14th International Symposium on Turbulence and Shear Flow Phenomena (TSFP14)
Heidelberg, Germany, July 28–31, 2026

# X - CAL: EXPLAINABLE FRAMEWORK FOR CAUSAL ANALYSIS OF LATENT REPRESENTATIONS IN SHEAR FLOWS

**Marcial Sanchis-Agudo**[1] **Ricardo Vinuesa**[2,1]
[1] FLOW, Engineering Mechanics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden
[2] Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan, USA
Corresponding author: sanchis@kth.se

## ABSTRACT

Flows past finite-height obstacles feature canonical vortices that govern transport and mixing. Classical reduced-order models (POD, DMD) capture energetic structures but not explicit causal links (Lumley, 1967; Schmid, 2010). We present **X - CAL**, a framework combining a $\beta$-variational autoencoder ($\beta$-VAE), the synergistic–unique–redundant decomposition (SURD) of causality, and Gradient–SHAP. DNS snapshots of the flow around a wall-mounted square cylinder at $Re_h = 2000$ are compressed into three near-orthogonal latents retaining >93% of fluctuation energy. Validation on synthetic systems (torus, Lorenz (Lorenz, 1963)) shows that the framework reduces redundancy without changing the dominant SURD ordering; in Lorenz it recovers the symmetry-switching inter-lobe transition through both state maps and SHAP fields. For the obstacle wake, the decoder and SHAP means separate three preferred mid-span sectors, and SURD links them through delays aligned with fractions of the physical shedding period. X - CAL thus provides a compact, interpretable route to analyze and control turbulent flows (Martínez-Sánchez *et al.*, 2024; Cremades *et al.*, 2024).

## INTRODUCTION

Coherent structures organize wall-bounded turbulence and wakes but are embedded in high-dimensional, non-linear interactions that are difficult to interpret causally (Jiménez, 2018). Linear ROMs (POD/DMD) are valuable yet limited for non-linear dependencies. Recent work has shown that deep compression and predictive architectures can be effective for reduced-order modelling and temporal prediction in turbulent flows, including $\beta$-VAE/transformer combinations (Solera-Rico *et al.*, 2024) and lightweight transformer attention mechanisms for temporal forecasting (Sanchis-Agudo *et al.*, 2025). In parallel, deep-learning-based surrogate closures have started to connect these ideas with practical modelling of turbulence (Eiximeno *et al.*, 2025). What these lines still lack is an explicit causal and physical interpretation layer. X - CAL addresses that gap by combining nonlinear compression with a $\beta$-VAE, explicit lagged causality analysis with SURD, and physical attribution with Gradient–SHAP plus percolation. This end-to-end pipeline links *who influences whom and when* in latent space with *where* the responsible support lies in the physical flow, thereby extending predictive/compressive models toward interpretable closure-oriented analysis (Vinuesa *et al.*, 2025).

## 1 Methodology

The present framework combines high-fidelity numerical data, information-theoretic causality, explainability tools, and non-linear compression into a single workflow. The numerical database is obtained from a direct numerical simulation (DNS) of the flow around a wall-mounted square cylinder at $Re_h = 2000$, performed with the Nek5000 spectral-element solver. This canonical geometry isolates key wake–wall interactions while remaining sufficiently rich to test causal analysis tools (Vinuesa *et al.*, 2015). The obstacle has a width-to-height ratio $b/h = 0.25$, and the flow develops over a laminar inflow boundary layer. The spectral-element mesh consists of 21.8 million points; a subdomain is interpolated onto a uniform grid of $(N_x, N_y, N_z) = (300, 100, 150)$. After discarding transients, 30,000 statistically stationary snapshots spanning 150 convective time units are extracted, each advanced with a time step $\Delta t = 0.005$. To emulate forward prediction, the dataset is split into training and testing subsets with a 5:1 ratio, preserving the temporal sequence. For interpretability, we restrict the analysis to the mid-span plane ($z/h = 0$), yielding two-dimensional fluctuation fields $\mathbf{U}^t = [u'(x,y,t), v'(x,y,t)]$. Consequently, the obstacle-side physical interpretation below is deliberately limited to mid-span wake signatures rather than a complete three-dimensional reconstruction of the wake topology.

The compression of these snapshots is performed with a convolutional $\beta$-Variational Autoencoder ($\beta$-VAE) (Solera-Rico *et al.*, 2024). For the obstacle case we use four hidden layers and a three-dimensional latent space, training for 300 epochs with batch size 64. The model encodes each field into a latent vector $\mathscr{L}(t) = [\ell_1, \ell_2, \ell_3]$ and decodes it back to a reconstruction $\tilde{\mathbf{U}}^t$. The training objective is the standard $\beta$-VAE functional, $\mathscr{J}_\beta = \|\mathbf{U}^t - \tilde{\mathbf{U}}^t\|_2^2 + \beta D_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|\mathbf{U}^t)\|p(\mathbf{z})\right]$, and latent quality is assessed through both reconstruction accuracy and the covariance of $\mathscr{L}$. This structure encourages the latent variables to be disentangled, near-orthogonal, and interpretable, so that each $\ell_i$ encodes a distinct dynamical mechanism rather than a mixture of features.

Once the latent space is obtained, we analyze the temporal evolution of the variables to uncover causal relations. The latent time series are examined with the synergistic–unique–redundant (SURD) decomposition of mutual information (Martínez-Sánchez *et al.*, 2024). Unlike correlation, which only measures linear dependencies, SURD partitions the information provided by sources, $\tilde{\ell}$ about a target, $\ell_j^{t+\Delta t}$ into unique, redundant, and synergistic contributions:

$$I\left(\ell_j^{t+\Delta t}; \tilde{\ell}\right) = \sum_{k=0}^{N-1} \Delta I_{k\to j}^{U} + \sum_{\mathbf{s}\in C} \Delta I_{\mathbf{s}\to j}^{R} + \sum_{\mathbf{s}\in C} \Delta I_{\mathbf{s}\to j}^{S}, \qquad (1)$$

Applied in a lagged manner, the quantity $I\left(\ell_j;\tilde{\ell}\right)$ in Eq. 1 measures how much the future state of a latent is informed by the present states of the others. By scanning integer delays $\Delta t \in [0,3500]$, the method identifies the characteristic lags associated with convection and the dominant shedding cycle.

The final stage is to connect these latent-space interactions back to physical structures. To achieve this, we compute Gradient–SHAP attribution maps (Lundberg & Lee, 2017; Cremades *et al.*, 2024; Shapley, 1953), which quantify the contribution of each pixel in the flow field to the value of a latent variable. In practice, this procedure follows a clear sequence: the $\beta$-VAE first encodes the physical fields into latent variables, in direct analogy to how POD or DMD produce temporal coefficients in linear frameworks; causality analysis is then performed in the latent space to determine directional relations such as $\ell_j \to \ell_i$; for each time step, Gradient–SHAP maps identify the regions of the physical domain that are most influential in the encoding of each latent variable; and finally, a percolation analysis is applied to these attribution fields, isolating coherent structures whose statistics can be directly associated with the latent-to-latent causal interactions. Denoting the encoder output by $\ell_i^t = \mathscr{E}_i(\mathbf{U}^t)$, we write the channel-wise attribution fields as $S_{i,u}^t, S_{i,v}^t = \mathrm{GradSHAP}(\mathscr{E}_i, \mathbf{U}^t; \mathscr{B})$, where the baseline ensemble $\mathscr{B}$ is approximated by 7000 reference snapshots. We then define the SHAP magnitude and its rms reference field by

$$A_i^t(x,y) = \sqrt{(S_{i,u}^t)^2 + (S_{i,v}^t)^2}, \quad \sigma_i(x,y) = \sqrt{(\overline{S_{i,u}})^2 + (\overline{S_{i,v}})^2}, \tag{2}$$

with the overbar denoting the empirical average over $\mathscr{B}$. The percolated support is

$$\Pi_i^t(x,y) = \mathbf{1}\left(A_i^t(x,y) > H_i\,\sigma_i(x,y)\right), \tag{3}$$

where the latent-dependent thresholds are fixed to $H = [1.14, 1.36, 1.14]$, selected as in the arXiv analysis to maximize the number of coherent connected structures recovered after percolation. The masked SHAP field is $\tilde{\mathbf{S}}_i^t = \Pi_i^t \odot \mathbf{S}_i^t$, and the corresponding structured turbulent kinetic energy is

$$STKE_i^t = \frac{1}{2}\left[(\tilde{S}_{i,u}^t)^2 + (\tilde{S}_{i,v}^t)^2\right]. \tag{4}$$

The unique fields shown below are obtained after removing pixels belonging to the intersection of the three latent supports, so that $STKE_i^t$ isolates the non-shared contribution of latent $i$.

Figure 1 summarizes this operational X - CAL chain. The causal graph is not prescribed a priori: for each target latent, SURD is scanned over lag, the dominant unique peaks are retained as directed edges, and those edges are then mapped back to both latent-state sectors and physical-space supports through the state maps and Gradient–SHAP fields.

Through this progression, the framework closes the loop from DNS data to latent compression, from causal inference to explainability, and from abstract information flows back to coherent physical structures in the turbulent wake.

## RESULTS

We validate X - CAL from controlled to realistic settings. On the *coupled torus*, the original SURD histogram shows that each target remains dominated by its own unique contribution, while the imposed cross-coupled effect stays subdominant in the causal budget (Fig. 2, left). Training $\beta$-VAEs with increasing regularization (Tab 1) progressively reduces redundancy while preserving that dominant unique ordering across the latent cases (Fig. 2, right); Case IV yields the cleanest separation without sacrificing reconstruction (Tab. 1). Thus, for this controlled benchmark, disentanglement sharpens the SURD decomposition without reassigning the dominant causal directions.

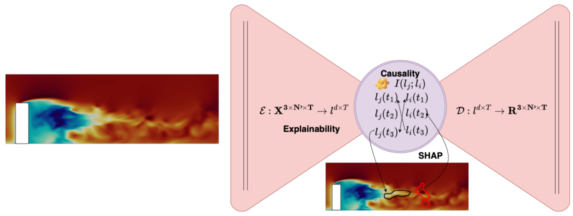


Figure 1. X - CAL workflow. DNS fluctuation fields are compressed into a low-dimensional latent space with a $\beta$-VAE, SURD scans lagged directed information to build the causal graph, and Gradient–SHAP plus percolation map each latent branch back to coherent physical supports.

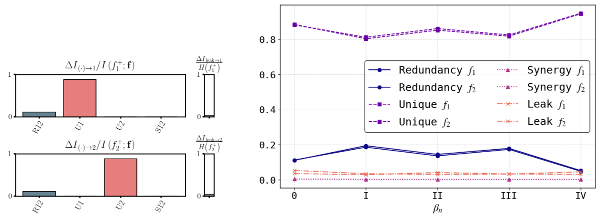


Figure 2. Torus validation of the causal claim. Left: SURD histogram for the original coupled torus, showing that the causal budget is dominated by the self-unique terms while the cross-coupled effect remains subdominant. Right: SURD summary across the four $\beta$-VAE cases in Table 1. Increasing disentanglement reduces redundancy while preserving the same dominant unique ordering.

| Case | Orthog. (%) | Recon. (%) | Latent | Ep. | $\beta_n$ |
|---|---|---|---|---|---|
| I | 88.53 | 98.98 | 2 | 300 | $5\times10^{-5}$ |
| II | 91.14 | 99.92 | 2 | 300 | $5\times10^{-4}$ |
| III | 99.69 | 99.89 | 2 | 300 | $5\times10^{-3}$ |
| IV | 95.79 | 99.89 | 2 | 800 | $5\times10^{-3}$ |

Table 1. Training results for the four torus cases with different hyperparameters.

For the *Lorenz* system (Lorenz, 1963), lifting $(x,y,z)$ to the original physical 2D field and training Case I and II with latent dimensions $d \in \{3,2\}$ recovers the known coupling: mutual-information maps align the learned latents with the $(x,y)$ pair and the $z$ coordinate (Fig. 3). The SURD state maps then localize the same transition sector in both the original Lorenz coordinates and the learned latent plane, concentrating the unique states near the inter-lobe switching corridor where the trajectory changes lobe and therefore changes symmetry sector (Fig. 4). Mean Gradient–SHAP and decoded fields averaged over the corresponding causal windows recover the familiar butterfly support of those transition states (Fig. 5). This is the key Lorenz result: X - CAL does not only preserve a low-dimensional embedding, it preserves the state-dependent symmetry transition that governs the dynamics, which is precisely the scientifically relevant target for low-dimensional control.

Applied to the *wall-mounted cylinder* at $Re_h = 2000$, the $\beta$-VAE attains near-orthogonal latents and $> 93\%$ reconstruc-

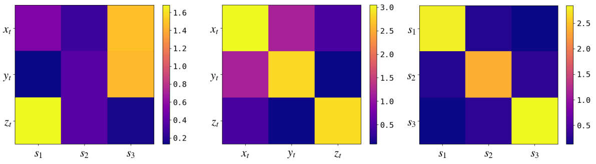

Figure 3. Lorenz mutual-information maps for Case I. Left: mutual information between original coefficients $(x_t, y_t, z_t)$ and learned latents $(s_1, s_2, s_3)$. Middle: mutual information between Lorenz coefficients. Right: mutual information between latent variables.

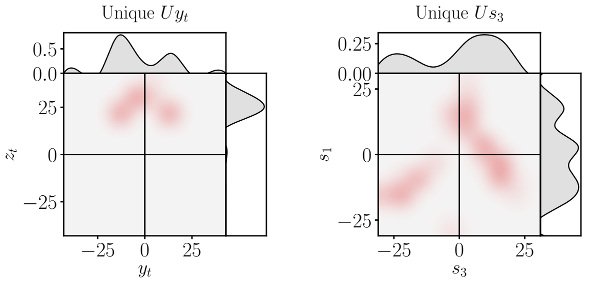


Figure 4. State-dependent SURD maps for Lorenz. Left: unique states in the original Lorenz coordinates, localized near the inter-lobe switching region. Right: the same transition sector represented in the learned latent plane, showing that the causal state localization is preserved across representations.

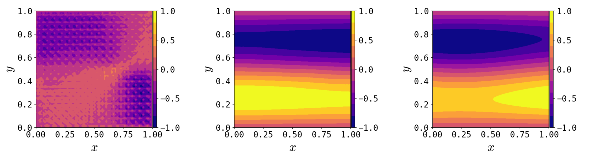

Figure 5. Mean fields over the Lorenz causal states of Case I. Left: unique Gradient–SHAP field associated with the selected transition states. Middle: mean decoded field $\overline{\mathscr{D}(s_1)}$. Right: mean physical field over the same states. Together they recover the butterfly switching support selected by SURD.

tion for $u'/v'$ (Fig. 6), with decoded modes capturing richer multi-scale content than POD. Before invoking causality, the mean decoder and mean unique SHAP fields already indicate that the representation separates three preferred mid-span wake locations (Figs. 7 and 8): $\ell_1$ is concentrated in the lower wake near the wall, $\ell_3$ occupies the upper wake connected with the free-end signature, and $\ell_2$ emphasizes a mid-wake shear-layer sector. SURD then adds the temporal organization of these sectors. It identifies informative lags at 381, 1569 and 3103 steps, consistent with convection over roughly $T_{\text{shed}}/9$, $T_{\text{shed}}/2$ and one shedding period. The corrected SURD unique-causality summary shows both the lag spectrum and the unique-term histograms at those dominant delays, making the main routes explicit: $\ell_3$ is the dominant unique driver of both $\ell_1$ and $\ell_2$, whereas the delayed feedback to $\ell_3$ is weaker and more distributed across the latent variables (Fig. 9).

These dominant routes are then summarized by a compact triadic graph: a short $\ell_3 \to \ell_1$ branch, a half-period $\ell_3 \to \ell_2$ branch, and the $\ell_1 \to \ell_3$ feedback that closes the cycle (Fig. 11). In this sense, the graph is built directly from the SURD decomposition: edge direction follows the source-target ordering, edge selection follows the dominant unique peaks at the selected lags, and the accompanying state maps

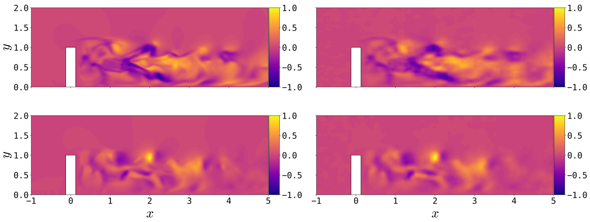

Figure 6. Reconstruction quality on held-out data for $u'$ (top) and $v'$ (bottom); left: ground truth, right: $\beta$-VAE reconstruction.

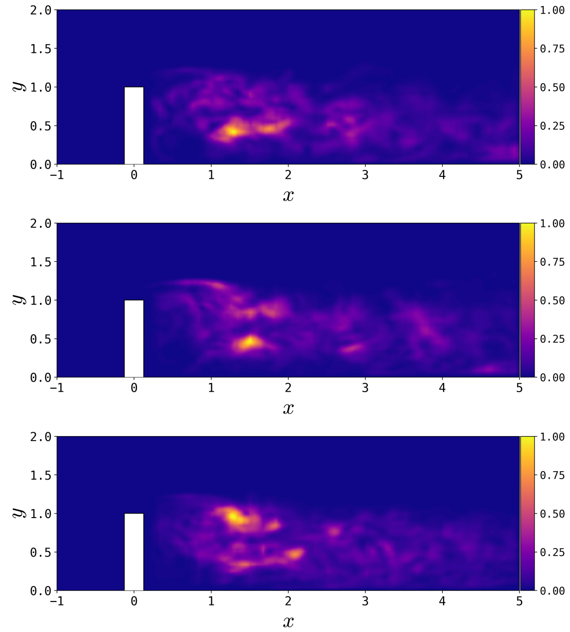

Figure 7. Mean decoded turbulent kinetic energy obtained by activating each latent independently. Top to bottom: $TKE_1$, $TKE_2$, and $TKE_3$.

indicate where each branch is activated in latent space. State-dependent SURD maps further show that these three dominant links occupy compact but distinct latent-space sectors, separating the two $\ell_3$-driven branches from the $\ell_1 \to \ell_3$ feedback route (Fig. 10). On the analyzed mid-span plane, the most robust physical-space evidence comes from the structured SHAP masks themselves: Fig. 12 reads top-to-bottom as the shared upper-wake source associated with $\ell_3$, the delayed lower-wake support at $\Delta t_1$, and the delayed mid-wake shear-layer support at $\Delta t_2$. More precisely, Fig. 12 does not show a reconstructed velocity or energy field; it shows the smoothed occupancy field $\widehat{\Pi}_i(x,y) = \sum_k w_k \mathbf{1}_{(x,y)\in U_i(t_k)} / \sum_k w_k$ built from the unique SHAP support $U_i(t_k)$ at the selected causal events and then normalized to $[0,1]$, so that bright regions mark where the attribution persists most consistently across the event ensemble. The obstacle result is therefore not only that three regions are detected, but that they cooperate over physical shedding periods in a consistent causal ordering. It is also the control-oriented reading of X-CAL: in latent space, the graph identifies which variable should be monitored or actuated and on which delay; in physical space, the SHAP support identifies the wake region where sensing or forcing should target the corresponding mechanism on the analyzed mid-span plane. We therefore keep the interpretation of $\ell_1$ conservative, as a lower-wake / near-wall response on the mid-span plane rather than a complete three-dimensional base-vortex topology, and we do not assign a separate physical topology to the synergistic term

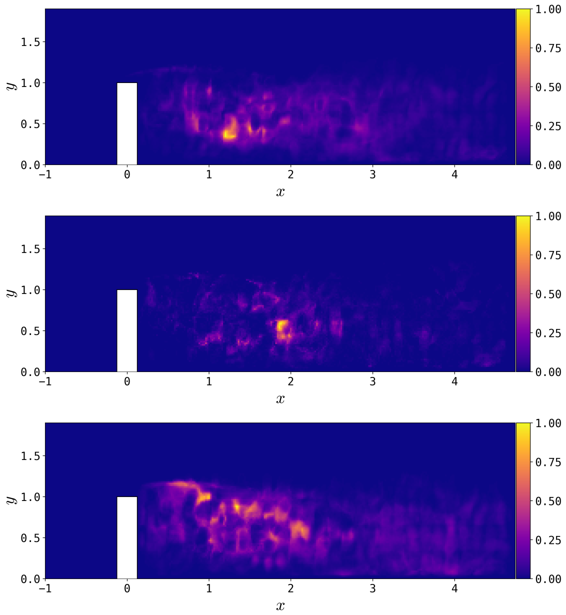


Figure 8. Percolated mean unique SHAP energy on the obstacle mid-span plane. Top to bottom: $STKE_1$, $STKE_3$, and $STKE_2$.

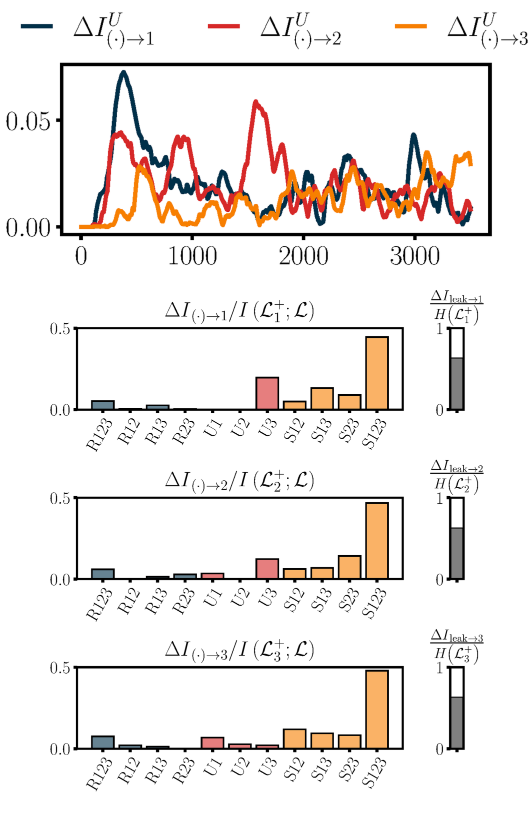


Figure 9. Obstacle SURD unique-causality summary. Top: total unique information versus lag. Bottom: unique histograms at $\tau_1 = 381$, $\tau_2 = 1569$, and $\tau_3 = 3103$.

here. The delayed $\ell_1 \to \ell_3$ feedback remains supported by the triadic graph and the obstacle state maps, but the clearest physical figure is the shared-source split of the two outgoing $\ell_3$ branches.

Across all cases, X-CAL reduces redundancy while retaining the dominant causal ordering, and SHAP provides a clear bridge from latent interactions to physical coherent struc-

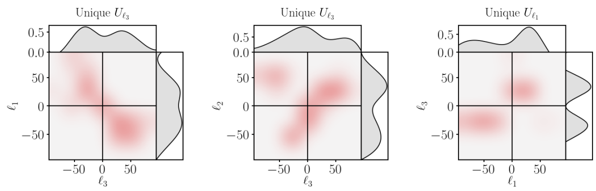

Figure 10. State-dependent SURD maps for the three dominant obstacle links. These raw state maps show that the unique obstacle interactions occupy compact but distinct latent-space sectors.

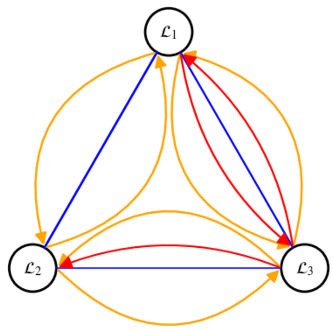


Figure 11. Global X-CAL graph for the obstacle case, built from the dominant interactions in the SURD decomposition. The red routes mark the dominant unique branches, while the remaining edges indicate the additional pathways retained in the global interaction graph.

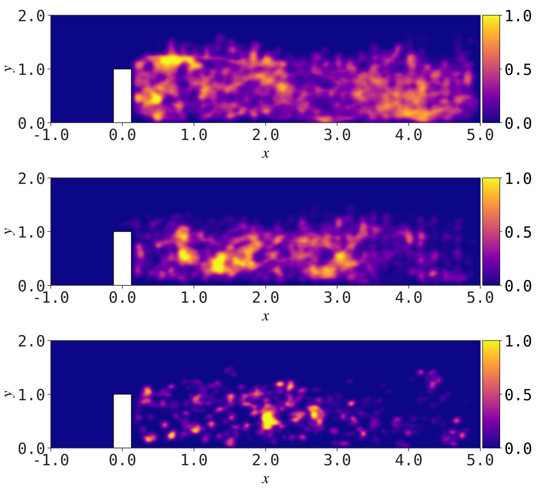


Figure 12. Smoothed, normalized occupancy of the unique SHAP support for the two outgoing unique branches from $\ell_3$; this is an attribution-support map, not a reconstructed velocity or TKE field. Top to bottom: shared source, delayed lower-wake support at $\Delta t_1 = 381$, and delayed mid-wake support at $\Delta t_2 = 1569$. Bright regions indicate the locations where the attribution remains most persistent across the selected events.

tures.

## CONCLUSIONS

X-CAL couples a convolutional $\beta$-VAE, SURD causal decomposition, and Gradient–SHAP/percolation into a workflow that turns flow snapshots into a causal graph with physically localized supports. Its main strength is that each stage is testable independently: reconstruction by the autoencoder, causal ordering by SURD, and physical interpretation by state maps and SHAP supports. On Lorenz, the framework recovers the inter-lobe symmetry-transition corridor in both latent and physical representations. On the wall-mounted cylinder, it separates three preferred mid-span wake sectors and links them through delays close to $T_{\mathrm{shed}}/9$, $T_{\mathrm{shed}}/2$, and $T_{\mathrm{shed}}$, yielding a compact control-oriented description in latent space and a spatially localized one in physical space. In this sense, X - CAL extends recent reduced-order prediction and closure efforts (Solera-Rico *et al.*, 2024; Eiximeno *et al.*, 2025) by adding causal ordering, compression with interpretation, and spatial explainability. The present evidence is limited to the mid-span plane; extending the analysis to full 3D wake fields is the natural next step.

## ACKNOWLEDGMENTS

R.V. acknowledges ERC grant 2021-CoG-101043998 (DEEPCONTROL) and the EU Doctoral Network MODELAIR. Computations used NAISS resources.

## REFERENCES

Cremades, Andrés, Hoyas, Sergio, Deshpande, Rahul, Quintero, Pedro, Lellep, Martin, Lee, W. J., Monty, Jason P., Hutchins, Nicholas, Linkmann, Moritz, Marusic, Ivan & Vinuesa, Ricardo 2024 Identifying regions of importance in wall-bounded turbulence through explainable deep learning. *Nature Communications* **15**, 3864.

Eiximeno, Benet, Sanchis-Agudo, Marcial, Miró, Arnau, Rodriguez, Ivette, Vinuesa, Ricardo & Lehmkuhl, Oriol 2025 On deep-learning-based closures for algebraic surrogate models of turbulent flows. *Journal of Fluid Mechanics* **1020**, A36.

Jiménez, Javier 2018 Coherent structures in wall-bounded turbulence. *Journal of Fluid Mechanics* **842**, P1.

Lorenz, Edward N. 1963 Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* **20** (2), 130–141.

Lumley, John L. 1967 The structure of inhomogeneous turbulent flows. In *Atmospheric Turbulence and Radio Wave Propagation* (ed. A. M. Yaglom & V. I. Tatarsky), pp. 166–178. Moscow: Nauka.

Lundberg, Scott M. & Lee, Su-In 2017 A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, , vol. 30, pp. 4765–4774.

Martínez-Sánchez, Álvaro, Arranz, Gonzalo & Lozano-Durán, Adrián 2024 Decomposing causality into its synergistic, unique, and redundant components. *Nature Communications* **15**, 9296.

Sanchis-Agudo, Marcial, Wang, Yuning, Arnau, Roger, Guastoni, Luca, Lim, Jasmin, Duraisamy, Karthik & Vinuesa, Ricardo 2025 Easy attention: A simple attention mechanism for temporal predictions with transformers. *APL Computational Physics* **1** (1), 016104.

Schmid, Peter J. 2010 Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics* **656**, 5–28.

Shapley, Lloyd S. 1953 A value for *n*-person games. In *Contributions to the Theory of Games* (ed. Harold W. Kuhn & Albert W. Tucker), , vol. 2, pp. 307–317. Princeton, NJ: Princeton University Press.

Solera-Rico, Alberto, Sanmiguel Vila, Carlos, Gómez-López, Marc, Wang, Yuning, Almashjary, Ahmad, Dawson, Scott T. M. & Vinuesa, Ricardo 2024 $\beta$-Variational autoencoders and transformers for reduced-order modelling of fluid flows. *Nature Communications* **15**, 1361.

Vinuesa, Ricardo, Cinnella, Paola, Rabault, Jean, Azizpour, Hossein, Bauer, Stefan, Brunton, Bingni W., Elofsson, Arne, Jarlebring, Elias, Kjellstrom, Hedvig, Markidis, Stefano, Marlevi, David, Garcia-Martinez, Javier & Brunton, Steven L. 2025 Decoding complexity: how machine learning is redefining scientific discovery.

Vinuesa, Ricardo, Schlatter, Philipp, Malm, Johan, Mavriplis, Catherine & Henningson, Dan S. 2015 Direct numerical simulation of the flow around a wall-mounted square cylinder under various inflow conditions. *Journal of Turbulence* **16** (6), 555–587.