THE DEGENERACY DISTILLERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Models are usually parameterized by scientific convention, not according to the combinations of parameters that the data can actually distinguish. When observations depend on only a few combinations of many parameters, simulation-based inference is unnecessarily high-dimensional: neural estimators spend expensive simulator calls learning redundant directions. Even at modest dimensionality, nonlinear degeneracies can make inference highly simulation-intensive and obscure the quantities actually being measured. We present the *degeneracy distillery*, which automatically discovers these combinations and expresses them in closed form, simultaneously reducing the effective parameter dimension and the simulation budget required for downstream inference.

From parameter–simulation pairs alone, without likelihood evaluations, simulator gradients, or an observed data realization, the method learns an information-aware coordinate map across the prior and distills it into compact symbolic expressions. A joint objective determines the number of identifiable coordinates while driving the transformed information metric toward the identity. The resulting coordinates are both low-dimensional and interpretable: they reveal how the original parameters combine to control the data and restrict downstream inference to the directions the data actually constrain.

Across scientific applications, the method recovers known structure from simulations alone: $R_0 \propto \beta/\gamma$ from noisy epidemic trajectories and chirp mass from gravitational-wave inspirals. For full inspiral–merger–ringdown waveforms, it instead discovers total mass and an asymmetric mass combination that yields better-conditioned and better-calibrated inference than conventional chirp-mass coordinates. In controlled scaling experiments, the method recovers rank-1 and rank-2 structure embedded in parameter spaces with up to 32 dimensions. Neural posterior estimation in the discovered coordinates tracks an oracle given the true parameterization, while inference in the original parameters degrades with ambient dimension. By removing redundant parameter directions before density estimation, the distilled coordinates substantially reduce the simulation budget required for simulation-based inference: 9–14$\times$ fewer simulations at matched validation log-probability and approximately 2$\times$ fewer at matched CRPS.

1 INTRODUCTION

Many scientific models are parametrised by variables whose effect on predictions is highly nontrivial: different parameter variations can produce nearly indistinguishable data, so the model responds primarily to particular *combinations* of parameters rather than to each independently. We call such directions *degeneracies* in parameter space.

Identifying these directions matters for two reasons. Statistically, aligning coordinates with directions of sensitivity improves conditioning, making posterior exploration and optimisation more efficient, and lets simulations be targeted along informative directions rather than spent along degeneracies that produce indistinguishable outputs. Scientifically, it reveals which combinations of parameters actually control observables rather than leaving that to an arbitrary coordinate choice, which in turn supports priors defined over meaningful combinations.

Such behaviour is naturally captured by the Fisher information, which defines a geometry over parameters that is typically highly anisotropic and ill-conditioned. In settings with analytic control, one can sometimes identify appropriate reparametrisations by hand. However, in more complex

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.

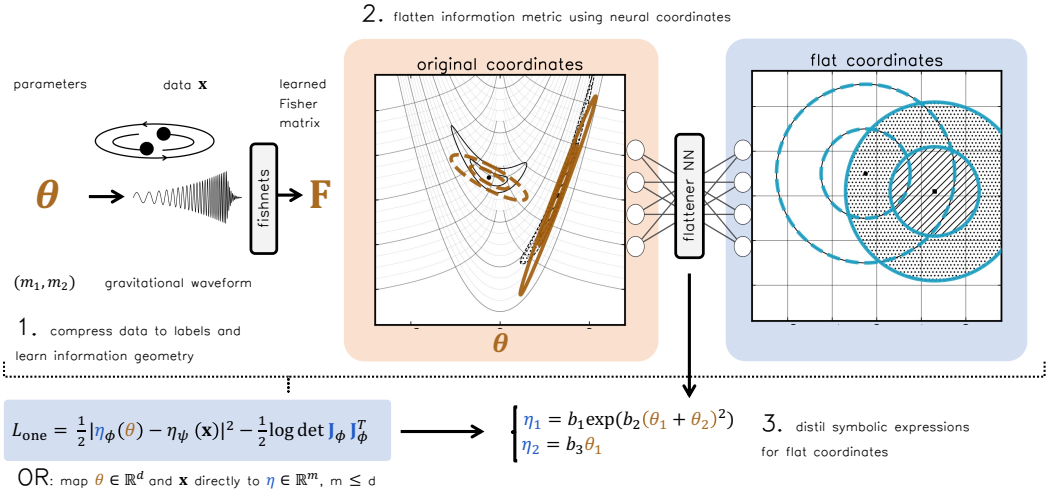

Figure 1: Schematic of the degeneracy distillery pipeline in three steps (illustration, not experimental results). (1) Parameter-data pairs $(\boldsymbol{\theta}, \mathbf{x})$ are passed to an ensemble of Fishnet networks to learn approximations to the Fisher matrix at each parameter value. (2) A "flattener" network is trained so that its Jacobian maps the learned metric to the identity as a function of $\boldsymbol{\theta}$. Alternatively, steps 1 and 2 can be combined into a joint neural objective (Eq. (6)). (3) Symbolic regression is performed on each output η coordinate to obtain short, nonlinear, degeneracy-resolving expressions.

or computationally intensive models this becomes infeasible, motivating the need for systematic, data-driven approaches. In this work, we develop a method to discover global reparametrisations that regularise this geometry, yielding coordinates in which the Fisher information is approximately isotropic across parameter space.

The *degeneracy distillery* maps the information geometry of a system from parameter-data pairs and discovers a symbolic parametrisation in terms of the system's original parameters for which the Fisher matrix is as close to the identity as possible. It consumes only parameter-simulation pairs: no likelihood evaluations, no analytic Fisher matrix, and no differentiable simulator. As outlined in Fig. 1, we achieve this in three main steps: (i) Estimate the information geometry everywhere with neural networks; (ii) Learn neural coordinates that flatten the Fisher geometry, unique up to a rigid motion that we fix by alignment; (iii) Find nonlinear symbolic expressions that approximate these neural coordinates and collect like terms. The output of the distillery is therefore a coordinate transformation that resolves the degeneracies of the original parametrisation.

Contributions. (i) We introduce a three-stage pipeline that infers the Fisher metric globally over parameter space from simulations alone, and uses it to learn a symbolic, invertible reparametrisation that flattens the Fisher information. We also link the first two steps into a unified loss formalism that can solve rank-deficient and dimensionality reduction problems. (ii) We provide ensemble estimators for both the Fisher matrix and the flattened coordinates, with an alignment scheme that fixes the residual constant offset and orthogonal symmetries of the flattening loss (see Appendix B). (iii) We validate the method on two synthetic problems with known geometry (Rosenbrock and a one-dimensional Gaussian), where analytic flat coordinates and geodesic normal coordinates serve as benchmarks. (iv) We apply the distillery to four scientific problems—SIR epidemic dynamics, gravitational-wave inspiral waveforms, weak-lensing cosmology from expensive dark matter simulations, and an industrial heater control problem—each chosen to probe a different regime of nonlinearity, dimensionality, and prior knowledge. (v) We demonstrate that the resulting coordinates lower the simulation budget required by downstream neural posterior estimation while preserving calibrated coverage.

2 Related Work

Sloppy models. "Sloppy" models depend strongly on a few parameter combinations and weakly on many others, giving a Fisher matrix whose eigenvalues span orders of magnitude and a "hyper-ribbon" model manifold (Gutenkunst et al., 2007; Transtrum et al., 2011; 2015). Gutenkunst et al. (2007) read the Fisher information as a Riemannian metric and use its eigenstructure to separate stiff from sloppy directions; near degeneracies are handled by reductions such as the manifold boundary approximation method, which collapses them at manifold boundaries (Transtrum & Qiu, 2014). **Local conformal autoencoders and intrinsic coordinates.** LOCA (Peterfreund et al., 2020) and Evangelou et al. (2022) also build intrinsic coordinates that remove redundancies from nonlinear observation maps, but without likelihood-based geometry: LOCA enforces local whitening in data space, producing coordinates invariant up to conformal factors, while Evangelou et al. (2022) separate predictive directions from level sets of the parameter-to-observable map. Both yield coordinates adapted to the model manifold; we instead regularise the Fisher geometry over parameter space. **Geometric variational inference and posterior reparametrisation.** Geometric variational inference (geoVI) (Frank et al., 2021) also uses the Fisher information, but locally: it constructs posterior-specific coordinates that flatten the geometry near the posterior, enabling Gaussian approximations. Our approach instead provides a single reparametrisation covering the prior support, estimated from simulations and independent of any particular observed data realisation. Similarly, Dacunha et al. (2022) learn global density models to derive the local Fisher geometry and decompose parameters into constrained and unconstrained directions. This is also posterior-specific, whereas our method is global; additionally, their coordinates are neural, while ours are symbolic (symbolic regression could also be applied to theirs). **Symbolic regression and information-geometry priors.** The symbolic component of our pipeline builds on OPERON (Burlacu et al., 2020; Kronberger et al., 2024) and on minimum-description-length selection criteria for symbolic regression (Bartlett et al., 2024; 2023). Ours is, to our knowledge, the first method to combine simulation-based Fisher estimation with a global, symbolic flattening of the resulting metric.

3 Theoretical Background

Consider a model producing $n_{\mathbf{x}}$ data points $\mathbf{x}$ from n_θ parameters $\boldsymbol{\theta}$, with likelihood $p(\mathbf{x}|\boldsymbol{\theta})$. The Kullback–Leibler divergence between two parameter values is neither symmetric nor a metric, but for a small displacement it reduces to a quadratic form (Rao, 1945),

$$D_{\mathrm{KL}}(\boldsymbol{\theta} \,||\, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \frac{1}{2} F_{ab}\, \delta\theta^a\, \delta\theta^b + \mathcal{O}(\|\delta\boldsymbol{\theta}\|^3), \tag{1}$$

whose coefficient is the *Fisher information matrix* $F_{ab} = -\,\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}[\partial_a \partial_b \log p(\mathbf{x}|\boldsymbol{\theta})]$ (Einstein summation throughout). Being positive semi-definite and transforming as a rank-2 tensor under smooth diffeomorphisms, $\mathbf{F}$ is a natural Riemannian metric that underpins the likelihood's *information geometry* (Amari, 2016; Nielsen, 2020). Crucially, it is not invariant to how we choose to parametrise the system. Suppose we have some parametrisation in terms of the parameters $\{\theta^a\}$, for which we have the Fisher matrix $\mathbf{F}$. If we now change to a new parametrisation in terms of $\{\eta^a\}$ then we have the new Fisher matrix

$$F'_{ab} = \frac{\partial\theta^c}{\partial\eta^a} \frac{\partial\theta^d}{\partial\eta^b}\, F_{cd}. \tag{2}$$

If all the parameters of a system acted completely independently, then we would have a diagonal Fisher matrix. However, in general, this is not the case, since some of the parameters are *degenerate*: changing the value of one of them changes the data in a similar way to changing another. In terms of the Fisher information, a statistical degeneracy is a parameter-space direction $\delta\boldsymbol{\theta}$ along which the metric has a small eigenvalue, $\delta\boldsymbol{\theta}^\top \mathbf{F}\, \delta\boldsymbol{\theta} \approx 0$. This is a statement about the model's *sensitivity* in that direction, and not about the intrinsic curvature of the information geometry.

Ideally, one would find a parametrisation in $\boldsymbol{\eta}$ that makes the Fisher matrix diagonal everywhere in parameter space, and hence all degeneracies would be removed. It is even more desirable to have a parametrisation in which $\mathbf{F}$ is equal to the identity, such that a change of unity in any of the $\boldsymbol{\eta}$ parameters has a similar effect on the likelihood.

Utilising the analogy with geometry, this corresponds to finding a coordinate system (parametrisation) in which the metric (Fisher matrix) is as close to the identity matrix as possible. This is achievable exactly, and everywhere, only where the Riemann tensor of $g_{ab} = F_{ab}$ vanishes (Lee, 2018).

Intrinsic curvature is a coordinate-independent invariant, so wherever it is non-zero no reparametrisation can make the components equal to the identity everywhere: Appendix E exhibits a Gaussian model with Ricci scalar $R = -1$, for which no globally flat coordinates exist. We therefore do not claim to remove curvature. What we optimise is the *coordinate representation* of the metric: given a prior, we seek a single map, global within that prior's support, that minimises the expected Frobenius distance of the transformed components from the identity. Our goal is to find such a (symbolic) transformation automatically from data–parameter pairs.

4 METHOD

The pipeline is summarised in Fig. 1. We separate the construction into three stages so that each can be assessed independently.

1. Mapping information geometry with Fishnets. Given a set of parameter-data pairs, $\{(\boldsymbol{\theta}, \mathbf{x})\}$, we first wish to estimate the Fisher matrix at all points in parameter space. Traditionally, this is done by running new simulations at perturbed values of the parameters and using finite differences, or by leveraging auto-differentiation. The former can be costly, and the latter is only possible when one has a differentiable pipeline. To circumvent these issues, we use the Fishnets formalism (Makinen et al., 2023), which we briefly review here. Given the data vector $\mathbf{x}$, we train a compression network to predict estimates of the parameters, $\hat{\boldsymbol{\theta}}(\mathbf{x})$, as well as the Fisher matrix, $\mathbf{F}(\mathbf{x})$. Both functions are controlled by a set of model weights, φ. To ensure a positive-definite matrix, the network does not directly output $\mathbf{F}$, but instead its Cholesky decomposition, $\mathbf{L}$, such that $\mathbf{F} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$.

Our optimisation objective is

$$L_{\mathrm{fish}}(\varphi) = \tfrac{1}{2}\Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{x})\Big)^{\top} \mathbf{F}(\mathbf{x}) \Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{x})\Big) - \tfrac{1}{2}\log\det \mathbf{F}(\mathbf{x}), \tag{3}$$

minimised over batches of pairs $(\boldsymbol{\theta}, \mathbf{x})$:

$$\min_{\varphi} L_{\mathrm{fish}} = \min_{\varphi} \int \mathrm{d}\mathbf{x}\, \mathrm{d}\boldsymbol{\theta}\, L_{\mathrm{fish}}\, p(\mathbf{x}, \boldsymbol{\theta}), \tag{4}$$

which can be seen as a variational minimisation over parameter space of Eq. (1). We show in Appendix A that minimising this objective yields the posterior mean and (inverse) covariance, which approaches the Fisher information only asymptotically, under a Gaussian (Laplace) approximation.

The Fishnets prediction is a noisy estimator of the underlying information, so we train an ensemble of 10–20 networks with randomised hidden widths and activations and combine them entrywise into a validation-weighted mean $\bar{\mathbf{F}} = \sum_i w_i \mathbf{F}^{(i)}$ and spread δF_{ab}, with $w_i \propto \exp(-L_i)$ for the validation loss L_i of member i and no covariance between entries (Appendix D). This ensemble size is what makes the sampling distribution of the entries of $\mathbf{F}$ approximately Gaussian, which the symbolic stage relies on.

2. Neural coordinates. Once we have an estimate for the Fisher matrix at all points in parameter space, we wish to find a new set of coordinates for which the Fisher is close to the identity everywhere. We parametrise this coordinate transformation with a neural network which takes the original coordinate values as input and outputs the transformed coordinates, $\boldsymbol{\eta} = \boldsymbol{\eta}_{\mathrm{NN}}(\boldsymbol{\theta})$. One can take the derivative of this network with respect to the inputs to obtain its Jacobian $\mathbf{J} = [\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}]$. We use the pseudo-inverse of $\mathcal{J}$ to transform $\bar{\mathbf{F}}$ to its value in the new coordinates, $\bar{\mathbf{F}}'$, using Eq. (2).

To train the network, we compute the Frobenius norm between $\bar{\mathbf{F}}'$ and the identity:

$$\ell(\mathbf{A}) = \big\|A_{ab} - \delta_{ab}\big\|^2_{\mathcal{F}} \equiv \mathrm{tr}\big((\mathbf{A}-\mathbf{1})(\mathbf{A}-\mathbf{1})^{\mathrm{T}}\big)\,. \tag{5}$$

We additionally penalise the inverse, so that the loss is $L(\boldsymbol{\eta}) = \ell(\bar{\mathbf{F}}') + \ell(\bar{\mathbf{F}}'^{-1})$. This loss is summed over all training points, averaging over the prior from which the simulations were drawn. Instead of locally flattening our coordinates around a fiducial point, this corresponds to making them approximately flat over all of space. We again ensemble: one network is fitted using $\bar{\mathbf{F}}$, copied, and each copy fine-tuned against a different member $\mathbf{F}^{(i)}$ to give $\boldsymbol{\eta}^{(i)}$. Once aligned (Appendix B) these are combined into $\bar{\eta}^a$ and $\delta\eta^a$ under the same validation weighting, so every coordinate carries an uncertainty for the symbolic stage. We provide several architecture options for the flattening operator, but use an inverse-conditioned, fully-connected MLP by default (see Appendix D).

Unified Loss Function. The first two steps can be combined into a single objective which provides advantages in cases where data is controlled by a much smaller set of latents $\boldsymbol{\eta} \in \mathbb{R}^m$ than input parameters, $\boldsymbol{\theta} \in \mathbb{R}^d$. Here the learned latent space is taken to be a unit Gaussian,

$$L_{\text{one}}[\phi, \psi] = \mathbb{E}_{\boldsymbol{\theta}, \mathbf{x}} \left[\tfrac{1}{2} \|\boldsymbol{\eta}_\phi(\boldsymbol{\theta}) - \hat{\boldsymbol{\eta}}_\psi(\mathbf{x})\|^2 - \tfrac{1}{2} \log \det \left(\mathbf{J}_\phi \mathbf{J}_\phi^\top \right) \right], \tag{6}$$

where the objective matches a function of parameters $\boldsymbol{\eta}_\phi(\boldsymbol{\theta})$ to a function of the data, $\hat{\boldsymbol{\eta}}_\psi(\mathbf{x})$, and is regularised by the reparameterisation $\mathbf{J}_\phi(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}$; the log-determinant term reduces to $\log |\det \mathbf{J}_\phi|$ when $m = d$. Appendix C relates this objective to Steps 1–2 at $m = d$. Replacing its profiled posterior covariance by the Jacobian approximation $\mathbf{G}^{-1}$, with $\mathbf{G} = \bar{\mathbf{F}}'$, gives $S_{\text{one}}^{\text{loc}} = \text{const} + \mathbb{E}_\pi D_{\text{KL}}(\mathcal{N}(0, \mathbf{G}^{-1}) \,\|\, \mathcal{N}(0, \mathbf{1}))$: Step 2 with a KL penalty in place of the Frobenius penalty $S_{\text{flat}} = \mathbb{E}_\pi \|\mathbf{G} - \mathbf{1}\|_{\mathcal{F}}^2$.

Theorem 1 (informal)**.** *(i) If exact flattening ($\mathbf{G} = \mathbf{1}$) is achievable in the admissible class, then S_{flat} (with or without the inverse penalty) and $S_{\text{one}}^{\text{loc}}$ have the same minimisers, unique up to a rigid motion. (ii) If $\min S_{\text{flat}} = \epsilon > 0$, the minimiser of S_{flat} is not stationary for $S_{\text{one}}^{\text{loc}}$: under rescaling its derivative is at least $d\epsilon/(d - \epsilon)$.*

Both parts carry over to Eq. (6) when it equals $S_{\text{one}}^{\text{loc}}$ up to a constant, as for affine maps; a neural class does not guarantee this (Theorem 2 and Proposition 2). In practice, the number of latents m is a hyperparameter that can be tuned to reduce inference dimensionality. For network diversity we vary the data function architecture $\hat{\boldsymbol{\eta}}_\psi$ along the same prescription as Fishnets, with each resulting $\boldsymbol{\eta}_\phi^{(i)}(\boldsymbol{\theta})$ weighted by validation loss.

3. Symbolic coordinates. Both routes to $\boldsymbol{\eta}$ determine it only up to a constant offset and a global orthogonal transformation, so we first align the ensemble to a common canonical frame (Appendix B) and then let the MDL criterion below pick the simplest expression. Isotropy alone thus does not single out interpretable axes; the selection principle is isotropy plus MDL. We seek a symbolic approximation to the aligned map. We run symbolic regression (Kronberger et al., 2024) with OPERON (Burlacu et al., 2020) separately on each output coordinate—and so assume sufficiently independent component errors—under a Gaussian loss using the ensemble mean and uncertainty. Once trained, $\boldsymbol{\eta}(\boldsymbol{\theta})$ is decoupled from the simulator, so we draw a fresh, dense set of parameters from the prior and pass them through the flattening ensemble to build the augmented $(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}, \delta\boldsymbol{\eta})$ training set used by SR. This augmentation was the most consistent way to improve convergence and costs no simulator calls ($\sim$2,000 prior draws against $\sim$250 Fishnet validation points on Rosenbrock); the OPERON configuration and per-component 120 s budget are in Appendices D and F. From the Pareto front we select using a minimum-description-length (MDL) criterion for symbolic regression (Bartlett et al., 2024), which penalises poor fit, complexity, and many or finely tuned parameters, and approximates the log probability of the function given the data under a prior on equation form (Bartlett et al., 2023).

As an alternative we select component-wise on Eq. (5) itself, substituting the symbolic gradient for the network's Jacobian column $\partial \eta_i / \partial \boldsymbol{\theta}$ and keeping the expression that changes the loss least. This does not penalise expression length, but usually yields a geometry closer to the identity. All models are then pruned (Appendix D.4).

5 EXPERIMENTS

We apply our method to both synthetic and scientific applications, demonstrating the physical insight the learned coordinates give and the efficiency gains degeneracy resolution brings to Bayesian inference, in particular neural posterior estimation (NPE): a conditional density estimator $q(\boldsymbol{\theta} \,|\, \mathbf{x})$ trained on simulated pairs, which amortises inference across observations and whose sample efficiency depends strongly on how well the target coordinates are conditioned (Papamakarios et al., 2021; Alsing et al., 2019). Each result below is one representative run. Appendix G reports 10 fully independent end-to-end trials per problem, for both the three-step and the one-step pipeline, together with a one-factor ablation of the SIR observation noise and prior width (Table 2).

5.1 SYNTHETIC VALIDATION

Rosenbrock. The Rosenbrock function is a two-dimensional test whose Fisher geometry maps to the identity in flat coordinates (Appendix F.1). Fig. 2 uses $\boldsymbol{\theta} = (\mu_0, \mu_1)$ and $\boldsymbol{\eta} = (\nu_0, \nu_1)$. The

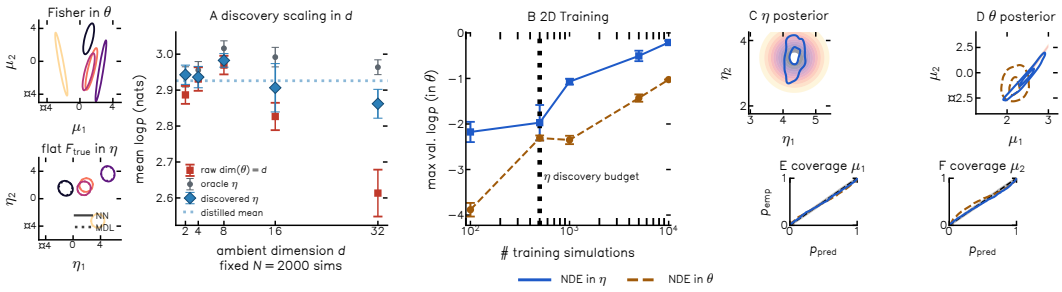

Figure 2: Rosenbrock validation case. *Left*: Learned Fishers (top) are flattened by learning η (bottom). *Middle*: Recovering a hidden 2D η in increasing ambient dimensions prevents downstream inference degradation. *Right*: Knowing the symbolic η B) makes learning posteriors easier even with increasing simulation budget C) yields faithful posteriors in target θ coordinates and D) is better calibrated than learning in degenerate θ space.

Fishnet ensemble gives an unbiased and calibrated estimate of the true Fisher geometry. Fig. 6 shows the Cholesky factors and the trace-normalised shape $M_{ij} = F_{ij}/\operatorname{tr}\mathbf{F}$. Flattening gives neural η and residuals $\delta\eta$ consistent with $\mathcal{N}(0,1)$, which suits the MDL selection. Symbolic regression recovers $\eta = [b_0\mu_0^2 + b_1\mu_1,\ b_2\mu_1]$ in 2D.

We train NPE ensembles on $\Sigma = \operatorname{diag}(4^2, 1^2)$ across simulation budgets. Learned η recovers the true posterior with better calibration than raw θ. At matched *maximum validation log-probability* inference in η reaches that of θ's score with 9–14$\times$ fewer simulations (panel B); scored instead by CRPS on the same held-out observations the gain at matched budget is $1.2\times$. We quote each factor with its metric, since the two differ by an order of magnitude. The baseline is NPE trained on the same simulations in the original parameters, with identical architecture, ensemble size, and seeds; only the coordinates the density is fitted in differ. Discovery is a one-time cost, reused across observations (500 simulations). Where a likelihood is available we claim no gain over tuned samplers such as HMC, only interpretability. **Low-dimensional latent discovery.** We also hide this 2D banana in d ambient coordinates (Appendix F.1). Two parameters carry it, and the other $d-2$ are nuisance, yielding a rank-2 Fisher at every d. This experiment uses the same one-step loss Eq. (6) and the same settings as Section 5.2. We recover the degeneracy *and its symbolic form* as d grows, and compare NPE inference performance in η versus θ. At fixed $N = 2000$, raw $\log p$ decreases for $d \geq 16$. Discovered η stays near the oracle up to $d = 32$ (panel A).

Gaussian with unknown variance. We now turn to an example which cannot be flattened exactly at all points in parameter space. We consider a one-dimensional Gaussian, parametrised by mean μ and standard deviation σ. Taking draws from this distribution, the Fisher matrix is $F_{ab} = n_{\mathrm{d}} \operatorname{diag}(1/\sigma^2, 2/\sigma^2)$. As discussed in Appendix E, the Riemann tensor for a metric with $g_{ab} = F_{ab}$ is non-zero in this case (Eq. (82)), confirming that one cannot flatten the Fisher matrix everywhere. Hence, we wish to find a choice of parameters which makes the Fisher as flat as possible, even if one cannot make it identical to the identity everywhere. Given a fiducial point, it is possible to enforce only quadratic deviations from flatness through the use of geodesic normal coordinates. These are constructed in Appendix E, and are used as a benchmark: when averaged over the prior, our method should construct coordinates with a lower loss than these analytic ones, although close to the fiducial point the analytic ones must perform better. We choose the fiducial point to be $(0.0, 1.0)$; simulation, prior, and network details are in Appendix F. **Results.** We obtain (i) neural coordinates and (ii) two sets of symbolic coordinates selected according to (a) MDL and (b) Frobenius criteria. We compare the "flatness" of our coordinates against *true*, analytic Fisher matrices as a function of geodesic distance β (see Appendix E), computed over test data. For comparison, we compute geodesic normal coordinates around the fiducial point $\theta^\star = (\mu^\star, \sigma^\star) = (0, 1)$, as well as "ad-hoc" coordinates typically used for Gaussian parameter inference, $\eta^{\text{ad-hoc}} = (\mu/\sigma, \log\sigma)$. Fig. 3 shows that geodesic coordinates obtain perfect flattening close to the fiducial point ($\beta \approx 0$), but fail exponentially quickly further away. Both neural and symbolic coordinates consistently flatten the true geometry across the parameter space.

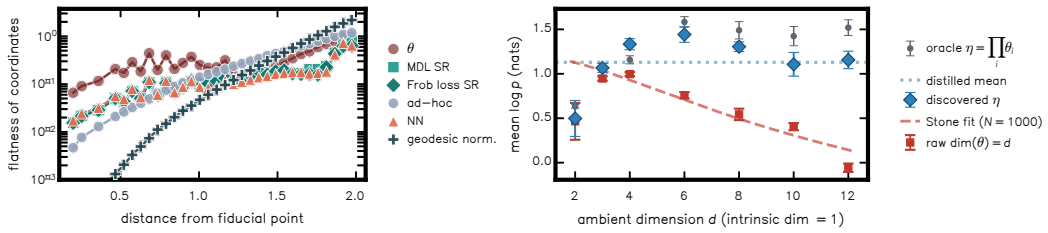

Figure 3: *Left:* Neural and symbolic regression components outperform (i) hand-picked (ad-hoc) coordinates and (ii) flatten geometry far from the fiducial point Eq. (107), without the need for an analytic Fisher matrix to derive geodesic normal coordinates. *Right:* Using Eq. (6), finding true latents prevents inference degradation at fixed training budget N. Raw $\log p$ in $\theta \in \mathbb{R}^d$ decreases with ambient dimension d. The dashed curve is a Stone fit at $\beta = 1$. Discovered η aligns with the oracle for $d \geq 6$.

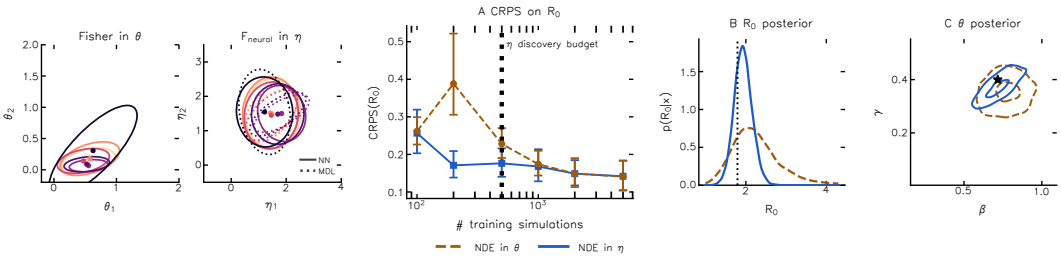

Figure 4: Epidemiology example. Distilled coordinates bring the learned Fisher matrices close to the identity across the prior (*left*). Amortised η make downstream NPE more efficient in $R_0 = \beta/\gamma$, measured by CRPS at fixed simulation budgets and higher noise levels (*middle*). *Right:* Posteriors sampled in η capture R_0 with higher fidelity and trace the degeneracy direction in (β, γ), shown for the $N_{\text{sim}} = 500$ case.

5.2 Scientific Applications

Industrial Heater control. Here we show that the distillery can recover a hidden low-dimensional latent parameter via a synthetic resistive heater experiment. Input parameters are a voltage and a current, $\theta = (V, I)$. The data are a 20-point temperature step response (Bergman et al., 2011),

$$y(t) = (V \cdot I)\,(1 - e^{-t/\tau}) + \varepsilon(t), \qquad \varepsilon(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{7}$$

The temperature depends on (V, I) only through the power $P = VI$. The Fisher matrix has rank 1 at every θ. The data manifold is 1-D. We extend the plant to a product of d inputs, where the intrinsic dimension is 1 at every d. The Fisher rank is 1 at every θ and every d,

$$y(t) = \Big(\prod_{i=1}^{d} \theta_i\Big)\,(1 - e^{-t/\tau}) + \varepsilon(t), \qquad \theta_i \sim \mathcal{U}[1, 2]. \tag{8}$$

We use the one-step loss Eq. (6) using 10^3 total simulations. This is the hidden-banana experiment of Section 5.1 at intrinsic rank 1 instead of rank 2. A screen ranks the four probe axes by held-out information and returns a rank $\hat{r}$. We refit a fresh network at that rank. We then align eight copies and pass them to symbolic regression. A trial counts as correct when $\hat{r} = 1$ and the recovered coordinate has absolute Spearman correlation at least 0.99 with $\prod_i \theta_i$. The coordinate is *discovered* at every d, not supplied analytically.

We then train three NPE mixture density networks for posteriors $p(\cdot|y)$ over raw θ, the oracle $\prod_i \theta_i$, and the discovered η. Fig. 3 plots the one-dimensional log density at the true axis value: raw $\log p$ decreases with d along a Stone fit at $\beta = 1$, while discovered η tracks the oracle for $d \geq 6$. Appendix F.5 defines the screen, the refit, and the score, at a fixed $N = 2000$ simulation budget.

Epidemic infection rates (SIR). The Susceptible–Infected–Recovered model is a standard model of epidemic dynamics (Kermack & McKendrick, 1927; Hethcote, 2000). Epidemic-outbreak dynamics can be modelled using coupled ordinary differential equations describing susceptible, infectious, and recovered populations, governed by infection (β) and recovery (γ) rates. A disease becomes epidemic when its basic reproduction number $R_0 = \beta/\gamma > 1$, corresponding to the case where one infected person infects more than one other person on average. **Data.** Epidemiologists usually only see noisy estimates of infection numbers over time. Here we ask whether a nonlinear reparameterisation of (β, γ) can be learned from noisy simulations alone. We focus on the *epidemic regime*, retaining simulations with $R_0 \geq 1.15$ and discarding a narrow band around the epidemic threshold (details in Appendix F.3). Each simulation vector is a normalised infectious trajectory $I(t)/N$ on a uniform grid of 30 time-points over $[0, 50]$ with additive Gaussian noise of standard deviation 0.01. To distil equations, we use 500 simulations in (β, γ). To demonstrate the usefulness of the learned coordinates, we extend our validation inference problem to include the initial number of infected individuals, I_0. The full parameter vector at inference is then $\theta = (\beta, \gamma, I_0/10)$. **Results.** We recover the symbolic coordinates

$$\eta_1 = \frac{0.64\beta}{1.41\gamma - 0.72} + 0.54, \quad \eta_2 = 0.55\beta + 0.67$$

whose first component follows the expected $R_0 \propto \beta/\gamma$ direction (mean gradient cosine 0.94 with ∇R_0, Table 2), flattening the learned Fisher matrices shown in Fig. 4. Downstream neural posterior estimation using amortised η coordinates gives a sharper R_0 posterior: scored by CRPS against held-out observations, inference in η at 500 simulations is matched by the θ baseline only at $\sim$970, a factor of $\sim$2, and the paired per-observation difference is significant at that budget. The gain closes by $\sim$2,000 simulations, which is the expected behaviour for a reparameterisation: it helps when simulations are scarce and noise is high, and stops mattering once the baseline has enough data to learn the degeneracy itself (details in Appendix F.3).

Gravitational-wave chirp mass. Here we apply our method to gravitational-wave (GW) data, with the goal of inferring the component masses $\theta = (m_1, m_2)$ of a compact binary black-hole system from full waveform data. We compare two simulation regimes with increasing complexity. In both cases, we discover coordinates with $N = 500$ simulations and benchmark against the conventional chirp-mass / mass-ratio combination $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$, $q = m_2/m_1$ used in GW astronomy.

Run 1. TaylorF2 frequency-domain inspiral waveforms (Buonanno et al., 2009; Blanchet, 2014) at 2PN order (non-spinning), whitened by the Advanced LIGO design sensitivity (Aasi et al., 2015) and PCA-compressed to 40 dimensions, over a uniform prior in $m_1 \geq m_2 \in [5, 50]\, M_\odot$. Heavy systems terminate at lower $f_{\mathrm{ISCO}} \propto M^{-1}$ and accumulate fewer in-band cycles, so no closed-form flat coordinates in (m_1, m_2) exist. **Results.** With masses rescaled to $[0.1, 1.0]$ the distillery returns $\eta_1 = 0.2\, m_1$, $\eta_2 = 5.5 \exp(-0.08(m_1 + m_2)^2)$, invertible over the prior studied: the second coordinate captures the inverse chirp-mass-like sensitivity along the heavy-mass direction, and together they form a bijection of $(\mathcal{M}_c, q)$ (Appendix F.6 and Fig. 10).

Run 2. Beyond the chirp-mass approximation. TaylorF2 only captures the inspiral phase of the binary signal; for moderately heavy systems the merger and ringdown carry most of the in-band signal-to-noise and the leading-order $\mathcal{M}_c \propto (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ scaling no longer dominates the waveform. We repeat the experiment with the phenomenological inspiral-merger-ringdown waveform IMRPHENOMD (Khan et al., 2016; Husa et al., 2016), keeping the prior and PCA pipeline of the TaylorF2 case unchanged. **Results.** The distillery now returns coordinates that do not contain the leading-order chirp-mass factor explicitly: the symbolic outputs are

$$\eta_1 = 0.34\,(m_1 + m_2)^2 - 1.9, \quad \eta_2 = -0.1\, m_1 - 0.3\, m_2 + 0.10\,(m_1 + m_2)^2,$$

i.e. a monotonic function of total mass $g(M)$ and the asymmetric combination $\Delta m = m_1 - m_2$, with no chirp-mass-like factor anywhere. Averaged over the prior, the discovered coordinates bring the metric substantially closer to the identity than $(\mathcal{M}_c, q)$ does for the IMRPHENOMD case, and the downstream posteriors are rounder and better calibrated under TARP (Lemos et al., 2023)—a test of whether credible regions attain their nominal coverage—than the chirp-mass baseline (Fig. 5; details in Appendix F.7). This result indicates that the distillery has identified a more informative, data-driven combination of parameters. Total mass is recovered on essentially every independent trial; the misses in Appendix G are threshold misses on the harder asymmetric direction, not failures to run.

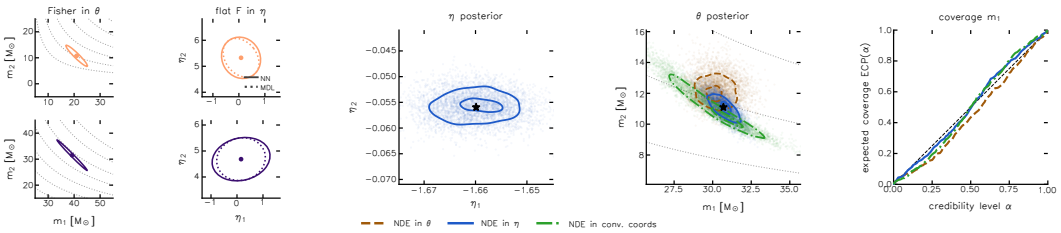

Figure 5: Main GW result (IMRPHENOMD). *Left*: Example Fisher matrices in (m_1, m_2) and in the learned η coordinates. *Middle*: posterior contours from neural posterior estimators trained in $\theta = (m_1, m_2)$, in the η basis, and in the conventional $(\mathcal{M}_c, q)$ basis. The η-trained posterior is round and centred on the truth, while both θ-space and $(\mathcal{M}_c, q)$ posteriors stretch into bananas along the $\mathcal{M}_c$ direction (grey contours). *Right*: TARP joint coverage; η is closest to the diagonal across the full credibility range, indicating the learned coordinates are the best calibrated of the three.

Weak gravitational lensing. Weak gravitational lensing (WL) alters the trajectories of photons as they pass through massive structures of visible and dark matter. The two-point correlation function of this observable is sensitive to two degenerate parameters, $(\Omega_{\rm m}, \sigma_8)$, which control the Universe's matter content and dark-matter clustering, respectively. The nonlinear parameter $S_8 = \sigma_8(\Omega_{\rm m}/0.3)^{0.5}$ is conventionally used to summarise the leading amplitude degeneracy to changes in the two-point function (Jain & Seljak, 1997; Bernardeau et al., 1997; Kilbinger, 2015). We apply our method to 2-point statistics from mock WL convergence images, computed from expensive dark matter simulations (see Appendix F.4), varied over wide priors in (Ω_m, σ_8) and initial conditions. No tractable likelihood exists for these maps: the pipeline sees only $1,000$ parameter-simulation pairs for discovery, and there is no analytic Fisher matrix to check against. The tractable examples above were chosen so the answer could be checked. **Results.** We recover flattened coordinates $\eta_1 = \Omega_m\sigma_8 - 0.9\,\Omega_m^{0.144}$ and $\eta_2 = 0.5(\Omega_m - 1.0)$. These coordinates are similar to the standard (Ω_m, S_8) parameterization. In particular, η_1 captures the dominant weak-lensing amplitude degeneracy: near typical values $\Omega_m \simeq 0.3$ and $\sigma_8 \simeq 0.8$, contours of constant η_1 have local slope $d\ln\sigma_8/d\ln\Omega_m \simeq -0.55$, close to the constant-$S_8$ slope of -0.5.

6 Discussion and Limitations

Across all experiments, the distillery captured information geometry and produced amortised, symbolic coordinates that make downstream inference more simulation-efficient and better calibrated in low- and noisy-data regimes, and provide physical insight about how observables are controlled by model parameters of interest.

Limitations. Our method has five main limitations. (i) Genuinely *multimodal* posteriors cannot be resolved by an invertible reparametrisation, since the Fisher information is a local, second-order object and is blind to global mode structure. (ii) The demonstrated regime is low-dimensional: the symbolic stage is shown at $n_\theta \leq 3$, and the dimensionality-reduction step that precedes it at ambient d up to 32 (Fig. 2) and 12 (Fig. 3). Symbolic regression scales poorly in the number of input variables, and stable Fisher estimation at $n_\theta \gtrsim 100$ would need substantially larger networks; we make no claim there. (iii) An exactly singular $\mathbf{F}$ cannot be resolved by an invertible square map (Transtrum & Qiu, 2014). The one-step loss Eq. (6) handles it by learning a rank-m map with $m < d$, as in the heater and hidden-banana experiments, but the difficulty then moves to selecting m, which we do by thresholding held-out information. (iv) The Fishnet stage is only as accurate as the underlying Gaussian-posterior assumption (Appendix A); for strongly non-Gaussian likelihoods, the recovered metric is the inverse of the posterior covariance rather than the Fisher. (v) The symbolic stage is a stochastic search under a fixed wall-clock budget (Appendix G).

The distillery turns parameter–simulation pairs into closed-form statements about which parameter combinations the data constrain. It recovers known degeneracies and hidden low-dimensional structure, finds better-conditioned alternatives to conventional coordinates, and focuses downstream inference on the informative directions, enabling simulation-efficient inference and physical insight.

Reproducibility Statement

All architectures, optimiser settings, and symbolic-regression hyperparameters are given in Appendix D; per-experiment simulators, priors, budgets, and downstream-inference settings in Appendix F; and the proofs supporting the Fisher estimator and the one-step objective in Appendices A and C. Appendix G reports 10 fully independent end-to-end trials per problem, with numerically evaluated success criteria for both neural pipelines.

AI Use Statement

In this work, we used generative AI tools to help formulate mathematical claims and write proofs (the analysis relating the one-step and Fisher-flattening objectives in Appendix C, summarised in Theorem 1), to implement methods, and to help interpret results. We have not used generative AI tools to design or give feedback on research methodology or experiments, and the remaining required-disclosure tasks (generating synthetic datasets, developing theoretical models, proposing hypotheses, translation, cleaning datasets, and qualitative data analysis) are not applicable to this work. Additionally, we used generative AI tools for code creation and for creating or modifying figures. We have reviewed all AI-assisted work: AI-assisted proofs were checked by two authors. Generated code was tested against analytic cases and author-written metrics and plotting scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

J. Aasi et al. Advanced LIGO. *Class. Quantum Grav.*, 32:074001, 2015. doi: 10.1088/0264-9381/32/7/074001.

Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*, 488(3):4440–4458, Sep 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz1960. URL http://dx.doi.org/10.1093/mnras/stz1960.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016. doi: 10.1007/978-4-431-55978-8.

Deaglan Bartlett, Harry Desmond, and Pedro Ferreira. Priors for symbolic regression. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, pp. 2402–2411. ACM, July 2023. doi: 10.1145/3583133.3596327. URL http://dx.doi.org/10.1145/3583133.3596327.

Deaglan J. Bartlett, Harry Desmond, and Pedro G. Ferreira. Exhaustive symbolic regression. *IEEE Transactions on Evolutionary Computation*, 28(4):950–964, August 2024. ISSN 1941-0026. doi: 10.1109/tevc.2023.3280250. URL http://dx.doi.org/10.1109/TEVC.2023.3280250.

Theodore L. Bergman, Adrienne S. Lavine, Frank P. Incropera, and David P. DeWitt. *Fundamentals of Heat and Mass Transfer*. Wiley, Hoboken, NJ, 7 edition, 2011.

Francis Bernardeau, Ludovic van Waerbeke, and Yannick Mellier. Weak lensing statistics as a probe of ω and power spectrum. *Astronomy & Astrophysics*, 322:1–18, 1997.

C. M. Biwer, Collin D. Capano, Soumi De, Miriam Cabero, Duncan A. Brown, Alexander H. Nitz, and V. Raymond. PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals. *Publ. Astron. Soc. Pac.*, 131:024503, 2019. doi: 10.1088/1538-3873/aaef0b.

Luc Blanchet. Gravitational radiation from post-Newtonian sources and inspiralling compact binaries. *Living Rev. Relativ.*, 17:2, 2014. doi: 10.12942/lrr-2014-2.

Alessandra Buonanno, Bala R. Iyer, Evan Ochsner, Yi Pan, and B. S. Sathyaprakash. Comparison of post-Newtonian templates for compact binary inspiral signals in gravitational-wave detectors. *Phys. Rev. D*, 80:084043, 2009. doi: 10.1103/PhysRevD.80.084043.

Bogdan Burlacu, Gabriel Kronberger, and Michael Kommenda. Operon c++: An efficient genetic programming framework for symbolic regression. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, GECCO '20, pp. 1562–1570, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371278. doi: 10.1145/3377929.3398099. URL https://doi.org/10.1145/3377929.3398099.

Tara Dacunha, Marco Raveri, Minsu Park, Cyrille Doux, and Bhuvnesh Jain. What does a cosmological experiment really measure? Covariant posterior decomposition with normalizing flows. *Physical Review D*, 105(6):063529, March 2022. doi: 10.1103/PhysRevD.105.063529.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Nikolaos Evangelou, Noah J. Wichrowski, George A. Kevrekidis, Felix Dietrich, Mahdi Kooshkbaghi, Sarah McFann, and Ioannis G. Kevrekidis. On the parameter combinations that matter and on those that do not: data-driven studies of parameter (non)identifiability. *PNAS Nexus*, 1(4): pgac154, September 2022. doi: 10.1093/pnasnexus/pgac154.

Philipp Frank, Reimar Leike, and Torsten A. Enßlin. Geometric Variational Inference. *Entropy*, 23 (7):853, July 2021. doi: 10.3390/e23070853.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Computational Biology*, 3(10):e189, October 2007. doi: 10.1371/journal.pcbi.0030189.

Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000. doi: 10.1137/S0036144500371907.

Alston S. Householder. Unitary Triangularization of a Nonsymmetric Matrix. *J. ACM*, 5(4):339–342, October 1958. ISSN 0004-5411. doi: 10.1145/320941.320947. URL https://dl.acm.org/doi/10.1145/320941.320947.

Sascha Husa, Sebastian Khan, Mark Hannam, Michael Pürrer, Frank Ohme, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. *Phys. Rev. D*, 93:044006, 2016. doi: 10.1103/PhysRevD.93.044006.

Bhuvnesh Jain and Uroš Seljak. Cosmological model predictions for weak lensing: Linear and nonlinear regimes. *The Astrophysical Journal*, 484(2):560–573, 1997. doi: 10.1086/304372.

W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, September 1976. ISSN 0567-7394. doi: 10.1107/S0567739476001873. URL https://journals.iucr.org/a/issues/1976/05/00/a12999/.

William Ogilvy Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118.

Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *Phys. Rev. D*, 93:044007, 2016. doi: 10.1103/PhysRevD.93.044007.

Martin Kilbinger. Cosmology with cosmic shear observations: a review. *Reports on Progress in Physics*, 78(8):086901, 2015. doi: 10.1088/0034-4885/78/8/086901.

Gabriel Kronberger, Bogdan Burlacu, Michael Kommenda, Stephan M. Winkler, and Michael Affenzeller. *Symbolic Regression*. Chapman & Hall / CRC Press, 2024.

John M. Lee. *Introduction to Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2 edition, 2018. ISBN 978-3-319-91754-2.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference, 2023. URL https://arxiv.org/abs/2302.03026.

Yin Li, Libin Lu, Chirag Modi, Drew Jamieson, Yucheng Zhang, Yu Feng, Wenda Zhou, Ngai Pok Kwan, François Lanusse, and Leslie Greengard. pmwd: A Differentiable Cosmological Particle-Mesh N-body Library, 2022. arXiv:2211.09958.

T. Lucas Makinen, Justin Alsing, and Benjamin D. Wandelt. Fishnets: Information-optimal, scalable aggregation for sets and graphs. *arXiv preprint arXiv:2310.03812*, Oct 2023.

T. Lucas Makinen, Alan Heavens, Natalia Porqueres, Tom Charnock, Axel Lapel, and Benjamin D. Wandelt. Hybrid summary statistics: neural weak lensing inference beyond the power spectrum. *Journal of Cosmology and Astroparticle Physics*, 2025(01):095, January 2025. ISSN 1475-7516. doi: 10.1088/1475-7516/2025/01/095. URL http://dx.doi.org/10.1088/1475-7516/2025/01/095.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL https://doi.org/10.7717/peerj-cs.103.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10):1100, October 2020. ISSN 1099-4300. doi: 10.3390/e22101100. URL https://www.mdpi.com/1099-4300/22/10/1100.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Erez Peterfreund, Ofir Lindenbaum, Felix Dietrich, Tom Bertalan, Matan Gavish, Ioannis G. Kevrekidis, and Ronald R. Coifman. Local conformal autoencoder for standardized data coordinates. *Proceedings of the National Academy of Science*, 117(49):30918–30927, December 2020. doi: 10.1073/pnas.2014627117.

Calyampudi Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

H. H. Rosenbrock. An Automatic Method for Finding the Greatest or Least Value of a Function. *The Computer Journal*, 3(3):175–184, March 1960. ISSN 0010-4620. doi: 10.1093/comjnl/3.3.175. URL https://doi.org/10.1093/comjnl/3.3.175.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. doi: 10.1214/aos/1176345206.

Mark K. Transtrum and Peng Qiu. Model reduction by manifold boundaries. *Phys. Rev. Lett.*, 113:098701, Aug 2014. doi: 10.1103/PhysRevLett.113.098701. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.098701.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, March 2011. doi: 10.1103/PhysRevE.83.036701.

Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.*, 143(1):010901, July 2015. doi: 10.1063/1.4923066.

BROADER IMPACT

This method is a tool for understanding and accelerating scientific inference. The most direct impact is on simulation-based inference for problems where simulation cost is the dominant bottleneck. By focusing learning from simulations along informative directions, the distillery can reduce the energy and compute cost of large inference campaigns and uncertainty quantification. The amortised, symbolic outputs additionally mitigates the opacity of black-box neural simulation-based inference. We do not foresee specific negative societal impacts beyond those generic to machine-learning tools deployed in scientific domains, although the epidemiology application could be misapplied if used for policy decisions without expert calibration.

APPENDIX

The appendices are organised as follows. Appendix A proves that the Fishnet objective recovers the posterior mean and covariance, and motivates its use as a Fisher estimator. Appendix B describes the alignment scheme used to fix the constant-offset and orthogonal symmetries of the flattening loss when ensembling. Appendix C relates the one-step neural objective to the Fisher-then-flatten approach. Appendix D details the Fishnet, flattener, and symbolic- regression architectures and training hyperparameters. Appendix E constructs geodesic normal coordinates for the Gaussian validation example used in Section 5.1. Appendix F collects per-experiment simulation, prior, and downstream-inference details. Appendix G reports independent-trial recovery rates for both pipelines and a one-factor ablation of the SIR noise and prior width.

A FORWARD-KL MINIMISATION YIELDS THE POSTERIOR MOMENTS

This appendix establishes that the Fishnet objective (4), when treated as a variational loss over a Gaussian family, is minimised by the posterior mean and posterior covariance of θ given X. The identification with the Fisher information then follows in the standard Laplace / Bernstein–von Mises regime.

Setup. Let X be observed data and $\theta \in \mathbb{R}^d$ parameters, with joint density $p(X, \theta)$ and posterior $p(\theta \mid X)$. Throughout we assume the posterior has finite first and second conditional moments

$$m(X) := \mathbb{E}_{p(\theta|X)}[\theta], \qquad S(X) := \mathrm{Cov}_{p(\theta|X)}[\theta] \in \mathbb{S}^d_{++}, \tag{9}$$

where $\mathbb{S}^d_{++}$ denotes the cone of $d \times d$ symmetric positive-definite matrices. The Fishnet model parametrises a Gaussian variational family

$$q_\varphi(\theta \mid X) = \mathcal{N}\big(\theta;\, \mu_\varphi(X),\, \Sigma_\varphi(X)\big), \quad \Sigma_\varphi(X) \in \mathbb{S}^d_{++}, \tag{10}$$

through the network outputs $\mu_\varphi(X) := \hat{\theta}(X)$ and $\Sigma_\varphi(X) := \mathbf{F}_\varphi(X)^{-1}$, where $\mathbf{F}_\varphi$ is the log-Cholesky-parametrised matrix returned by the network. Writing $\mathcal{L}$ for the loss (4), the per-sample loss is

$$\ell(\varphi; X, \theta) = \tfrac{1}{2}\big(\theta - \mu_\varphi(X)\big)^\top \mathbf{F}_\varphi(X)\big(\theta - \mu_\varphi(X)\big) - \tfrac{1}{2}\log\det \mathbf{F}_\varphi(X), \tag{11}$$

and the population loss is $\mathbb{E}_{p(X,\theta)}[\ell(\varphi; X, \theta)]$.

A.1 MAIN RESULT

Proposition 1 (Forward-KL Gaussian moment matching)**.** *Fix X. Within the Gaussian family* (10)*, the conditional forward KL divergence $D_{\mathrm{KL}}\big(p(\theta \mid X) \,\big\|\, q_\varphi(\theta \mid X)\big)$ admits a unique minimum over $(\mu_\varphi, \mathbf{F}_\varphi) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$ at*

$$\mu^\star_\varphi(X) = m(X), \qquad \Sigma^\star_\varphi(X) = \mathbf{F}^\star_\varphi(X)^{-1} = S(X), \tag{12}$$

the posterior mean and posterior covariance of θ given X.

Proof. Direct expansion of the Gaussian density in (10) gives, for any $(\mu_\varphi, \mathbf{F}_\varphi)$,

$$\mathbb{E}_{p(\theta|X)}\big[\ell(\varphi; X, \theta)\big] = D_{\mathrm{KL}}\big(p(\theta \mid X) \,\big\|\, q_\varphi(\theta \mid X)\big) + H\big(p(\theta \mid X)\big) - \tfrac{d}{2}\log(2\pi), \tag{13}$$

where $H(\cdot)$ denotes differential entropy. Both correction terms are independent of φ, so minimising (11) in expectation and minimising the conditional KL are equivalent.

Decompose $\theta - \mu_\varphi = (\theta - m) + (m - \mu_\varphi)$ in the quadratic form. The cross-term vanishes because $\mathbb{E}_p[\theta - m] = 0$, and the trace identity $\mathbb{E}_p\big[(\theta - m)^\top A(\theta - m)\big] = \mathrm{tr}(AS)$, valid for any symmetric A, reduces (11) to

$$\mathbb{E}_p\big[\ell(\varphi; X, \theta)\big] = \underbrace{\tfrac{1}{2}(m - \mu_\varphi)^\top \mathbf{F}_\varphi (m - \mu_\varphi)}_{\text{(I)}\,\geq 0} + \underbrace{\tfrac{1}{2}\mathrm{tr}(\mathbf{F}_\varphi S) - \tfrac{1}{2}\log\det \mathbf{F}_\varphi}_{\text{(II)}}. \tag{14}$$

Term (II) does not depend on μ_φ, and term (I) is a non-negative quadratic form that vanishes if and only if $\mu_\varphi = m$ (since $\mathbf{F}_\varphi \succ 0$). For any $\mathbf{F}_\varphi \in \mathbb{S}^d_{++}$, replacing μ_φ by m therefore strictly decreases (14) unless $\mu_\varphi = m$ already; any minimiser must satisfy $\mu^\star_\varphi = m$.

Substituting $\mu_\varphi = m$ leaves the precision objective

$$g(\mathbf{F}) := \tfrac{1}{2}\mathrm{tr}(\mathbf{F}S) - \tfrac{1}{2}\log\det \mathbf{F}, \qquad \mathbf{F} \in \mathbb{S}^d_{++}. \tag{15}$$

The map $\mathbf{F} \mapsto -\log\det \mathbf{F}$ is strictly convex on $\mathbb{S}^d_{++}$ and $\mathbf{F} \mapsto \mathrm{tr}(\mathbf{F}S)$ is linear, so g is strictly convex on the (convex) cone $\mathbb{S}^d_{++}$. Standard matrix calculus gives $\nabla_{\mathbf{F}}\, g(\mathbf{F}) = \frac{1}{2}\big(S - \mathbf{F}^{-1}\big)$, which vanishes uniquely at $\mathbf{F}^\star = S^{-1}$, hence $\Sigma^\star_\varphi = S$. The pair $(\mu^\star_\varphi, \mathbf{F}^\star_\varphi) = (m, S^{-1})$ is the unique global minimiser of (14), and by (13) also of the conditional KL. □

Corollary 1 (Fishnets recover the posterior moments)**.** *Suppose the variational class $\{(\mu_\varphi, \mathbf{F}_\varphi) : \varphi \in \Phi\}$ is sufficiently expressive to represent the conditional moment maps $X \mapsto m(X)$ and $X \mapsto S(X)^{-1}$, and that the training distribution covers $p(X)$. Then any global minimiser of the population Fishnet loss $\mathbb{E}_{p(X,\theta)}[\ell(\varphi; X, \theta)]$ satisfies, for $p(X)$-almost every X,*

$$\hat\theta(X) = \mathbb{E}_{p(\theta|X)}[\theta], \qquad \mathbf{F}_\varphi(X) = \mathrm{Cov}_{p(\theta|X)}[\theta]^{-1}. \tag{16}$$

Proof. By the tower property, $\mathbb{E}_{p(X,\theta)}[\ell] = \mathbb{E}_{p(X)}\big[\mathbb{E}_{p(\theta|X)}[\ell]\big]$. Applying Proposition 1 pointwise in X and using the identity (13), the conditional minimum is attained at (12); expressivity of the variational class allows this minimum to be realised simultaneously for $p(X)$-almost every X. □

Remark (Posterior covariance versus Fisher information). Corollary 1 identifies $\mathbf{F}_\varphi(X)$ with the inverse *posterior* covariance, not the likelihood Fisher information

$$\mathcal{I}(\theta) := -\mathbb{E}_{p(X|\theta)}\big[\nabla^2_\theta \log p(X \mid \theta)\big].$$

Under standard regularity, the local Laplace approximation gives $p(\theta \mid X) \approx \mathcal{N}\big(\hat\theta_{\mathrm{MAP}}, H(\hat\theta_{\mathrm{MAP}})^{-1}\big)$ with posterior curvature $H(\theta) = \mathcal{J}(\theta; X) - \nabla^2_\theta \log p(\theta)$, where $\mathcal{J}(\theta; X) := -\nabla^2_\theta \log p(X \mid \theta)$ is the observed information. In the Bernstein–von Mises regime of n i.i.d. observations, $H(\theta_0) \to n\,\mathcal{I}(\theta_0)$ in probability and the prior contribution becomes negligible, so $\mathbf{F}_\varphi(X)$ asymptotically recovers $n\,\mathcal{I}(\theta_0)$. Outside this regime $\mathbf{F}_\varphi$ should be read as an inverse posterior covariance, which is the relevant object for the flattening pipeline of Section 4: the metric (2) transforms identically under $\theta \to \eta$ whether one substitutes $\mathcal{I}(\theta)$ or $\mathrm{Cov}_{p(\theta|X)}[\theta]^{-1}$.

B Obtaining a unique solution in flattened coordinates

The flattening loss we minimise (Eq. (5)) has two important symmetries:

1. $L(\boldsymbol{\eta})$ is invariant if $\eta^a \to \eta^a + c^a$, for some constant vector c^a (this is trivial from the invariance of F_{ab} to this).
2. $L(\boldsymbol{\eta})$ is invariant under global orthogonal transformations. To prove this, consider new coordinates $\{\tilde\phi^a\}$ such that $\eta^a = R^a{}_b \tilde\phi^b$ where $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{1}$. The new Fisher is

$$\tilde{\mathbf{F}}(\phi) = \mathbf{R}^{\mathrm{T}} \tilde{\mathbf{F}}(\theta) \mathbf{R}. \tag{17}$$

Using the cyclic property of the trace, one sees that $L(\tilde\phi) = L(\eta)$.

Optimising the Frobenius norm loss does not, therefore, yield a unique set of parameters, but one that is known only up to a constant offset and an orthogonal transformation. Isotropy alone therefore defines an equivalence class of coordinate systems rather than a frame; the canonicalisation below fixes the frame, and the minimum-description-length criterion of Step 3 selects the simplest representative within it. This is particularly problematic when we train an ensemble of networks, since each member of the ensemble will have a different offset and orthogonal transformation relative to some fiducial network.

To deal with this, we employ a two-step rotation scheme to align ensemble members $\{\boldsymbol{\eta}\}$. First, we choose a reference ensemble member and employ a Householder reflection (Householder, 1958) to rotate its learned coordinates at a fiducial value $\eta^* = \eta(\theta^*)$ such that η^* points along the first eigenvalue of the variance-weighted Fisher matrix:

$$\mathbf{F}^*/\delta\mathbf{F}v = \lambda v, \quad R^H\eta^* = \lambda_1. \tag{18}$$

The Householder reflection is given as

$$\text{ref}_{\mathbf{n}}(\mathbf{u}) = \mathbf{u} - 2\frac{\mathbf{n}\mathbf{n}^\top\mathbf{u}}{\mathbf{n}^\top\mathbf{n}} \tag{19}$$

Where $\mathbf{n}$ is the normal to the reflection plane and $\mathbf{n}^\top\mathbf{n}$ is the squared norm $\|\mathbf{n}\|^2$. This can be cast as two successive reflections from $\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\hat{\mathbf{y}} = \frac{\mathbf{y}}{\|\mathbf{y}\|}$, one across the hyperplane defined by the sum of the two vectors, and one across the target vector:

$$S = I - 2\frac{(\hat{\mathbf{y}} + \hat{\mathbf{x}})(\hat{\mathbf{y}} + \hat{\mathbf{x}})^\top}{(\hat{\mathbf{y}} + \hat{\mathbf{x}})^\top(\hat{\mathbf{y}} + \hat{\mathbf{x}})} \tag{20}$$

$$R = S - 2\frac{\hat{\mathbf{y}}\hat{\mathbf{y}}^\top S}{\hat{\mathbf{y}}^\top\hat{\mathbf{y}}}, \tag{21}$$

where R satisfies

$$R\hat{\mathbf{x}} = \hat{\mathbf{y}}. \tag{22}$$

Once this reference point is established, we employ a Kabsch rotation (Kabsch, 1976) to re-align each ensemble member's coordinates. We wish to align each ensemble member's output centroid-subtracted coordinates $P = [\eta_j^i - \langle\eta^i\rangle] \in \mathbb{R}^{N\times n_p}$, where N is the number of samples and n_p is the number of parameters, to the reference ensemble member's centroid-subtracted coordinates $P = [\eta_j^0 - \langle\eta^0\rangle] \in \mathbb{R}^{N\times n_p}$. The Kabsch algorithm finds an optimal rotation R^* which minimises

$$\mathcal{L}_{\text{kabsch}} = \frac{1}{2}\sum_{i=0}^{N} ||Q - RP||^2. \tag{23}$$

The optimal rotation can be found from the Singular Value Decomposition (SVD) of the point cloud covariance matrix, $C = P^TQ$:

$$C = U\Sigma V^T, \quad R^* = VBU^T, \tag{24}$$

where $d = \text{sign}(\det VU^T)$ and $B = \text{diag}(1, 1, d)$ fixes rotations to be right-handed.

Jacobian Procrustes alignment. Coordinate Kabsch matches ensemble members in the $\boldsymbol{\eta}$-space, but the loss (5) cares only about the Jacobian $\boldsymbol{\mathcal{J}} = \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}$ through the metric pullback. As an option to users we also provide another alignment scheme to highlight nonlinear flattened components along directions of high variance in the network Jacobian. We perform an orthogonal Procrustes fit performed directly on the batch of Jacobians: given source Jacobians $\{\mathbf{J}_b^{(i)}\}_{b=1}^B$ for ensemble member i and reference Jacobians $\{\mathbf{J}_b^{(0)}\}_{b=1}^B$, the rotation acts on the $\boldsymbol{\eta}$-axis (left index) of $\mathbf{J}$, $\mathbf{J}'_b = \mathbf{R}\,\mathbf{J}_b^{(i)}$, and the alignment problem

$$\mathbf{R}^* \;=\; \arg\min_{\mathbf{R}\in O(n_p)} \sum_{b=1}^{B} w_b \left\|\mathbf{R}\,\mathbf{J}_b^{(i)} - \mathbf{J}_b^{(0)}\right\|_{\mathcal{F}}^2 \tag{25}$$

admits the closed-form solution $\mathbf{R}^* = \mathbf{U}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ is the SVD of the cross-covariance $\mathbf{M} = \sum_b w_b\,\mathbf{J}_b^{(0)}\big(\mathbf{J}_b^{(i)}\big)^\top$. The sample weights w_b optionally reflect the validation weighting from

Eq. (4). Compared with coordinate Kabsch, this scheme uses every sample and unit-magnitude derivatives instead of noisy coordinates, so it is markedly more stable on near-degenerate axes whose $\boldsymbol{\eta}$-displacements are dominated by ensemble noise. We allow improper rotations ($\det \mathbf{R}^* = -1$) per member when this strictly reduces (25): the flatness loss $\left\|\mathbf{R}\bar{\mathbf{F}}'\mathbf{R}^{\mathrm{T}} - \mathbf{1}\right\|_{\mathcal{F}}^2$ is invariant under arbitrary $\mathbf{R} \in O(n_p)$, so the only requirement is that all members land in a common frame, which the Procrustes solution achieves.

Member-independent canonical basis. After Procrustes alignment the ensemble shares a single orthogonal frame, but the choice of frame is still arbitrary. We fix it once on the reference member by composing two operations. First, a *nonlinearity-separating rotation* takes the SVD of the sample-centred Jacobian stack $\Delta\mathbf{J}_b = \mathbf{J}_b^{(0)} - \bar{\mathbf{J}}^{(0)}$ (with optional column rescaling by the prior widths $\Delta\theta_a = \theta_a^{\max} - \theta_a^{\min}$ to put all parameters on a dimensionless footing): reshaping $\Delta\mathbf{J} \in \mathbb{R}^{B\times n_p\times n_p}$ to $\Omega \in \mathbb{R}^{n_p\times Bn_p}$ and writing $\Omega = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$, the rotation $\mathbf{R}_{\mathrm{nl}} = \mathbf{U}^{\mathrm{T}}$ orders the $\boldsymbol{\eta}$-axes by decreasing nonlinearity "energy" (singular values of Ω): axes whose Jacobian rows are sample-independent (i.e. η_a is a linear function of $\boldsymbol{\theta}$) collect into the trailing rows. Second, a sign and (optionally) permutation fix-up $\mathbf{R}_{\mathrm{canon}}$ uses the eigenstructure of the mean prior-normalised $\boldsymbol{\theta}$-Fisher $\bar{F}_{ab}/(\Delta\theta_a\,\Delta\theta_b)$ to flip per-axis signs so that $\langle(\mathbf{R}_{\mathrm{nl}}\,\mathbf{J})_a, v_a\rangle \geq 0$ for the corresponding Fisher eigenvector v_a. Composing $\mathbf{R}_{\mathrm{basis}} = \mathbf{R}_{\mathrm{canon}}\mathbf{R}_{\mathrm{nl}}$ once on the reference and applying $\mathbf{R}_{\mathrm{total}}^{(i)} = \mathbf{R}_{\mathrm{basis}}\,\mathbf{R}^{*\,(i)}$ to every member guarantees the same axis ordering and orientation across the ensemble, which keeps the dispersion of $\boldsymbol{\eta}^{(i)}$ tight on the near-degenerate (least-nonlinear) axes.

C One-step learning and Fisher flattening

This appendix asks when one-step learning and Fisher flattening select the same coordinates. We compare the flattening loss (Eq. (5)) with a local approximation to the one-step loss (Eq. (6)). We express their relationship in terms of ϵ, the attained minimum of the prior-averaged squared Frobenius distance to the identity. If exact flattening is possible ($\epsilon = 0$), these two objectives have the same optimal coordinate maps. If $\epsilon > 0$, no map minimizes both objectives, provided small increases and decreases in overall scale are allowed. Such residual distortion can arise when nonzero curvature prevents exact flattening. Near exact flattening, the two penalties agree to leading quadratic order, up to a factor of four. Replacing the Frobenius penalty with a matching Kullback–Leibler (KL) penalty makes the local objectives equal up to a constant. These conclusions extend to the exact profiled one-step loss when the profiling identity holds for every allowed coordinate map.

C.1 Exact one-step action and pairwise variation

Let $p(\theta, x) = \pi(\theta)p(x \mid \theta)$ on an open $\Omega \subseteq \mathbb{R}^d$, with a smooth positive normalized prior π. The coordinate map $\eta : \Omega \to \eta(\Omega)$ is a C^2 diffeomorphism with square Jacobian $J = D\eta$; $\hat{\eta}(x)$ is a separate predictor. We write $\mathcal{A}$ for the set of coordinate maps allowed in the optimization. It may contain all such diffeomorphisms, or a smaller family, such as affine maps or the smooth invertible maps represented by a fixed architecture. The term *admissible* means belonging to this set, including any boundary or normalization constraints. Assume all displayed expectations and variations are well defined. Writing S for the loss functional, the one-step action is

$$S_{\mathrm{one}}[\eta, \hat{\eta}] = \tfrac{1}{2}\mathbb{E}_{\theta,x}\|\eta(\theta) - \hat{\eta}(x)\|^2 - \mathbb{E}_\pi \log|\det J|. \tag{26}$$

For each fixed η, minimizing over $\hat{\eta}$ gives the optimal predictor $\hat{\eta}^\star(x) = \mathbb{E}[\eta(\theta) \mid x]$, assuming finite conditional second moments and an attainable conditional mean. The superscript prof denotes *profiling* over the predictor: substituting its optimum leaves an objective depending only on η,

$$\begin{aligned} S_{\mathrm{one}}^{\mathrm{prof}}[\eta] &:= \min_{\hat{\eta}} S_{\mathrm{one}}[\eta, \hat{\eta}] = S_{\mathrm{one}}[\eta, \hat{\eta}^\star] \\ &= \tfrac{1}{2}\mathbb{E}_x \operatorname{tr}\operatorname{Cov}[\eta(\theta) \mid x] - \mathbb{E}_\pi \log|\det J|. \end{aligned} \tag{27}$$

Here the predictor class is assumed to contain this mean for every $\eta \in \mathcal{A}$.

Given x, draw θ, θ' independently from $p(\theta \mid x)$, and write $\mathbb{E}_{\mathrm{pair}} = \mathbb{E}_x\mathbb{E}_{\theta,\theta'|x}$. The conditional variance identity gives the exact symmetric action

$$S_{\mathrm{one}}^{\mathrm{prof}}[\eta] = \tfrac{1}{4}\mathbb{E}_{\mathrm{pair}}\|\eta(\theta) - \eta(\theta')\|^2 - \mathbb{E}_\pi \log|\det D\eta|. \tag{28}$$

The pairing distribution is fixed under variations of η.

Perturb $\eta_t = \eta + tv$, so $\delta J = Dv$: the map and its Jacobian cannot be varied independently. To derive an interior Euler–Lagrange equation, we assume that $\eta + tv$ remains in $\mathcal{A}$ for every smooth, compactly supported v and sufficiently small $|t|$. If $\mathcal{A}$ is smaller, stationarity requires zero derivative only for changes that stay inside $\mathcal{A}$; the unrestricted equation need not hold. For matrices, the colon denotes the Frobenius inner product, $A : B = \sum_{a,i} A_{ai} B_{ai} = \operatorname{tr}(A^\top B)$. In particular, $T : Dv = \sum_{a,i} T_{ai} \partial_{\theta_i} v_a$ is a scalar. Divergence acts row by row: $(\operatorname{Div} T)_a = \sum_i \partial_{\theta_i} T_{ai}$.

Define the posterior-averaging operator

$$\mathcal{M}_\eta(\theta) = \mathbb{E}_{x|\theta} \mathbb{E}_{\theta'|x} \eta(\theta') = \mathbb{E}_{x|\theta} \hat{\eta}^\star(x). \tag{29}$$

With $\Delta f = f(\theta) - f(\theta')$, symmetry of the two conditional draws gives

$$\delta S_{\text{pair}}[v] = \tfrac{1}{2} \mathbb{E}_{\text{pair}}[\Delta\eta \cdot \Delta v] = \mathbb{E}_{\text{pair}}[\Delta\eta \cdot v(\theta)] = \int_\Omega \pi(\eta - \mathcal{M}_\eta) \cdot v \, d\theta, \tag{30}$$

$$\delta S_{\log}[v] = -\int_\Omega \pi J^{-\top} : Dv \, d\theta = \int_\Omega \operatorname{Div}(\pi J^{-\top}) \cdot v \, d\theta. \tag{31}$$

The coefficient $1/4$ in Eq. (28) accounts for variation of both posterior draws. Combining the variations yields

$$\boxed{\pi(\eta - \mathcal{M}_\eta) + \operatorname{Div}(\pi J^{-\top}) = 0.} \tag{32}$$

The operator $\mathcal{M}_\eta$ depends on the map across posterior distributions, so this equation is nonlocal. For a uniform prior in one dimension it reduces to $\eta'' = (\eta - \mathcal{M}_\eta)(\eta')^2$.

C.2 Sequential learning and local coordinate actions

The first stage minimizes the Gaussian negative log-likelihood

$$S_{\text{fish}}[\hat{\theta}, P] = \mathbb{E}_{\theta,x} \left[\tfrac{1}{2} (\theta - \hat{\theta}(x))^\top P(x) (\theta - \hat{\theta}(x)) - \tfrac{1}{2} \log\det P(x) \right], \tag{33}$$

with $P(x) \succ 0$. Let $\mu = \mathbb{E}[\theta \mid x]$ and $\Sigma = \operatorname{Cov}(\theta \mid x) \succ 0$. Its conditional stationarity conditions are $P(\hat{\theta} - \mu) = 0$ and $\Sigma + (\hat{\theta} - \mu)(\hat{\theta} - \mu)^\top - P^{-1} = 0$, giving $\hat{\theta}^\star = \mu$ and $P^\star = \Sigma^{-1}$. For coordinate training, assume a smooth positive-definite parameter-space field $F(\theta)$ has been supplied and frozen, with $\mathbb{E}_\pi |\log\det F| < \infty$. Representing the data-dependent precision as $F(\mu(x)) = \Sigma(x)^{-1}$, together with any extension to all of Ω, is an additional modeling assumption. The Gaussian loss alone does not identify likelihood Fisher information.

With $G = J^{-\top} F J^{-1}$, coordinate training minimizes $S_{\text{flat}}[\eta; F] = \mathbb{E}_\pi \|G - I\|_{\text{F}}^2$. For comparison, introduce the local one-step action

$$S^{\text{loc}}_{\text{one}}[\eta; F] = \mathbb{E}_\pi \left[\tfrac{1}{2} \operatorname{tr}(J F^{-1} J^\top) - \log|\det J| \right]. \tag{34}$$

The local action replaces the true transformed posterior covariance by the Jacobian-based approximation $J F^{-1} J^\top$. Both coordinate actions use the same prior and the same allowed maps $\mathcal{A}$.

When the local and exact losses agree. Results for $S^{\text{loc}}_{\text{one}}$ also apply to the exact profiled loss if

$$S^{\text{prof}}_{\text{one}}[\eta] = S^{\text{loc}}_{\text{one}}[\eta; F] + C, \qquad \text{for every } \eta \in \mathcal{A}, \tag{35}$$

where C is independent of η. We call this the *profiling identity*. It is an additional assumption about $\mathcal{A}$ and the supplied field F; choosing a restricted $\mathcal{A}$ does not establish it. The log-determinant terms of Eqs. (27) and (34) coincide, so Eq. (35) is a statement about covariances alone:

$$\mathbb{E}_x \operatorname{tr} \operatorname{Cov}[\eta(\theta) \mid x] = \mathbb{E}_\pi \operatorname{tr}\big(J F^{-1} J^\top\big) + 2C. \tag{36}$$

Thus the identity requires the two covariance terms to agree on average, up to the same constant for every allowed map. A first-order Taylor expansion of η about the posterior mean gives $\operatorname{Cov}[\eta(\theta) \mid x] \approx J(\mu)\Sigma(x)J(\mu)^\top = J(\mu)F(\mu)^{-1}J(\mu)^\top$ when $F(\mu) = \Sigma(x)^{-1}$. Replacing the average over posterior means $\mu(x)$ by a prior average is a further approximation. Neither step follows from Gaussian training, and their errors can depend on η. Approximate agreement of loss values alone does not establish agreement of their derivatives or minimizers.

An exact example: affine maps. Let $\overline{\Sigma} = \mathbb{E}_x \operatorname{Cov}[\theta \mid x]$ be finite and positive definite, choose the constant field $F = \overline{\Sigma}^{-1}$, and take $\mathcal{A} = \{\eta(\theta) = M\theta + b : \det M \neq 0\}$. Since $J = M$, covariance propagation is exact:

$$\mathbb{E}_x \operatorname{Cov}[M\theta + b \mid x] = M\overline{\Sigma}M^\top = \mathbb{E}_\pi[JF^{-1}J^\top].$$

Hence Eq. (35) holds with $C = 0$ throughout this class. For nonlinear maps, covariance depends on the map across each posterior region, while J describes only local changes. The identity therefore requires a separate justification for a learned field and nonlinear map family; no such justification for a general neural class is established here.

Why the class must be restricted for exact agreement. Equation (35) cannot hold on a neighborhood allowing all smooth compactly supported perturbations. Indeed, comparing the coefficients of t^2 in Eq. (36) for $\eta + tv$ would require

$$\mathbb{E}_x \operatorname{tr} \operatorname{Cov}[v(\theta) \mid x] = \mathbb{E}_\pi \operatorname{tr}(DvF^{-1}Dv^\top) \quad \text{for every such } v. \tag{37}$$

The left side is at most $\mathbb{E}_\pi\|v\|^2$. For $v_n(\theta) = \chi(\theta)\sin(n\theta_1)e_1$, with a nonzero smooth bump χ supported inside Ω, this bound stays finite uniformly in n, whereas the right side grows as n^2: π is bounded below and F^{-1} is uniformly positive definite on the compact support. Therefore the identity cannot hold for all smooth invertible maps, or even throughout a neighborhood allowing arbitrary smooth local perturbations. On a smaller class where it does hold, it equates derivatives only along changes that remain in that class.

Moreover, with $\Delta[\eta] = S^{\mathrm{prof}}_{\mathrm{one}}[\eta] - S^{\mathrm{loc}}_{\mathrm{one}}[\eta; F]$, the covariance terms give $\Delta[a\eta] = a^2\Delta[\eta]$. Consequently, if Eq. (35) holds at both η and an admissible $a\eta$ with $a > 0$, $a \neq 1$, its constant must satisfy $C = a^2C$, hence $C = 0$ for the losses defined here.

Local coordinate equations. The local one-step variation is

$$\delta S^{\mathrm{loc}}_{\mathrm{one}}[v] = \int_\Omega \pi T_{\mathrm{one}} : Dv \, d\theta, \qquad T_{\mathrm{one}} = JF^{-1} - J^{-\top} = (G^{-1} - I)J^{-\top}. \tag{38}$$

For flattening, the induced metric variation and the resulting first variation are

$$\delta G = -J^{-\top}(Dv)^\top G - G(Dv)J^{-1}, \tag{39}$$

$$\delta S_{\mathrm{flat}}[v] = 2\int_\Omega \pi(G - I) : \delta G \, d\theta = \int_\Omega \pi T_{\mathrm{flat}} : Dv \, d\theta, \qquad T_{\mathrm{flat}} = 4G(I - G)J^{-\top}. \tag{40}$$

Integration by parts therefore gives the respective coordinate equations

$$\boxed{\operatorname{Div}(\pi T_{\mathrm{one}}) = 0, \qquad \operatorname{Div}(\pi T_{\mathrm{flat}}) = 0.} \tag{41}$$

These equations apply when arbitrary smooth local perturbations are allowed. The profiling identity on a smaller $\mathcal{A}$ does not make the first equation equivalent to the exact Eq. (32). Since F is frozen after stage one, $\delta_\eta S_{\mathrm{fish}} = 0$, so the second is also the coordinate equation for $S_{\mathrm{fish}} + S_{\mathrm{flat}}$ in the sequential construction. A joint variation of F through both losses would be a different procedure.

C.3 Residual distortion and failure of equivalence

Although $T_{\mathrm{flat}} = 4G^2T_{\mathrm{one}}$, the factor G^2 generally varies with θ. Differentiating it introduces an extra term, so the two divergences need not vanish together. Writing $R_{\mathrm{one}} = \operatorname{Div}(\pi T_{\mathrm{one}})$ and $R_{\mathrm{flat}} = \operatorname{Div}(\pi T_{\mathrm{flat}})$, with e_i the ith standard basis vector, the product rule gives

$$\boxed{R_{\mathrm{flat}} = 4G^2R_{\mathrm{one}} + 4\pi\sum_i (\partial_{\theta_i}G^2)T_{\mathrm{one}}e_i.} \tag{42}$$

The extra term generally survives when residual distortion varies across the domain. At an exact flattening, $G = I$ and both stress tensors vanish.

Proposition 2 (Positive-loss obstruction)**.** *Suppose η_f minimizes S_{flat}, with $S_{\mathrm{flat}}[\eta_f; F] = \epsilon > 0$, the relevant expectations below are finite, and $e^t\eta_f$ is admissible for both signs of sufficiently small t. Then η_f is not stationary for $S^{\mathrm{loc}}_{\mathrm{one}}$. Under Eq.* (35)*, it is not stationary for $S^{\mathrm{prof}}_{\mathrm{one}}$ either.*

Proof. At η_f, set $A = \mathbb{E}_\pi \operatorname{tr} G^2$, $B = \mathbb{E}_\pi \operatorname{tr} G$, and $D = \mathbb{E}_\pi \operatorname{tr} G^{-1}$. Rescaling $\eta_t = e^t \eta_f$ gives $G_t = e^{-2t} G$, so

$$0 = \frac{d}{dt} S_{\text{flat}}[\eta_t; F]\Big|_0 = 4(B - A), \qquad \epsilon = A - 2B + d = d - B. \tag{43}$$

Cauchy–Schwarz, applied jointly to θ and eigenvalue index, gives $BD \geq d^2$. Since $F \succ 0$ and J is invertible, $G \succ 0$ pointwise, so $B = \mathbb{E}_\pi \operatorname{tr} G > 0$; hence $\epsilon = d - B < d$, and

$$\boxed{\frac{d}{dt} S^{\text{loc}}_{\text{one}}[\eta_t; F]\Big|_0 = D - d \geq \frac{d^2}{B} - d = \frac{d\epsilon}{d - \epsilon} > 0.} \tag{44}$$

A sufficiently small negative t therefore lowers the local one-step action, and also the profiled action if Eq. (35) holds along this path. □

Fixed-scale or boundary constraints can exclude this rescaling without making Eq. (41) equivalent. Nonzero Riemann curvature precludes exact flattening on a region; the proposition additionally assumes an attained positive minimum.

Best scales along rescaling. The proof uses only scale stationarity, $A = B$, not global optimality of η_f. Suppose $a\eta \in \mathcal{A}$ for every $a > 0$, with A, B, D evaluated at η. Since $G \mapsto a^{-2} G$, the scale-dependent losses are $S_{\text{flat}}[a\eta; F] = a^{-4} A - 2a^{-2} B + d$ and $S^{\text{loc}}_{\text{one}}[a\eta; F] = \frac{1}{2} a^2 D - d \log a + \text{const}$, with best scales

$$a^2_{\text{flat}} = A/B, \qquad a^2_{\text{one}} = d/D. \tag{45}$$

A map stationary for both losses along this family has $A = B$ and $D = d$; Eq. (44) then forces $\epsilon = S_{\text{flat}}[\eta; F] = 0$, i.e. $G = I$ almost everywhere. Thus, in a class that allows every positive rescaling of each map, a common minimizer of S_{flat} and $S^{\text{loc}}_{\text{one}}$ requires exact flattening. When it is unavailable and both minima are attained with the stated finiteness assumptions, no map minimizes both losses. This conclusion does not exclude optimizers related by a constant rescaling; for a class consisting of one scaling family, such a relation is unavoidable. Whether the optimal maps agree up to rescaling requires a separate argument. The same claims for $S^{\text{prof}}_{\text{one}}$ require Eq. (35) on the comparison class.

The realizable limit. The bound in Eq. (44) is a quantitative obstruction to shared optima that vanishes in the realizable limit:

$$\frac{d\epsilon}{d - \epsilon} = \frac{\epsilon}{1 - \epsilon/d} \;\to\; 0 \quad \text{as } \epsilon \to 0, \tag{46}$$

This lower bound measures a failure of stationarity, not a distance between optimal maps. Its vanishing alone does not establish equality of minimizers. Theorem 2 in Sec. C.5 proves agreement for the local action directly, and conditionally for the profiled loss. Agreement at the attained endpoint $\epsilon = 0$ does not itself imply convergence of minimizing maps as $\epsilon \to 0$.

Normalized obstruction. Dividing by d gives a per-dimension measure. With $r = \epsilon/d$, define

$$\gamma := \frac{1}{d} \frac{d}{dt} S^{\text{loc}}_{\text{one}}[e^t \eta_f; F]\Big|_0 = \frac{D}{d} - 1 \;\geq\; \frac{r}{1 - r}. \tag{47}$$

Here γ measures the response of the local one-step loss to an overall rescaling at the Frobenius optimum. The lower bound does not ensure similar loss values or coordinate maps: eigenvalues close to zero on small prior regions can make D large despite small ϵ. Here $r = \mathbb{E}_{\theta,j}(\lambda_j(G) - 1)^2$, with j uniform over eigenvalue index, measures average squared distortion per dimension, not intrinsic curvature. Using the bound requires scale stationarity, $A = B$.

C.4 A MATCHING DISTORTION ACTION

The local one-step action admits an exact representation as a distortion penalty on G. Define

$$\Phi(G) = \tfrac{1}{2}[\operatorname{tr} G^{-1} + \log\det G - d], \qquad K_F = \tfrac{d}{2} - \tfrac{1}{2}\mathbb{E}_\pi \log\det F. \tag{48}$$

Since $G^{-1} = JF^{-1}J^\top$ and $\log\det G = \log\det F - 2\log|\det J|$,

$$S_{\text{one}}^{\text{loc}} = K_F + \mathbb{E}_\pi \Phi(G). \tag{49}$$

For every eigenvalue $\lambda > 0$,

$$\lambda^{-1} + \log\lambda - 1 \geq 0, \qquad \text{with equality only at } \lambda = 1. \tag{50}$$

Thus $\|G - I\|_{\text{F}}^2$ and $\Phi(G)$ are nonnegative and vanish precisely at $G = I$; the uncentered one-step action need not be nonnegative.

The corresponding KL distortion is

$$S_{\text{flat}}^{\text{KL}}[\eta; F] := \mathbb{E}_\pi \Phi(G) = \mathbb{E}_\pi D_{\text{KL}}\big(\mathcal{N}(0, G^{-1}) \,\big\|\, \mathcal{N}(0, I)\big) \tag{51}$$

The identity $S_{\text{one}}^{\text{loc}} = K_F + S_{\text{flat}}^{\text{KL}}$ is exact. If the profiling identity additionally holds on $\mathcal{A}$,

$$\boxed{S_{\text{one}}^{\text{prof}} = C + K_F + S_{\text{flat}}^{\text{KL}}.} \tag{52}$$

Thus replacing the Frobenius penalty by this KL penalty gives exactly the local one-step objective, up to a constant. The two have identical variations and global coordinate minimizers, whether or not exact flattening exists. Under Eq. (35), the profiled action shares their minimizers on $\mathcal{A}$ and their variations along admissible paths in $\mathcal{A}$.

C.5 Shared global minimizers in the realizable case

When exact flattening is available ($\epsilon = 0$), the Frobenius action itself selects the same coordinates as the local one-step action. Let

$$\mathcal{E}_{\text{fl}} = \{\eta \in \mathcal{A} : \; G = I \text{ everywhere on } \Omega\}. \tag{53}$$

Theorem 2 (Realizable equivalence)**.** *Let F be the frozen stage-one field, continuous and positive definite on Ω with $\mathbb{E}_\pi|\log\det F| < \infty$, and suppose $\mathcal{E}_{\text{fl}} \neq \emptyset$. Then*

$$\boxed{\arg\min_{\eta\in\mathcal{A}} S_{\text{flat}}[\eta; F] = \arg\min_{\eta\in\mathcal{A}} S_{\text{one}}^{\text{loc}}[\eta; F] = \mathcal{E}_{\text{fl}}.} \tag{54}$$

If, additionally, the predictor class attains $\hat\eta^\star$ for every $\eta \in \mathcal{A}$ and Eq. (35) *holds throughout $\mathcal{A}$, then $\arg\min_{\eta\in\mathcal{A}} S_{\text{one}}^{\text{prof}}[\eta] = \mathcal{E}_{\text{fl}}$ as well.*

Proof. By Eq. (49), $S_{\text{one}}^{\text{loc}}[\eta; F] = K_F + \mathbb{E}_\pi\Phi(G)$, with K_F finite. By Eq. (50), $S_{\text{one}}^{\text{loc}} \geq K_F$ and $S_{\text{flat}} \geq 0$, with either bound attained exactly when $G = I$ π-almost everywhere. Since $\mathcal{E}_{\text{fl}} \neq \emptyset$, both bounds are attained on $\mathcal{A}$, so each set of minimizers consists of the admissible maps with $G = I$ almost everywhere. Because J is C^1 and F is continuous, G is continuous; since $\pi > 0$ on Ω, almost-everywhere equality upgrades to equality everywhere, and both sets equal $\mathcal{E}_{\text{fl}}$. Under the additional profiling identity, adding the constant C does not change the minimizers on $\mathcal{A}$, proving the final assertion. □

Under the additional hypotheses for the profiled loss, for each optimal coordinate map the optimal one-step predictor is $\hat\eta^\star(x) = \mathbb{E}[\eta(\theta) \mid x]$. Sequential coordinate training does not itself train this predictor; fitting it afterwards recovers the same optimal pairs $(\eta, \hat\eta^\star)$.

Optimality within a class. At every $\eta \in \mathcal{E}_{\text{fl}}$, $G = I$ gives $T_{\text{one}} = T_{\text{flat}} = 0$, so both equations in Eq. (41) hold. Equality of the profiled and local actions on a restricted $\mathcal{A}$ implies only that allowed changes have zero derivative; it does not imply the unrestricted Eq. (32). Away from $\mathcal{E}_{\text{fl}}$, stationarity for the two local objectives is not generally equivalent, as Eq. (42) and Proposition 2 show.

For example, take $d = 1$, $\theta \sim \text{Uniform}(-1, 1)$, x independent of θ, $F = 3$, and $\mathcal{A} = \{a\theta + b : a \neq 0\}$. The profiling identity holds with $C = 0$, and $\eta = \sqrt{3}\,\theta$ minimizes both losses on $\mathcal{A}$ with $G = 1$. Nevertheless, $\eta - \mathcal{M}_\eta = \sqrt{3}\,\theta$ and $\text{Div}(\pi J^{-\top}) = 0$, so this map does not satisfy the unrestricted exact Euler–Lagrange equation: optimality among affine maps does not imply optimality among nonlinear maps.

Uniqueness up to rigid motions. The relation $G = I$ is equivalent to $J^\top J = F$. If Ω is connected, any two such maps differ by a constant Euclidean rigid motion: the transition map has orthogonal Jacobian, and differentiating this identity and using equality of mixed partials forces its Jacobian to be constant. Thus translation, rotation, and reflection remain unidentifiable when allowed by $\mathcal{A}$. In one dimension, the solutions reduce to $\eta(\theta) = \pm \int^\theta \sqrt{F(t)}\, dt + c$. For sufficiently smooth F, vanishing Riemann curvature guarantees local flattening but does not by itself provide a global admissible diffeomorphism. Thus $\mathcal{E}_{\mathrm{fl}} \neq \emptyset$ is a global realizability assumption.

C.6 Quantitative similarity near the identity

Remove the constant from the local action and write $Q[\eta] = \mathbb{E}_\pi \Phi(G) = S_{\mathrm{one}}^{\mathrm{loc}} - K_F$. Under Eq. (35), this also gives the profiled loss after subtracting its constant, $Q = S_{\mathrm{one}}^{\mathrm{prof}} - C - K_F$. Suppose every eigenvalue stays uniformly close to one: almost everywhere under the prior,

$$1 - \delta \leq \lambda_j(G(\theta)) \leq 1 + \delta \quad \text{for all } j, \qquad 0 \leq \delta < 1, \tag{55}$$

Then

$$\boxed{\frac{S_{\mathrm{flat}}[\eta; F]}{(1+\delta)^2} \;\leq\; 4Q[\eta] \;\leq\; \frac{S_{\mathrm{flat}}[\eta; F]}{(1-\delta)^2}.} \tag{56}$$

To see this, define $\phi(\lambda) = \frac{1}{2}(\lambda^{-1} + \log\lambda - 1)$. For $h = \lambda - 1 > -1$, integrating $\phi'(\lambda) = (\lambda - 1)/(2\lambda^2)$ gives

$$4\phi(1+h) = 2h^2 \int_0^1 \frac{t}{(1+th)^2}\, dt.$$

Bounding the denominator, summing over eigenvalues, and averaging gives Eq. (56), hence $4Q = S_{\mathrm{flat}}[1 + O(\delta)]$. For $G = I + H$, the expansion $\Phi(I + H) = \frac{1}{4}\|H\|_{\mathrm{F}}^2 + O(\|H\|_{\mathrm{F}}^3)$ shows that the penalties agree to leading quadratic order.

These bounds also compare the distortion achieved by the two optimizers. If η_f minimizes S_{flat} and η_o minimizes Q over the same class, and both satisfy Eq. (55) with the same δ, optimality and Eq. (56) imply

$$\epsilon \leq S_{\mathrm{flat}}[\eta_o; F] \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \epsilon, \qquad \epsilon = S_{\mathrm{flat}}[\eta_f; F]. \tag{57}$$

This bounds the Frobenius distortion of the one-step surrogate's optimizer. Bounding the distance between the maps themselves requires an additional stability argument and a fixed choice of translation, rotation, and reflection. At $\epsilon = 0$ both sides vanish, consistent with Theorem 2.

The eigenvalue bounds also make $r = \epsilon/d$ informative about the scale derivative. At a map whose Frobenius loss is stationary under rescaling, $B = d - \epsilon$ gives the exact identity

$$D - d = \epsilon + \mathbb{E}_\pi \operatorname{tr}\left[(G - I)^2 G^{-1}\right]. \tag{58}$$

Consequently, under Eq. (55),

$$\left(1 + \frac{1}{1+\delta}\right) r \leq \gamma \leq \left(1 + \frac{1}{1-\delta}\right) r, \qquad \gamma = 2r + O(\delta r). \tag{59}$$

Without a positive lower bound on the eigenvalues, small r gives no such upper bound on γ. The quantities r, the observed spectral range of G, and γ provide empirical diagnostics; finite validation samples alone do not certify a uniform spectral bound.

Scope of the population results. Finite data, restricted model capacity, and incomplete optimization introduce further estimation, approximation, and optimization errors. These are separate from curvature and from failure of Eq. (35), which more data alone does not remove. These population statements do not imply convergence of neural-network training or equal practical performance.

C.7 Relation to likelihood scores

The conditional profiled-loss conclusion of Theorem 2 does not require Gaussian posteriors. Up to $\frac{d}{2}\log(2\pi)$, and when $\eta(\Omega)$ has full Gaussian measure, S_{one} is the expected negative log density of

the model $q(\theta \mid x) = \mathcal{N}\big(\eta(\theta); \hat{\eta}(x), I\big)\,|\det J(\theta)|$. If, additionally, the *true* transformed posterior is $p(u \mid x) = \mathcal{N}(\hat{\eta}^\star(x), I)$ for $u = \eta(\theta)$, define the residual $\xi = \hat{\eta}^\star(x) - \eta(\theta)$. At points where the densities are positive and differentiable, change of variables and Bayes' theorem give

$$\begin{aligned} &\nabla_u \log p(u \mid x) = \xi, \\ &\nabla_\theta \log p(x \mid \theta) = J^\top \xi + \nabla_\theta \log|\det J| - \nabla_\theta \log \pi(\theta), \\ &\nabla_u \log p(x \mid u) = \xi - \nabla_u \log p(u), \qquad p(u) = \pi(\theta)/|\det J(\theta)|. \end{aligned} \tag{60}$$

Writing $s_u = \nabla_u \log p(u)$, a small s_u makes ξ approximately the likelihood score in learned coordinates. In parameter coordinates,

$$\begin{aligned} \nabla_\theta \log p(x \mid \theta) - J^\top \xi &= -J^\top s_u, \\ \|\nabla_\theta \log p(x \mid \theta) - J^\top \xi\| &\le \|J\|_{\mathrm{op}} \|s_u\|. \end{aligned} \tag{61}$$

Thus a small absolute error there requires $J^\top s_u$ to be small; a large Jacobian can amplify a small transformed-prior score. The prior and Jacobian corrections may cancel even when neither is individually small. This Gaussian assumption concerns the true posterior and does not follow merely from choosing the model density q.

D Architecture details

D.1 Fisher Information Neural Networks

We train an ensemble of neural networks (Fishnets) to predict Fisher information matrices $F(\theta)$ directly from simulation data $\mathbf{x}$. Each network comprises a residual MLP followed by a Fisher prediction layer. The Fisher matrices are parameterised via log-Cholesky decomposition to ensure positive-definiteness:

$$F = LL^\top, \quad L_{ii} = \mathrm{softplus}(\ell_i) + 10^{-4}, \tag{62}$$

where ℓ_i are learned diagonal parameters and off-diagonal elements are unconstrained.

D.1.1 SoftSquelch activation

The Fisher prediction branch employs a learned sparse activation function:

$$\mathrm{SoftSquelch}(x) = x \cdot \sigma\big(s(|x| - t)\big), \tag{63}$$

where σ is the sigmoid function, and threshold t and sharpness s are hyperparameters. This suppresses small activations whilst preserving large ones, promoting sparsity.

D.1.2 Ensemble hyperparameter priors

Ensemble diversity is achieved through randomised architectural choices, which contribute to pushing the sampling distribution of learned Fisher matrix entries and downstream flattened coordinates towards Gaussian. Hidden layer widths are sampled uniformly from a user-specified range (default $[10, 300)$ neurons), and activation functions are drawn from a discrete distribution:

$$p(\text{activation}) = \begin{cases} 3/23 & \text{ReLU} \\ 4/23 & \text{Leaky ReLU} \\ 3/23 & \text{Swish} \\ 2/23 & \text{Mish} \\ 3/23 & \text{smooth_leaky} \\ 8/23 & \text{GELU} \end{cases} \tag{64}$$

where Mish is defined as $\mathrm{Mish}(x) = x \tanh(\mathrm{softplus}(x))$, and smooth_leaky is defined as:

$$\mathrm{smooth_leaky}(x) = \alpha x + (1 - \alpha) x \log \sigma(x), \quad \alpha = 0.2, \tag{65}$$

with $\log \sigma(x) = -\log(1 + e^{-x})$ ensuring no discontinuities in derivatives. The SoftSquelch hyperparameters are sampled from independent Gaussians:

$$s \sim \mathcal{N}(5.0, 0.7^2), \tag{66}$$

$$t \sim \mathcal{N}(1.0, 0.7^2), \tag{67}$$

where s controls the sharpness of the transition and t sets the suppression threshold.

D.1.3 TRAINING OBJECTIVE

Networks minimise the Kullback–Leibler divergence between predicted and true parameter distributions:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2}\mathbb{E}\Big[\mathrm{tr}\big(F(\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top\big) - \log\det F\Big], \tag{68}$$

where $\hat{\theta}$ denotes the predicted parameter estimate.

D.2 GEOMETRIC FLATTENING

We learn a coordinate transformation $\eta = f(\theta; w)$ that renders the Fisher metric approximately Euclidean. The network architecture is a residual MLP with inverse whitening:

$$\eta = W^{-1} \cdot \mathrm{MLP}(\theta), \tag{69}$$

where $W^{-1} = F_{\mathrm{mean}}^{1/2}$ is a fixed inverse whitening matrix computed from the global mean Fisher.

D.2.1 FLATTENING LOSS

The transformation is learned by minimising deviations of the pullback metric from the identity:

$$\mathcal{L}_{\mathrm{flat}} = \Big\| J^{-\top} F J^{-1} - I \Big\|_F + \Big\| (J^{-\top} F J^{-1})^{-1} - I \Big\|_F, \tag{70}$$

where $J = \partial\eta/\partial\theta$ is the Jacobian. To speed up progress on hard points, each ℓ is rescaled by an adaptive factor

$$r(\mathcal{L}) \;=\; \frac{\lambda\,\mathcal{L}}{\ell + \mathrm{e}^{-\alpha\mathcal{L}}}, \qquad \alpha = \alpha(\lambda, \varepsilon),$$

and the training objective is the batch mean of $\log\big(r(\mathcal{L})\,\mathcal{L}\big)$. Larger λ makes the map $r(\cdot)$ more aggressive about emphasizing large-residual samples relative to easy ones.

D.2.2 WHITENING

The inverse whitening layer implicitly transforms the training objective. With $\eta = W^{-1}\eta_{\mathrm{raw}}$ and $J = W^{-1}J_{\mathrm{raw}}$, the loss becomes:

$$\mathcal{L}_{\mathrm{flat}} = \Big\| J_{\mathrm{raw}}^{-\top}(WFW)J_{\mathrm{raw}}^{-1} - I \Big\|_F, \tag{71}$$

effectively training on pre-whitened Fishers without modifying the data.

D.3 SYMBOLIC COORDINATE DISCOVERY

We apply symbolic regression (OPERON, (Burlacu et al., 2020)) to the flattened coordinates η to discover interpretable expressions. Each component $\eta_i(\theta)$ is fitted independently using genetic programming with objectives balancing accuracy (R^2) and complexity (equation length). The Pareto frontier yields candidate expressions.

SR input augmentation. The trained flattening ensemble defines a deterministic map $\bar{\eta}(\theta)$ with attached uncertainty $\delta\eta(\theta)$ that is independent of the simulations originally used to fit the Fisher–flattener stack. Rather than feeding only the (small) Fishnet validation set into SR, we draw a fresh dense sample from the prior support and evaluate $(\bar{\eta}, \delta\eta)$ on those points; the resulting triples are used as the Gaussian-loss SR targets. We find this materially improves SR convergence in every experiment we have tried (cleaner Pareto fronts, more reliable recovery of the analytic coordinates on Rosenbrock) and is essentially free since it requires no additional simulator calls; see Appendix F for the per-experiment sample counts.

D.3.1 MODEL SELECTION

We rank expressions by multiple criteria:

- **Minimum description length (MDL):** Bayesian model comparison penalising complexity (Bartlett et al., 2023; 2024).
- **Frobenius loss:** Flattening quality $\|J_{\text{symb}}^{-\top} F J_{\text{symb}}^{-1} - I\|_F$ on held-out data.
- R^2 **score:** Prediction accuracy on test set.

The symbolic Jacobian J_{symb} is computed via automatic differentiation with Sympy (Meurer et al., 2017). Final coordinates are selected to minimise either the MDL or the Frobenius loss whilst maintaining interpretability.

D.4 Postprocessing Symbolic Expressions

Let $\ell_{\text{flat}}(\theta) = \frac{1}{B}\sum_{b=1}^{B} \big\|\widetilde{F}_b(\theta) - I\big\|_F$ be the mean (over samples b) Frobenius distance of the flattened Fisher $\widetilde{F}_b$ from the identity, for linear coefficients θ in the symbolic coordinates. Fix $\ell_{\text{ref}} = \ell_{\text{flat}}(\theta_0)$ at the current parameterization.

Importance (importance sampling / local sensitivity): for each scalar coefficient c with $c \neq 0$, use a one-sided finite difference step ε and set

$$w_c = |c| \left| \frac{\ell_{\text{flat}}(\theta_0 + \varepsilon e_c) - \ell_{\text{ref}}}{\varepsilon} \right|,$$

where e_c perturbs only that coefficient.

Pruning: sort coefficients by increasing w_c. Greedily try setting coefficients to zero (optionally in batches); accept a proposed sparsification θ' only if the *relative* flattening increase stays below a tolerance τ,

$$\frac{\ell_{\text{flat}}(\theta') - \ell_{\text{ref}}}{\ell_{\text{ref}}} < \tau. \tag{72}$$

D.5 Invertible network mapping: Real NVP

We optionally parameterize the reparameterization $\eta(\theta)$ with a Real NVP normalizing flow Dinh et al. (2017), implemented in JAX/Flax as a stack of affine coupling layers Dinh et al. (2017); Papamakarios et al. (2021). Each layer fixes half the coordinates (alternating even/odd dimensions across layers) and applies an elementwise affine map to the remaining coordinates,

$$y_i = x_i \odot \exp\big(s(x_{\text{fix}})\big) + t(x_{\text{fix}}),$$

where s and t are MLPs of one hidden layer with width `hidden_dims`; the s-head uses a final $\tanh$ for bounded logits, and scales use $\exp$ by default so Jacobians stay tractable. The forward pass accumulates $\log|\det J|$ over layers; the inverse applies the couplings in reverse order. For training, θ is min–max scaled to $[0, 1]$, shifted by one (so inputs lie in $[1, 2]$), passed through the flow, and optionally left-multiplied by a fixed inverse-whitening matrix $\mathbf{W}^{-1}$ on η_{raw} (`WhitenedRealNVP`); the inverse reverses these steps in order.

D.6 Forward–backward MLP

As a lightweight alternative to the RealNVP flow, we pair the forward MLP $\eta = f_\phi(\theta)$ with a separate *reverse-path* MLP $\hat{\theta} = g_\psi(\eta)$ that approximates the inverse, and add a mean-square cycle-consistency penalty $\|\theta - g_\psi(f_\phi(\theta))\|^2$ to the flattening loss with weight λ_{inv}, so that the total objective is $\mathcal{L}_{\text{flat}} + \lambda_{\text{inv}}\, \mathbb{E}_\theta \|\theta - g_\psi(f_\phi(\theta))\|^2$; this softly enforces invertibility on the prior support without constraining the architecture as a normalising flow.

E Geodesic normal coordinates for Gaussian example

E.1 Motivation

The aim of this paper is to construct coordinates which make the Fisher information as flat as possible. In Section 5.1, we considered the example of a Gaussian, parametrised by mean μ and standard

deviation σ. We will write these as coordinates $y^a = (\mu, \sigma)$. The Fisher information will be our metric, which is $F_{ab} = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$, where we have ignored the multiplicative factor of n_d for simplicity.

In this appendix, we wish to find an analytic coordinate transformation to compare against the one obtained by our method. As we will show, there does not exist a coordinate transformation which can make the Fisher equal to the identity everywhere, and thus we construct geodesic normal coordinates to find the coordinates which make the metric appear as flat as possible in the vicinity of some point.

Suppose we were at some special point p with coordinates $y^a_\star = (\mu_\star, \sigma_\star)$ and wanted to make the metric flat in the vicinity of that point. For simplicity, let us begin by rescaling μ and σ by functions of $\sigma_\star$ to make the metric equal to the identity at p. We define

$$u \equiv \frac{\mu}{\sigma_\star}, \quad v \equiv \sqrt{2}\frac{\sigma}{\sigma_\star}, \tag{73}$$

so that the corresponding line element for a metric $g_{ab} = F_{ab}$ becomes

$$\mathrm{d}s^2 = \left(\frac{\sigma_\star}{\sigma}\right)^2 \left(\mathrm{d}u^2 + \mathrm{d}v^2\right) = \frac{2}{v^2}\left(\mathrm{d}u^2 + \mathrm{d}v^2\right), \tag{74}$$

i.e. the metric in these coordinates is

$$g_{ab} = \frac{2}{v^2}\delta_{ab}. \tag{75}$$

We will use $x^a = (u, v)$. The coordinates of the point p are then

$$u_\star = \frac{\mu_\star}{\sigma_\star}, \quad v_\star = \sqrt{2}. \tag{76}$$

Although the metric is Euclidean at p in these coordinates, we would see that as we moved away from p, the leading order terms for the metric would be linear in our coordinates. A better set of coordinates to use is geodesic normal coordinates, since these give quadratic corrections to the metric at leading order:

$$\tilde{g}_{ab} = \delta_{ab} - \frac{1}{3}R_{acbd}(x_\star)\left(x^c - x^c_\star\right)\left(x^d - x^d_\star\right) + \mathcal{O}\left(\left|x - x_\star\right|^3\right), \tag{77}$$

where R_{abcd} is the Riemann tensor. The aim of this appendix is to construct these coordinates for a Gaussian. Of course, the flatness is only valid for some region of space. This does, however, represent a good benchmark to see whether our ML method can provide something better over all space than this construction.

E.2 Series expansion for geodesic normal coordinates

Before explicitly constructing geodesic normal coordinates, let us first obtain a series expansion for the metric in these coordinates, so that we have something to verify our result against.

Let us first construct the Christoffel symbols for this metric, given by

$$\Gamma^a_{bc} = \frac{1}{2}g^{ad}\left(\partial_c g_{db} + \partial_b g_{dc} - \partial_d g_{bc}\right). \tag{78}$$

For our metric, these are

$$\Gamma^0_{ab} = \begin{pmatrix} 0 & -\frac{1}{v} \\ -\frac{1}{v} & 0 \end{pmatrix}, \quad \Gamma^1_{ab} = \begin{pmatrix} \frac{1}{v} & 0 \\ 0 & -\frac{1}{v} \end{pmatrix}. \tag{79}$$

Now we must compute the Riemann tensor

$$R^d_{cab} = \partial_a \Gamma^d_{bc} - \partial_b \Gamma^d_{ac} + \Gamma^d_{ae}\Gamma^e_{bc} - \Gamma^d_{be}\Gamma^e_{ac}. \tag{80}$$

Since we are in two dimensions, this has only one independent component, and is thus fully determined by the Ricci scalar, R. For our metric, we find that

$$R = -1, \tag{81}$$

which yields the non-zero components of the Riemann tensor

$$R_{0110} = R_{1001} = -R_{1010} = -R_{0101} = \frac{1}{v^2}. \tag{82}$$

Since the Riemann tensor is non-zero, we cannot find global coordinates which flatten the metric. Instead, we are now in a position to write down the metric expansion in geodesic normal coordinates using Eq. (77).

$$\tilde{g}_{ab} \approx \mathbf{1} + \frac{1}{3v_\star^2} \begin{pmatrix} (v-v_\star)^2 & -(u-u_\star)(v-v_\star) \\ -(u-u_\star)(v-v_\star) & (u-u_\star)^2 \end{pmatrix} + \mathcal{O}\left(|x-x_\star|^3\right). \tag{83}$$

E.3 Geodesic Equations

To construct geodesic normal coordinates, we first need to compute the geodesics on this manifold. Let us use t as our affine parameter, and use a dot to denote a derivative with respect to t. The geodesic equation is defined to be

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}t^2} + \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}t} \frac{\mathrm{d}x^c}{\mathrm{d}t} = 0, \tag{84}$$

hence our equations are

$$\ddot{u} - \frac{2}{v}\dot{u}\dot{v} = 0, \qquad \ddot{v} + \frac{\dot{u}^2}{v} - \frac{\dot{v}^2}{v} = 0. \tag{85}$$

The equation for u is easy to integrate once. If we multiply by $1/v^2$, we note that this can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{u}}{v^2}\right) = 0 \quad \implies \quad \dot{u} = Av^2, \tag{86}$$

where A is a constant. Similarly, we note that we can write the equation for v as

$$v\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{v}}{v}\right) + \frac{\dot{u}^2}{v} = v\frac{\mathrm{d}^2 \log v}{\mathrm{d}t^2} + \frac{\dot{u}^2}{v} = 0. \tag{87}$$

We can then substitute in our result for $\dot{u}$ to obtain an equation for v only. This becomes easier to solve if we define $w = \log v$, since this is

$$\ddot{w} + A^2 e^{2w} = 0, \tag{88}$$

which has the solution

$$\dot{w} = \pm\left(B - A^2 e^{2w}\right)^{\frac{1}{2}}, \tag{89}$$

where B is a constant. This can then be integrated to obtain

$$\pm(t - t_0) = \frac{-1}{\sqrt{B}} \tanh^{-1}\left(\sqrt{1 - \frac{A^2}{B}e^{2w}}\right), \tag{90}$$

where t_0 is an integration constant. Rearranging (squaring will remove the $\pm$), one finds

$$v(t) = \frac{\sqrt{B}}{|A|} \operatorname{sech}\left(\sqrt{B}\,(t - t_0)\right), \tag{91}$$

where we take the positive square root, since we will always have $v > 0$ for a Gaussian. This can then be used to obtain an equation for $\dot{u}$

$$\dot{u} = Av^2 = \frac{B}{A} \operatorname{sech}^2\left(\sqrt{B}\,(t - t_0)\right), \tag{92}$$

which gives

$$u(t) = \frac{\sqrt{B}}{A} \tanh\left(\sqrt{B}\,(t - t_0)\right) + C, \tag{93}$$

where C is again a constant.

We will redefine our constants such that $\beta = \sqrt{B}$, and we will assume that the geodesic goes through the point $(u_\star, v_\star)$ at $t = 0$. This gives us

$$u(t) = u_\star + v_\star \cosh(\beta t_0) \left[\tanh\left(\beta\left(t - t_0\right)\right) + \tanh(\beta t_0)\right], \tag{94}$$

$$v(t) = v_\star \cosh(\beta t_0) \operatorname{sech}\left(\beta\left(t - t_0\right)\right). \tag{95}$$

It is useful to rewrite this with the identity $\tanh x + \tanh y = \operatorname{sech} x \operatorname{sech} y \sinh(x + y)$, to give

$$u(t) = u_\star + v_\star \sinh(\beta t_0) \operatorname{sech}\left(\beta\left(t - t_0\right)\right), \tag{96}$$

$$v(t) = v_\star \cosh(\beta t_0) \operatorname{sech}\left(\beta\left(t - t_0\right)\right). \tag{97}$$

E.4 Geodesic Normal Coordinates

Now we have found the geodesics which pass through the point p, we are in a position to construct geodesic normal coordinates about this point. These are constructed as follows. Suppose one has a vector V^a which lives in the tangent space at p, then one can parallel transport this to the point X, which one defines as the point on the geodesic with affine parameter $t = 1$ (the exponential map). In this case (provided the metric is Euclidean at p) the geodesic normal coordinates of X are V^a.

Hence, to construct the geodesic normal coordinates, we need to know the tangent vector of our geodesic at the point p. If we define $V^a = (\xi, \eta)$, then these are

$$\xi \equiv \dot{u}(0) = \beta v_\star \operatorname{sech}(\beta t_0), \quad \eta \equiv \dot{v}(0) = \beta v_\star \tanh(\beta t_0). \tag{98}$$

We can therefore relate the remaining two free constants (β and t_0) in our geodesic equation to the geodesic normal coordinates ξ and η.

It is useful to use the identity $\cosh(x - y) = \cosh x \cosh y - \sinh x \sinh y$ to split up the $\operatorname{sech}(\beta(t - t_0))$ term into terms which are functions of only βt or βt_0. Doing this, and recalling the definition that (u, v) is at the point $t = 1$, we obtain the following expressions to relate (u, v) to (ξ, η)

$$u = u_\star + v_\star \frac{\xi \sinh\beta}{\beta v_\star \cosh\beta - \eta \sinh\beta}, \qquad v = v_\star \frac{\beta v_\star}{\beta v_\star \cosh\beta - \eta \sinh\beta}, \tag{99}$$

for

$$\beta = \frac{1}{v_\star}\sqrt{\xi^2 + \eta^2}. \tag{100}$$

This gives expressions for u and v in terms of the geodesic normal coordinates ξ and η. One can see by Taylor expansion, that the coordinates $\xi = \eta = 0$ indeed corresponds to the point $x^a = (u_\star, v_\star)$. For small perturbations, we find that

$$u - u_\star \approx \xi + \mathcal{O}(\text{quadratic}), \quad v - v_\star \approx \eta + \mathcal{O}(\text{quadratic}), \tag{101}$$

which will be useful later.

E.5 Inverting the coordinate transformation

In the previous section we found u and v as a function of ξ and η. In this section, we find the opposite: ξ and η as a function of u and v. We begin by taking $u - u_\star$ and dividing by our expression for v, one can obtain the relation

$$\frac{\mu - \mu_\star}{\sigma} = \xi \frac{\sinh\beta}{\beta}. \tag{102}$$

The left hand side of this expression is reminiscent of the z-score, or standardised Gaussian variable. For small β, $\sinh\beta/\beta \approx 1$, and thus ξ is approximately equal to this standardised variable when one is close to the fiducial point. The rest of the derivation requires us to consider separate cases, depending on the values of ξ and β, which we outline below before summarising the result.

Case 1: $\xi \neq 0$, $\beta \neq 0$ If we rearrange our expression for v and assume $\xi \neq 0$, and substitute in this result to remove $\sinh\beta$, we obtain

$$\beta v_\star \cosh\beta - \frac{\eta\beta}{\xi}\frac{\mu - \mu_\star}{\sigma} = \frac{\beta v_\star^2}{v} = \beta v_\star \frac{\sigma_\star}{\sigma}. \tag{103}$$

Assuming $\beta \neq 0$, this can be rearranged to obtain

$$\cosh\beta = \frac{\sigma_\star}{\sigma} + \frac{\mu - \mu_\star}{v_\star \sigma}\frac{\eta}{\xi}. \tag{104}$$

These can be combined using a hypergeometric identity to obtain

$$\frac{\eta}{\xi} = \frac{\frac{1}{2}(\mu - \mu_\star)^2 + \sigma^2 - \sigma_\star^2}{\sigma_\star(\mu - \mu_\star)\sqrt{2}}, \tag{105}$$

where we have used that $v_\star = \sqrt{2}$. We can now substitute this into our earlier expression for $\sinh\beta$

$$\sinh\beta = \frac{\beta}{\xi}\frac{\mu - \mu_\star}{\sigma} = \frac{\mathrm{sgn}(\xi)}{v_\star}\sqrt{1 + \frac{\eta^2}{\xi^2}}\frac{\mu - \mu_\star}{\sigma}. \tag{106}$$

Since $\beta > 0$, we require the condition $\mathrm{sgn}(\xi) = \mathrm{sgn}(\mu - \mu_\star)$. But given that $\mathrm{sgn}(\sinh\beta/\beta) = 1$, from the earlier formula, we know that this must be true. We therefore find that we can obtain β from

$$\sinh\beta = \frac{\sqrt{\left(\frac{1}{2}(\mu - \mu_\star)^2 + (\sigma - \sigma_\star)^2\right)\left(\frac{1}{2}(\mu - \mu_\star)^2 + (\sigma + \sigma_\star)^2\right)}}{2\sigma_\star\sigma}, \tag{107}$$

and thus we now have a way of computing $(|\xi|, |\eta|)$ from (μ, σ). We note that the signs of the coordinates can be determined using

$$\mathrm{sgn}(\xi) = \mathrm{sgn}(\mu - \mu_\star), \quad \mathrm{sgn}(\eta) = \mathrm{sgn}\left(\frac{1}{2}(\mu - \mu_\star)^2 + \sigma^2 - \sigma_\star^2\right). \tag{108}$$

Case 2: $\xi = 0$**,** $\beta \neq 0$ If $\xi = 0$, then automatically we have that $\mu = \mu_\star$ and that $\beta v_\star = |\eta|$. We therefore obtain

$$v = v_\star \frac{1}{\cosh\beta - \mathrm{sgn}(\eta)\sinh\beta} = v_\star \exp\left(\mathrm{sgn}(\eta)\beta\right), \tag{109}$$

and thus (given that $\beta \geq 0$)

$$\beta = \left|\log\left(\frac{\sigma}{\sigma_\star}\right)\right|, \tag{110}$$

where $\eta > 0$ for $\sigma > \sigma_\star$ and $\eta < 0$ for $\sigma < \sigma_\star$. Note that this is identical to Eq. (107) for $\mu = \mu_\star$.

Case 3: $\beta = 0$ It looks like there are two remaining cases to consider: $\xi \neq 0$ with $\beta = 0$, and $\xi = 0$, $\beta = 0$, but from the definition of β, the first of these is impossible. We also note that the second of these means that $\xi = \eta = 0$. We already noted that this corresponds to the point $\mu = \mu_\star$ and $\sigma = \sigma_\star$, so there is nothing else to add here.

Summary In summary, to obtain (μ, σ) from (ξ, η):

1. If $\mu = \mu_\star$ and $\sigma = \sigma_\star$, then $\xi = \eta = 0$.
2. Otherwise, β can be obtained from evaluating Eq. (107). The signs of ξ and η are given by Eq. (108). To obtain the magnitudes of ξ and η, one can use the results:
   (a) If $\mu = \mu_\star$, then $|\eta| = v_\star\beta$ and $\xi = 0$.
   (b) Otherwise, use Eq. (105)

E.6 Metric and Jacobian for Geodesic Normal Coordinates

Now we have functions for u and v as a function of ξ and η, let us find the metric in these new coordinates. We first find the Jacobian matrix although we do not write the explicit components here, as they are not particularly illuminating. One important property, however, is its determinant, which is

$$\det J = \frac{v_\star^2\beta\sinh\beta}{(\beta v_\star\cosh\beta - \eta\sinh\beta)^2} = \frac{\mu - \mu_\star}{\sigma_\star}\frac{\sigma}{\sigma_\star}\frac{1}{\xi}. \tag{111}$$

We see here that, unless $\mu = \mu_\star$, $\xi = 0$ or $\xi \to \infty$, we will definitely have a finite and non-zero determinant for finite values of μ and σ (provided $\sigma \neq 0$ and $\sigma_\star \neq 0$, but we always have this case). Let us investigate the case $\mu = \mu_\star$ to see if this is problematic. This case corresponds to $\xi = 0$ (so our first two problematic cases are identical), and thus we must consider the limit of $(u - u_\star)/\xi$ as $\xi \to 0$. In this limit $\beta = \eta/v_\star$, and hence

$$\lim_{\xi \to 0} \frac{u - u_\star}{\xi} = \frac{v_\star \sinh(\eta/v_\star)}{\eta \cosh(\eta/v_\star) - \eta \sinh(\eta/v_\star)} = \left[\frac{\eta}{v_\star} \left(\coth\left(\frac{\eta}{v_\star} \right) - 1 \right) \right]^{-1}. \tag{112}$$

The function $y = (x(\coth x - 1))^{-1}$ has no roots and is always positive, hence we find that the Jacobian is always positive at the point $u = u_\star$. Therefore, unless ξ, $\mu - \mu_\star$ or σ is infinite, then we always have a finite and non-zero Jacobian, and thus we can locally invert the coordinate transformation.

E.6.1 Metric

We can use the Jacobian to obtain the metric (i.e. the Fisher) in our new coordinates, which is

$$\tilde{F}_{ab} = \tilde{g}_{ab} = \frac{1}{2\beta^4} \begin{pmatrix} \eta^2 \sinh^2\beta + \xi^2\beta^2 & \eta\xi\left(\beta^2 - \sinh^2\beta\right) \\ \eta\xi\left(\beta^2 - \sinh^2\beta\right) & \eta^2\beta^2 + \xi^2 \sinh^2\beta \end{pmatrix}, \tag{113}$$

where we have used that $v_\star = \sqrt{2}$. To verify that this has the desired behaviour, let us expand the metric around the point $\xi = \eta = 0$. In this case, we obtain

$$\tilde{g}_{ab} \approx \mathbf{1} + \frac{1}{6} \begin{pmatrix} \eta^2 & -\eta\xi \\ -\eta\xi & \xi^2 \end{pmatrix} + \mathcal{O}\left(|x|^3\right), \tag{114}$$

so indeed, in these coordinates, the metric is flat at zeroth and linear order, with the correction occurring at quadratic order. Using the series expansion of Eq. (101), this can be written as

$$\tilde{g}_{ab} \approx \mathbf{1} + \frac{1}{6} \begin{pmatrix} (v - v_\star)^2 & -(u - u_\star)(v - v_\star) \\ -(u - u_\star)(v - v_\star) & (u - u_\star)^2 \end{pmatrix} + \dots \tag{115}$$

which is identical to Eq. (83) (since $v_\star = \sqrt{2}$), as required.

E.7 Loss function for neural network

In the neural network search for our coordinates, we compute the Frobenius norm (Eq. (5)) between the metric and the identity. For the geodesic normal coordinates we find that this takes a particularly nice form

$$\left\| \tilde{g}_{ab} - \delta_{ab} \right\|_{\mathcal{F}}^2 = \left(\frac{\sinh^2\beta}{\beta^2} - 1 \right)^2 \approx \frac{\beta^4}{9} + \mathcal{O}(\beta^6), \tag{116}$$

where the first equality is exact, and the approximation is valid for $\beta \ll 1$. The quartic behaviour in the square of the Frobenius norm is consistent with geodesic normal coordinates having quadratic deviations from flatness near the fiducial point.

F Experiment Details

Default Distillery setup. The distillery comes with controls that users can tune for specific problems. We calibrated our setup on the Rosenbrock problem, and made slight modifications (such as changing network size or depth) for each application. A fixed simulation budget is chosen in advance. The simulations are 50–50 split into Fishnets training and validation datasets. Once the Fisher matrices are learned, the flattening network is fit on the *validation* set of true parameters and ensemble of learned Fishers $(\theta, \{\mathbf{F}\}_j)$ (the train split is discarded). **SR data augmentation.** Once the flattening ensemble is trained, the map $\boldsymbol{\eta}(\boldsymbol{\theta})$ no longer depends on the underlying simulator; we therefore draw a fresh, dense, uniform sample from the prior support (typically $\sim$5–10$\times$ the Fishnet validation count, e.g. $\sim$2,000 Rosenbrock draws against $\sim$250 validation simulations) and pass it through the ensemble to obtain the augmented triple $(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}, \delta\boldsymbol{\eta})$ used as SR inputs. Empirically this dense, well-conditioned coverage of the prior is the single most reliable lever for clean Pareto fronts

and recovery of the analytic coordinates—using only the $\sim$250 original validation points routinely yields noisier expressions and slower convergence. Each ensemble member is aligned to a reference member using the Procrustes alignment scheme. Symbolic regression is carried out for 120 seconds per component, and expressions are pruned with an importance-weighted Frobenius loss threshold of 0.1.

Default one-step setup. The one-step arm drops the Fishnet and the separate flattening network. Two maps are trained jointly: $\eta_\phi(\boldsymbol{\theta})$, a GELU MLP of widths $(128, 128, 128)$ with a linear skip from the box-scaled parameters, and $\eta_\psi(\boldsymbol{x})$, which mean-pools the replicates and then applies a GELU MLP of widths $(32, 32)$, again with a linear skip. The loss is $\frac{1}{2}|\eta_\phi(\boldsymbol{\theta}) - \eta_\psi(\boldsymbol{x})|^2 - \frac{1}{2}\log\det(JJ^\top)$, with Adam at 10^{-3} and batch size 512. A probe with four latent axes is trained for 20,000 steps; the rank is the number of axes whose held-out information exceeds 0.5 nats. A new network is then trained from scratch at that rank for another 20,000 steps (the probe weights are discarded). Four copies form the ensemble, which is aligned and passed to the same symbolic-regression stage as above. Fifteen percent of the training simulations are held out for the rank rule; the rest of the budget is used for the fit.

Downstream neural posterior estimation. For each of the four scientific applications and the Rosenbrock validation, we train a Masked Autoregressive Flow (MAF) as a Neural Posterior Estimator (NPE) using `ltu-ili` with the `lampe` backend (Papamakarios et al., 2021). We use 5 MAF transforms, 50 hidden features per transform, two MAF repeats, learning rate 10^{-4}, and up to 1 000 epochs, with early stopping by smoothed validation log-probability over a 10-epoch moving window. The simulation-budget sweeps span $\{100, 500, 1\,000, 5\,000, 10\,000\}$ training simulations with corresponding batch sizes $[8, 22, 32, 32, 32]$, and average over three random seeds. Crucially, the simulations are generated uniformly in θ-space and, for the η-coordinate inference, converted via the discovered coordinate transformation. This ensures that prior volumes do not influence the inference results. Log-losses in η-space are converted to θ-space using the mean $\log\left|\det\left(\partial\theta/\partial\eta\right)\right|$ over the prior, estimated by central finite differences on 512 samples.

F.1 Rosenbrock

Consider a Gaussian in two dimensions controlled by means $\boldsymbol{\eta} = (\nu_0, \nu_1)$, which we would like to infer from data, and fixed diagonal covariance $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$. Assuming a uniform prior on θ, the posterior distribution can be expressed as

$$\log p(\boldsymbol{\eta}|\{\mathbf{x}^i\}) = -\frac{1}{2}\frac{(\nu_0 - \langle x_0\rangle)^2}{(\sigma_0/\sqrt{n_d})^2} - \frac{1}{2}\frac{(\nu_1 - \langle x_1\rangle)^2}{(\sigma_1/\sqrt{n_d})^2}, \tag{117}$$

where $\langle a\rangle = \frac{1}{n_d}\sum_{i=1}^{n_d} a^i$, which is also Gaussian, and n_d is the number of data points. This likelihood has a known Fisher information matrix: $\mathbf{F}' = n_d\,\mathrm{diag}(1/\sigma_0^2, 1/\sigma_1^2)$. Hence, in this coordinate system, by a simple rescaling we can make the Fisher information the identity.

Let us now consider a new coordinate system for which the Fisher information is not so clearly related to the identity. Let us define $\boldsymbol{\theta} = (\mu_0, \mu_1)$ such that

$$\mu_0 = \nu_0, \quad \mu_1 = \nu_1 + \nu_0^2, \quad \Leftrightarrow \quad \nu_0 = \mu_0, \quad \nu_1 = \mu_1 - \mu_0^2. \tag{118}$$

Applying the Jacobian of this transformation (which has determinant 1), the posterior in the new coordinates is given as

$$\log p(\boldsymbol{\theta}|\{\mathbf{x}^i\}) = -\frac{1}{2}\frac{(\mu_0 - \langle x_0\rangle)^2}{(\sigma_0/\sqrt{n_d})^2} - \frac{1}{2}\frac{(\mu_1 - \mu_0^2 - \langle x_1\rangle)^2}{(\sigma_1/\sqrt{n_d})^2}, \tag{119}$$

which is a non-convex Rosenbrock-like function (e.g. Rosenbrock, 1960). In these coordinates, the Fisher matrix is given by

$$\mathbf{F} = n_d \begin{pmatrix} \frac{1}{\sigma_0^2} + \frac{4\mu_0^2}{\sigma_1^2} & \frac{-2\mu_0}{\sigma_1^2} \\ \frac{-2\mu_0}{\sigma_1^2} & \frac{1}{\sigma_1^2} \end{pmatrix}. \tag{120}$$

We have therefore shown that a Rosenbrock-like posterior (Eq. (119)), which has a complicated Fisher matrix, can be transformed into a Gaussian posterior (Eq. (117)) with a simple Fisher matrix.

Since this result is exact, we use it as a benchmark to determine whether our methodology can find the coordinates $\boldsymbol{\eta}$ given simulations performed in the coordinates $\boldsymbol{\theta}$.

To test the algorithm, we generate 500 simulations over a fixed uniform prior $(\mu_0, \mu_1) \sim \mathcal{U}[-3, 3] \times [-3, 3]$, splitting 50-50 into train and test sets. These coordinates are then transformed to (ν_0, ν_1) via Eq. (118). We then generate $n_{\mathrm{d}} = 50$ two-dimensional data points from a multivariate normal $\{\mathbf{x}\}_{i=1}^{n_d} \sim \mathcal{N}([\nu_0, \nu_1], \boldsymbol{\Sigma})$ with fixed, diagonal covariance $\boldsymbol{\Sigma} = \mathrm{diag}(1.0, 2.0)$. For inference scaling we investigate $\boldsymbol{\Sigma} = \mathrm{diag}(4.0, 1.0)$, which exacerbates the "banana" shape of the posterior, deviating from a Gaussian approximation.

Hidden rank-2 banana in d dimensions. The dimensionality-reduction panel of Fig. 2 uses the banana map Eq. (118) in an ambient space that is mostly nuisance. We draw $\boldsymbol{\theta} \sim \mathcal{U}[-3, 3]^d$. The trial seed selects two indices (i_0, i_1). For $d = 2$ these indices are $(0, 1)$. The simulator mean is the banana in those two coordinates,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \big(\theta_{i_0},\ \theta_{i_1} - \theta_{i_0}^2\big). \tag{121}$$

The observation is 8 replicates of $\boldsymbol{\mu}$. Each replicate adds independent noise with standard deviations $(0.25, 0.5)$. The other $d - 2$ parameters do not enter the simulator. The Fisher matrix therefore has rank 2 at every $\boldsymbol{\theta}$ and at every d. The parameter domain is a d-dimensional box.

We use a budget of $N = 1000$ simulations for η discovery and $N = 2000$ simulations for posterior inference for all pieces of the comparison (oracle, discovered, raw θ). For the discovered coordinate inference **simulations used in discovery are reused for downstream inference**.

This experiment is the rank-2 counterpart of the rank-1 heater in Appendix F.5, and it uses the same stages. The one-step loss Eq. (6) trains a probe with $m = 4$ latent axes. A screen ranks the axes by held-out information and returns $\hat{r}$. We refit a new network from scratch at $m_{\mathrm{fit}} = \hat{r}$. We then align eight networks and run the same symbolic search.

One condition is harder here than in the heater. A trial is correct only if $\hat{r} = 2$. The two axes with the most information must also be the hidden pair $\{i_0, i_1\}$. The screen must therefore find *which* coordinates carry the signal, and not only how many. We sweep $d \in \{2, 4, 8, 16, 32\}$ with 10 trials for each d at $N_{\mathrm{sim}} = 2000$. We run the three-step arm only for $d \leq 16$.

The downstream comparison trains three MAF posterior estimators for each trial at the same budget. The *raw* arm uses the full $\boldsymbol{\theta} \in [-3, 3]^d$. The *discovered* arm uses the symbolic stack of that trial. The *oracle* arm uses the true coordinates $(\theta_{i_0},\ \theta_{i_1} - \theta_{i_0}^2)$. We standardise the last two arms on the training draws. Every arm uses the same estimator class, so the comparison shows the effect of the coordinates and not the effect of the density estimator.

We regenerate the simulations from the discovery seed, so all three arms train on identical data. The score extends the scalar axis score of the heater to k axes. We draw posterior samples for each validation observation. We project the samples onto the shared axes and fit a k-dimensional Gaussian. We report the log density of that Gaussian at the true axis value. We compare raw with oracle on the two true banana axes, and raw with discovered on the m_{fit} discovered axes.

F.2 Gaussian with unknown variance

We generate 10 000 simulations via $\mu \sim \mathcal{U}[-5, 5]$ and $\sigma \sim \mathcal{U}[0.2, 7.0]$, and generate $x \sim \mathcal{N}(\mu, \sigma)$ with $\dim(x) = 50$. We adopt the same network setup calibrated on the Rosenbrock example. **Results.** Using the same distillery settings as the Rosenbrock example, we recover the MDL expressions

$$\eta_1 = \mu\,(0.4\,\sigma - \exp(0.14\,\sigma))\,(0.06\,\mu^2 - 0.02\,\exp(0.95\,\sigma) - 3.0) + 7.0,$$
$$\eta_2 = 0.8\,\sigma + (0.4\,\mu^2 + 1.0\,\sigma)^{\exp(-0.13\,\sigma)} - 0.9$$

and Frobenius expressions

$$\eta_1 = \big(\mu\,(0.06\,\mu^2 - 3.0)\,(0.02\,\sigma\,\exp(0.28\,\sigma) - 1.0) + (0.3\,\mu + 7.0)\,\exp(0.28\,\sigma)\big)\,\exp(-0.28\,\sigma),$$
$$\eta_2 = 0.8\,\sigma + (0.4\,\mu^2 - 0.04\,\mu + \sigma)^{\exp(-0.12\,\sigma)} - 0.9.$$

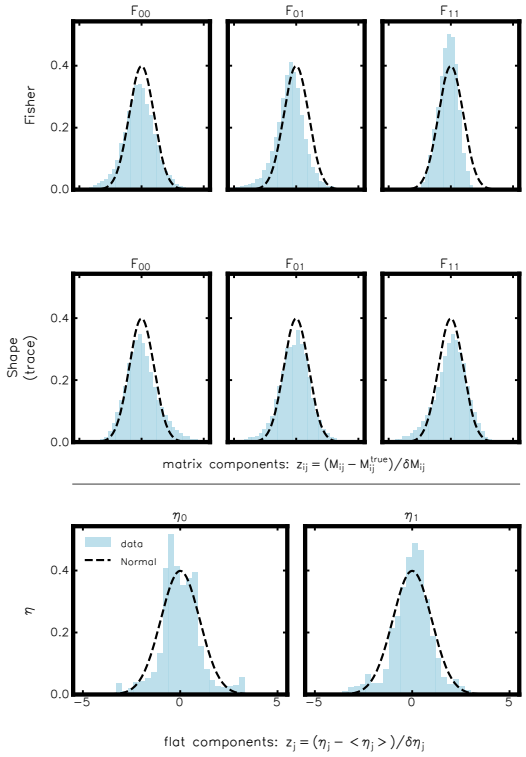

Figure 6: *Top panels:* Fishnets provides an unbiased estimate of a) unique Cholesky factors and b) normalised shape of the true Rosenbrock Fisher Matrix. *Bottom panel:* Learned flat coordinates η and their ensembled-estimated uncertainties $\delta\eta$ approximately follow a Gaussian sampling distribution.

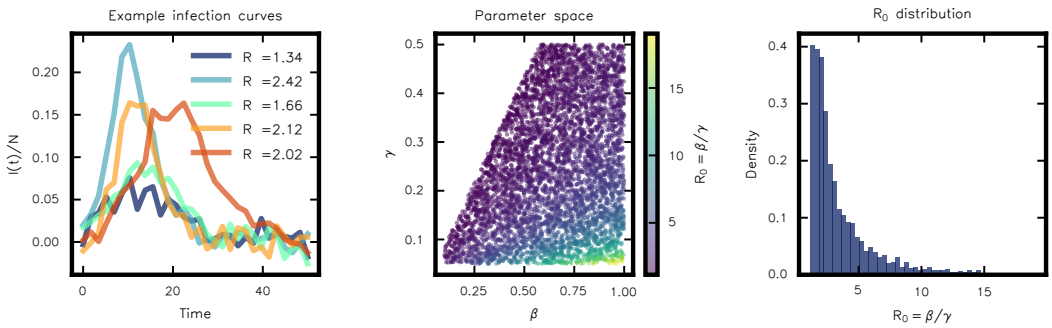

Figure 7: Example SIR epidemic ($R_0 > 1.0$) time-series data and parameter values, coloured by R_0.

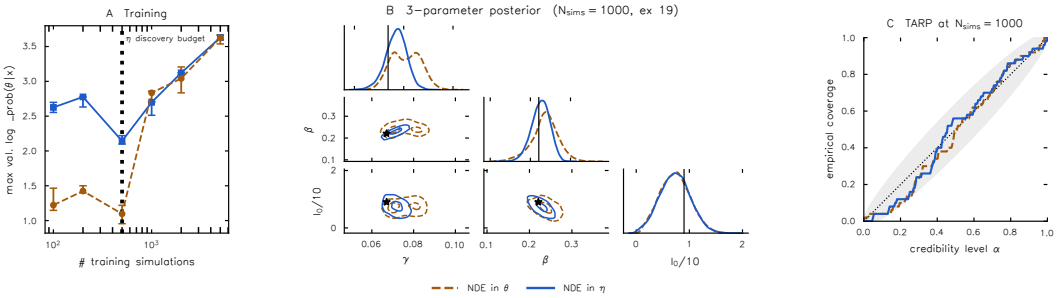

Figure 8: Supplementary details for the SIR experiment. *Left:* Discovered coordinates require fewer simulations in noisier settings to achieve the same log-probability scores in downstream NDE posterior estimation in θ space. *Middle:* 3D posteriors (here for $N_{\text{sim}} = 1000$) show good agreement with simulation truths and tighter η constraints. *Right:* both η and θ posteriors indicate underconfident (conservative) coverage across test simulations.

F.3 SIR Simulation Details

We use the classical susceptible–infectious–removed (SIR) epidemic model (Kermack & McKendrick, 1927; Hethcote, 2000), in which a closed population of size N is partitioned into susceptible $S(t)$, infectious $I(t)$, and removed/recovered $R(t)$ compartments:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\frac{SI}{N}, \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \beta\frac{SI}{N} - \gamma I, \qquad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I, \tag{122}$$

where β is the transmission rate, γ the recovery rate, and the basic reproduction number is $R_0 = \beta/\gamma$. Fig. 8 shows supplementary downstream results.

Simulations use $N = 5000$ with initial conditions $I(0) \sim \text{Pois}(\lambda = 8)$, $S(0) = N - I(0)$, $R(0) = 0$. We observe the normalised infectious trajectory $I(t)/N$ at 30 evenly spaced times over $t \in [0, 50]$ with additive Gaussian noise of standard deviation $\sigma_{\text{obs}} = 0.01$. Parameters are sampled independently as $\beta \sim \mathcal{U}(0.1, 1.0)$, $\gamma \sim \mathcal{U}(0.05, 0.5)$, with the rescaled initial condition $I_0/10$ entering as a third inference coordinate; we generate 2 500 training and 2 500 test simulations. For the epidemic-regime analysis we retain samples with $R_0 \geq 1 + \delta$ with $\delta = 0.15$, and discard the band $0.85 \leq R_0 < 1.15$; the corresponding non-epidemic subset is $R_0 < 0.85$. We display example simulations and parameters in Fig. 7.

To demonstrate information capture in the identified R_0 direction, we grade posterior values of R_0, computed from posterior chains in (β, γ). For each test observation $x^{(i)}$ with true reproduction number $R_0^{(i)} = \beta^{(i)}/\gamma^{(i)}$, we score the posterior $q(R_0 \mid x^{(i)})$ via the continuous ranked probability score (CRPS)

$$\text{CRPS}\Big(q,\, R_0^{(i)}\Big) \;=\; \mathbb{E}_{R\sim q}\,\big|R - R_0^{(i)}\big| \;-\; \tfrac{1}{2}\,\mathbb{E}_{R,R'\sim q}\,\big|R - R'\big|, \tag{123}$$

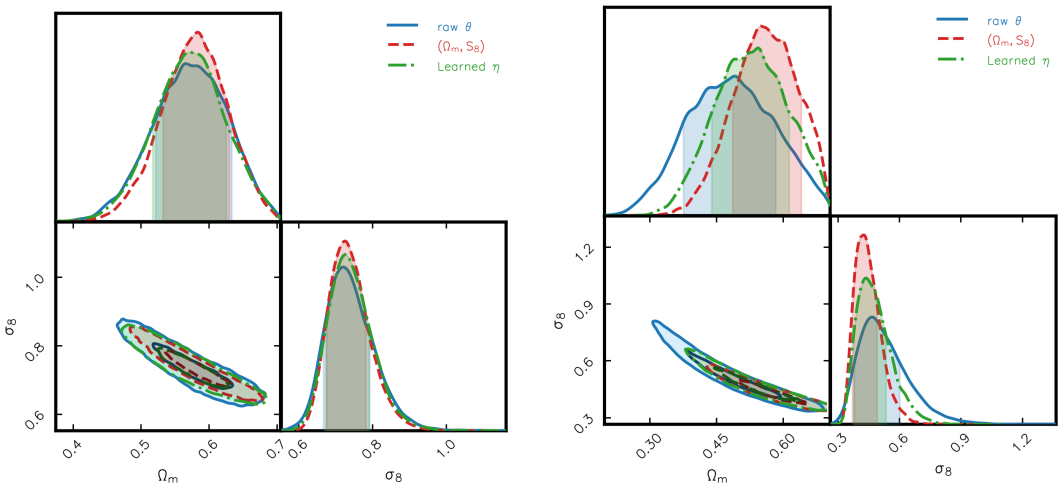

Figure 9: Example NPE posteriors using learned versus conventional weak lensing coordinates. Learned coordinates trace the conventional coordinate constraints, and outperform raw θ sampling.

and report the mean over the test set, $\overline{\mathrm{CRPS}}(R_0) = N^{-1}\sum_i \mathrm{CRPS}(q, R_0^{(i)})$. CRPS is a strictly proper scoring rule with units of R_0: it reduces to the absolute error when q is a point mass, and it is uniquely minimized in expectation when q matches the true posterior over R_0 (Gneiting & Raftery, 2007). We estimate (123) from M posterior samples $\{R_j\}_{j=1}^{M}$ using the unbiased U-statistic

$$\widehat{\mathrm{CRPS}} = \frac{1}{M}\sum_j \left|R_j - R_0^{(i)}\right| - \frac{1}{2M(M-1)}\sum_{j\neq k}\left|R_j - R_k\right|, \tag{124}$$

with $R_j = \beta_j/\gamma_j$ obtained from posterior draws $(\beta_j, \gamma_j) \sim q(\cdot \mid x^{(i)})$.

F.4 Weak Gravitational Lensing

For our weak lensing example we adopt the 5 000 noisy tomographic WL simulations from Makinen et al. (2025), generated with the `pmwd` particle-mesh code (Li et al., 2022) and collected into four redshift bins to form convergence-image data of shape $(128, 128, 4)$. Simulations span a wide uniform prior $p(\Omega_m, S_8) = \mathcal{U}([0.15, 0.7] \times [0.35, 1.52])$ and additionally vary the initial-condition seed; following Makinen et al. (2025) we histogram all auto- and cross-power spectra per redshift bin into a 60-component summary per simulation. Additive shape noise with variance $[0.00045021, 0.00087473, 0.00134725, 0.00183411]$ per tomographic bin is injected for downstream neural posterior estimation (Fig. 9) The *same* underlying dark matter simulations are used in the distillery and inference steps.

F.5 Heater experiment

This appendix expands Section 5.2. We repeat the simulator and the Stone cost. The screen, the symbolic search, and the score follow. We use a budget of $N = 1000$ simulations for η discovery and $N = 2000$ simulations for posterior inference for all methods. For the discovered coordinate inference **simulations used in discovery are reused for downstream inference**.

The plant is the resistive-heater step response Eq. (7). The inputs are voltage and current, $\theta = (V, I)$. The temperature depends on (V, I) only through the power $P = VI$. The Fisher matrix has rank 1 at every θ, $F_\theta(\theta) = (\|k\|^2/\sigma^2)\,(\theta_2, \theta_1)^\top(\theta_2, \theta_1)$. The parameter domain is a 2-D box. The data manifold is 1-D.

Density estimation needs more simulations as the ambient dimension d grows Stone (1980). An error ε needs $N(d, \varepsilon) \sim \varepsilon^{-(2\beta+d)/(2\beta)}$ simulations. If the intrinsic dimension is $d^* < d$, training on

that coordinate needs only $N(d^*, \varepsilon)$. The ratio of the two costs is

$$\frac{N(d, \varepsilon)}{N(d^*, \varepsilon)} = \varepsilon^{-k/(2\beta)}, \qquad k \equiv d - d^*. \tag{125}$$

For $\beta = 1$ and $\varepsilon = 10^{-2}$, the ratio equals $10^{k/2}$. The chain-product heater fixes $d^* = 1$ and varies d.

We rotate η and $\hat{\eta}(x)$ into the prior principal-component basis of η. v_{prior} is the prior variance on one axis. v_{resid} is the mean squared error of $\hat{\eta}(x)$ on that axis. The information on an axis is $\frac{1}{2}\log(v_{\text{prior}}/v_{\text{resid}})$. If the data add nothing, $v_{\text{resid}} = v_{\text{prior}}$ and the information is 0 nats. The rank $\hat{r}$ counts axes above 1 nat, and we cap it at d.

We train a new network from random weights with $m_{\text{fit}} = \hat{r}$ axes. If the next axis is below one tenth of the previous axis, $m_{\text{fit}} = \hat{r} + 1$. We cap m_{fit} at $\min(4, d)$.

We fit eight networks at m_{fit} and align them to one origin, and then calculate standard errors. Operon searches with add, multiply, divide, square root, constants, and variables. We keep the candidate with the lowest one-step loss. A correct trial has $\hat{r} = 1$. Its absolute Spearman correlation with $\prod_i \theta_i$ is at least 0.99. The screen returns rank 1 for at least 9 of 10 trials at each d.

On the $d = 2$ heater, the same search returns two affine coordinates. The nuisance coordinate is $\eta_{\text{nuis}} \approx 0.10(V - I) + 0.4$. The identifiable coordinate is $\eta_{\text{iden}} \approx -0.6(V + I) + 6.0$. These coordinates separate the identifiable direction from the nuisance direction.

For each validation y, we draw posterior samples and project them onto a scalar axis. We fit a Gaussian to those samples. The score is the Gaussian log density at the true axis value. The raw and discovered networks use the discovered axis. The oracle uses the true product. Discovered η aligns with the oracle performance for $d \geq 6$.

The prior is $\theta_i \sim \mathcal{U}[1, 2]$. We draw $N_{\text{sim}} = 10^3$ training simulations and 10^3 validation simulations. The time constant is $\tau = 1$, with 20 samples on $t \in [0, 4]$. The noise standard deviation is $\sigma = 0.2$. We train each posterior network as a 4-component mixture density network. The learning rate is 10^{-4}, the batch size is 64, and training stops by 400 epochs.

F.6 Gravitational-wave inspiral waveforms (TaylorF2 warm-up)

We use TaylorF2 inspiral waveforms with non-spinning 2PN phase. The frequency grid runs from $f_{\text{low}} = 20\,\text{Hz}$ to the lightest-system ISCO frequency $f_{\text{ISCO}} = (6^{3/2}\pi(m_1 + m_2)M_\odot)^{-1}$ in steps of $\Delta f = 0.5\,\text{Hz}$, capped at $1\,024\,\text{Hz}$. Waveforms are whitened with the analytic Advanced-LIGO design PSD $S_n(f) = 10^{-49}(x^{-4.14} + 2 + 2x^2)$, $x = f/215\,\text{Hz}$, for $f \geq 10\,\text{Hz}$, with whitening factor $\sqrt{4\,\Delta f / S_n(f)}$. Real and imaginary parts of the whitened waveform are concatenated, and PCA is fit on a bank of $1\,000$ noiseless whitened waveforms with 40 retained components. Unit Gaussian noise is added in the PCA-compressed space. Mass priors are $m_1, m_2 \sim \mathcal{U}(5, 50)\,M_\odot$ subject to $m_1 \geq m_2$. Fig. 10 shows the results.

Conventional baseline. For comparison we also train a neural posterior estimator in the conventional $(\mathcal{M}_c, q)$ coordinates,

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \qquad q = m_2/m_1,$$

with their analytic inverse used to convert posterior samples back to (m_1, m_2) for evaluation.

Benchmarking results. For masses rescaled to $[0.1, 1.0]$, the distillery returns the symbolic coordinates of Section 5.2, $\eta_1 = 0.2\,m_1$, $\eta_2 = 5.5\,\exp(-0.08\,(m_1 + m_2)^2)$, with analytic inverse $m_1 = 5\eta_1$, $m_2 = -5\eta_1 + 2.5\sqrt{3.41 - 2\log|\eta_2|}$ used for downstream sampling. Posterior estimation in η-space matches or exceeds the chirp-mass coordinates at all simulation budgets tested, with the largest gains at low budgets.

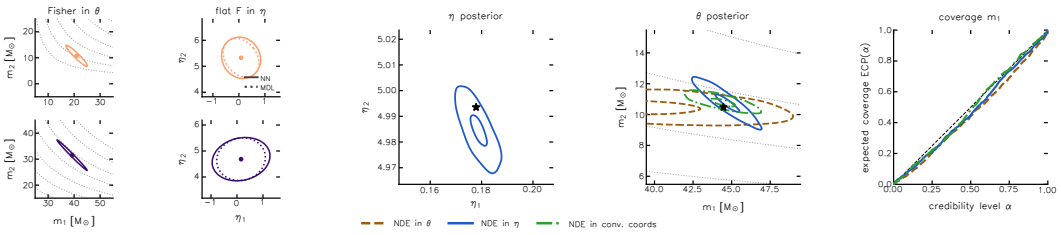

Figure 10: Benchmark GW result, TAYLORF2 (inspiral-only, 2PN, non-spinning). Same panel layout as Fig. 5. The learned coordinates collapse onto a smooth bijection of the conventional chirp-mass / mass-ratio combination $(\mathcal{M}_c, q)$; posterior contours and PP-plots in η and $(\mathcal{M}_c, q)$ are consistent with the analytic argument that the inspiral phase is dominated by $\mathcal{M}_c$ in this mass range.

F.7 GRAVITATIONAL-WAVE INSPIRAL–MERGER–RINGDOWN (IMRPHENOMD)

For the main GW result we replace the inspiral-only TaylorF2 model with the phenomenological inspiral–merger–ringdown waveform IMRPHENOMD (Khan et al., 2016; Husa et al., 2016), generated through `pycbc` (Biwer et al., 2019) on the same uniform (m_1, m_2) prior and Advanced-LIGO design PSD as the TaylorF2 warm-up. The frequency grid is extended to $f_{\max} = 2\,048$ Hz to capture the merger and ringdown of the heaviest systems; the lower cutoff $f_{\text{low}} = 20$ Hz, frequency step $\Delta f = 0.5$ Hz, whitening, and PCA pipeline (40 components fit on 5 000 noiseless waveforms, unit Gaussian noise added in the PCA basis) are identical to Appendix F.6.

Why the chirp-mass scaling fails. At the heavy end of the prior, the merger amplitude and ringdown frequency ($\propto 1/(m_1 + m_2)$) carry a substantial fraction of the in-band signal-to-noise, so the $\mathcal{M}_c \propto (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ scaling that dominates inspiral-only Fisher information is no longer the principal degeneracy direction. The distillery exploits this empirically: the symbolic outputs are functions of $M = m_1 + m_2$ and $\Delta m = m_1 - m_2$ with no $\mathcal{M}_c$ factor, and we verified that running SR on $(M, \mathcal{M}_c)$ inputs reproduces a similar functional form, confirming the chirp-mass dependence is genuinely absent from the post-merger Fisher rather than hidden by a coordinate symmetry.

Calibration. The IMR sweep ($N_{\text{sims}} \in \{500, 1\,000, 2\,500, 5\,000, 10\,000\}$) shows the learned η coordinates achieve TARP coverage within the 1σ bootstrap band across the full $\alpha \in [0, 1]$ range, whereas both (m_1, m_2) and $(\mathcal{M}_c, q)$ NPEs under-cover at high α.

G ROBUST SYMBOLIC DISCOVERY

This appendix compares two discovery pipelines on four problems: The two-step approach fits a compressor and then flattens the Fisher matrix in θ. The one-step objective fits $\eta(\theta)$ and $\hat{\eta}(x)$ together under the joint loss Eq. (6). Once an ensemble of aligned latent coordinates η are found, symbolic regression is run to find short expressions.

We ran 10 fully independent end-to-end trials per problem per arm, at 500 training simulations each. Each trial varies the training simulations and the symbolic-regression search together, not only the network initialisation. Success criteria are evaluated numerically on held-out points. We compare coordinate functions and their Jacobians up to permutation, sign, scaling, and algebraic equivalence. String matching is not used. Algebraically equivalent forms and residual orthogonal freedom make literal comparison meaningless.

Recovered counts trials in which every part of the criterion is met. Alignment is $|\text{Pearson}\, r|$ on held-out points against the expected nonlinear coordinates: the Rosenbrock valley, $R_0 = \beta/\gamma$, chirp mass $\mathcal{M}_c$, or total mass M. We quote the median over trials, with the interquartile range in brackets. Rosenbrock also requires a complementary linear alignment of 0.5. IMRPhenomD also requires a complementary asymmetric mass combination at 0.5. Table 1 uses a uniform correlation bar of 0.6. That bar was chosen after the results were seen.

Problem	N_{sims}	Three-step		One-step	
		recovered	alignment	recovered	alignment
Rosenbrock	500	10/10	0.929 (0.814, 0.960)	10/10	0.999 (0.999, 1.000)
SIR	500	10/10	0.674 (0.636, 0.741)	10/10	0.667 (0.637, 0.779)
GW TaylorF2	500	10/10	0.972 (0.966, 0.979)	10/10	0.983 (0.967, 0.983)
GW IMRPhenomD	500	8/10	0.997 (0.988, 0.999)	6/10	0.991 (0.984, 0.995)

Table 1: Recovery of the expected coordinate, 10 independent trials per problem per arm. Alignment is median $|\mathrm{Pearson}\, r|$ (IQR) on held-out points.

Recovered means extracting coordinates that align with the ground-truth flat coordinates, except for GW IMRPhenomD, where no ground-truth exists. For that case we check that we extract the total-mass coordinates presented in the main text.

IMRPhenomD is the only problem below $10/10$. The misses are threshold misses rather than failures to run. Success there requires two coordinates: the total mass, and a complementary asymmetric mass combination. Both pipelines completed all 10 trials. The total mass is recovered essentially every time (median alignment 0.997 and 0.991). The misses are on the asymmetric combination, which is the harder and more variable direction.

The one-step information rule sets the fit rank to 1 on every IMRPhenomD trial. The second probe axis carries between -0.06 and $+0.002$ nat. One coordinate cannot pass both mass gates. A run with the rank fixed at 2 recovers 10 of 10 trials at this bar.

SIR noise and prior width. We vary one factor at a time about the frozen SIR configuration: observation-noise standard deviation over a five-fold range, and the β and γ prior widths by $0.5\times$ and $2\times$. Each setting uses three independent trials at 500 training simulations, against the ten-trial baseline. Alignment is $|\mathrm{Pearson}\, r|$ against $R_0 = \beta/\gamma$ on held-out points. Gradient cosine is the agreement between the coordinate gradient and ∇R_0.

Setting	Recovered	Alignment (mean, min)	Grad cosine
baseline (frozen)	10/10	0.717, 0.615	0.936
noise sd 0.02	3/3	0.737, 0.596	0.945
noise sd 0.10	3/3	0.668, 0.615	0.938
prior width $0.5\times$	3/3	0.612, 0.584	0.899
prior width $2\times$	2/3	0.525, 0.436	0.955

Table 2: SIR one-factor ablation about the frozen configuration. Alignment is mean and minimum $|\mathrm{Pearson}\, r|$ against R_0. Gradient cosine is the mean agreement with ∇R_0.

The recovered structure is stable under the noise model: $3/3$ recovery at both ends of a five-fold change, and gradient cosine varying by under 0.01. Under a two-fold wider prior the linear correlation drops to 0.525 while the gradient cosine rises to 0.955. The direction the coordinate encodes is not degrading. What degrades is the linearity of the relationship. R_0 is strongly nonlinear in the parameters over that range, and $|\mathrm{Pearson}\, r|$ penalises that while the gradient comparison does not. We report both statistics for this reason.

Widening the prior also shifts the fraction of draws that survive the cut $\beta/\gamma \geq 1.15$ (0.87 at $0.5\times$, 0.76 at baseline, 0.72 at $2\times$). The $2\times$ row is therefore not a clean rescale of the frozen prior, and part of its drop may come from that cut rather than from width alone.