-10 training images with random horizontal flips; learned mask phases use Adam (learning rate 10^{-2}) and start from uniform random phases, from the quadratic masks, or from zero phases (flat). The fourth-order regularizer is the mean flatness defect of the difference spectra (Theorem 3) with $\lambda = 1$, applied to the flat initialization. Evaluation uses 2,000 held-out images per test set. Fine-tuned decoders start from the trained decoder and are fine-tuned for 1,000 steps (learning rate 0.03) on a disjoint split of the target set (8,000 images for CIFAR-10 and its variants, 6,000 for STL-10, 8,000 for Fashion-MNIST) before evaluation on the held-out 2,000; we fine-tune on five target sets (CIFAR-10, STL-10, Fashion-MNIST, fully scrambled CIFAR-10 and Gaussian-noise CIFAR-10-C at severity 5). The CRB is computed exactly for the phase parametrization ($d - 1 = 960$ identifiable parameters) on 30 images for each of 13 test sets (CIFAR-10, STL-10, SVHN, Fashion-MNIST, the three scrambling levels, and six CIFAR-10-C corruptions at severity 5: Gaussian noise, defocus blur, contrast, fog, pixelate and JPEG); the CIFAR-10-C bound in Figure 4 averages these six. We run three seeds per mask set.

Configuration map. Table 3 lists every configuration used in the prediction analysis of Section 5.2, its row in Table 1, the number of seeds and the training angle. The 12 configurations trained at $\theta = 0$ are the ones entering the 60 runs and the configuration-level correlations of Table 4; the rotation-augmented sensor is trained on the whole family and is therefore reported but excluded from those correlations.

Table 3: **Configuration map.** Every synthetic configuration, its row in Table 1, seeds, and training angle. "In the 12" marks the configurations entering the correlations of Table 4.

Configuration	Penalty or control	Table 1 row	Seeds	Trained at
Random phases	fixed design	yes	5	n/a
Quadratic (4 MUBs)	fixed design	yes	5	n/a
Learned	none	yes	5	$\theta = 0$
Learned + tight-frame penalty	$\lambda_2 = 1$	yes	5	$\theta = 0$
Learned + phase noise	$\mathcal{N}(0, 0.3^2)$	yes	5	$\theta = 0$
Learned + proximity	weight 1	yes	5	$\theta = 0$
Learned, quadratic init	none	yes	5	$\theta = 0$
Learned + 4th-order	$\lambda = 0.03, 0.1, 0.3, 1, 3$	yes (5 rows)	5	$\theta = 0$
Learned, rotation-augmented	none	yes	3	$\theta \sim U[0°, 90°]$

E ADDITIONAL RESULTS

The matched family (Section 5.3). Each frame is a union of four orthonormal bases of $\mathbb{C}^{31}$, so $\Phi_2(A) = I/d$ to machine precision (second-order gap at most 3×10^{-15}) and, by Proposition 2, the mean row energy is $\mathrm{tr}(\Sigma)/d$ on every subspace, so the aligned second-order statistic is exactly 0 dB at every θ. TF_j has j of the four bases containing an orthonormal basis of the training subspace V_0 (a different one in each, so no row repeats) and $4 - j$ Haar-random bases. At $j = 4$ every basis is block diagonal with respect to $V_0 \oplus V_0^{\perp}$, no measurement mixes the two blocks, and the local bound is infinite; $j = 3$ is the largest identifiable alignment. Within one frame the sensor is literally unchanged along the θ sweep, so $\Phi_2(A)$, $\mathrm{FP}(A)$ and $\Delta(A)$ are frozen exactly ($\mathrm{FP}/\mathrm{FP}^\star = 4.75$, 4.75 and 5.08 for TF_0, TF_1 and TF_3); only the direction of $T(A)$ relative to $M(\mathcal{D}_\theta)$ varies. Decoder-free values are computed on the same 200 signals and 150 orientations as the rest of the study, and target-trained decoders use the protocol above.

Decoder-free statistics as predictors. Table 4 compares five statistics as predictors of the sensor-level loss, over the 60 runs and over the 12 configuration means, which removes the dependence between seeds of one configuration; the bootstrap intervals resample configurations. Three of them rank the configurations equally well: the separation bet, the aligned second-order statistic, and a purely second-order bet that needs neither the training subspace nor the fourth moment, built by

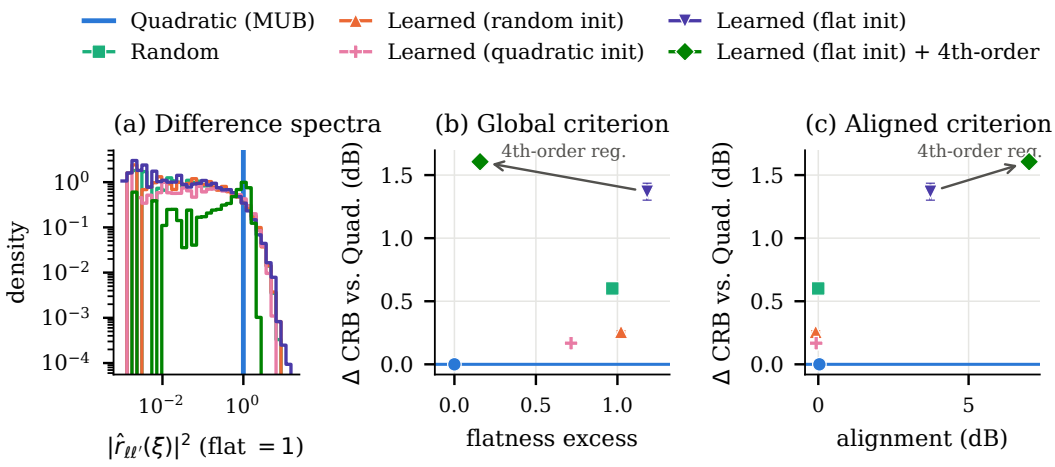


Figure 4: **Coded diffraction imaging: the two criteria disagree.** (a) Distribution of $|\hat{r}_{\ell\ell'}(\xi)|^2$ over all mask pairs (seed 0); the quadratic masks are exactly flat (vertical line). Phase-object bound on CIFAR-10 relative to the quadratic masks, against (b) the mask-dependent flatness excess of Theorem 3 and (c) the alignment $10\log_{10}(\mathcal{S}_A/\mathcal{S}^\star)$ of Corollary 1; the arrow marks the design-defect penalty, which moves left in (b) and right in (c) while the bound worsens. Mean over 3 seeds, bars are seed standard deviations.

replacing $\mathcal{S}_A$ with the mean row energy $\bar{e}$ in the same median-over-orientations construction. The local bet is close behind and the global design defect much further back: the defect is a budget and not a prediction. That the second-order bet does as well is the reason for the matched family of Section 5.3: on this grid nothing forces the fourth-order reading, because energy concentration and fourth-order alignment arrive together. Both second-order statistics are identically zero on tight frames, where the separation bet is not. The mean-based bet orders the loss as the median-based one does (Spearman 0.95 over runs for the local bet). Adding the 3 rotation-augmented runs, whose 90° shift lies inside their training family, lowers the local bet's correlation to 0.87 over 63 runs.

Table 4: **Decoder-free statistics as predictors of the sensor-level loss** (target-trained NMSE at 90° minus 0°). Spearman correlation over all runs and over configuration means, with a 95% bootstrap interval over configurations (10,000 resamples).

Statistic	over 60 runs	over 12 configurations	bootstrap interval
Separation bet β_A (Section 4)	0.98	0.99	[0.89, 1.00]
Local bet (median CRB minus training CRB)	0.95	0.97	[0.82, 1.00]
Aligned second-order statistic (needs V_0)	0.98	0.99	[0.89, 1.00]
Data-driven second-order bet (mean row energy)	0.98	0.99	[0.89, 1.00]
Global design defect $\Delta(A)$	0.71	0.69	[0.12, 0.96]

Regime grid. Table 5 lists every noise level and measurement budget of Section 5.4; the $m/d = 4$ rows are means over 5 seeds, paired over seeds, and the $m/d \in \{2, 8\}$ rows are seed 0. The learned sensor loses information at 90° in every cell where the bound is well conditioned. At $m/d = 2$ the sensor has only one measurement more than the $2d - 1$ real unknowns: the Fisher matrix is nearly singular and the bound, 2 to 38 dB above the signal energy, is dominated by a few near-singular signals, so we do not compare it. The target-trained decoder difference changes sign with the noise level at every m. At $\sigma = 0.003$ the target-trained decoders are far above the bound (15.7 and 4.5 dB for the quadratic and the learned sensor at $m/d = 4$, 19.3 and 11.8 dB at $m/d = 8$), while at $\sigma \geq 0.03$ they fall below it, because they exploit the Gaussian prior and the bound concerns unbiased estimators.

Rotation augmentation. Co-training on θ drawn uniformly from $[0°, 90°]$ (three seeds) gives target-trained decoder errors of -5.93 and -5.90 dB at 0° and 90°, 1.61 and 1.61 dB better than the

Table 5: **Regime grid.** Target-trained decoder NMSE (dB) of the learned and quadratic sensors at $\theta = 0°$ and $90°$, and the learned-minus-quadratic difference at $90°$ for target-trained decoders and for the Cramér-Rao bound (positive: learned is worse; differences computed before rounding, paired over seeds where more than one was run).

			Learned		Quadratic		Learned − Quad. at 90°	
m/d	σ	seeds	0°	90°	0°	90°	decoder	bound
2	0.003	1	−12.87	−7.36	−4.94	−4.86	−2.50	n/a
2	0.03	1	−5.56	−2.37	−3.17	−3.16	+0.78	n/a
4	0.003	5	−14.80	−9.19	−6.87	−6.86	−2.32	+8.94
4	0.01	5	−10.89	−6.46	−6.27	−6.26	−0.20	+5.17
4	0.03	5	−6.90	−3.52	−4.31	−4.29	+0.78	+3.92
4	0.1	5	−3.85	−0.95	−1.77	−1.77	+0.81	+9.34
8	0.003	1	−15.71	−10.57	−8.90	−8.81	−1.76	+5.80
8	0.03	1	−8.10	−4.53	−5.57	−5.58	+1.05	+4.15

quadratic design, with no loss inside the family (0.03 ± 0.30 dB). Its rows concentrate on the family subspace $\mathrm{span}(V_0, V_1)$, with a frame potential as large as the unconstrained sensor's (Table 1); it bets 2.15 ± 0.28 dB on the family and pays 3.04 ± 0.07 dB in the orientation-averaged bound, and even at $\theta = 0$ its bound (-1.62 dB) is above the quadratic design's (-2.56 dB). The Cramér-Rao bound, which weights all directions equally, does not register this sensor-level advantage: the sensor measures the high-variance family subspace that prior-driven decoders exploit. Augmentation moves the bet from one orientation to a known family of orientations.

Decoder budget. At $\sigma = 0.003$ and $m = 4d$ (seed 0), training sensors and decoders for 20,000 instead of 6,000 steps lowers the target-trained error at $90°$ from -6.85 to -7.81 dB for the quadratic design, from -6.90 to -7.82 dB for random phases and from -9.08 to -10.39 dB for the learned sensor: all decoders were still improving, and the learned sensor stays ahead of the quadratic design at $90°$ (2.22 and 2.58 dB).

Decoder capacity. With target-trained decoders of three hidden layers of 1,024 units (seed 0, $\sigma = 0.03$, $m = 4d$), every sensor improves in distribution (by 0.44, 0.36 and 0.39 dB for the learned, quadratic and $\lambda = 1$ sensors) and the picture is unchanged: the learned sensor is 2.69 dB better than the quadratic design at $0°$ and 1.00 dB worse at $90°$ (2.60 and 0.80 dB with the two-layer decoder), the $\lambda = 1$ sensor is 0.81 dB better and 0.12 dB worse, and random phases lose 0.07 dB between $0°$ and $90°$.

Coded diffraction: all results and generic signals. Table 6 lists the co-trained and fine-tuned decoder errors and the phase-object bounds of all mask sets on every test set, and, in its last block, the budget these sets have and the part of it the data collect. The six mask sets differ in design defect by 0.12% and realize at most 7×10^{-4} of the Cauchy-Schwarz bound of Corollary 1, so the uniform bound is about 1400 times larger than any alignment measured here and cannot separate the mask sets; the penalized masks, which have the smallest defect, are the ones that collect most of it. For generic complex signals $x \sim \mathcal{CN}(0, I)$, whose law is invariant under every unitary, the orientation-averaged bound equals $\mathbb{E}_x \mathrm{CRB}_A(x)$; computed exactly on 200 random signals shared by every mask set and seed, so the comparison is paired, the 6 mask sets agree within 0.01 dB and lie 6.1 dB above the unattainable 2-design value $\mathrm{CRB}^\star$; on a single signal the spread across mask sets is 0.12 dB (95% bootstrap interval over signals $[0.11, 0.13]$), so the agreement of the averages is not an artefact of a small sample. The differences between mask sets in Figure 4b,c therefore come from the structure of phase objects, which the Haar average of Theorem 2 does not see.

Table 6: **Coded diffraction imaging, all mask sets** (means over three seeds per mask set; seed standard deviations at most 0.08 dB for the bound, 0.06 dB for co-trained and 0.05 dB for fine-tuned decoders). The first three blocks are in dB: co-trained decoder NMSE, fine-tuned decoder NMSE and the Cramér-Rao bound for phase objects. The fourth block is dimensionless and reports what Corollary 1 bounds: the mean flatness excess $\bar{E}$ of the difference spectra, the design defect $\Delta(A)$ computed exactly from Theorem 3, its relative spread across mask sets, and the fraction of the Cauchy-Schwarz bound the data actually collect, over the 13 test sets. Differences in the text use unrounded means. "C-10-C s3/s5": CIFAR-10-C averaged over its 15 corruptions at severity 3/5; "scr." or "s.": pixel scrambling of a fraction ρ of CIFAR-10 pixels.

Co-trained decoder NMSE (dB)

Masks	CIFAR-10	STL-10	SVHN	F-MNIST	C-10-C s3	C-10-C s5	scr. 25%	scr. 50%	scr. 100%
Quadratic (MUB)	-11.28	-9.22	-16.49	-5.84	-12.54	-13.01	-7.47	-5.83	-4.97
Random	-11.97	-10.49	-17.52	-7.60	-13.22	-13.71	-7.57	-5.89	-4.99
Learned (random init)	-12.09	-10.55	-17.71	-7.50	-13.34	-13.84	-7.62	-5.93	-5.03
Learned (quadratic init)	-12.00	-10.39	-17.53	-7.32	-13.25	-13.73	-7.63	-5.94	-5.05
Learned (flat init)	-11.61	-10.05	-17.08	-7.14	-12.88	-13.42	-7.50	-5.85	-4.96
Learned (flat init) + 4th-order	-11.51	-10.05	-16.73	-7.32	-12.72	-13.22	-7.40	-5.77	-4.91

Fine-tuned decoder NMSE (dB)

Masks	CIFAR-10	STL-10	F-MNIST	C-10-C noise s5	scr. 100%
Quadratic (MUB)	-11.33	-9.73	-10.11	-9.84	-7.86
Random	-12.04	-10.73	-10.69	-10.18	-7.61
Learned (random init)	-12.16	-10.80	-10.74	-10.27	-7.72
Learned (quadratic init)	-12.07	-10.66	-10.65	-10.23	-7.77
Learned (flat init)	-11.69	-10.36	-10.43	-9.97	-7.53
Learned (flat init) + 4th-order	-11.58	-10.33	-10.41	-9.87	-7.34

Cramér-Rao bound for phase objects (dB, 30 images per set; C-10-C at severity 5)

Test set	Quad.	Random	L. random	L. quad.	L. flat	L. flat + 4th
CIFAR-10	-16.49	-15.89	-16.23	-16.32	-15.12	-14.88
STL-10	-16.61	-15.89	-16.25	-16.35	-15.16	-14.90
SVHN	-16.98	-15.89	-16.42	-16.56	-15.17	-14.71
Fashion-MNIST	-16.03	-15.89	-16.04	-16.03	-15.01	-14.93
C-10-C Gaussian noise	-16.45	-15.89	-16.21	-16.29	-15.18	-14.93
C-10-C defocus blur	-16.57	-15.89	-16.28	-16.39	-15.09	-14.81
C-10-C contrast	-17.49	-15.88	-16.61	-16.82	-15.29	-14.52
C-10-C fog	-16.86	-15.89	-16.42	-16.55	-15.16	-14.72
C-10-C pixelate	-16.52	-15.89	-16.25	-16.35	-15.11	-14.86
C-10-C JPEG	-16.49	-15.89	-16.24	-16.32	-15.12	-14.88
scrambled 25%	-16.53	-15.89	-16.22	-16.31	-15.34	-14.99
scrambled 50%	-16.56	-15.88	-16.22	-16.30	-15.48	-15.06
scrambled 100%	-16.57	-15.89	-16.22	-16.29	-15.56	-15.08

Budget and how much of it the data collect (dimensionless)

Masks	flatness excess $\bar{E}$	$\Delta(A)$	$\Delta/\Delta_{\min}-1$	$\lvert\langle T,\widetilde{M}\rangle\rvert/(\Delta\lVert\widetilde{M}\rVert_F)$, $\times 10^{-4}$
Quadratic (MUB)	0.00	1.8585×10^{-2}	0.000%	0.00 to 0.17
Random	0.97	1.8604×10^{-2}	0.101%	0.00 to 0.02
Learned (random init)	1.02	1.8605×10^{-2}	0.107%	0.00 to 0.05
Learned (quadratic init)	0.72	1.8599×10^{-2}	0.075%	0.00 to 0.05
Learned (flat init)	1.18	1.8608×10^{-2}	0.123%	0.21 to 2.50
Learned (flat init) + 4th-order	0.16	1.8589×10^{-2}	0.016%	0.04 to 7.37