Supplementary Information for *From compression to discovery: representations whose structure reveals the governing equations*

This document collects the data-generation protocols, the architecture and loss details, the discovery and identification algorithms, and the supplementary results and ablations that support the main text. Cross references of the form S2 are internal to this document, while unprefixed references point to the main article. Throughout, "Phase A", "Phase B", and "Phase C" denote the three stages of the reduction pipeline, whereas "Stage 1" through "Stage 4" denote the four-stage weight schedule that runs inside Phase A and Phase C.

Contents

S1 Data generation and CFD setup

This section reports the high-fidelity simulations that generate the training data, together with the mesh and sampling settings. The three two-dimensional configurations are computed on meshes whose refinement follows the regions of strongest gradient; since the pipeline consumes each mesh purely as a graph, the element type plays no role downstream. A supplementary three-dimensional cylinder wake at Re = 300 tests the pipeline at higher dimension under a constrained computational budget. After statistical stationarity is confirmed, each two-dimensional case provides 8,000 snapshots, which are split chronologically in a 4:1 train-test ratio so that the test segment shares no temporal overlap with the training set. Unless stated otherwise, reconstruction errors are averaged over 2,000 randomly sampled snapshots; the latent trajectories used for spectral analysis and sparse identification are extracted on the held-out test segment, within which the final fifth serves as the validation split. Snapshots are saved every 0.02 s for the BFS, every 0.01 s for the cylinder pair, and at every solver step (2.5×10^{-3} s) for the airfoil, so that the held-out test segments span approximately 32, 16, and 4 s respectively.

Turbulent backward-facing step. The backward-facing step is computed as a two-dimensional unsteady Reynolds-averaged Navier-Stokes simulation with the realizable k-ε closure and scalable wall functions. The working fluid is air ($\rho = 1.225$ kg m^{-3}, $\mu = 1.789 \times 10^{-5}$ Pa s), and a uniform stream at $U_b = 0.5$ m/s enters through a velocity inlet of height equal to the step height $h = 0.2$ m, which gives $\mathrm{Re}_h = U_b h/\nu = 6{,}846$. The domain spans $7.5h \times 2h$, with the step located $1.5h$ downstream of the inlet and the outflow section extending $6h$ beyond it. The lower boundary, comprising the upstream floor, the step face, and the downstream floor, is a no-slip wall, while the upper and downstream boundaries are constant-pressure outlets, so that the upper boundary admits entrainment rather than confining the stream. The pressure-based segregated solver uses SIMPLE pressure-velocity coupling, second-order upwind discretization of the momentum and turbulence equations, and second-order pressure interpolation.

Cylinder-pair wake at $\mathrm{Re}_D = 1 \times 10^5$. The high-Reynolds-number case is the flow past a pair of identical circular cylinders of common diameter $D = 0.2$ m in side-by-side arrangement, separated by a gap of one diameter ($G/D = 1$, center-to-center spacing $T/D = 2$), and is computed as a two-dimensional unsteady Reynolds-averaged Navier–Stokes simulation with the SST k-ω closure. The working fluid is water with properties rounded to $\rho = 1000$ kg m^{-3} and $\mu = 10^{-3}$ Pa s, so that the uniform inlet velocity $U_\infty = 0.5$ m/s (turbulence intensity 5%, eddy-viscosity ratio 10) gives $\mathrm{Re}_D = U_\infty D/\nu = 1 \times 10^5$. The domain extends $4D$ upstream of the cylinder centers, $21D$ downstream, and $\pm 4D$ cross-stream; the lateral boundaries are no-slip walls, so that the pair is confined in a plane channel at a solid blockage of 25%, and the downstream boundary is a constant-pressure outlet. The pressure-based segregated solver uses SIMPLE pressure-velocity coupling, second-order upwind discretization of the momentum and turbulence equations, second-order pressure interpolation, and first-order implicit time integration.

Pitching NACA 0018 airfoil. The airfoil case is an unsteady Reynolds-averaged Navier–Stokes calculation with the k-ω SST closure and no transition model. For a NACA 0018 section of chord $c = 0.08$ m, the effective angle of attack is prescribed by oscillating the free-stream direction at the pressure-far-field boundaries as $\alpha(t) = \alpha_0 + \alpha_1 \sin(\omega t)$ with $\alpha_0 = 15°$, $\alpha_1 = 10°$, and $\omega = 5$ rad/s ($f_\alpha \approx 0.80$ Hz), which sweeps the incidence through $[5°, 25°]$. The free stream is ideal-gas air at $M = 0.074$ and 300 K ($U_\infty \approx 25.7$ m/s), giving a chord Reynolds number $\mathrm{Re}_c \approx 1.4 \times 10^5$. Time integration uses a pressure-based SIMPLEC scheme with second-order implicit stepping at $\Delta t = 2.5 \times 10^{-3}$ s.

Three-dimensional cylinder wake at $\mathrm{Re}_D = 300$**.** The supplementary three-dimensional case is computed under a constrained computational budget, which lowers reconstruction accuracy relative to the two-dimensional cases and leaves the identified dynamics less clean, as discussed in the main text. The wake of a single circular cylinder of diameter $D = 0.01$ m is computed as a three-dimensional unsteady laminar simulation with water at its default properties ($\rho = 998.2\ \mathrm{kg\,m^{-3}}$, $\mu = 1.003 \times 10^{-3}\ \mathrm{Pa\,s}$) and a uniform inlet velocity $U_\infty = 0.0302$ m/s, which gives $\mathrm{Re}_D = U_\infty D/\nu \approx 300$. The domain extends $10D$ upstream of the cylinder axis, $25D$ downstream, $\pm 10D$ in the cross-stream direction, and $8D$ along the span; the cross-stream and spanwise boundaries are symmetry planes, the downstream boundary is a constant-pressure outlet, and the cylinder surface is a no-slip wall. The pressure-based solver uses PISO pressure-velocity coupling, second-order upwind momentum discretization, second-order pressure interpolation, and second-order implicit time integration at $\Delta t = 0.01\ \mathrm{s} = 0.030\, D/U_\infty$, with at most 35 inner iterations per time step.

S2 Network architecture and dimension allocation

Hierarchical encoder. The encoder applies three pooling stages at a fixed ratio r, so that the retained node count after stage ℓ is $r^\ell N$, which yields four mesh resolutions in total. Each stage applies two graph-attention convolutions Velickovic et al. (2017), written GAT below, with a residual projection that matches feature dimensions across changes in width,

$$\mathbf{x}^{(\ell)} = \mathrm{GAT}_2^{(\ell)}\big(\mathrm{GAT}_1^{(\ell)}(\mathbf{x}^{(\ell-1)})\big) + W_{\mathrm{proj}}^{(\ell)} \mathbf{x}^{(\ell-1)}. \tag{S1}$$

The retained features after the final stage are flattened and passed through a fully connected stack to the shared encoding $\mathbf{h} \in \mathbb{R}^{d_h}$.

Graph connectivity augmentation. Because CFD meshes are strongly anisotropic, graph-topological distance is a poor proxy for Euclidean distance, so a fixed-radius operator may either truncate local interactions in refined regions or over-connect coarse regions to distant neighbors. We address this in two ways. During training, and following Gladstone et al. Gladstone et al. (2024), we add random edges at a fixed fraction of the existing edge count by drawing node pairs uniformly without replacement, where each added edge carries a weight equal to the Euclidean node distance. These shortcuts enlarge the effective receptive field without extra layers and are never removed. After each pooling stage, where the original edge set no longer matches the retained nodes, we discard the pooled edges and rebuild a k-nearest-neighbor graph on the retained node positions. This combination stabilizes message passing at every resolution and removes a class of degenerate topologies that otherwise appears after successive pooling.

Spatially uniform pooling. The pooling operator partitions the bounding box of the current node set into a $g \times g$ grid with $g = \lfloor \sqrt{rN_\ell} \rfloor$, and from each occupied cell it keeps the node nearest the cell center. If fewer than rN_ℓ nodes are selected, the deficit is filled at random from the unselected nodes, which keeps spatial coverage uniform. We prefer this operator to attention-based selectors such as TopK pooling Gao and Ji (2019); Cangea et al. (2018), because those collapse the retained nodes into a few contiguous regions and thereby degrade full-field reconstruction Barwey et al. (2025). The selection mask is cached at the first forward pass and reused, which holds the coarse graph fixed across the dataset.

Subspace branches and decoder routing. Each subspace is produced by an independent three-layer perceptron $f_k : \mathbb{R}^{d_h} \to \mathbb{R}^{2d_k}$ that shares no pathway downstream of $\mathbf{h}$, which prevents leakage at the architectural level. The output is split into a mean $\boldsymbol{\mu}_k$ and a log-variance $\log \boldsymbol{\sigma}_k^2$, and the sample is drawn through the reparameterization trick as $\mathbf{z}_k = \boldsymbol{\mu}_k + \boldsymbol{\sigma}_k \odot \boldsymbol{\epsilon}_k$. The decoder uses five sub-decoders, namely $D_{\text{mf}}, D_\omega, D_{\text{turb}} \in \mathbb{R}^{N'\times 2}$ for velocity, $D_E \in \mathbb{R}^{N'\times 1}$ for pressure, and $D_{\text{res}} \in \mathbb{R}^{N'\times 3}$ for residual corrections, where $N' = r^3 N$. The fields are assembled through learnable sigmoid gates,

$$(\hat{u}, \hat{v}) = \sigma(w_{\text{mf}})D_{\text{mf}} + \sigma(w_\omega)D_\omega + \sigma(w_{\text{turb}})D_{\text{turb}} + \sigma(w_{\text{res}})D_{\text{res}}[:, 0{:}2], \tag{S2}$$

$$\hat{p} = \sigma(w_E)D_E + \sigma(w_{\text{res}})D_{\text{res}}[:, 2]. \tag{S3}$$

This routing encodes two physical priors directly, since pressure is reconstructed from the energy subspace alone, which reflects the pressure-kinetic-energy coupling of incompressible flow, while the residual supplies cross-field corrections rather than an independent quantity. The coarse output is lifted to the original mesh through three unpooling stages, each followed by two graph-attention convolutions that interpolate fine-scale detail.

Per-case dimensions. Subspace dimensions are set from a proper-orthogonal-decomposition estimate of the intrinsic complexity of each physical field, with the per-case values listed in Table S1.

Table S1: Per-case architecture settings, where d_z is the total latent dimension, "hidden" is the encoder hidden width, "ratio" is the pooling ratio r per stage, and k is the post-pooling neighborhood size. For the three-dimensional wake the node features comprise four channels (u, v, w, p) rather than three.

Case	d_{mf}	d_ω	d_{turb}	d_E	d_{res}	d_z	hidden	ratio	k
BFS	12	12	12	20	15	71	64	0.4	8
Cylinder	7	12	7	13	8	47	64	0.5	6
NACA0018	9	5	9	17	6	46	64	0.5	8
Cylinder 3D	12	24	42	120	36	234	64	0.5	12

Dimension allocation by POD. For each scalar field Q_k we form the training data matrix $\mathbf{Q}_k \in \mathbb{R}^{N\times M_{\text{train}}}$ and read its singular values $\sigma_1^{(k)} \geq \sigma_2^{(k)} \geq \cdots$. The effective rank at cumulative energy τ is

$$r_\tau^{(k)} = \min\Big\{r : \frac{\sum_{i=1}^{r}(\sigma_i^{(k)})^2}{\sum_i(\sigma_i^{(k)})^2} \geq \tau\Big\}. \tag{S4}$$

The allocated dimensions, as reported in Table S2 for the BFS case, exceed the linear ranks $r_{0.99}^{(k)}$ by roughly 25 to 40% in order to accommodate the nonlinear interactions that a linear basis cannot capture Benner et al. (2015). The energy subspace is over-provisioned more aggressively, because it must additionally encode pressure variation that couples nonlinearly to velocity through continuity, and the residual dimension is chosen near the mean of the four physics-assigned dimensions, which gives enough capacity to absorb cross-quantity coupling without dominating the budget.

S3 Composite loss function

The full objective combines reconstruction, a variational term, physics supervision, the cascaded mutual-information loss, and a temporal regularizer,

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda_{\text{phys}}\,\mathcal{L}_{\text{phys}} + \lambda_{\text{cmi}}\,\mathcal{L}_{\text{cmi}} + \lambda_{\text{temp}}\,\mathcal{L}_{\text{temp}}, \tag{S5}$$

with weights set by the schedule of Section S5, and with $\mathcal{L}_{\text{cmi}}$ given in full in Section S4.

Table S2: POD-derived effective ranks at 99% cumulative energy for the BFS case, where the final column lists the dimension allocated to the corresponding subspace.

Quantity Q_k	Subspace	$r^{(k)}_{0.99}$	Allocated d_k
$\lvert\mathbf{u}\rvert$	Mean flow	$\sim$9	12
ω_z	Vorticity	$\sim$10	12
TI	Turbulence	$\sim$9	12
k (TKE)	Energy	$\sim$14	20
Residual	Residual	—	15
		Total d_z	**71**

Reconstruction. An area-weighted mean-squared error accommodates the heterogeneity of unstructured meshes,

$$\mathcal{L}_{\text{recon}} = \alpha_{\text{vel}} \sum_i w_i \sum_{q\in\{u,v\}} \big(\hat{q}^{(i)} - q^{(i)}\big)^2 + \alpha_p \sum_i w_i \big(\hat{p}^{(i)} - p^{(i)}\big)^2 + \alpha_{\text{grad}}\, \mathcal{L}_{\text{grad}}, \qquad w_i = \frac{A_i}{\sum_j A_j}, \tag{S6}$$

where A_i is the dual-cell area of node i. The gradient-preservation term penalizes discrepancies in field differences along mesh edges,

$$\mathcal{L}_{\text{grad}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \sum_{q\in\{u,v,p\}} \big|(\hat{q}^{(j)} - \hat{q}^{(i)}) - (q^{(j)} - q^{(i)})\big|^2, \tag{S7}$$

which sharpens the shear layers and vortical cores that an area-weighted MSE alone tends to smooth. We use $\alpha_{\text{vel}} = 1.0$, $\alpha_p = 1.0$, and $\alpha_{\text{grad}} = 0.1$.

Variational term. A Kullback-Leibler divergence with a per-subspace capacity constraint, in the β-VAE spirit Higgins et al. (2017); Burgess et al. (2018), prevents any one subspace from monopolizing the budget,

$$\mathcal{L}_{\text{KL}} = -\tfrac{1}{2} \sum_{j=1}^{d_z} \big(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\big) + \eta \sum_k \max\big(0,\, D_{\text{KL}}^{(k)} - C\big), \tag{S8}$$

with per-subspace capacity $C = 5.0$ nats and $\eta = 0.1$, so that the hinge activates only when a subspace exceeds its nat budget.

Physics supervision. Each physical subspace is anchored to its designated scalar by a diagnostic head P_k through a Huber loss, to which a sign-consistency term is added so that training stays stable for signed quantities such as ω_z,

$$\mathcal{L}_{\text{phys}} = \frac{1}{|\mathcal{K}_{\text{phys}}|} \sum_{k\in\mathcal{K}_{\text{phys}}} \Big[\text{Huber}_\delta\big(P_k(\mathbf{z}_k), Q_k^*\big) + 0.1\, \text{ReLU}\big(-P_k(\mathbf{z}_k)\cdot Q_k^*\big)\Big], \tag{S9}$$

with $\delta = 1.0$ and $\mathcal{K}_{\text{phys}} = \{\text{mf}, \omega, \text{turb}, E\}$, where the targets $Q_k^* \in \{|\mathbf{u}|, \omega_z, \text{TI}, k\}$ are precomputed from the CFD ground truth and normalized to zero mean and unit variance. The ReLU term activates only when the predicted value carries the wrong sign, which supplies a gradient where the Huber term has already become small.

Temporal regularization. A three-point stencil penalizes the first and second temporal differences of each subspace trajectory,

$$\mathcal{L}_{\text{temp}} = w_1 \sum_k \left\| \mathbf{z}_k(t) - \mathbf{z}_k(t-1) \right\|^2 + w_2 \sum_k \left\| \mathbf{z}_k(t+1) - 2\mathbf{z}_k(t) + \mathbf{z}_k(t-1) \right\|^2, \tag{S10}$$

where the first-order term suppresses jumps and the second-order term favors locally linear dynamics, which the low-order library of Section S8 relies on. We use $w_1 = 10^{-2}$ and $w_2 = 5 \times 10^{-3}$, chosen so that this term contributes no more than 5% of the total loss at the end of Stage 4.

S4 Cascaded mutual-information loss in full

HSIC and bandwidth. We quantify dependence with the Hilbert-Schmidt independence criterion (HSIC), which is a nonparametric statistic that vanishes if and only if its arguments are independent and that requires no auxiliary network Gretton et al. (2005). For two subspaces over a batch of n samples,

$$\text{HSIC}(\mathbf{z}_a, \mathbf{z}_b) = \frac{1}{(n-1)^2} \operatorname{tr}\big(K_a H K_b H\big), \tag{S11}$$

where K_a and K_b are Gaussian kernel matrices with bandwidth set per batch by the median heuristic, and $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ centers them. The estimator is differentiable in both arguments and is clamped at zero to absorb floating-point negativity, and the batch loop accumulates samples across the temporal window so that the estimate stays stable.

Topology-aware terms. The loss reads a coupling topology that is undirected in Phase A, where only the parent-child pairing matters, and directed in Phase C, where the discovered edges carry orientation. It combines three terms,

$$\mathcal{L}_{\text{cmi}} = w_{\text{enc}} \sum_{(a,b)\in\mathcal{C}} \text{softplus}\big(-s\,\text{HSIC}_{ab}\big) + w_{\text{sup}} \sum_{(a,b)\in\mathcal{S}} \text{HSIC}_{ab} + w_{\text{res}} \sum_{p\notin\mathcal{P}_{\text{res}}} \text{HSIC}_{\text{res},p}, \tag{S12}$$

where $\mathcal{C}$ holds the retained cascade edges and $\mathcal{S}$ holds the remaining physical pairs. The first term encourages mutual information along the cascade through a softplus that supplies a soft ceiling, so that the reward saturates once dependence is established rather than growing without bound. The second term suppresses dependence between unconnected subspaces, and the third keeps the residual independent of the physical channels that the topology leaves unconnected to it.

Scale-aware variant. Because the dissipative range is dominated by the energy-containing scales, a plain HSIC alignment lets the largest scales dominate at the expense of the smaller scales on which cross-scale coupling is informative. The scale-aware criterion SA-CMI corrects this by decomposing each subspace trajectory into K à-trous wavelet sub-bands, formed as successive differences of replicate-padded moving averages with window 2^k, and by computing the criterion within each sub-band before aggregating by a geometric-mean weight,

$$\text{HSIC}_{\text{SA}}(a,b) = \frac{\sum_k w^{(k)}\,\text{HSIC}\big(c_a^{(k)}, c_b^{(k)}\big)}{\sum_k w^{(k)}}, \qquad w^{(k)} = \sqrt{p_a^{(k)}\,p_b^{(k)}}, \tag{S13}$$

where $c_a^{(k)}$ is the sub-band of subspace a and $p_a^{(k)}$ its normalized band energy. The effective depth is clamped at $K_{\text{eff}} = \min(K, \lfloor \log_2 T \rfloor)$ for a window of length T, and we use $K = 4$. A soft ordering term additionally encourages the spectral centroid $\mu_a = \sum_k k\,p_a^{(k)}$ to decrease along the cascade, so that larger structures sit at lower frequency.

Directional term. In Phase C the discovered edges carry orientation, which we enforce on the temporal window. For a causal edge $a \to b$, a small perceptron predicts $\mathbf{z}_b(t+1)$ from $\mathbf{z}_a(t)$ while a second predicts $\mathbf{z}_a(t+1)$ from $\mathbf{z}_b(t)$, and the forward prediction is rewarded while the reverse is penalized,

$$\mathcal{L}_{\mathrm{dir}} = \frac{1}{|\mathcal{E}_{\mathrm{dir}}|} \sum_{(a,b)\in\mathcal{E}_{\mathrm{dir}}} \Big[\mathrm{MSE}\big(\hat{\mathbf{z}}_b(t+1), \mathbf{z}_b(t+1)\big) + \frac{1}{1+\mathrm{MSE}\big(\hat{\mathbf{z}}_a(t+1), \mathbf{z}_a(t+1)\big)} \Big]. \tag{S14}$$

The reverse term approaches one as the backward prediction improves and zero as it worsens, which suppresses backward coupling. Edges that point into the residual sink receive only the reverse-suppression term, since the residual is defined as what the physical channels leave behind and need not be predictable from them.

S5 Staged training and weight schedule

Pipeline. The reduction runs Phase A, then Phase B, then Phase C. Phase A trains the full architecture under the undirected CMI loss together with a weak prior on the parent-child pairing, which yields a disentangled representation without committing the coupling to a fixed orientation. Phase B operates on the frozen Phase A trajectories and recovers the directed topology by the sparse-regression test of Section S7, updating no network weight, after which a consensus across seeds is formed. Phase C then retrains the representation with the directional CMI term active along the discovered topology, which is why the orientation term appears only in the Phase C schedule.

Four-stage schedule. Both Phase A and Phase C follow the same four-stage weight schedule, as summarized in Table S3. Stage 1 optimizes reconstruction with a minimal variational term and produces the shared checkpoint from which all later stages start, so that every run begins from the same reconstruction pathway. Stage 2 warms up the variational prior and the suppression and residual terms, Stage 3 engages the physics supervision and the full CMI terms, and Stage 4 raises the coupling weights to their peak. Within each stage, every newly activated weight is warmed up linearly over the first 30% of that stage. Because every run starts from the same Stage-1 checkpoint and optimizes the same reconstruction objective, with a schedule that is identical between Phase A and Phase C, differences in the learned representation are attributable to the auxiliary objectives rather than to the reconstruction starting point.

Table S3: Four-stage weight schedule, shown as the peak weight reached in each stage. The same schedule runs inside Phase A and Phase C, the directional term is active only in Phase C, and Stage 1 produces the shared checkpoint that all runs load.

Term	Stage 1	Stage 2	Stage 3	Stage 4
Epochs	100	30	150	200
Reconstruction	on	on	on	on
β (KL)	10^{-5}	10^{-3}	10^{-3}	5×10^{-2}
Physics supervision	0	0	0.6	0.8
CMI encourage	0	0	0.01	0.05
CMI suppress	0	0.01	0.05	0.10
CMI residual	0	0.01	0.05	0.10
CMI directional (Phase C only)	0	0	0.01	0.05
Temporal	0	on	on	on

Optimization. We use Adam with a ReduceLROnPlateau scheduler that halves the learning rate after stalled validation, gradient-norm clipping at unity, and early stopping within each phase. To remove the dependence of the reconstruction starting point on the seed, Stage 1 is trained once per case at an initial learning rate of 10^{-3} and saved as a shared checkpoint; every Phase A and Phase C run then loads this checkpoint, repeats the diagnostic-head pretraining, and trains Stages 2 to 4 at an initial learning rate of 5×10^{-4}. The reference hyperparameters for the BFS case are listed in Table S4.

Table S4: Reference hyperparameters for the BFS case. Per-case latent and graph dimensions are listed separately in Table S1.

Component	**Value**
Graph architecture	
Number of pooling stages	3
GAT heads	4
Random edge augmentation fraction	0.20
Physics and diagnostic heads	
Head depth	$d_k \to 64 \to 32 \to 1$
Activation / normalization	GELU / LayerNorm
Loss coefficients	
$(\alpha_{\text{vel}}, \alpha_p, \alpha_{\text{grad}})$	$(1.0, 1.0, 0.1)$
KL capacity C, weight η	5.0 nats, 0.1
Huber threshold δ	1.0
Sign-consistency coefficient	0.1
Temporal weights (w_1, w_2)	$(10^{-2}, 5 \times 10^{-3})$
Optimization	
Optimizer	Adam
Batch size	8 snapshots
Temporal window for $\mathcal{L}_{\text{temp}}$	3 consecutive snapshots
LR scheduler	ReduceLROnPlateau (factor 0.5, patience 15)
Early-stopping patience	20 epochs
Sparse identification	
Savitzky-Golay order / window	3 / 11
STRidge ridge α / threshold λ / iterations	0.05 / 0.01 / 20
Extrapolation horizon	20% of trajectory length

S6 Interpretability validation protocols

Diagnostic heads and post-training audit. Four diagnostic heads $P_k : \mathbb{R}^{d_k} \to \mathbb{R}$ map each physical subspace to its designated quantity, and each is a three-layer perceptron with layer normalization and a smooth activation. During training, cross-subspace independence is enforced by the suppression term of the CMI loss (Section S4), which drives the HSIC between unconnected subspaces toward zero; no adversarial or interference predictor contributes to the training objective.

Perturbation-based sensitivity. For the sensitivity analysis reported in the main text, the raw sensitivity matrix $\mathbf{S}_{\text{raw}} \in \mathbb{R}^{d_z \times 4}$ is normalized column-wise before display,

$$S_{j,q} = \frac{S_{\text{raw},j,q} - \min_i S_{\text{raw},i,q}}{\max_i S_{\text{raw},i,q} - \min_i S_{\text{raw},i,q}}, \tag{S15}$$

so that each physical quantity spans $[0, 1]$. The perturbation magnitude $\epsilon = 0.1$ is expressed in units of the per-dimension posterior standard deviation, which keeps it dimensionally consistent across latent dimensions, and the matrix is averaged over 500 test anchors and three training runs.

Spatial activation mapping. Within-subspace activation maps are obtained by jointly perturbing all dimensions of subspace k, as $\mathbf{z}_k \to \mathbf{z}_k + \epsilon\mathbf{1}$ with $\epsilon = 0.1$, and measuring the nodewise reconstruction change,

$$A_i^{(k)} = \left\| \hat{\mathbf{u}}_\delta^{(k)}(\mathbf{x}_i) - \hat{\mathbf{u}}(\mathbf{x}_i) \right\|_2, \tag{S16}$$

normalized to $[0, 1]$ over the mesh. Applying the same map on each coarsened mesh produces the multi-resolution view in the main text, which confirms that spatial selectivity persists across scales rather than emerging at a single resolution.

S7 Topology discovery

Test. For each child subspace we estimate time derivatives with a Savitzky-Golay filter and fit $\dot{\mathbf{z}}_{\text{child}} = \Theta\,\Xi$ by sequentially thresholded ridge regression Rudy et al. (2017). The baseline library is the polynomial of the child alone, while the parent-augmented library adds the parent terms and the child-parent bilinears. We retain a directed edge when the gain

$$\Delta R^2_{\text{child}\leftarrow\text{parent}} = R^2_{\text{child}\leftarrow\text{parent}} - R^2_{\text{base}} \tag{S17}$$

exceeds a fixed threshold, where the sparsity threshold is selected on a validation split by a Pareto rule that takes the sparsest model within a small tolerance of the best validation R^2, and the per-dimension scores are aggregated by their mean. A baseline whose score falls below a floor is treated as unfittable, so its incoming edges are not tested.

Physical pruning. Before any fit we restrict the candidate parents by two rules drawn from the physics. Because the energy subspace is a second moment of the momentum field and is passively set in an incompressible isothermal flow, edges from energy to any momentum-carrying subspace are forbidden, while an edge from energy into the residual remains admissible, since the residual is a statistical sink rather than a dynamical quantity, and because the residual is a sink, it takes no outgoing edge. These rules remove the corresponding regressions, and a post-threshold pass zeroes any forbidden edge that survives, which should not occur if the restriction is correct. We also drop near-zero variance dimensions before fitting, which reduces estimator noise.

Consensus and defaults. We run the discovery across three seeds per case and form a consensus by majority vote over edges, retaining an edge that is present in at least the fraction of seeds listed in Table S5. The consensus adjacency is the directed topology passed to Phase C, and the seed-level edge presence is reported in the main-text topology figure. For the airfoil we additionally run a relaxed pass that permits the residual as a parent with a lower threshold, which tests whether the external forcing embedded in the residual drives the physical subspaces, and we form a separate consensus for that hypothesis.

Table S5: Topology-discovery defaults, shared across cases unless noted.

Setting	Value
Polynomial order	2
ΔR^2 threshold δ	0.10
Aggregation over dimensions	mean
Maximum parents per child	3
Baseline floor R^2_{base}	-1.0
Zero-variance cutoff	10^{-3}
Consensus vote fraction	0.6

S8 Cascade-aware equation distillation

Derivative estimation. Time derivatives of the latent trajectories are estimated with a Savitzky-Golay filter Savitzky and Golay (1964) of polynomial order 3 and window length 11 samples. These settings were selected by sweeping orders $\{2, 3, 4\}$ and windows $\{7, 9, 11, 13, 15\}$ against a synthetic trajectory with a known analytical derivative, where the chosen pair gave the lowest error.

Cascade-coupled library. For each latent dimension z_i in a subspace with parent set $\mathcal{P}$ fixed by the topology, the library combines an intra-subspace polynomial block, a parent-injection block, a cross-scale bilinear block, and a frequency block built from the dominant peaks of the latent spectra,

$$\mathcal{L}_i = \{z_i^k\}_{k\leq 2} \,\cup\, \{p_j\}_{p_j\in\mathcal{P}} \,\cup\, \{z_i p_j\}_{p_j\in\mathcal{P}} \,\cup\, \big\{\sin/\cos(2\pi f_k t),\, z_i \sin/\cos(2\pi f_k t)\big\}_k. \tag{S18}$$

The bilinear product $z_i p_j$ is the latent signature of a Galerkin projection of the advection term onto the learned basis, and it carries identifiable variance because adjacent cascade pairs retain mutual information by design. The library level is assigned by where the subspace sits in the cascade, so that the mean-flow and residual subspaces use a polynomial level, the vorticity subspace adds the parent injection and bilinears, and the turbulent and energy subspaces further add the bounded nonlinearities described in Section S12. The frequency block retains peaks above a tenth of the local spectral maximum, subject to a minimum separation.

Coefficient estimation. Both the library and the target derivatives are standardized column-wise before regression, and coefficients are obtained by sequentially thresholded ridge regression Rudy et al. (2017) with ridge weight $\alpha = 0.05$, threshold $\lambda = 0.01$, and at most 20 iterations. The validation R^2 is reported on the final fifth of the identification segment, and a dimension is declared active when it satisfies $R^2_{\text{val}} > 0$ and retains at least one nonzero term.

Cascade-rollout integration. Identified systems are integrated over a window equal to 20% of the trajectory length, with subspaces processed one at a time in an order obtained by a topological sort of the directed edges, and ties broken by the canonical subspace ordering. In the diagnostic mode each child is forced with the true parent trajectory, which isolates the quality of the child equation, whereas in the cascade-rollout mode each child is forced with the predicted parent if it has already been integrated, and otherwise with the true trajectory as a fallback, which tests the closed system. Reciprocal edges are not solved jointly, since a strict orientation is impossible there, so they are broken by the canonical ordering, with the earlier member using the later member's true trajectory and the later member optionally using the earlier member's prediction. We try several stiff and non-stiff solvers in turn and accept the first that converges. Alongside the trajectory R^2, we report the time-averaged relative error,

$$\bar{\varepsilon}_{\mathrm{rel}} = \frac{1}{|\mathcal{T}_{\mathrm{extr}}|} \sum_{t \in \mathcal{T}_{\mathrm{extr}}} \frac{\|\tilde{\mathbf{z}}_k(t) - \mathbf{z}_k(t)\|_2}{\|\mathbf{z}_k(t)\|_2 + \varepsilon}, \tag{S19}$$

where $\tilde{\mathbf{z}}_k$ is the integrated trajectory and $\mathbf{z}_k$ the CFD-derived one.

S9 Evaluation metrics

The velocity-magnitude error is the NRMSE normalized by the field L^2 norm,

$$\varepsilon_{|\mathbf{U}|} = \frac{\big\| \, |\hat{\mathbf{U}}| - |\mathbf{U}| \, \big\|_2}{\big\| \, |\mathbf{U}| \, \big\|_2}, \qquad |\mathbf{U}| = \sqrt{u^2 + v^2}, \tag{S20}$$

while the pressure error is range-normalized, because the L^2 norm is ill-conditioned for a pressure field whose mean is not separated from zero,

$$\varepsilon_p = \frac{\sqrt{\frac{1}{N}\sum_i (\hat{p}^{(i)} - p^{(i)})^2}}{p_{\max} - p_{\min}}. \tag{S21}$$

Both are averaged over 2,000 randomly sampled snapshots and reported as mean and standard deviation. Disentanglement is assessed through the perturbation-based sensitivity analysis of Section S6, and identification quality is reported as the per-subspace validation R^2 together with the forward-extrapolation R^2 over the 20% window.

S10 Reconstruction accuracy and baseline comparison

Across all three two-dimensional cases the velocity NRMSE stays between 3.0% and 4.3% and the pressure error between 0.9% and 4.3%, as reported in Table S6, and the pointwise errors concentrate near sharp gradients rather than degrading globally. The gap relative to reconstruction-optimal baselines such as POD, DMD, and AE-POD reflects the hierarchical graph pooling rather than the disentanglement objective, since those baselines either operate in a fixed global modal basis or regress its coefficients, whereas our encoder coarsens the computational mesh across several pooling stages so as to match the spatial morphology of the flow. The Graph β-VAE, which shares our pooling but enforces all-pairs independence, reconstructs markedly worse, which indicates that the gap is set by the pooling design and not by the physical decomposition.

S11 Spectral analysis of the latent trajectories

We compute the power spectral density of each latent dimension with a Hann-windowed periodogram over the held-out test segment of the latent trajectory, and we mark a peak as dominant when it rises above a tenth of the local maximum, subject to a minimum separation. The peak frequencies that seed the frequency block of Eq. (S18) are extracted by this protocol, from the measured latent spectra rather than from the prescribed kinematics, so that no case-specific prior enters the identification. Given test-segment durations of approximately 32, 16, and 4 s, the periodogram resolves frequencies on grids of spacing $\Delta f \approx 0.031$, 0.063, and 0.25 Hz for the BFS, cylinder-pair, and airfoil cases respectively, and the reported peak frequencies are positions on these grids rather than continuous estimates. The prescribed airfoil pitching frequency $f_\alpha \approx 0.80$ Hz (Section S1) therefore registers at $f_p = 0.75$ Hz and its second harmonic at 1.5 Hz, and the same quantization explains why higher harmonics need not sit at

Table S6: Reconstruction errors across reduced-order baselines on the three benchmark cases, evaluated over 2,000 randomly sampled snapshots (mean ± std, in percent). Here $\varepsilon_{|\mathbf{U}|}$ is the velocity-magnitude NRMSE and ε_p the range-normalized pressure RMSE. The number of POD modes equals the latent dimension of PID-GraphVAE per case, namely 71 for BFS, 47 for the cylinder pair, and 46 for NACA0018.

Method	**BFS**		**Cylinder**		**NACA0018**	
	$\varepsilon_{|\mathbf{U}|}$	ε_p	$\varepsilon_{|\mathbf{U}|}$	ε_p	$\varepsilon_{|\mathbf{U}|}$	ε_p
PID-GraphVAE	3.06 ± 0.87	4.24 ± 1.88	3.67 ± 0.20	0.96 ± 0.14	4.23 ± 0.42	1.07 ± 0.19
Graph β-VAE	14.44 ± 1.82	11.11 ± 0.83	14.20 ± 0.00	4.35 ± 0.00	10.07 ± 0.01	2.73 ± 0.01
AE-POD	0.93 ± 0.51	0.68 ± 0.37	0.88 ± 0.37	0.20 ± 0.09	0.38 ± 0.18	0.10 ± 0.05
POD	0.23 ± 0.08	0.43 ± 0.13	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.01	0.01 ± 0.00
DMD	0.29 ± 0.25	0.81 ± 0.84	0.02 ± 0.00	0.02 ± 0.01	0.05 ± 0.03	0.01 ± 0.01

exact integer multiples of the displayed fundamental, the cylinder harmonics registering at 1.44 and 2.12 Hz on the 0.69 Hz ladder. For the BFS case, the dominant peaks map onto Strouhal numbers $\mathrm{St}_h = f\,h/U_b$ using the step height h and the inlet bulk velocity U_b, as listed in Table S7, where the fundamental corresponds to $\mathrm{St}_h \approx 0.12$, in agreement with the reattachment flapping reported at comparable Reynolds numbers Driver and Seegmiller (1985); Kasagi and Matsunaga (1995).

Across cases, the spectra separate the three regimes: the BFS broadens toward the turbulent channel, every cylinder subspace concentrates on a single harmonic ladder built on the shedding frequency, the spectral counterpart of a low-dimensional global oscillator, and the airfoil splits between pitching-locked and higher-frequency content: the mean-flow and vorticity subspaces lock to the pitching ladder, the energy subspace combines its pitching-locked component (the f_p and $2f_p$ peaks of Section S14) with higher-frequency content, and the residual absorbs the broadband remainder.

Table S7: Dominant latent frequencies for the BFS case and their Strouhal numbers, where $\mathrm{St}_h = f\,h/U_b$ is anchored to the fundamental at $\mathrm{St}_h \approx 0.12$.

Peak	f_1 (fundamental)	f_2 (first harmonic)	f_3 (higher harmonic)
Frequency (Hz)	0.31	0.59	0.91
St_h	0.12	0.24	0.36

S12 Failure modes and cumulative ablation

The library design and the scale-aware representation of the main text were reached through a sequence of failure modes on the backward-facing step, which we document with the cumulative ablation of Table S8, where each row adds one component to the configuration above it, so that the resulting change in the per-subspace validation R^2, in the largest identified coefficient $|\hat{\beta}|_{\max}$, and in the library condition number κ isolates the contribution of that component.

The polynomial baseline already extracts useful structure from the upper cascade, where the vorticity subspace activates the cross-subspace bilinear products that carry the signature of advective coupling, as row (i) lists. The lower cascade exposes the first failure mode, since a polynomial ansatz cannot represent dissipative-range dynamics, in which Kolmogorov scaling involves fractional powers and the energy budget operates as a near-cancellation between pro-

Table S8: Cumulative ablation of the cascade-aware identification pipeline on the BFS configuration. Rows correspond to (i) CMI with a polynomial library; (ii) the addition of saturating and gated nonlinearities; (iii) the addition of mutual-exclusion groups; (iv) the addition of runtime collinearity gates ($|\rho| > 0.98$, $\kappa > 10^6$); (v) the SA-CMI representation ($K = 4$); and (vi) the mf $\to E$ topology correction. Columns report the median validation R^2 per subspace, the largest absolute coefficient, the condition number of the retained library, the mean active terms per fittable dimension, and the extrapolation stability across the non-residual subspaces (• stable, ○ unstable). All quantities are medians across three seeds.

	R^2 (median)							
Config.	MF	Vo	Tu	E	$\lvert\hat{\beta}\rvert_{\max}$	κ	Act./dim	Stab.
(i)	0.32	0.41	0.07	0.24	2.55	—	6.2	○
(ii)	0.36	0.48	0.17	0.54	$\sim 10^2$	—	10.1	•
(iii)	0.38	0.36	0.49	0.16	$\sim 10^1$	—	17.6	○
(iv)	0.50	0.45	0.50	0.16	1.83	3.6×10^2	9.4	•
(v)	0.37	0.29	0.29	0.20	0.94	4.7×10^1	4.7	•
(vi)	0.40	0.34	0.42	0.19	1.02	4.7×10^1	6.3	•

duction and dissipation, so the turbulent and energy scores reach only 0.07 and 0.24. Extending the library with bounded saturating nonlinearities, gated cascade products, and absolute-value terms raises these scores to 0.17 and 0.54 and stabilizes the integrator, as reported in row (ii), and the energy-subspace gain indicates that the ceiling there is set by the library rather than by the representation.

The second failure mode constrains any pipeline that pairs a kernel-based independence criterion with an overcomplete library, because the Hilbert-Schmidt criterion shapes the joint distribution away from product form yet imposes no functional structure, so within the training amplitude range the regression cannot separate functionally distinct terms that happen to be near-collinear. The symptom appears in row (ii), where the largest coefficient inflates to order 10^2 as the solver splits weight between collinear pairs that cancel inside the window but diverge outside it. We address this in two layers, first by declaring mutual-exclusion groups at library construction so that enabling one member disables its synonym, as row (iii) confirms, and then by a runtime correlation and condition-number screen, as row (iv) confirms, after which the largest coefficient returns to order unity. Because the construction-stage step removes the source of the cancellation rather than its effect, it carries the substantive load.

These layers contain the symptom but not its origin, which is that the criterion is applied to entire trajectories at once, so its gradient is dominated by the energy-containing range while the smaller informative scales contribute negligibly. The scale-aware variant of Eq. (S13) corrects this, and applying the same library to the scale-aware trajectories yields row (v), in which the largest coefficient falls to 0.94 and the condition number contracts while most subspace scores decline in step. This trade is the intended one, since the scale-aware representation replaces stiff, high-magnitude fits with smoother, lower-magnitude fits that integrate stably and generalize, as the forward-extrapolation scores of main-text extrapolation table make quantitative. Row (vi) adds the mf $\to E$ correction, under which the mean-flow and turbulent subspaces recover to 0.40 and 0.42, an improvement that holds in-window but not in forward extrapolation, a tension that the data-driven topology resolves by retaining turb $\to E$.

S13 Per-case readings of the data-driven topologies

Backward-facing step. The mf $\to \omega$ edge of the a priori cascade survives on every seed with its direction intact, while the ω–turb link survives only as a reciprocal pair in which neither direction outvotes the other; three further edges route mean-flow, turbulent, and energy information into the residual. Under this consensus the residual, fed by three parents, extrapolates stably, whereas the vorticity and turbulent subspaces extrapolate poorly because they sit on the reciprocal $\omega \leftrightarrow$ turb loop, which resists forward integration even under true-parent forcing (Table 1 of the main text). The verdict is therefore graded rather than binary: the data confirm the direction of the cascade at its root, retain its middle link only as an undirected coupling, and reject its sparsity outright, the two subspaces on the loop paying for the added density in forward extrapolation. That the consensus retains turb $\to E$ while recovering the cleanest mean-flow extrapolation also explains, at the representation level, why the manual mf $\to E$ correction improved the in-window fit but not the forward extrapolation.

Cylinder pair. The consensus comprises four edges: $\omega \to$ mf on every seed, $\omega \to$ turb and turb $\to$ mf on two of three seeds, and one edge routing vorticity into the residual. The energy subspace is left fully disconnected, the only energy edges seen on any single seed failing the majority vote. This is the representation-level counterpart of the routing insensitivity reported in the main text: the two a priori topologies scored comparably on the cylinder energy subspace not because the two routes were equivalently active, but because neither was active to begin with. Once the energy subspace is integrated as autonomous, its extrapolation rises by more than 0.2 over the better a priori configuration. On the third seed the Floquet bank activates only partially: three of the thirteen energy modes survive the thresholding, carrying weak, mode-specific frequencies rather than the common shedding period, consistent with the seed-level sensitivity visible in Table 1 of the main text.

Pitching airfoil. The consensus consists of a single mf $\to \omega$ edge carried by two of the three seeds, so that no subspace beyond vorticity has an incoming edge. A relaxed re-discovery in which residual out-edges were admitted to the candidate set and the acceptance threshold halved returned no residual-to-physical edge on any seed, corroborating the star reading in which the external forcing, embedded in the residual, reaches each physical subspace in parallel rather than along the cascade.

S14 Candidate parametric (Mathieu-type) resonance in the NACA0018 energy equations

Although the NACA0018 configuration violates the cascade assumption and its dominant dynamics relocate to the residual subspace, the data-driven energy equations on one seed exhibit a recurring template across several energy dimensions,

$$\dot{z}_{E,k} \;\approx\; \alpha_k \, z_{E,2}^2 + \beta_k \, z_{E,2} \cos(2\pi f_p t), \qquad k \in \{0, 1, 4, 5, 12\}, \tag{S22}$$

with $f_p = 0.75$ Hz the dominant pitching-locked latent frequency. The identified pairs (α_k, β_k) are listed in Table S9, where the amplitudes span more than an order of magnitude across dimensions, yet the ratio β_k/α_k stays tightly clustered, which fixes the relative strength of the quadratic self-interaction and the parametric forcing by the pitching kinematics rather than by the latent embedding.

Equation (S22) is the latent-space signature of a multiplicative (Mathieu-type) parametric forcing, in which the angle-of-attack oscillation drives the energy modulus multiplicatively rather than additively, consistent with classical Mathieu-Hill analyses of parametric instability Nayfeh and Mook (2024). Because the master coordinate $z_{E,2}$ is itself locked to the pitching frequency, both terms rectify energy to the second harmonic, and a peak at $2f_p = 1.5$ Hz appears in the energy subspace on all three seeds. We report this as a candidate phenomenology rather than

Table S9: Identified coefficient pairs (α_k, β_k) of Eq. (S22) on seed 123, together with the conserved ratio. Seed 456 reproduces the template on a single dimension with the same sign pattern and ratio 0.65, and seed 42 yields no surviving terms in the energy subspace under thresholding. The cross-seed mean of the per-seed medians, $r = 0.66 \pm 0.02$, is the value quoted in the main-text invariants table.

Dimension k	α_k	β_k	β_k/α_k
0	−5.52	−3.76	0.68
1	−7.06	−4.58	0.65
4	4.85	3.43	0.71
5	6.62	4.49	0.68
12	31.66	20.40	0.64
		Median	0.68

an established structural result, for three reasons. First, the template is robust within one seed but only partial on a second and absent on a third, so the cross-seed reproducibility is intermediate. Second, the interpretation of $z_{E,2}$ as the dominant pitching-coupled energy mode has not been verified by the perturbation-based sensitivity analysis of Section S6, which targeted the BFS configuration alone. Third, the library admits several near-collinear parameterizations of multiplicative forcing that the present gates do not distinguish, so a targeted identification with a weak-form Mathieu library is left to future work.

References

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Rini Jasmine Gladstone, Helia Rahmani, Vishvas Suryakumar, Hadi Meidani, Marta D'Elia, and Ahmad Zareei. Mesh-based gnn surrogates for time-independent pdes. *Scientific reports*, 14(1):3394, 2024.

Hongyang Gao and Shuiwang Ji. Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR, 2019.

Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. Towards sparse hierarchical graph classifiers. *arXiv preprint arXiv:1811.01287*, 2018.

Shivam Barwey, Hojin Kim, and Romit Maulik. Interpretable a-posteriori error indication for graph neural network surrogate models. *Computer Methods in Applied Mechanics and Engineering*, 433:117509, 2025.

Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM review*, 57(4):483–531, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

David M Driver and H Lee Seegmiller. Features of a reattaching turbulent shear layer in divergent channelflow. *AIAA journal*, 23(2):163–171, 1985.

Nobuhide Kasagi and Akio Matsunaga. Three-dimensional particle-tracking velocimetry measurement of turbulence statistics and energy budget in a backward-facing step flow. *International journal of heat and fluid flow*, 16(6):477–485, 1995.

Ali H Nayfeh and Dean T Mook. *Nonlinear oscillations*. John Wiley & Sons, 2024.