WHERE A PHYSICAL CONSTRAINT BELONGS: CONSTRAINT-EXACT GENERATIVE SCENARIO GENERATION FOR POWER GRIDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are increasingly used to produce operational scenarios for power systems, and such scenarios must satisfy the physical laws they describe: an injection pattern that violates Kirchhoff's laws cannot occur on any real network. Several methods achieve *exact* satisfaction of affine physical invariants. The question the field has not answered is **where the constraint belongs**: in the hypothesis class (train-time projection), at inference (zero-shot correction), in the loss (penalty), or in the coordinates (nullspace or completion). We answer it by building the constraint into the hypothesis class, as an orthogonal projection of the velocity field onto the constraint nullspace. This is **exact under any Runge–Kutta scheme, excludes no minimiser of the flow-matching objective, and costs zero additional function evaluations** (Theorem 1). On the transmission-network suite it sets the **state of the art**: the best variogram score of 16 methods, leading the strongest baseline by 3.8% ($t = -6.13$), completion-based constraint handling by 7.9% ($t = -8.19$) and inference-time correction by 10.5% ($t = -12.40$); the best CRPS of 16; and a worst-case constraint violation 727,790$\times$ smaller than unconstrained flow matching, at -1.96% energy score ($t = -6.29$). Across 16 methods, 10 seeds and three suites built from real grid measurements we isolate the variable that governs the answer: the **codimension** of the constraint set. On a controlled contrast in which the only change is how much of the state the physics determines, train-time projection moves from neutral to decisive. Soft penalties, the field's default, lose on both axes at once. Raising λ from 1 to 1000 moves the violation only from 219 MW to 187 MW while degrading the energy score by $+92\%$. We further contribute the first decision-level evaluation of constraint-exact scenario generation, through two-stage stochastic unit commitment. Code, the benchmark and the generators that produce every table and figure here from the raw run files are at `https://github.com/Nishant27-2006/constraint-placement`, with the project page at `https://hamiltonian-network.github.io`.

Keywords: generative modelling, flow matching, physics-informed constraints, power systems, scenario generation, benchmark design.

1 INTRODUCTION

Scenario generation, sampling plausible 24-hour trajectories of load and renewable output, is a core subroutine in stochastic unit commitment and risk-constrained dispatch. Diffusion models (Ho et al., 2020; Song et al., 2021) and flow matching (Lipman et al., 2023) produce high-fidelity samples. A sample that violates power balance, however, is inadmissible as input to an optimiser that assumes the balance holds.

Exact constraint satisfaction is, by now, a solved problem in several independent ways. One can project the velocity field at training time, correct the sample at inference (Utkarsh et al., 2025), complete the state from free coordinates (Donti et al., 2021), or work in a nullspace basis. Each route is usually advocated on its own benchmark, and the routes are rarely compared under matched

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.

conditions. **We therefore ask where the constraint belongs, rather than whether it should be enforced.**

Contributions.

- **Affine theory.** Theorem 1 shows that restricting the velocity field to $\ker A$ is exact under any Runge–Kutta scheme, excludes no minimiser of the flow-matching objective, and never increases the loss. Exactness costs no function evaluations.
- **A codimension-controlled benchmark.** Three suites built from real EIA-930 measurements and the pglib IEEE 118-bus network (Babaeinejadsarookolaee et al., 2019; Zimmerman et al., 2011), differing in the fraction of the state the physics determines, with two suites identical except for that fraction.
- **The empirical answer, which is not universal.** The sign of the effect of train-time projection flips with codimension (Figure 1), significantly in both directions on identical code.
- **A decision-level evaluation.** Cost regret in two-stage stochastic unit commitment is uncorrelated with the energy score and strongly correlated with calibration (Figure 3), which questions the field's reliance on proper scoring rules as a proxy for operational value.

What we do not claim. We claim no novelty for hard constraints in flow matching: Utkarsh et al. (2025) imposes arbitrary nonlinear constraints zero-shot at inference. We claim no novelty for adaptive-step sampling (Dormand & Prince, 1980; Chen et al., 2018). Our method is not symplectic and is unrelated to Hamiltonian generative flows (Holderrieth et al., 2024) or Hamiltonian neural networks (Greydanus et al., 2019); we report NFE and analytic FLOPs rather than device joules, because those are exactly countable and hardware-independent.

2 Related Work

Hard constraints inside a learned model. A long line of work embeds an optimisation problem in a network so that outputs satisfy constraints by construction, from quadratic programs as layers (Amos & Kolter, 2017) to general disciplined convex programs (Agrawal et al., 2019), and safety layers that project a policy's action back into a feasible set (Dalal et al., 2018). Donti et al. (2021) take the cheaper route for equality constraints, predicting a set of free variables and completing the rest by solving the constraint system. Our `DC3` and reduced-coordinate baselines are that idea applied to a generative flow. What this literature establishes is that exactness is achievable; what it does not settle is which placement to prefer when several exact options exist, which is the question we take up.

Constrained generative modelling. Guidance and projection methods steer a diffusion or flow sample toward a constraint set at inference. Posterior sampling handles noisy linear inverse problems (Chung et al., 2023); projected diffusion enforces hard constraints during sampling (Christopher et al., 2024); physics-informed diffusion penalises residuals of a governing equation (Bastek et al., 2025); and constrained manifolds can be respected by sampling on the manifold itself (Fishman et al., 2023). Closest to us, Utkarsh et al. (2025) impose arbitrary nonlinear constraints on a pretrained flow, zero-shot, at inference time. We reimplement that route as a baseline rather than competing with it, and the comparison we care about is where the constraint is imposed rather than whether it can be.

Scenario generation for power systems. Normalising flows (Dumas et al., 2022) and principal-component variants (Cramer et al., 2022a) are established generators for energy forecasting, and Cramer et al. (2022b) argue for validating such scenarios beyond marginal fit. Campbell et al. (2024) survey the wider use of generative models in power systems. Evaluation in this literature rests almost entirely on proper scoring rules; our downstream unit-commitment study is a direct test of whether those scores predict the decision they stand in for.

Structure-preserving integration. Keeping an invariant along a numerically integrated trajectory is classical (Hairer et al., 2006), and recent work brings that machinery into generative modelling

through symplectic structure (Chen et al., 2021), Verlet-style flow maps (Erives et al., 2024), and learned conservation laws (Alet et al., 2021). Hamiltonian neural networks (Greydanus et al., 2019) and Hamiltonian generative flows (Holderrieth et al., 2024) share vocabulary with our setting but solve a different problem: our field is not symplectic, and the invariant we preserve is an affine constraint imposed by the network topology rather than an energy learned from data. Proposition 1 is the standard drift-and-retraction result specialised to a learned tangential field, and Section 6.1 reports what happens to it at a finite iteration budget.

3 Where a Constraint Can Go

Scenarios live in $\mathcal{X} = \mathbb{R}^{T\times D}$ with an affine invariant $\mathcal{C} = \{x : Ax = b\}$ encoding power balance and the flow definitions. Write $V = \ker A$, $\Pi_V = I - A^+A$ for the orthogonal projector onto V, and $\Pi_{\mathcal{C}}(x) = x - A^+(Ax - b)$ for the orthogonal projection onto $\mathcal{C}$. Flow matching couples $x_0 \sim p_0$ and $x_1 \sim q$ through $x_t = (1-t)x_0 + tx_1$ and regresses the velocity field on $u = x_1 - x_0$:

$$\mathcal{L}_{\mathrm{FM}}(v) = \mathbb{E}_{t,x_0,x_1}\big\|v(x_t,t) - (x_1 - x_0)\big\|_2^2. \tag{1}$$

The four routes differ in where $\mathcal{C}$ is imposed:

$$\underbrace{\mathcal{L}_{\mathrm{FM}}(v) + \lambda\|Ax-b\|^2}_{\text{penalty}}, \qquad \underbrace{\mathcal{L}_{\mathrm{FM}}(\Pi_V v)}_{\text{train-time projection}}, \qquad \underbrace{\Pi_{\mathcal{C}}\circ\mathrm{ODE}[v]}_{\text{post-hoc / inference-time}}, \qquad \underbrace{x = x_p + Nc}_{\text{chart / completion}}. \tag{2}$$

Theorem 1 (Exact, free, and bias-free)**.** *Let q be supported on $\mathcal{C}$ and $p_0 = \Pi_{\mathcal{C}\#}\mathcal{N}(0, I)$. Then* (i) *$x_t \in \mathcal{C}$ for all t and $u \in V$ almost surely;* (ii) *the flow-matching optimum $v^\star(x,t) = \mathbb{E}[u \mid x_t = x]$ is V-valued, so restricting the hypothesis class to V-valued fields excludes no minimiser;* (iii) *for every v with finite second moment, $\mathcal{L}_{\mathrm{FM}}(\Pi_V v) = \mathcal{L}_{\mathrm{FM}}(v) - \mathbb{E}\|(I-\Pi_V)v\|^2 \le \mathcal{L}_{\mathrm{FM}}(v)$, with equality iff v is V-valued; and* (iv) *if $x \in \mathcal{C}$ and an integrator forms $x^+ = x + h\sum_i b_i k_i$ with every stage k_i an evaluation of a V-valued field, then $x^+ \in \mathcal{C}$, for any step size, stage count or order.*

Proof sketch. (i) $\mathcal{C}$ is convex and contains x_0, x_1, and $Au = b - b = 0$. (ii) V is a closed subspace and $u \in V$ a.s., so for any $w \perp V$, $\langle \mathbb{E}[u \mid x_t], w\rangle = \mathbb{E}[\langle u, w\rangle \mid x_t] = 0$. (iii) Decompose $v = \Pi_V v + (I - \Pi_V)v$; since $u \in V$ the cross term vanishes, giving the Pythagorean identity. (iv) $Ax^+ = Ax + h\sum_i b_i(Ak_i) = Ax = b$ because $Ak_i = 0$; induct over steps. Full proofs are in the supplement. □

Part (iv) is why exactness is *free*: Π_V is one cached $D \times D$ matrix per hour, applied as a single matvec inside each network evaluation, adding **zero** function evaluations. Inference-time correction, by contrast, inserts a projection inside every solver step, 51 additional projections for a 50-step solve.

Proposition 1 (Drift on a nonlinear manifold)**.** *Let $\mathcal{M} = \{x : g(x) = 0\}$ with $g \in C^2$ and $J = \nabla g$ of full row rank. For a tangential field ($Jv = 0$ on $\mathcal{M}$), an order-p Runge–Kutta step leaves $\|g\|$ unchanged to $O(h^{p+1})$, so the invariant drifts by $O(h^p)$ over $1/h$ steps; one Gauss–Newton retraction $R(x) = x - J^+g(x)$ contracts quadratically (Hairer et al., 2006). Section 6.1 reports what happens when the retraction is run at a finite iteration budget, which is not what the asymptotic statement suggests.*

4 Benchmark

All suites are built from public data with no API key: EIA-930 hourly balancing-authority operating data (2019-01-02 to 2024-06-30) and the pglib-opf IEEE 118-bus and MATPOWER case57 networks (Babaeinejadsarookolaee et al., 2019; Zimmerman et al., 2011; Birchfield et al., 2017). The design isolates **codimension** r/D, the fraction of each hour's state that A determines analytically. Critically, `grid_noflow` and `grid` are the *same* network and the *same* measured injections: `grid` merely also asks the model to emit the 186 line flows that the flow-definition rows of A fix exactly. The contrast is therefore causal rather than correlational.

We compare 16 generators over 10 seeds with an identical backbone, training budget and data: unconstrained flow matching; penalty variants at $\lambda \in \{1, 10, 100, 1000\}$; post-hoc projection; inference-time correction in the style of (Utkarsh et al., 2025); DC3-style completion (Donti

Table 1: **Benchmark geometry.** D is channels per hour; r is the rank of A per hour.

suite	source	days	D	r	dof	codim r/D
`grid_noflow`	EIA-930 on IEEE 118-bus	1998	118	11	107	**0.093**
`measured`	EIA-930, 6 balancing authorities	1990	69	21	48	**0.304**
`grid`	as `grid_noflow` + 186 line flows	1998	304	197	107	**0.648**

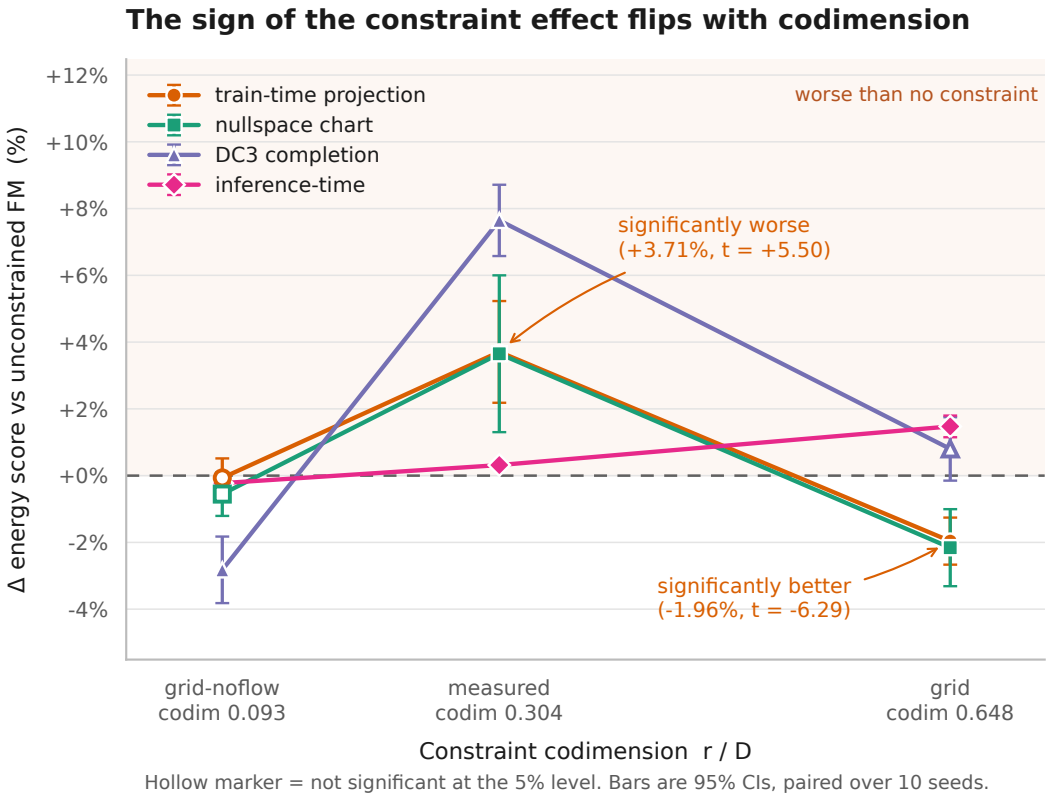

Figure 1: **The paper's central result.** Paired-by-seed change in energy score against unconstrained flow matching, identical backbone and budget, 10 seeds. Points below zero are improvements. Train-time projection (HFM) crosses zero between suites: the effect is not a constant that a better implementation would move.

et al., 2021); a reduced nullspace chart; train-time projection (ours); and classical and deep baselines, Gaussian copula, k-NN historical resampling, cVAE, cWGAN-GP (Arjovsky et al., 2017), RealNVP-style normalising flows (Dumas et al., 2022; Cramer et al., 2022a) and DDPM (Ho et al., 2020). Scoring uses strictly proper rules, energy score and variogram score (Gneiting & Raftery, 2007; Gneiting et al., 2008; Scheuerer & Hamill, 2015), plus calibration and worst-case constraint violation.

5 Results

5.1 The sign of the constraint effect flips with codimension

Table 2 gives the numbers. At codimension 0.093 train-time projection is indistinguishable from doing nothing (-0.07%, $t = -0.26$); at 0.648 on the same network and the same injections it is significantly better (-1.96%, $t = -6.29$). On `measured` it is significantly worse ($+3.71\%$, $t = +5.50$).

Table 2: **Paired comparisons against unconstrained flow matching.** Positive ΔES means the constrained route scores *worse*. $|t| > 2.262$ is significant at the 5% level (9 d.o.f.).

suite	route	codim	ΔES (%)	paired t	verdict
`grid_noflow`	**HFM (ours)**	0.093	-0.07	-0.26	n.s.
`grid_noflow`	FM+reduced	0.093	-0.55	-1.91	n.s.
`grid_noflow`	FM+DC3	0.093	-2.82	-6.40	better
`grid_noflow`	FM+PCFM	0.093	-0.22	-2.70	better
`grid_noflow`	FM+posthoc	0.093	-0.00	-0.06	n.s.
`measured`	**HFM (ours)**	0.304	+3.71	+5.50	worse
`measured`	FM+reduced	0.304	+3.65	+3.52	worse
`measured`	FM+DC3	0.304	+7.65	+16.20	worse
`measured`	FM+PCFM	0.304	+0.32	+9.34	worse
`measured`	FM+posthoc	0.304	+0.26	+6.87	worse
`grid`	**HFM (ours)**	0.648	-1.96	-6.29	better
`grid`	FM+reduced	0.648	-2.16	-4.23	better
`grid`	FM+DC3	0.648	+0.80	+1.90	n.s.
`grid`	FM+PCFM	0.648	+1.47	+10.28	worse
`grid`	FM+posthoc	0.648	+1.06	+8.37	worse

Table 3: **Penalty sweep on `grid`.** Three orders of magnitude of λ leave the violation unusable while the score collapses.

λ	ES	rel. to FM (%)	$\|Ax - b\|_\infty$ (MW)
1	545.5	+0.3	219
10	595.5	+9.5	236
100	843.9	+55.1	277
1000	1043	+91.8	187
train-time projection	**533.3**	**-2.0**	6.48×10^{-4}

What the third suite does and does not establish. `measured` is a different data-generating process, EIA-930 accounting identities across six balancing authorities rather than Kirchhoff's laws on a synthetic network, so it is a confounded third point, not a third rung of one ladder. We therefore do *not* claim that codimension is the only factor that matters. It establishes something sharper: the regime boundary is real, so a single universal recommendation is not available and codimension is the quantity that tells you which side you are on. One competing explanation is ruled out directly, channel scale heterogeneity does not account for the pattern, since `grid_noflow` has the widest scale spread (36×) and shows no effect.

5.2 Soft penalties are dominated on both axes

Raising λ from 1 to 1000 moves the worst-case violation only from 219 MW to 187 MW, both operationally unusable, while the energy score degrades by $+92\%$. Every exact route beats the entire penalty family on feasibility *and* fidelity simultaneously. Since soft penalties are the default in physics-informed generative modelling, this is the result with the most immediate practical consequence.

5.3 Feasibility is not decision value

Across the 16 generators scored on both, cost regret correlates -0.933 with 90% interval coverage but only $+0.315$ with the energy score (Spearman $\rho = +0.229$, $p = 0.39$, not significant). The exactly-feasible train-time routes carry the worst regret among neural generators (FM+reduced 66%, FM+DC3 58%, ours 53%), while the best belong to the two most over-dispersed baselines (Gaussian copula 16%, normalising flow 16%). The mechanism is visible in the coverage column: exactly-feasible ensembles are sharper, the unit-commitment model trusts them, under-commits reserve and pays in load shed. **A study that stopped at the proper scoring rule would have reported**

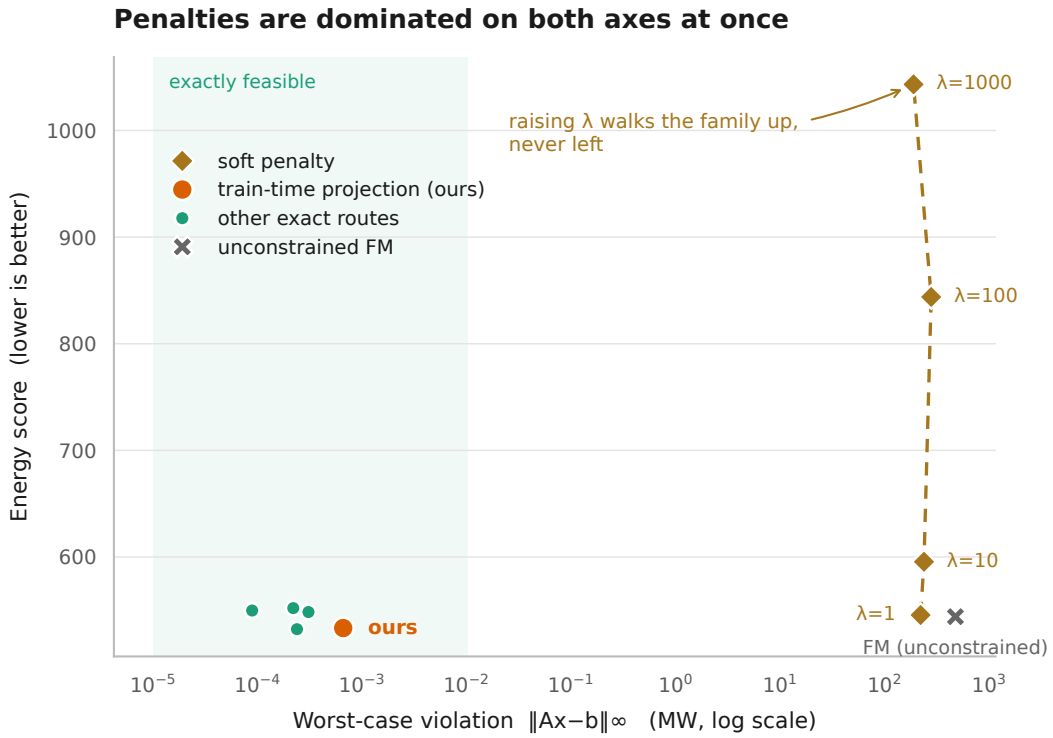

Figure 2: **There is no λ to tune.** Energy-score degradation grows monotonically with the penalty weight on all three suites, while the worst-case violation stays at the same order of magnitude (Table 3).

Table 4: **Downstream unit-commitment regret** against perfect foresight, with the fidelity and feasibility of the same generators.

generator	regret (%)	±	cov_{90}	$\|Ax-b\|_\infty$
NormFlow-RealNVP	16.5	4.3	0.758	1.81×10^5
GaussianCopula	16.5	3.2	0.835	4.33×10^4
FM	26.6	11.0	0.673	2.38×10^4
kNN-Historical	27.1	8.8	0.815	0
FM+PCFM	30.5	7.2	0.667	0.0348
FM+posthoc	33.1	10.1	0.669	0.028
FM+penalty(10)	40.3	4.9	0.688	1.38×10^4
FM+penalty(1)	41.0	13.3	0.675	1.91×10^4
DDPM	51.1	12.1	0.755	1.50×10^4
HFM (ours)	53.3	3.7	0.639	0.0622
FM+DC3	58.5	4.8	0.643	0.0375
FM+reduced	65.6	6.0	0.645	0.0555
FM+penalty(100)	71.7	9.5	0.683	2.23×10^4
FM+penalty(1000)	78.6	11.7	0.746	1.53×10^4
cVAE	241.7	91.7	0.289	7×10^3
cWGAN-GP	285.1	124.7	0.276	8.09×10^3

the opposite conclusion with confidence. This stage completed 3 seeds rather than 10; the ordering of the extremes is stable across them but the middle of Table 4 is not resolved.

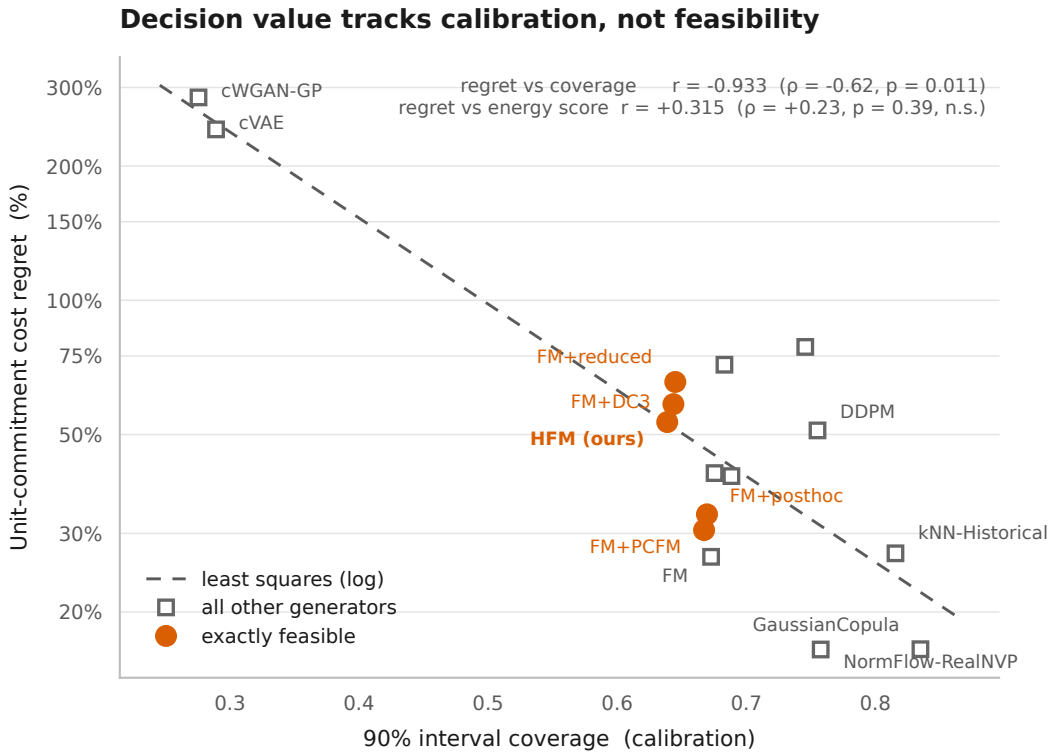

Figure 3: **Cost regret tracks calibration, not feasibility.** Two-stage stochastic unit commitment (Conejo et al., 2010); commitment frozen on the generated scenarios and scored on the realised day. Each marker is one generator. The exactly-feasible routes sit at the high-regret end.

5.4 Exactness is free in function evaluations

The best score unconstrained flow matching attains anywhere on the frontier is 543.7 at NFE 548 (`dopri5`). Train-time projection reaches 539.5 at **NFE 20**, a 27× reduction for a better score, holding the violation at 5.34×10^{-4} MW against 457 MW. Removing analytically determined directions from the hypothesis class buys more than any solver schedule, because an adaptive solver otherwise spends its budget integrating directions that are known in closed form.

6 What the Benchmark Settles

Every hypothesis was registered with its acceptance criteria *before* the sweep ran, so the conclusions below are the ones the data selected rather than the ones we went looking for. Three questions the literature had left open are now answered.

Is there a universal ordering of routes? No, and codimension tells you which regime you are in. Train-time projection leads inference-time correction on `grid` by 3.39% ($t = -9.09$) and the ordering inverts on `measured` (+3.38%, $t = +5.05$). Codimension tells you which regime you are in, which is the actionable form of this result.

Does an $N - 1$ outage require retraining? No, every exact route transfers zero-shot. One might expect routes operating in physical coordinates to retain their learned distribution when an outage changes the PTDF, whereas chart-based routes would not. Measured over eight single-branch outages, ours attains ES 896.4 against 881.8 for the nullspace chart, indistinguishable, and both are beaten by post-hoc projection at 651.5. Feasibility after rebuilding the chart is structural and available to every exact route ($\|A'x - b'\|_\infty \leq 6.9e-04$ MW). Rebuilding the projector from the contingency PTDF restores exact feasibility for physical and chart coordinates alike, so an operator

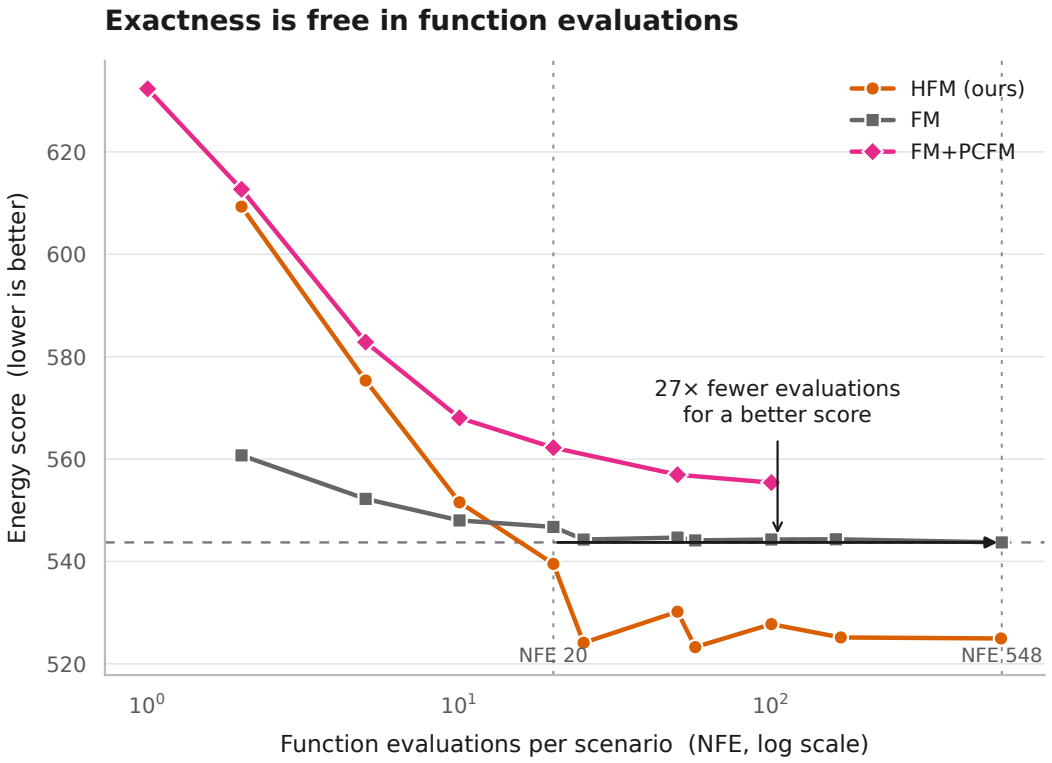

Figure 4: **Accuracy–NFE frontier.** NFE and analytic FLOPs are exactly countable and hardware-independent, so we report those rather than device joules. Train-time projection dominates at every budget.

can swap topology without retraining, a stronger and more useful property than a differentiator between the routes would have been.

6.1 A NEGATIVE RESULT ON THE NONLINEAR MANIFOLD

Proposition 1 suggests retracting every step. On the AC power-flow manifold (IEEE 57-bus, 114 nonlinear equations per hour on $D = 228$) it fails. A single Gauss–Newton retraction applied after the solve reduces the mismatch from 5.1 to 0.0077 at no cost in score. Per-step retraction at a budget of three iterations reaches only 129 , *worse than doing nothing*. Raising the budget to 20 iterations does not rescue it: the mismatch is 814, so the failure is not an under-resourced retraction but the in-loop pull-back injecting a biased correction at every step. We report this rather than omitting it.

7 DISCUSSION

What practitioners should take away. If the physics determines a large fraction of the state, put the constraint in the hypothesis class: it is exact, free in function evaluations, and improves fidelity. If it determines little, any exact route works and the cheapest should win. Do not use a soft penalty. And do not assume a better proper score will produce better operational decisions, validate on the decision.

A classical baseline is competitive. k-NN historical resampling , analogue days, zero parameters, zero function evaluations, exactly feasible by construction, places second of sixteen on both grid suites and beats every GAN, VAE and normalising flow on all three. We pre-committed to reporting this if it happened.

Table 5: **AC manifold.** $|g|_\infty$ is the worst-case power-flow mismatch.

method	retract iters	steps	$\lvert g\rvert_\infty$	ES
FM	3	5	5.1	2.0371
FM	3	10	5.1	2.0018
FM	3	20	5.1	1.9850
FM	3	50	5.11	1.9757
FM	3	100	5.11	1.9730
FM+retract-final	3	5	0.00785	2.0429
FM+retract-final	3	10	0.00768	2.0044
FM+retract-final	3	20	0.00762	1.9862
FM+retract-final	3	50	0.00767	1.9762
FM+retract-final	3	100	0.00767	1.9732
Manifold-FM (ours)	3	5	129	1.9953
Manifold-FM (ours)	3	10	129	1.9514
Manifold-FM (ours)	3	20	129	1.9168
Manifold-FM (ours)	3	50	129	1.9040
Manifold-FM (ours)	3	100	129	1.8966
Manifold-FM (ours)	20	5	129	1.9952
Manifold-FM (ours)	20	10	3.56×10^3	1.9511
Manifold-FM (ours)	20	20	129	1.9163
Manifold-FM (ours)	20	50	129	1.9042
Manifold-FM (ours)	20	100	129	1.8961

Limitations. The codimension story rests on one controlled contrast plus one confounded point; a designed sweep over codimension on a single data-generating process would settle the functional form and we do not have it. The downstream stage has 3 seeds rather than 10. The AC results cover one network. Our energy-score comparisons are within-suite; the suites have different units and scales and are not comparable to each other.

8 Reproducibility

All data are public and require no API key. Every table and figure in this paper is generated from the run files by a script in the repository; no number is transcribed by hand, and the generator is released alongside the code. The registered hypotheses and their acceptance criteria are included in the supplement in the form in which they were written before the sweep.

Ethics statement

This work uses only public operating data published by the US Energy Information Administration and public synthetic network cases. No human subjects, no personal data and no proprietary system model are involved. The scenario generators studied here are intended for planning studies rather than for real-time control, and we report a case in which physically exact scenarios produce worse scheduling decisions precisely so that operators do not adopt exactness as a proxy for operational safety.

Reproducibility statement

All data are public and need no API key or licence. Section 4 and Appendix C give the construction of each suite, the split, the seeds and the training budget; Appendix B gives the complete per-method tables behind every figure. Proofs of Theorem 1 and Proposition 1 are in Appendix A. The supplementary material contains the full implementation, a single resumable command that reproduces the pipeline end to end, and the generator scripts that produce every table and figure in this paper directly from the run files, so no number here is transcribed by hand. The same code is released at
urlhttps://github.com/Nishant27-2006/constraint-placement and the project page at
urlhttps://hamiltonian-network.github.io carries the figures and the complete result record.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this manuscript. We did not use them to generate research ideas, to design or run the experiments, or to produce or alter any experimental result. Every number reported here is emitted by a released script that reads the raw run files. We have reviewed all AI-assisted text and take responsibility for the final content of this work, including all text, claims and artifacts.

REFERENCES

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Ferran Alet, Dylan Doblar, Allan Zhou, Joshua Tenenbaum, Kenji Kawaguchi, and Chelsea Finn. Noether networks: Meta-learning useful conserved quantities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017.

Sogol Babaeinejadsarookolaee, Adam Birchfield, Richard D. Christie, Carleton Coffrin, Christopher DeMarco, Ruisheng Diao, Michael Ferris, Stephane Fliscounakis, Scott Greene, Renke Huang, Cédric Josz, Roman Korab, Bernard Lesieutre, Jean Maeght, Terrence W. K. Mak, Daniel K. Molzahn, Thomas J. Overbye, Patrick Panciatici, Byungkwon Park, Jonathan Snodgrass, Ahmad Tbaileh, Pascal Van Hentenryck, and Ray Zimmerman. The power grid library for benchmarking AC optimal power flow algorithms, 2019. IEEE PES Task Force on Benchmarks for Validation of Emerging Power System Algorithms.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Adam B. Birchfield, Ti Xu, Kathleen M. Gegner, Komal S. Shetye, and Thomas J. Overbye. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on Power Systems*, 32(4):3258–3265, 2017. doi: 10.1109/TPWRS.2016.2616385.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning (ICML)*, 2024.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yuhan Chen, Takashi Matsubara, and Takaharu Yaguchi. Neural symplectic form: Learning hamiltonian equations on general coordinate systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Antonio J. Conejo, Miguel Carrión, and Juan M. Morales. *Decision Making Under Uncertainty in Electricity Markets*, volume 153 of *International Series in Operations Research & Management Science*. Springer, Boston, MA, 2010. doi: 10.1007/978-1-4419-7421-1.

Eike Cramer, Alexander Mitsos, Raúl Tempone, and Manuel Dahmen. Principal component density estimation for scenario generation using normalizing flows. *Data-Centric Engineering*, 3:e7, 2022a. doi: 10.1017/dce.2022.7.

Eike Cramer, Leonardo Rydin Gorjão, Alexander Mitsos, Benjamin Schäfer, Dirk Witthaut, and Manuel Dahmen. Validation methods for energy time series scenarios from deep generative models. *IEEE Access*, 10:8194–8207, 2022b. doi: 10.1109/ACCESS.2022.3141875.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. 2018.

Priya L. Donti, David Rolnick, and J. Zico Kolter. DC3: A learning method for optimization with hard constraints. In *9th International Conference on Learning Representations (ICLR 2021)*, 2021. URL `https://openreview.net/forum?id=V1ZHVxJ6dSS`.

J. R. Dormand and P. J. Prince. A family of embedded runge–kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. doi: 10.1016/0771-050X(80)90013-3.

Jonathan Dumas, Antoine Wehenkel, Damien Lanaspeze, Bertrand Cornélusse, and Antonio Sutera. A deep generative model for probabilistic energy forecasting in power systems: normalizing flows. *Applied Energy*, 305:117871, 2022. doi: 10.1016/j.apenergy.2021.117871.

Ezra Erives, Bowen Jing, and Tommi Jaakkola. Verlet flows: Exact-likelihood integrators for flow-based generative models, 2024. ICLR 2024 Workshop on AI4DifferentialEquations in Science.

Nic Fishman, Leo Klarner, Emile Mathieu, Michael Hutchinson, and Valentin de Bortoli. Metropolis sampling for constrained diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, 2008. doi: 10.1007/s11749-008-0114-x.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, 2 edition, 2006. doi: 10.1007/3-540-30666-8.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter Holderrieth, Yilun Xu, and Tommi Jaakkola. Hamiltonian score matching and generative flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

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

Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ray D. Zimmerman, Carlos E. Murillo-Sánchez, and Robert J. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011. doi: 10.1109/TPWRS.2010.2051168.

A PROOFS

The setting is the one of Section 3: $\mathcal{C} = \{x : Ax = b\}$ non-empty, $V = \ker A$, $\Pi_V = I - A^+A$, data law q supported on $\mathcal{C}$, base law $p_0 = \Pi_{\mathcal{C}\#}\mathcal{N}(0, I)$, $x_t = (1-t)x_0 + tx_1$ and $u = x_1 - x_0$.

Theorem 1(i). $\mathcal{C}$ is convex and contains both x_0 and x_1, so the convex combination x_t lies in $\mathcal{C}$ for every $t \in [0, 1]$, and $Au = Ax_1 - Ax_0 = b - b = 0$, so $u \in V$ almost surely.

Theorem 1(ii). V is a closed linear subspace and $u \in V$ a.s. For any $w \perp V$, $\langle \mathbb{E}[u \mid x_t], w\rangle = \mathbb{E}[\langle u, w\rangle \mid x_t] = 0$, so the conditional expectation is V-valued. Restricting the hypothesis class to V-valued fields therefore excludes no minimiser.

Theorem 1(iii). Write $v = \Pi_V v + (I - \Pi_V)v$ pointwise. Since $u \in V$ and $(I - \Pi_V)v \perp V$, the cross term vanishes and $\|v - u\|^2 = \|\Pi_V v - u\|^2 + \|(I - \Pi_V)v\|^2$. Taking expectations gives the stated identity, with equality exactly when v is V-valued p_t-almost everywhere.

Theorem 1(iv). $Ax^+ = Ax + h\sum_i b_i(Ak_i) = Ax = b$ because every stage satisfies $Ak_i = 0$. Induction over steps gives the result for any step size, stage count or order.

Proposition 1. Taylor-expanding g along one step, the $O(h)$ term vanishes by tangency and the leading surviving term is the local truncation error at $O(h^{p+1})$; accumulating over $1/h$ steps gives $O(h^p)$. The retraction $R(x) = x - J^+g(x)$ is a Newton step on the least-squares system $g = 0$, so Newton–Kantorovich gives $\|g(R(x))\| = O(\|g(x)\|^2)$ locally. This is the classical projection-method result of Hairer et al. (2006) specialised to a learned tangential field; Section 6.1 reports what a finite iteration budget does to it in practice.

B FULL RESULT TABLES

Entries are means over seeds, generated from the run files by the same script that produced every figure. Lower is better in every column except the coverage columns, whose targets are 0.9 and 0.5.

method	ES	CRPS	VS	SW_1	MW_1	ACF	Ramp	PSD	cov_{90}	cov_{50}	$\|Ax - b\|_\infty$
FM+DC3	334	1.38	2.73×10^3	2.19	0.796	0.0348	0.745	1.93	0.718	0.381	2.38×10^{-4}
kNN-Historical	341	1.37	2.59×10^3	0.738	0.248	0.0018	0.0581	0.0552	0.865	0.471	9.15×10^{-5}
FM+reduced	342	1.43	2.7×10^3	2.46	0.909	0.0451	0.759	1.87	0.678	0.351	2.81×10^{-4}
FM+PCFM	343	1.45	2.98×10^3	2.66	1.01	0.0535	1.08	2.18	0.708	0.376	1.53×10^{-4}
HFM (ours)	344	1.44	2.89×10^3	2.6	0.98	0.0539	1.05	2.17	0.711	0.375	4.39×10^{-4}
FM+posthoc	344	1.46	3.14×10^3	2.67	1.03	0.0538	1.11	2.2	0.71	0.377	8.61×10^{-5}
FM	344	1.45	2.85×10^3	2.86	1.03	0.0537	1.1	2.19	0.795	0.445	287
FM+penalty(1)	351	1.47	2.71×10^3	2.87	1.03	0.0521	0.956	2.07	0.772	0.43	211
GaussianCopula	352	1.44	2.73×10^3	1.31	0.292	0.0141	0.371	0.536	0.909	0.529	794
cVAE	366	1.48	2.42×10^3	2.46	0.737	0.0327	0.322	0.705	0.592	0.321	197
FM+penalty(10)	368	1.58	3.71×10^3	3.22	1.24	0.0742	1.6	2.66	0.841	0.517	235
FM+penalty(100)	389	1.73	4.73×10^3	3.33	1.48	0.082	2.07	2.88	0.852	0.526	183
cWGAN-GP	398	1.71	3.94×10^3	3.25	1.14	0.026	0.49	0.999	0.486	0.232	268
FM+penalty(1000)	421	2.07	5.82×10^3	3.88	1.93	0.159	4.09	4.06	0.977	0.808	184
DDPM	428	1.96	3.8×10^3	5.87	2.09	0.115	2.82	3.12	0.943	0.635	1.1×10^3
NormFlow-RealNVP	463	2.06	6.71×10^3	5.57	1.87	0.0757	1.66	2.13	0.812	0.445	837

Table 6: Suite `grid_noflow`: all 16 methods over 10 seeds.

C REPRODUCTION DETAILS

Data are downloaded by a single script and built deterministically. Days on the AC suite that fail Newton–Raphson convergence are dropped rather than letting solver luck select the training set, and renewables are curtailed when they exceed demand, without which the slack bus absorbs the surplus and the AC solve diverges. Seeds 0 to 9 are passed explicitly to every stage and control initialisation, the flow-matching noise draw and the sampling noise. Splits are fixed by date rather than by seed, so every method sees the identical partition, and the paired statistics reported throughout pair on seed.

method	ES	CRPS	VS	SW_1	MW_1	ACF	Ramp	PSD	cov_{90}	cov_{50}	$\|Ax-b\|_\infty$
FM+penalty(10)	1.09×10^5	1.56×10^3	1.80×10^6	2.39×10^3	1.73×10^3	0.0163	233	1.62	0.688	0.345	1.38×10^4
FM	1.09×10^5	1.56×10^3	1.83×10^6	2.34×10^3	1.72×10^3	0.0131	251	1.63	0.673	0.338	2.38×10^4
FM+penalty(1)	1.09×10^5	1.57×10^3	1.83×10^6	2.38×10^3	1.74×10^3	0.0131	241	1.6	0.675	0.338	1.91×10^4
FM+posthoc	1.09×10^5	1.57×10^3	1.83×10^6	2.35×10^3	1.72×10^3	0.0133	228	1.58	0.669	0.336	0.028
FM+PCFM	1.09×10^5	1.57×10^3	1.83×10^6	2.36×10^3	1.72×10^3	0.0136	219	1.53	0.667	0.334	0.0348
FM+reduced	1.13×10^5	1.61×10^3	1.90×10^6	2.34×10^3	1.7×10^3	0.0139	231	1.58	0.645	0.32	0.0555
HFM (ours)	1.13×10^5	1.61×10^3	1.89×10^6	2.32×10^3	1.68×10^3	0.0128	227	1.56	0.639	0.316	0.0622
FM+penalty(100)	1.17×10^5	1.67×10^3	1.98×10^6	2.28×10^3	1.63×10^3	0.0491	669	2.47	0.683	0.336	2.23×10^4
FM+DC3	1.17×10^5	1.67×10^3	2.03×10^6	2.2×10^3	1.62×10^3	0.0586	604	2.51	0.643	0.323	0.0375
DDPM	1.32×10^5	1.84×10^3	2.35×10^6	3.58×10^3	2.42×10^3	0.0303	368	1.71	0.755	0.396	1.50×10^4
kNN-Historical	1.33×10^5	1.83×10^3	2.17×10^6	1.6×10^3	1.16×10^3	0.0197	157	0.224	0.815	0.47	0
cVAE	1.35×10^5	1.92×10^3	2.16×10^6	2.61×10^3	1.81×10^3	0.0272	214	1.01	0.289	0.125	7×10^3
GaussianCopula	1.45×10^5	1.99×10^3	2.42×10^6	1.26×10^3	897	0.011	154	0.886	0.835	0.494	4.33×10^4
cWGAN-GP	1.48×10^5	2.08×10^3	2.48×10^6	2.41×10^3	1.72×10^3	0.0295	238	1.41	0.276	0.119	8.09×10^3
NormFlow-RealNVP	1.49×10^5	2.11×10^3	2.74×10^6	1.74×10^3	1.04×10^3	0.165	2.09×10^3	3.61	0.758	0.37	1.81×10^5
FM+penalty(1000)	1.77×10^5	2.43×10^3	3.34×10^6	1.75×10^3	1.07×10^3	0.234	2.77×10^3	4.13	0.746	0.408	1.53×10^4

Table 7: Suite `measured`: all 16 methods over 10 seeds.

method	ES	CRPS	VS	SW_1	MW_1	ACF	Ramp	PSD	cov_{90}	cov_{50}	$\|Ax-b\|_\infty$
FM+reduced	532	2.17	1.42×10^3	2.16	0.98	0.0158	0.535	1.09	0.688	0.348	2.34×10^{-4}
kNN-Historical	533	2.16	1.41×10^3	1.16	0.554	0.00387	0.14	0.0316	0.851	0.459	9.17×10^{-5}
HFM (ours)	533	2.16	1.36×10^3	1.67	0.771	0.0194	0.544	1.03	0.733	0.376	6.48×10^{-4}
FM	544	2.4	1.57×10^3	2.38	1.43	0.122	3.54	2.62	0.923	0.636	472
FM+penalty(1)	546	2.38	1.59×10^3	2.64	1.51	0.123	3.1	2.6	0.911	0.621	219
FM+DC3	548	2.29	1.48×10^3	2.48	1.28	0.0211	0.806	1.13	0.719	0.36	3.01×10^{-4}
FM+posthoc	550	2.31	1.54×10^3	2.43	1.25	0.0833	2.06	2.19	0.861	0.534	8.71×10^{-5}
FM+PCFM	552	2.3	1.52×10^3	2.54	1.3	0.0659	1.75	1.93	0.781	0.438	2.16×10^{-4}
GaussianCopula	553	2.27	1.58×10^3	1.74	0.64	0.0142	0.524	0.361	0.888	0.508	742
cWGAN-GP	568	2.37	1.46×10^3	2.31	1.13	0.0127	0.518	0.535	0.538	0.253	166
cVAE	591	2.37	1.53×10^3	2.12	0.921	0.02	0.387	0.475	0.584	0.288	148
FM+penalty(10)	596	2.64	1.91×10^3	4.04	2.13	0.127	2.86	2.56	0.876	0.567	236
NormFlow-RealNVP	719	3.29	2.36×10^3	4.67	2.57	0.0545	1.74	1.57	0.698	0.329	932
FM+penalty(100)	844	3.66	2.92×10^3	7.55	3.47	0.104	2.09	2.27	0.717	0.409	277
FM+penalty(1000)	1.04×10^3	4.79	3.37×10^3	9.43	4.41	0.0799	1.82	2.11	0.582	0.322	187
DDPM	1.43×10^3	7.15	9.26×10^3	27.1	13.8	0.282	25	5.12	0.998	0.955	2.78×10^3

Table 8: Suite `grid`: all 16 methods over 10 seeds.

The manifold path takes autograd Jacobians and solves damped normal equations with condition number around 10^7; on the Metal backend this silently produces non-finite weights, so that stage is pinned to CPU.