Reconstructing Physical Observables from Sparse Data

A diffusion-based generative framework for reconstructing physical observables from sparse and noisy data

Herzallah Alharazin — in collaboration with



Evgeny Epelbaum



Julia Panteleeva



Bao-Dong Sun



Sparse and noisy data — an ill-posed inverse problem

The unknown is a function on a grid of L points:

$$x_0 = (f(t_1), \cdots, f(t_L)) \in \mathbb{R}^L$$

We know only a few of its values, and only with errors:

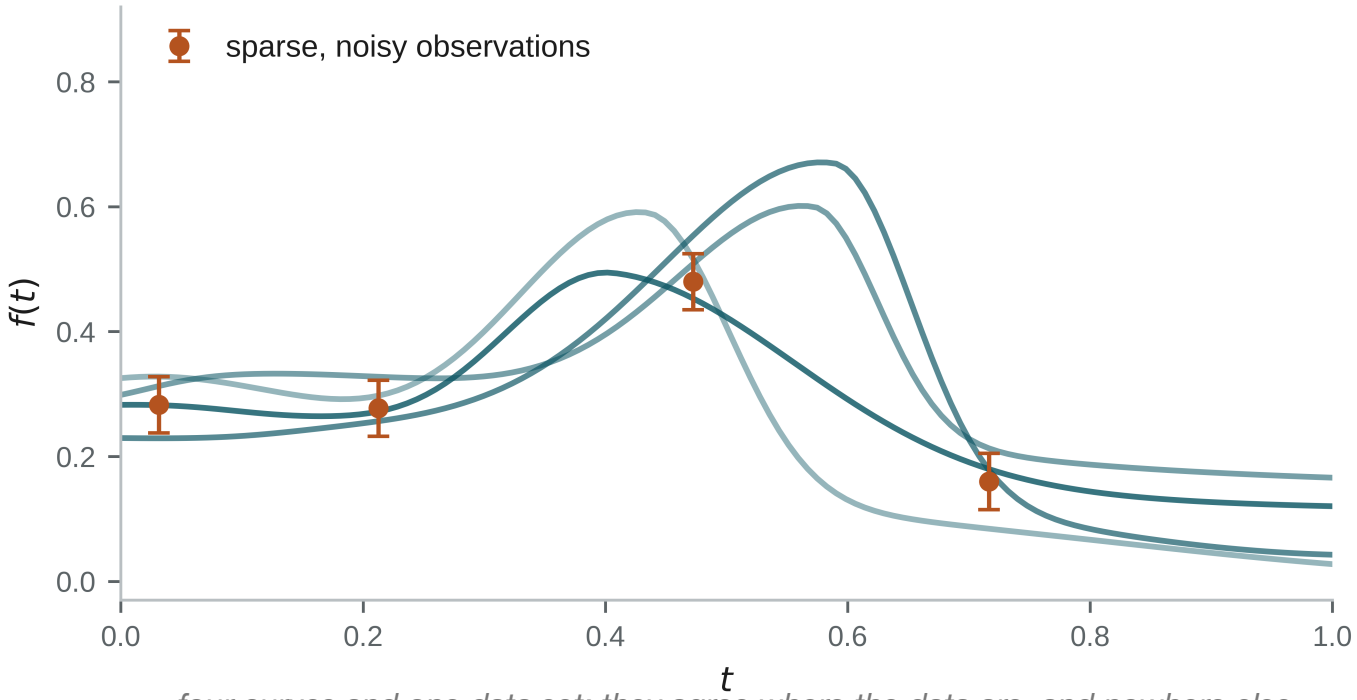
$$y_k = f(t_k) + \eta_k, \quad k \in \{1, \cdots K\}, \quad k < L$$

Infinitely many curves are compatible with the data.

The honest object to look for is not a curve but the distribution of curves compatible with the data:

$$p(x_0 \mid y) \quad \propto \quad p(y \mid x_0) \quad p(x_0)$$

posterior $\propto$ *likelihood* $\times$ *prior*

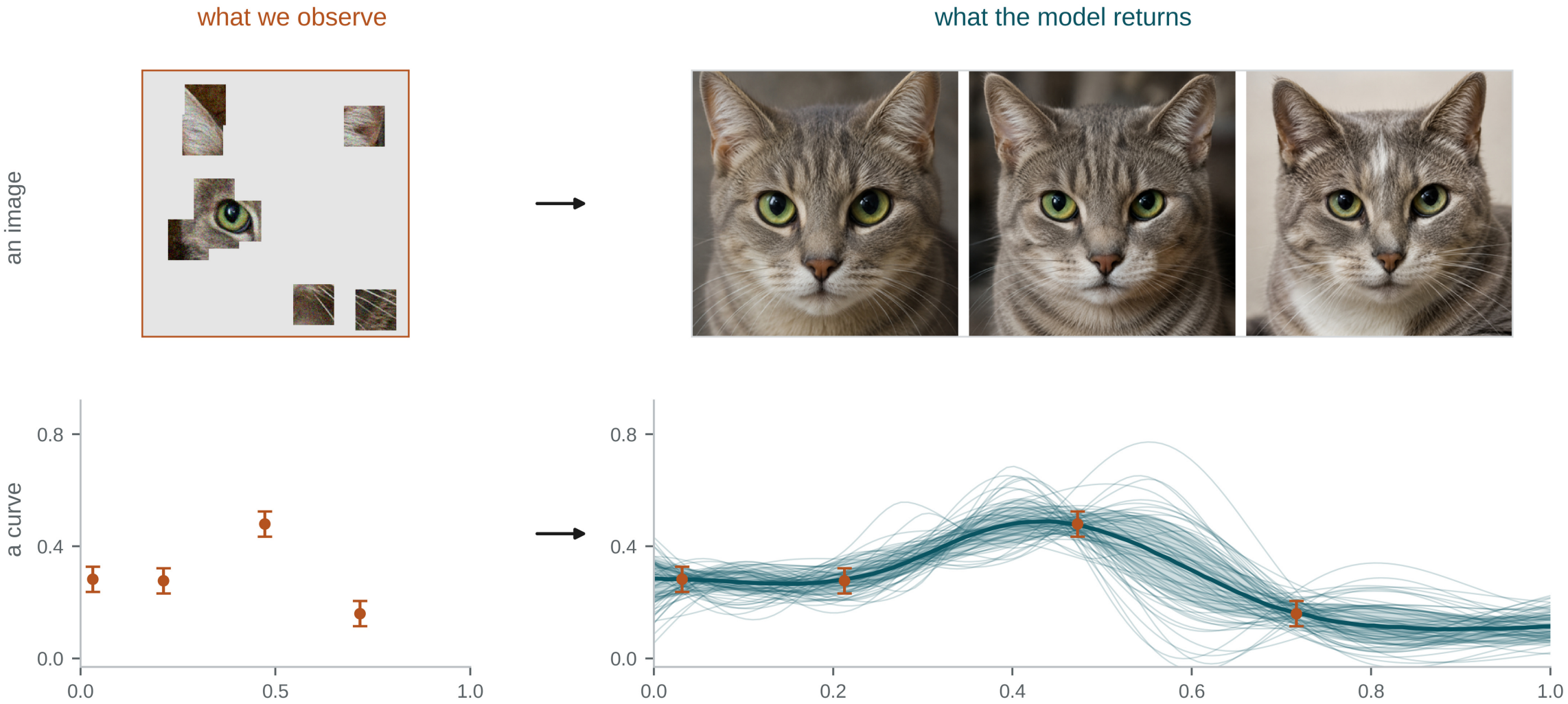


four curves and one data set: they agree where the data are, and nowhere else

The hard part is the prior: what does a plausible curve even look like?

Connection to generative Diffusion Models

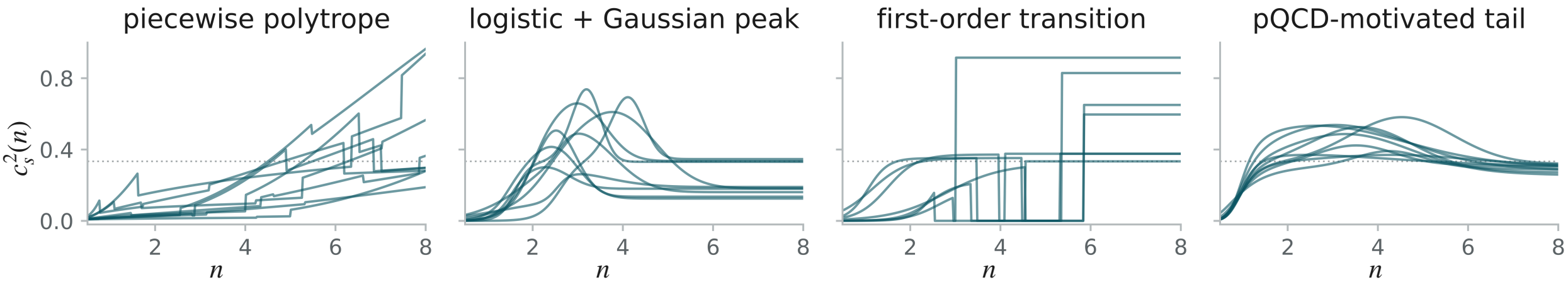
The model supplies the structure the observation lacks, because it has learned what such objects look like.



A generative model returns draws from the set of objects consistent with what you supplied, and to do that it has to have learned, beforehand, what such objects look like.

Where the prior comes from

For most observables the literature already holds several physically motivated families $F_i = \{ f(\,.\,;\theta) : \theta \in M_i \}$, such as:



each panel is one class, sampled over its whole parameter range; all physically motivated classes build the prior

In this approach, we do not choose one family and fit θ. We draw a family, draw its parameters over their admissible range, and evaluate, not a fit. Convex mixtures populate the region between the families too (wherever meaningful).

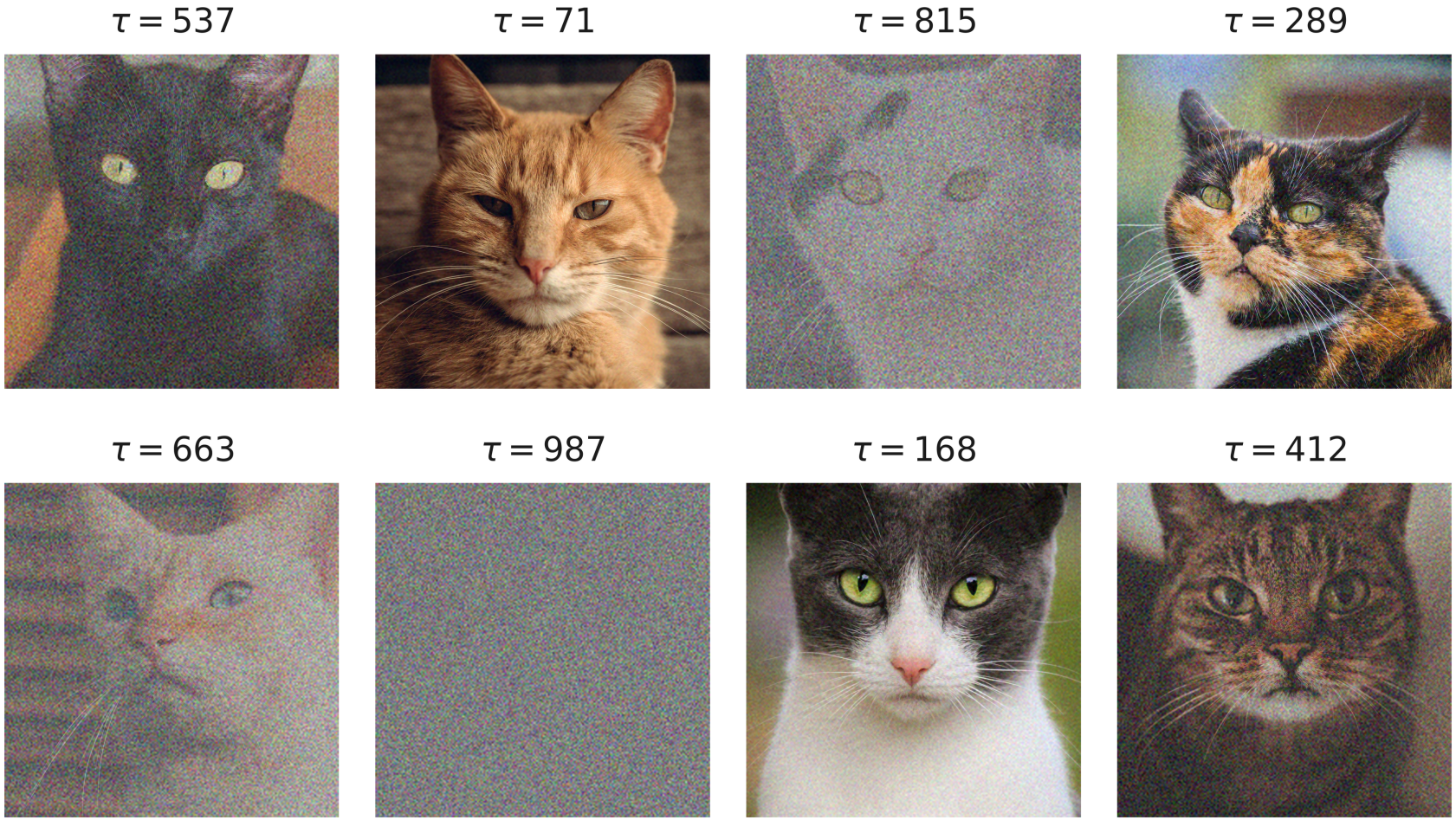
The diffusion network is trained on draws from this ensemble. It learns all the families and their mixtures at once. What it generates afterwards belongs to **no single** one of them.

Learning the prior: corrupt, then learn to undo

- During training, each clean curve in the batch is independently noised at a randomly sampled noise level as follows:

$$x_\tau = \sqrt{\bar{\alpha}_\tau}\; x_0 + \sqrt{1-\bar{\alpha}_\tau}\; \epsilon\,,$$

$$\epsilon \sim N(0, I) \qquad \bar{\alpha}_T \approx 0\,, \qquad \bar{\alpha}_0 = 1$$



- Then we ask the network to predict the following quantity:

$$v_\tau = \sqrt{\bar{\alpha}_\tau}\; \epsilon - \sqrt{1-\bar{\alpha}_\tau}\; x_0$$

- That is, the entire training objective is

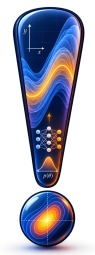
$$\mathcal{L} = E_{x_0,\tau,\epsilon,(m,c)} \left| NN(x_\tau, \tau, m, c) - v_\tau \right|^2$$

The network also receives a randomly selected subset of values from the clean curve, together with a mask indicating their locations (m, c).

- So at any noise level we get an estimate of the clean object

$$\hat{x}_0 = \sqrt{\bar{\alpha}_\tau}\; x_\tau - \sqrt{1-\bar{\alpha}_\tau}\; \hat{v}_\tau \quad \rightarrow \textit{ the engine of the sampler.}$$

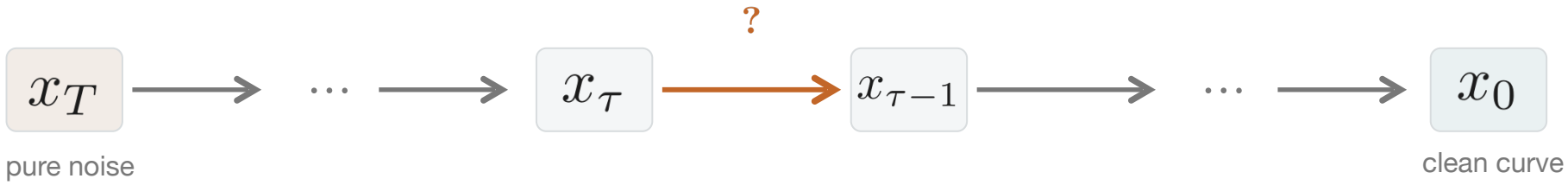
No likelihood, no real measurements anywhere in training.
The network only ever learns what plausible curves look like - it is the prior's sampler, nothing else.



forward process: Sohl-Dickstein+ '15 · Ho+ '20 — cosine schedule (T = 1000, s = 0.008): Nichol & Dhariwal '21 — v-prediction: Salimans & Ho '22

Generating: run the chain backwards

The network is trained. To reconstruct curves from available data we start from pure noise and walk the chain back, one noise level at a time.



backwards: from noise to a curve, every arrow is one random draw

1. Start from noise: $x_T \sim \mathcal{N}(0, I)$

2. At level τ the network sees the noisy state, the level, and the known values (m, c), and returns its estimate of the clean curve:

$$\hat{x}_0 = \sqrt{\bar{\alpha}_\tau}\ x_\tau - \sqrt{1-\bar{\alpha}_\tau}\ \hat{v}_\tau$$

Ho+ '20 — the anchors' full treatment (frozen draws, correlated errors, exact insertion) follows once the step is understood.

On estimating x_0

The training joint, at fixed τ and mask m: $x_0 \sim p_{\text{tr}},\ c = m \odot x_0,\ u := c^K,\ x_\tau = \sqrt{\bar\alpha_\tau}\, x_0 + \sqrt{1-\bar\alpha_\tau}\,\epsilon,\ v_\tau = \sqrt{\bar\alpha_\tau}\,\epsilon - \sqrt{1-\bar\alpha_\tau}\, x_0$

① **Condition the joint on the network's inputs,** x_τ and c are independent given x_0, so Bayes' theorem factorizes:

$$p(x_0 \mid x_\tau, u) \ \propto\ \underbrace{\mathcal{N}\left(x_\tau;\ \sqrt{\bar\alpha_\tau}\, x_0,\ (1-\bar\alpha_\tau)\, I\right)}_{\text{measurement of } x_0 \text{ --- known by construction}} \quad \underbrace{\delta\left(u - x_0^K\right)}_{\text{anchor constraint}} \quad \underbrace{p_{\text{tr}}(x_0)}_{\text{prior}}$$

② **The MSE optimum is the conditional mean, and x_0 is affine in v,** so the ideal output is exactly that posterior's mean:

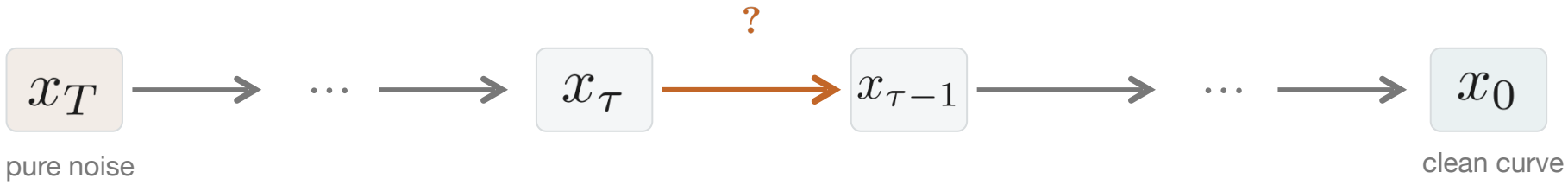
$$f^\star = \arg\min_f\ \mathbb{E}\left\| f(X) - v_\tau \right\|^2 = \mathbb{E}\left[v_\tau \,\middle|\, X \right], \qquad X = (x_\tau, \tau, m, c)$$

What the diffusion network is, precisely:

f_θ approximates $f^\star$, so its output (after the inversion) is the posterior mean $\mathbb{E}_{\text{tr}}\left[x_0 \,\middle|\, x_\tau,\ x_0^K = u \right]$ - the average of all training curves that could have produced its inputs; not a sample, not the prior. In the sampler its input is a reverse-process iterate rather than a forward draw; the function does not know the difference and reads x_τ as if it were one.

Generating: run the chain backwards

The network is trained. To reconstruct curves from available data we start from pure noise and walk the chain back, one noise level at a time.



backwards: from noise to a curve, every arrow is one random draw

1. Start from noise: $x_T \sim \mathcal{N}(0, I)$

2. At level τ the network sees the noisy state, the level, and the known values (m, c), and returns its estimate of the clean curve:

$$\hat{x}_0 = \sqrt{\bar{\alpha}_\tau}\; x_\tau - \sqrt{1-\bar{\alpha}_\tau}\; \hat{v}_\tau$$

3. Take one step down by drawing one noise level lower, steered by that estimate:

$$x_{\tau-1} \sim q(x_{\tau-1} \mid x_\tau)$$?

4. Repeat down to τ = 1. The final step yields the complete reconstruction of the curve.

But what exactly is $x_{\tau-1}$?

Ho+ '20 — the anchors' full treatment (frozen draws, correlated errors, exact insertion) follows once the step is understood.

Generating: run the chain backwards

- The exact reverse step $q(x_{\tau-1}\,|\,x_\tau)$ is intractable: $q(x_{\tau-1}\,|\,x_\tau) = \int dx_0 \;\; q(x_{\tau-1}\,|\,x_\tau, x_0) \;\; p(x_0)$

  unknown

- If we condition on the current clean curve as well, i.e. we sample $x_{\tau-1}$ from $q(x_{\tau-1}\,|\,x_\tau, x_0)$, then Bayes' theorem plus the Markov property give it in closed form:

$$q(x_{\tau-1} \mid x_\tau, x_0) \;=\; q(x_\tau \mid x_{\tau-1}) \, \frac{q(x_{\tau-1} \mid x_0)}{q(x_\tau \mid x_0)}$$

known from →

$$x_\tau = \sqrt{\bar{\alpha}_\tau}\; x_0 + \sqrt{1-\bar{\alpha}_\tau}\;\epsilon$$
$$x_\tau = \sqrt{\alpha_\tau}\; x_{\tau-1} + \sqrt{1-\alpha_\tau}\;\epsilon$$

- All three factors are known Gaussians, hence the result is Gaussian: $q(x_{\tau-1} \mid x_\tau, x_0) \;=\; \mathcal{N}\big(x_{\tau-1};\; \tilde{\mu}_\tau(x_\tau, x_0),\, \tilde{\beta}_\tau I\big)$,

$$\tilde{\mu}_\tau \;=\; a_\tau\, x_0 \;+\; b_\tau\, x_\tau, \qquad a_\tau = \frac{\sqrt{\bar{\alpha}_{\tau-1}}\;\beta_\tau}{1-\bar{\alpha}_\tau}, \qquad b_\tau = \frac{\sqrt{\alpha_\tau}\,(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}, \qquad \tilde{\beta}_\tau = \frac{(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\,\beta_\tau$$

- That is, $x_{\tau-1} \;=\; \tilde{\mu}_\tau\big(x_\tau, \hat{x}_0\big) \;+\; \sqrt{\tilde{\beta}_\tau}\; z\,,\;\; z \sim \mathcal{N}(0, I)$

Ho+ '20 — the anchors' full treatment (frozen draws, correlated errors, exact insertion) follows once the step is understood.

Generating: run the chain backwards

- To summarize: the full loop is given in four steps. For each sample:

1. Sample a realization of the known values, held fixed along its whole trajectory:

$$c = y_A + L\,z\,,\quad LL^{\mathsf{T}} = \Sigma_A,\ \text{(or if uncorrelated)}\ c_k = y_{A,k} + \sigma_k\,z_k,\quad z_k \sim \mathcal{N}(0,1)$$

frozen for each sample, because the T level-analyses update one and the same Bayes problem; the ensemble over draws performs the marginalization over the anchor law

2. Draw random and pure noise: $x_T \sim \mathcal{N}(0, I)$.

3. Then start the denoising loop from $\tau = T$ until $\tau = 1$ as follows:

   1. At every step τ, the known entries are overwritten by a forward-noised copy of c at the noise level of the state they are replacing:

$$x_\tau \leftarrow m \odot \left(\sqrt{\bar{\alpha}_\tau}\,c + \sqrt{1-\bar{\alpha}_\tau}\,\varepsilon'\right) + (1-m)\odot x_\tau, \qquad \varepsilon' \sim \mathcal{N}(0, I)$$

   2. The network receives the noisy state, the noise level, and what is already known. Then it predicts the clean estimation at the corresponding noise level:

$$\hat{x}_0 = \sqrt{\bar{\alpha}_\tau}\,x_\tau - \sqrt{1-\bar{\alpha}_\tau}\,\hat{v}_\tau\,,\quad \hat{v}_\tau = f_\theta\left(x_\tau, \tau, m, c\right)$$

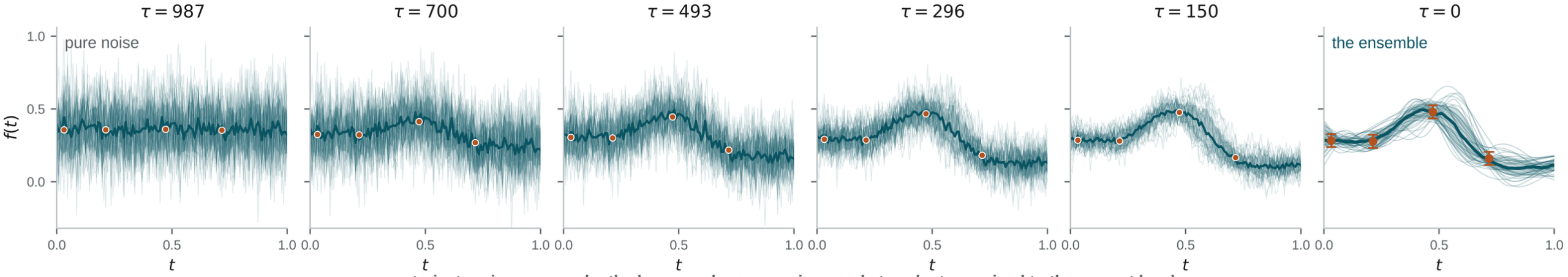
   3. Take one reverse step:

$$x_{\tau-1} = a_\tau\,\hat{x}_0 + b_\tau\,x_\tau + \sqrt{\tilde{\beta}_\tau}\,z\,,\quad z \sim \mathcal{N}(0, I)$$

4. The final clean sample: $x_0 \leftarrow m \odot c + (1-m)\odot x_0$

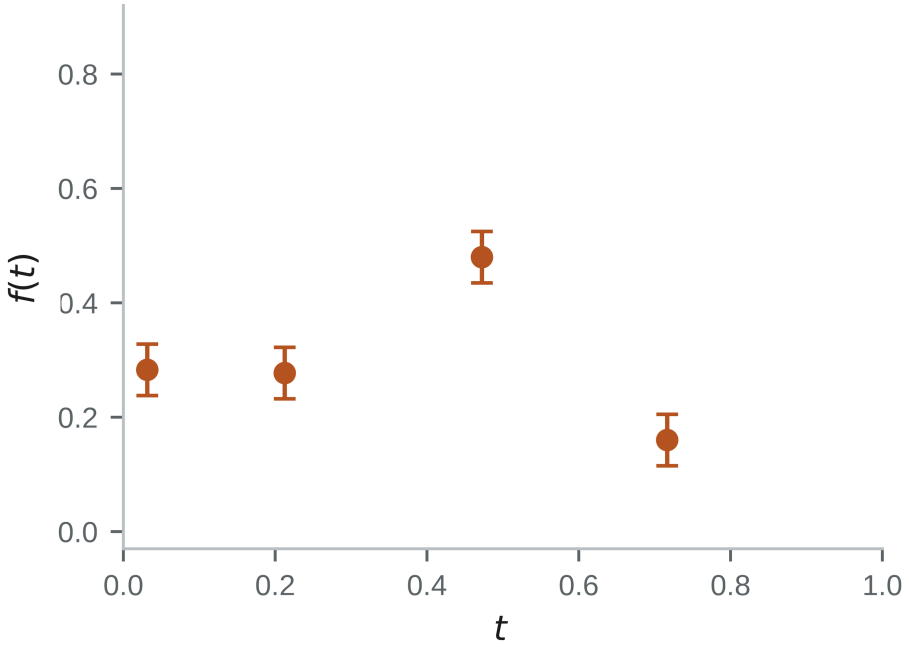
- $m \in \{0,1\}^L$ binary mask, unity at the K known (anchor) indices
- $c \in R^L$ conditioning channel: the anchor values at those indices, zero elsewhere
- $y_A \in R^K$ the measured anchor values
- Σ_A their covariance ($\Sigma_A = diag(\sigma_1^2, \ldots, \sigma_K^2)$ if uncorrelated)
- $\odot$ elementwise product

Watch it happen

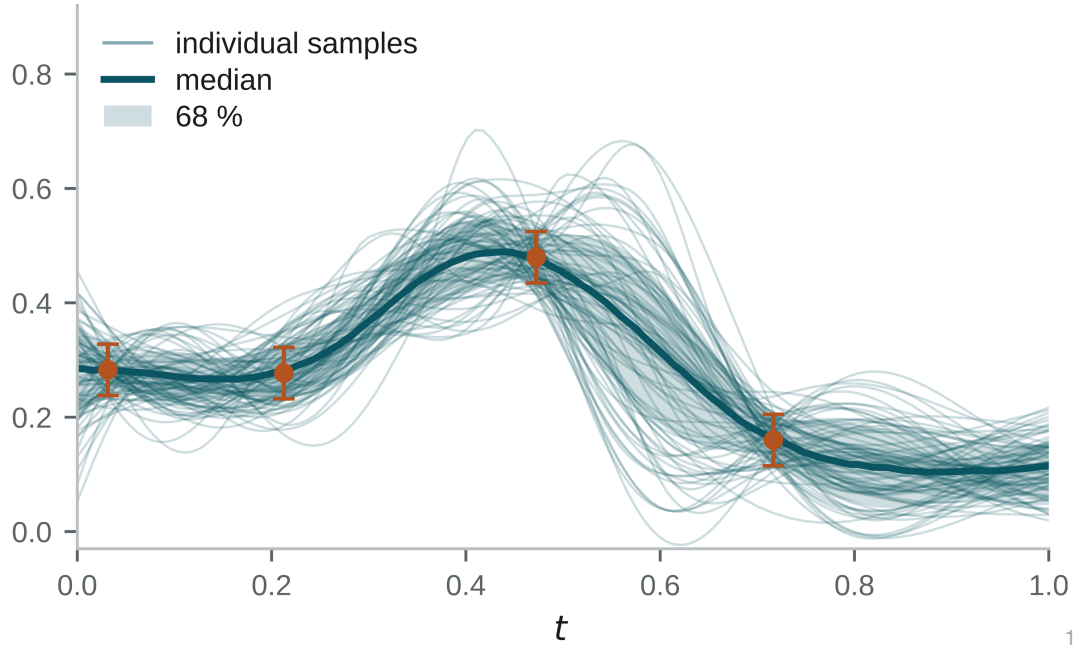


every trajectory is one sample; the known values are re-imposed at each step, noised to the current level

what we know



what the model reconstructs



What generation cannot impose, weights can

Setting

$x_i \in \{x_1, \cdots, x_N\}$ a generated curve, $p(x)$ their distribution

d data that exist only as functionals of the finished curve,

$\mathscr{L}(x_i) := p(d \mid x_i)$ the likelihood of d,

$g(x_i)$ a quantity computed from a x_i

1 **Bayes' theorem:** $p(x_i \mid d) = \dfrac{\mathscr{L}(x_i)\, p(x_i)}{p(d)}$, $p(d) = \displaystyle\int dx\, \mathscr{L}(x) p(x)$.

$$w_n = \frac{\mathscr{L}\left(x_n\right)}{\sum_{m=1}^{N} \mathscr{L}\left(x_m\right)}, \qquad \sum_n w_n = 1$$

2 **For any expectation of $g(x)$ on $p(x \mid d)$ we have:** $\mathbb{E}\left[g \mid d\right] = \displaystyle\int dx\, g(x) p(x \mid d) = \frac{\int dx\, g(x)\, \mathscr{L}(x)\, p(x)}{\int dx'\, \mathscr{L}(x')\, p(x')} \approx \sum_{n=1}^{N} w_n\, g\left(x_n\right),$

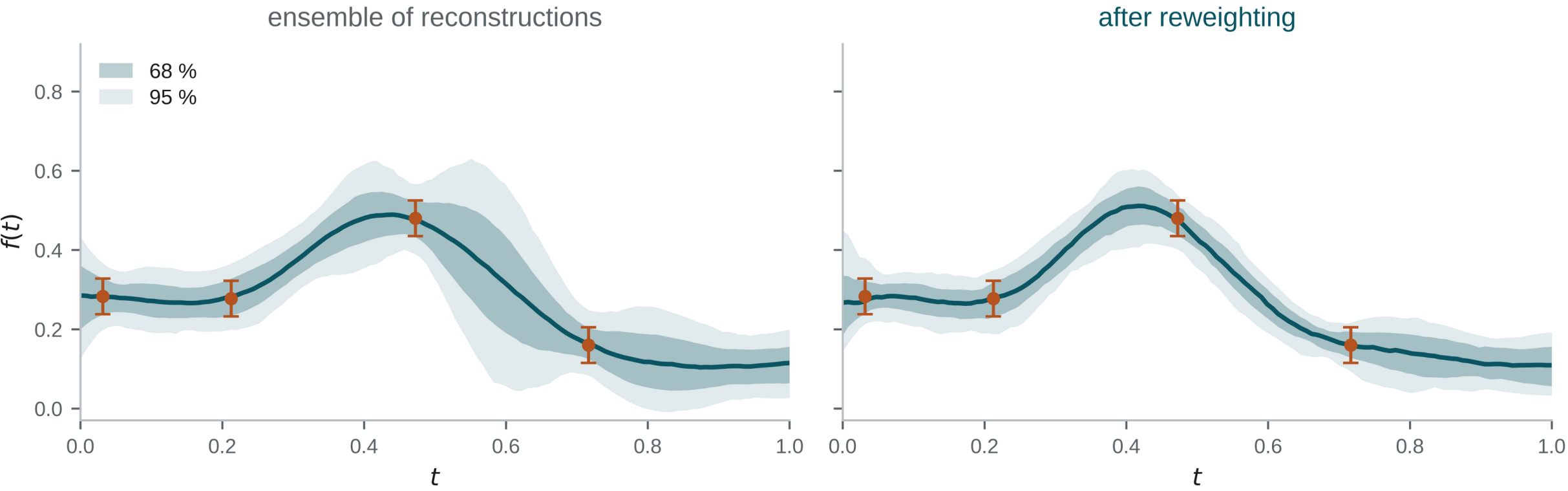
3 **Finally, we can obtain the credible bands q_j^α (weighted quantiles) pointwise on the grid:**

$$F_j(t) = \mathbb{E}\left[\, \mathbf{1}\{\mathbf{x}_j \le t\} \;\middle|\; d \right] \simeq \sum_{n=1}^{N} w_n\, \mathbf{1}\left\{x_{n,j} \le t\right\}$$

$$q_j^\alpha = \min\left\{\, t \in \mathbb{R} \;:\; F_j(t) \ge \alpha \,\right\}, \qquad \alpha \in \{0.16,\, 0.50,\, 0.84\}$$

median curve $q_j^{0.50}$, 68 % band $\left[q_j^{0.16}, q_j^{0.84}\right]$, $j = 1, \ldots, L$

What generation cannot impose, weights can



$$F_j(t) \;=\; \mathbb{E}\big[\,\mathbf{1}\{x_{0,j} \le t\} \;\big|\; d\,\big] \;\simeq\; \sum_{n=1}^{N} w_n\,\mathbf{1}\big\{x_j^{(n)} \le t\big\}$$

$$q_j^{\alpha} \;=\; \min\big\{\, t \in \mathbb{R} \;:\; F_j(t) \ge \alpha \,\big\}, \qquad \alpha \in \{0.16,\, 0.50,\, 0.84\}$$

median curve $q_j^{0.50}$, 68 % band $\big[q_j^{0.16}, q_j^{0.84}\big]$, $j = 1,\ldots,L$

Applications

Neutron-star matter — equation of state

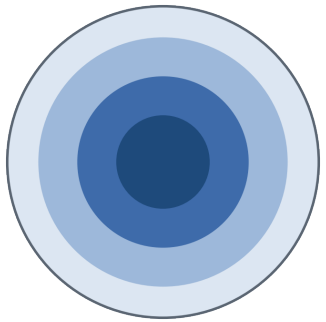
object	$c_s^2(n)$ on a 200-point grid, 0.5 – 8 n_{sat}
prior	10 physically motivated classes → 10^6 curves
anchors	5 χEFT points at low density, with correlated covariance
Reweighting	NICER [R(M)], GW170817 $\tilde{\Lambda}$ [Λ(M)], J1614–2230 [M_max] , pQCD [endpoint] - one likelihood each.

J. Panteleeva, H. Alharazin, E. Epelbaum — arXiv:2608.17457 (2026)

Hadron structure — gravitational form factors

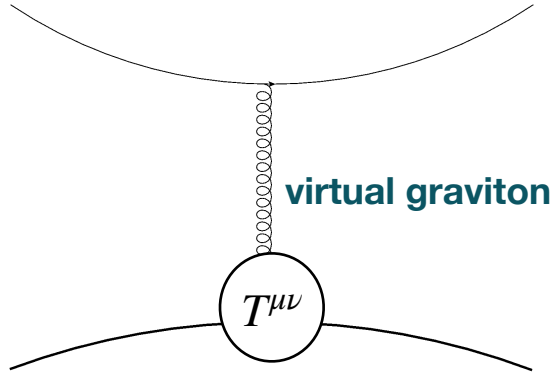
object	F(t) on a 200-point grid, $0 \le -t \le 2$ GeV²
prior	8 theory-rooted families + 2 mixture classes → 6×10^5 curves
anchors	a few lattice-QCD points + exact Poincaré constraints $A(0) = 1$, $J(0) = ½$
Reweighting	physics sign constraints — e.g. $D(0) < 0$

H. Alharazin, J. Panteleeva — Phys. Rev. D 113 (2026)



sound speed

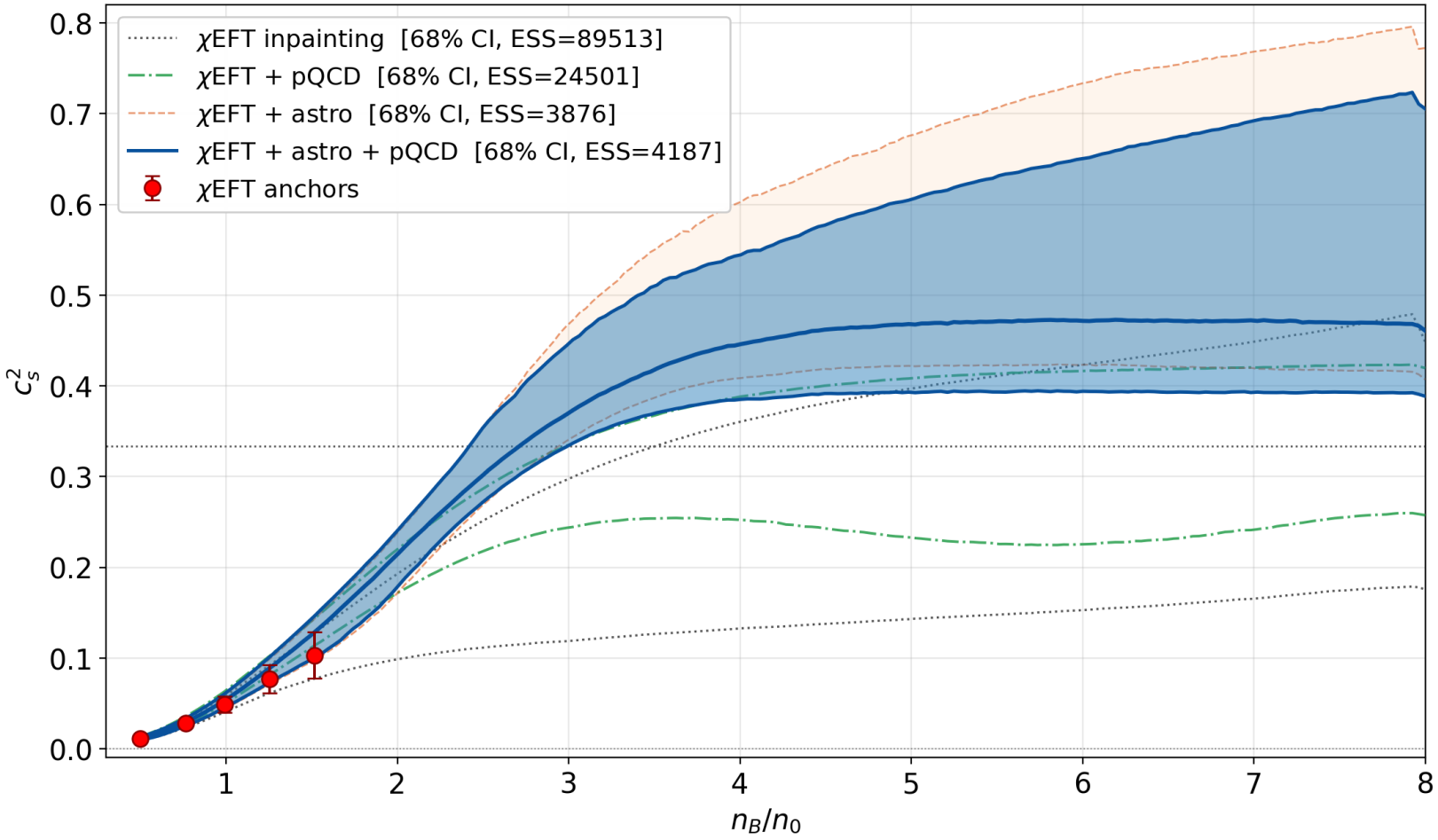
$$c_s^2 = \frac{dP}{d\varepsilon}$$



Same machinery in both: a one-dimensional ResNet–attention denoiser (~19M parameters), v-prediction, random-mask training.

The reconstructed sound speed

one inpainted ensemble, reweighted four ways — the ablation isolates what each constraint does



χEFT inpainting alone

above ~2 n_{sat} the band is wide open — the anchored prior (89,513 samples, unit weights)

+ pQCD

bites above ~3 n_{sat}: removes EoS too soft to reconnect to the perturbative limit — statistically cheap (ESS = 24,501)

+ astrophysics

two-solar-mass stars demand stiffness — the strongest contraction (ESS = 3,876)

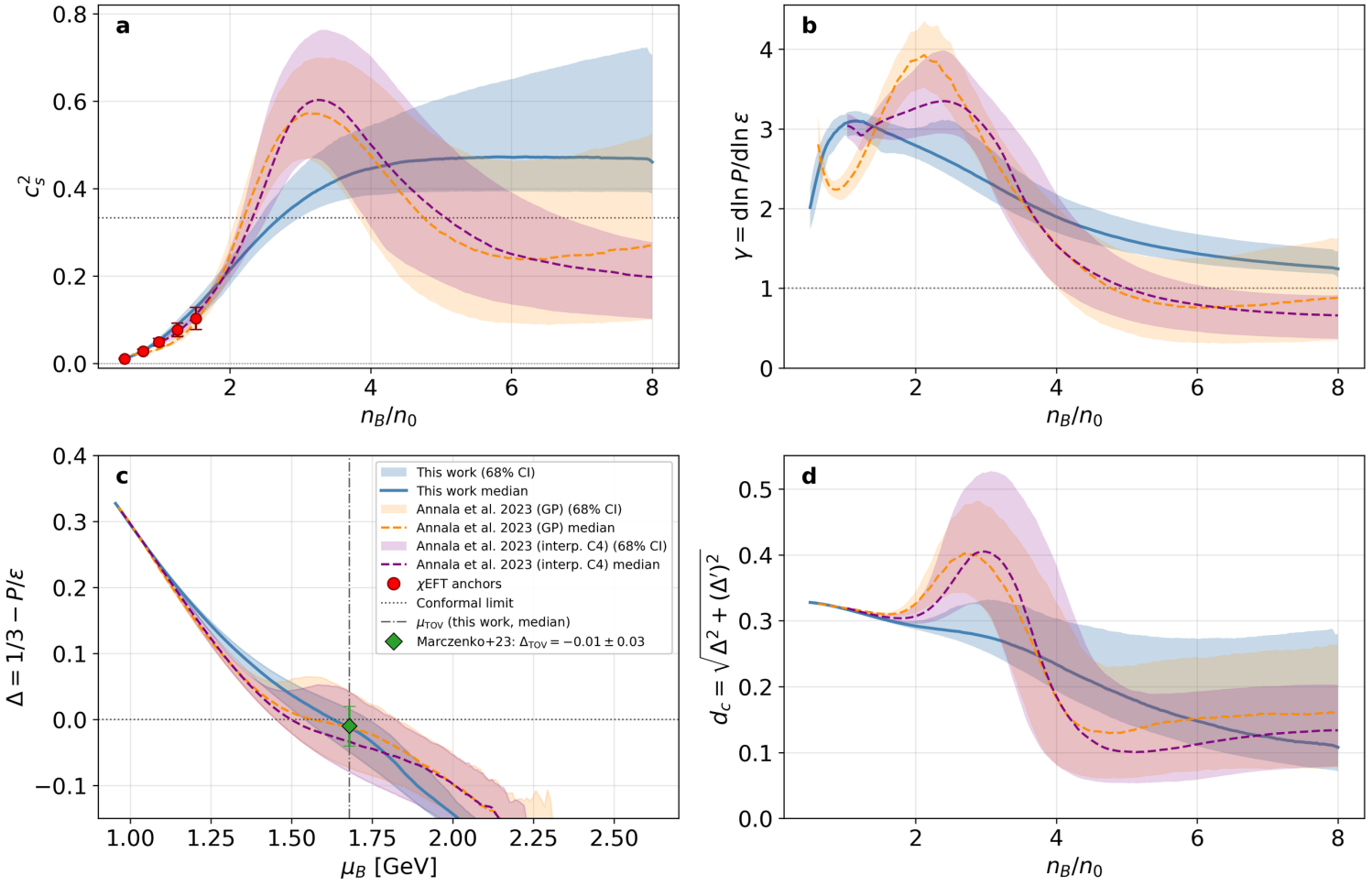
full posterior

astro fixes the stellar range; pQCD trims the residual stiff tail back toward 1/3 (ESS = 4,187)

$c_s^2 \approx 0.47$ at the highest densities: the posterior settles well above the conformal 1/3 — stiff matter to the end of the stable branch.

bands: pointwise equal-tailed 68% credible intervals; red points: the five χEFT anchors — the reconstruction is conditional on them by construction.

Stiff, yet near-conformal — a crossover



At the centre of the heaviest stable star (n_{TOV} = 6.4 n_{sat}), three independently proposed diagnostics agree:

$\gamma = 1.39\ ^{+0.20}_{-0.10}$ — below the quark-matter criterion 1.75

$\Delta = -0.01\ ^{+0.04}_{-0.09}$ — trace anomaly consistent with zero

$d_c = 0.13\ ^{+0.14}_{-0.03}$ — below the near-conformality threshold 0.2

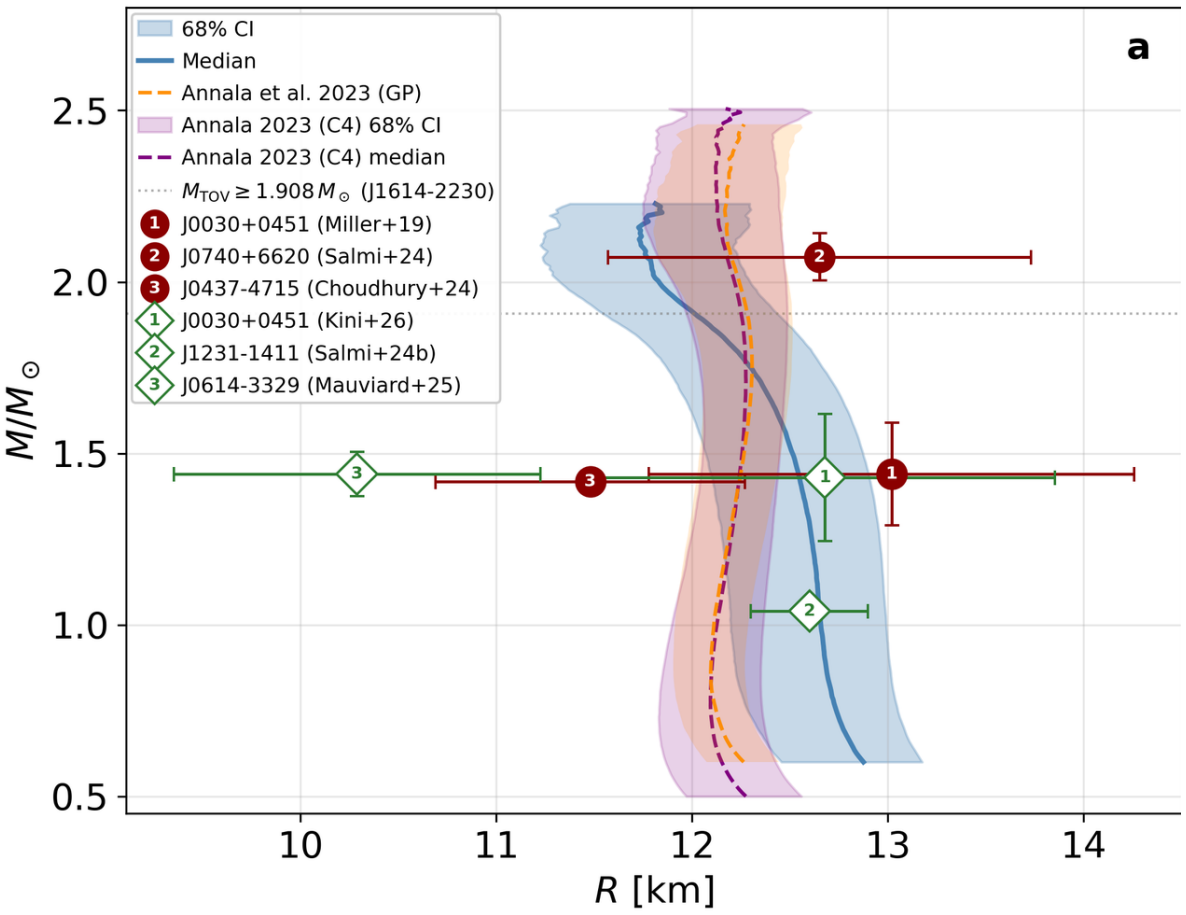
…while $c_s^2 \approx 0.47$ stays far above 1/3. The bulk thermodynamics conformalizes; the stiffness does not — Δ tracks P/ε, c_s^2 its slope, and the two need not conformalize together.

GP and four-segment reconstructions develop a sharp peak, then decline. Ours does not need one: **for 45% of the posterior weight c_s^2 is still rising at the stellar centre.** A turnover must exist — the data do not place it inside a star. A fixed ansatz does, by construction.

Stiff yet near-conformal: the hallmark of a gradual hadron–quark crossover.

benchmarks: the released GP and C4 ensembles of Annala+ '23 — same χEFT anchoring, comparable data; the difference is the prior. Independent Δ(TOV) determination overlaid: Marczenko+ '23.

Stars — first-order transition



filled: conditioning pulsars · open: withheld cross-checks

12.6 km
R(1.4 M⊙)

469
Λ(1.4 M⊙)

≈ 2.06 M⊙
M_{TOV}

The three conditioning pulsars pull in different directions; the posterior settles between them, within 1.2σ of each. Of three withheld NICER posteriors, two land on the band unprompted (the third, J0614–3329, prefers a smaller radius). The data move the prior radius by only half a kilometre (12.0 → 12.6 km): the result is a joint consequence of the χEFT-anchored prior and the observations.

A strong first-order transition? Its necessary signature — c_s^2 dropping below 0.1 above its peak — loses weight under the data: the probability falls 0.088 → 0.014,

$$\mathrm{BF}(\mathrm{PT\ vs.\ no\text{-}PT}) = \frac{\mathcal{P}^{\mathrm{post}}/(1-\mathcal{P}^{\mathrm{post}})}{\mathcal{P}^{\mathrm{prior}}/(1-\mathcal{P}^{\mathrm{prior}})} = 0.14$$

odds ≈ 7 : 1 against — substantial, not decisive; stable (0.04 – 0.24) across deliberately broadened priors, and shape-level questions like this survive reweighting because the raw samples do.

And back to the laboratory: the symmetry-energy slope L = 60.5 $^{+7.5}_{-7.6}$ (stat) ± 3.0 (sys) MeV — betwe the leading multimessenger and experiment-driven determinations.

comparisons: Annala+ '23 (GP, C4) · Legred+ '21 · Pang+ '21 · Koehn+ '25 · Rutherford+ '24 — independent transition analysis, different prior and criterion: Huth+ '22, BF ≈ 0.42

The proton's gravitational form factors

Gravity couples to matter through the energy–momentum tensor. Its nucleon matrix element defines three form factors — the gravitational analogues of charge and magnetic form factors:

A(t)
momentum — how mass is distributed

J(t)
angular momentum — how spin is composed

D(t)
internal forces — pressure and shear inside the proton

$$A(0) = 1\,, \qquad J(0) = \tfrac{1}{2}$$ — *exact, from the Poincaré algebra: conditioning data for free*

D(0) — the D-term — is fixed by no symmetry: a global property of the nucleon on a par with its mass and charge, and still unknown. It must be extrapolated from data that stop short of t = 0.

lattice QCD	full t-range, but unphysical pion masses — and noisy exactly where D(t) matters, near t = 0
ChPT	controlled at low |t|, but with unknown low-energy constants c_8, c_9
experiment (DVCS)	indirect and model-dependent

Extrapolating sparse, noisy points to t = 0 through a fitted ansatz is precisely the problem this machinery was built for.

Kobzarev & Okun '62 · Pagels '66 · Ji '97 · Polyakov '03 — lattice input below: Hackett, Pefkou & Shanahan, PRL 132, 251904 (2024)

Reconstruction at the lattice pion mass

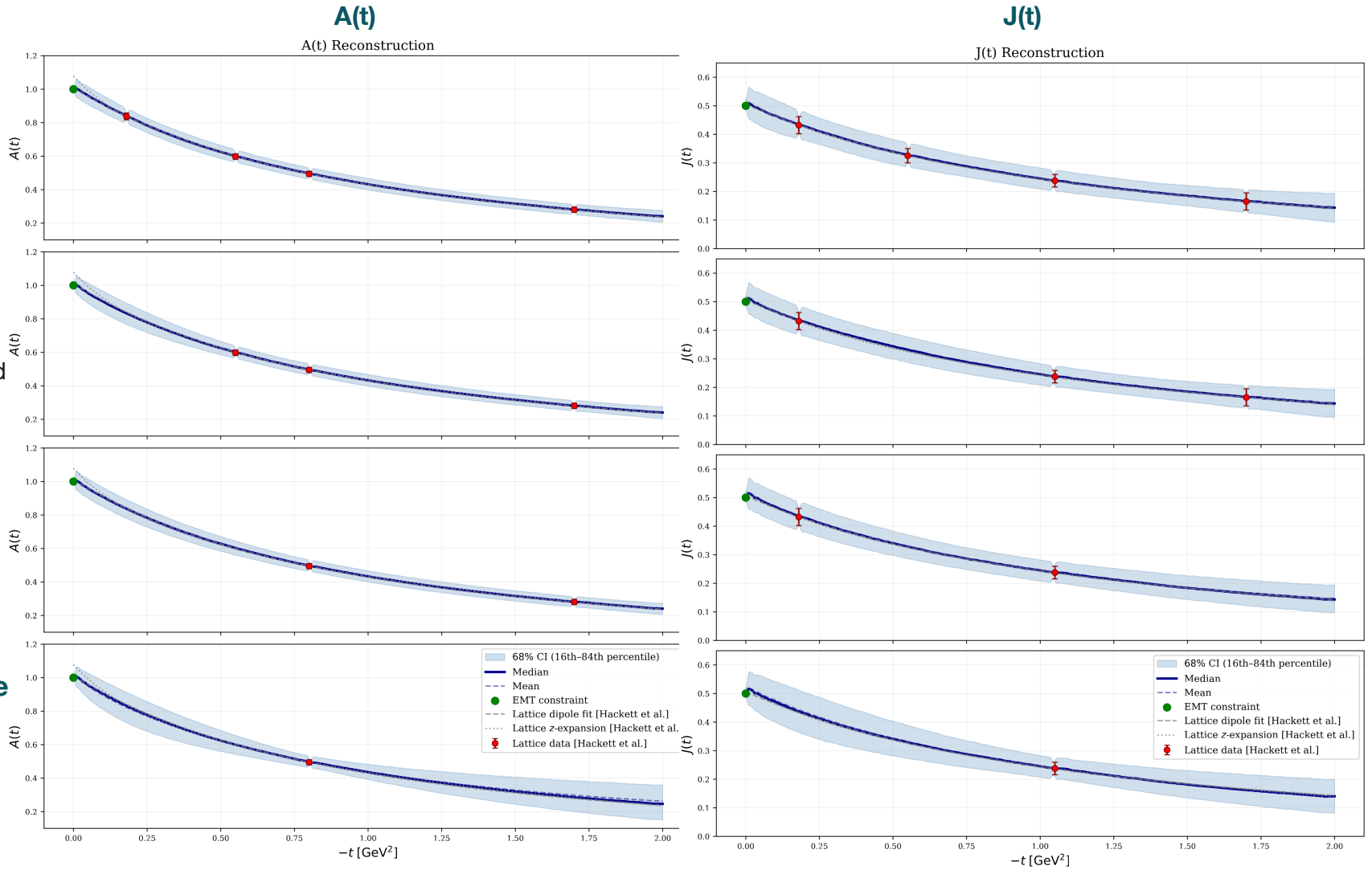
Conditioning: a handful of the most precise lattice points (red) + the exact values at t = 0 (green).

The ablation protocol: remove lattice points row by row, down to a single datum plus the constraint.

— the median barely moves, and keeps tracking the dipole and z-expansion fits it never saw;

— the band widens exactly where data were removed: a data-driven extrapolation uncertainty, where a fit hides the same uncertainty inside its ansatz.

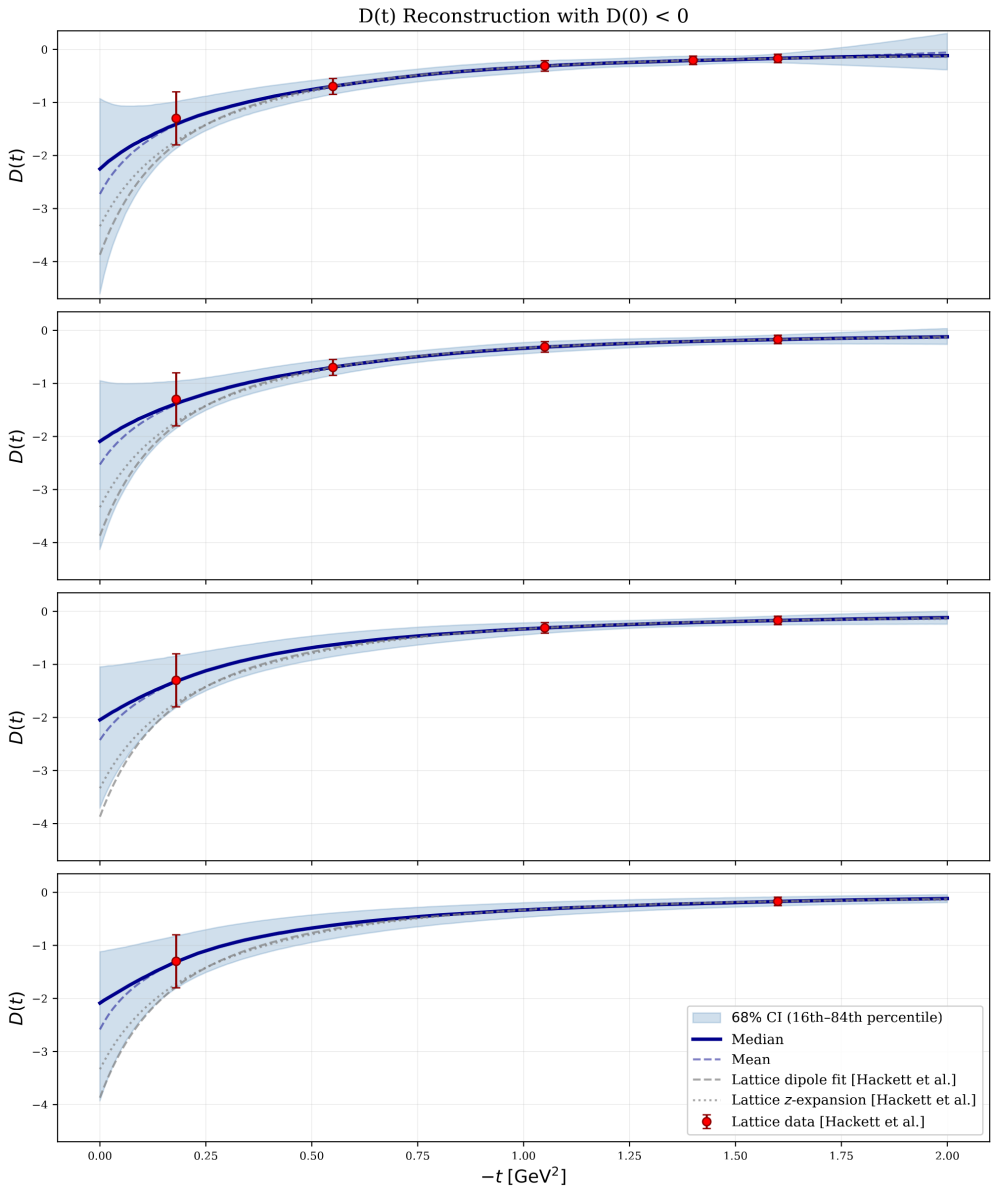
One point and one symmetry already pin the curve



lattice data, dipole and z-expansion fits: Hackett, Pefkou & Shanahan '24 — conditioning points chosen among the most precise, so the test is not aided by large error bars.

The D-term: no symmetry to lean on

D(t), sign-constrained: D(0) < 0



No sum rule fixes D(0), and the low-|t| lattice points are the noisiest — the ansatz-sensitivity problem in its purest form.

Physics enters as a sign. Mechanical stability of the proton demands D(0) < 0 — imposed by filtering the finished ensemble: reweighting with an indicator likelihood, the generation untouched.

The prior already disfavours zero crossings: requiring D(0) < 0 alone keeps the median and mean of D(t) negative across the whole grid; enforcing D(t) < 0 everywhere changes almost nothing — a non-trivial statement about the learned shape space.

Where the leverage is: one precise point at $-t \lesssim 0.25$ GeV² anchors normalisation and slope; one more at intermediate −t stabilises the interpolation. High-|t| points add little — concrete guidance for where future lattice effort pays off.

Four conditioning scenarios, one stable forward region: the quantity of physical interest survives the ablation.

sign: Polyakov '03 (stability), supported by ChPT, chiral quark–soliton models, and the data — lattice: Hackett, Pefkou & Shanahan '24

GFFs on the physical pion mass

The output is a continuous curve with full covariance — so ChPT can be matched directly to the posterior (generalised least squares at low |t|), with no parametric fit in between:

$c_9 = -0.61 \pm 0.19$ GeV^{-1}

from A(t) at O(p^4); J(t) cross-check agrees

$c_8 = -4.6 \pm 0.8$ GeV^{-1}

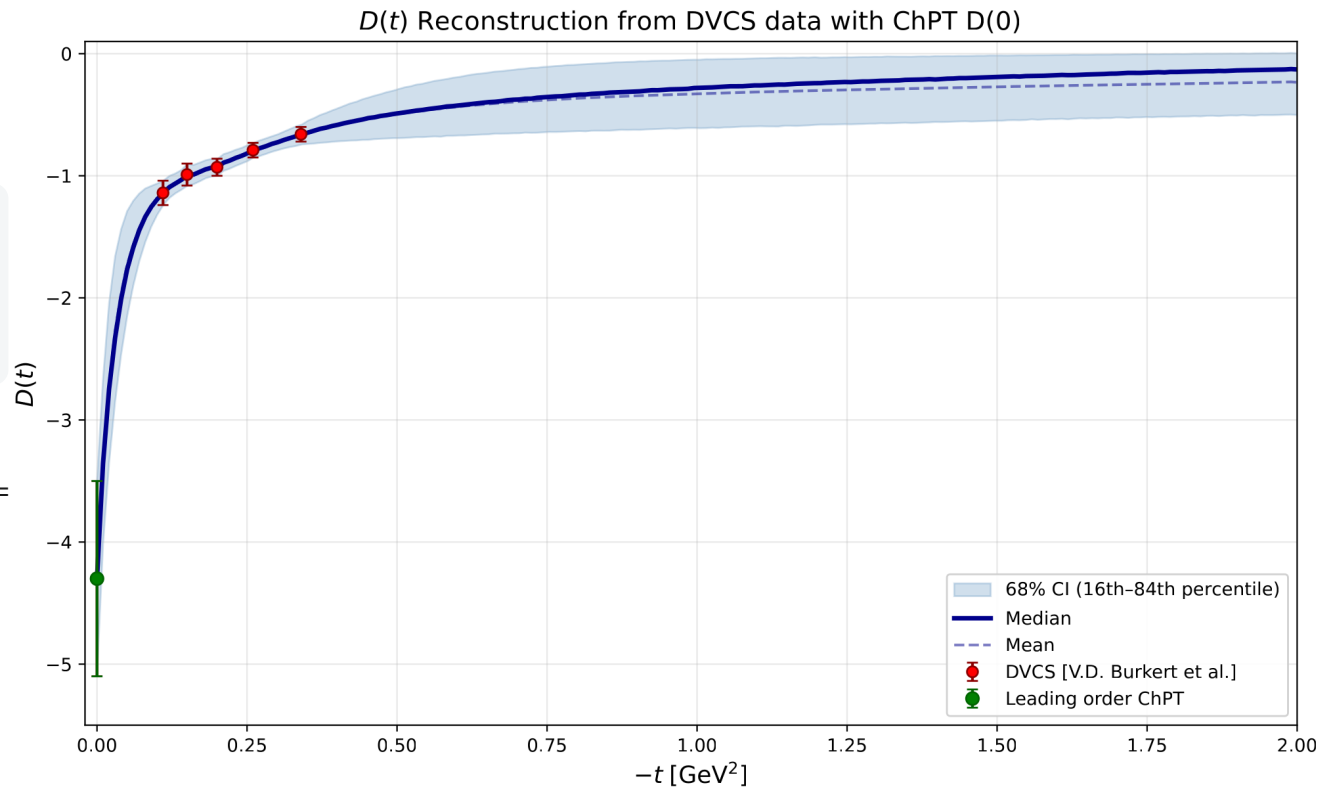
from D(t) at O(p^2)

An entirely different route agrees: the dispersive analysis of Cao et al. — ππ/KK̄ spectral functions from Roy–Steiner solutions, at the physical pion mass — gives $c_8 = -4.28(38)$, $c_9 = -0.68(6)$ GeV^{-1}.

Evaluating the chiral expressions at the physical pion mass (139 MeV) then yields the D-term of the physical proton:

$$D(0) = c_8\, m_N = -4.3 \pm 0.8$$

consistent with the lattice fits (−3.87 ± 0.97 dipole, −3.35 ± 0.58 z-exp.), the dispersive value $-3.38\ ^{+0.34}_{-0.35}$ — and comfortably inside the ChPT bound $D(0)/m_N \le -1.1(1)$ GeV^{-1}.



LECs: GLS over 5 points, $-t \in [0.02, 0.10]$ GeV² (c_9) and 4 points, $-t \in [0.02, 0.08]$ GeV² (c_8); stat ⊕ ChPT-truncation ⊕ fit-range in quadrature — dispersive: Cao, Yao, Zou '24