Symmetry-Regularized Learning of Continuous Attractor Dynamics

Editors: List of editors' names

Abstract

Neural population dynamics exhibit rich geometric structure, yet prevailing computational models often overlook this by primarily accounting for variability in the data. We show that incorporating prior knowledge about dynamical symmetries yields efficient and interpretable models. Focusing on ring attractor dynamics—canonical circuits that are approximately equivariant under planar rotations—we introduce a symmetry-regularized variational state space model. Our method augments the standard variational objective with a symmetry penalty, encouraging the learned dynamical system to respect rotational invariance. We demonstrate that this regularization preserves predictive performance while yielding parsimonious models with interpretable latent dynamics. This framework establishes a principled approach for embedding symmetry priors into neural dynamical system learning, highlighting how exploiting geometric structure can improve both scientific insight and model generalization.

Keywords: Neural dynamics, continuous attractors, symmetry regularization, Lie brackets, variational inference

1. Introduction

Understanding neural population dynamics (in the dynamical system sense) is central to computational neuroscience, as these dynamics underlie the brain's ability to process, integrate, and maintain information over time. One striking example of neural population dynamics is the maintenance of continuous representations—ranging from head direction to spatial location—over time (Kim et al., 2017, 2019; Stringer et al., 2002; Gardner et al., 2022). Similarly, working memory also relies on the brain's ability to sustain information over seconds (Wimmer et al., 2014; Seeholzer et al., 2019). To explain how these representations can be maintained, theorists have proposed continuous attractor networks as a unifying framework. In these idealized models, activity patterns evolve along a continuum of fixed points, allowing the network to store continuous variables (Zhang, 1996; Seung, 1996; Wu et al., 2008; Fung et al., 2010).

A concrete instantiation of continuous attractor dynamics is found in ring attractor networks, which encode angular variables—such as head direction—through rotationally symmetric connectivity (Burak and Fiete, 2009; Hulse and Jayaraman, 2020; Noorman et al., 2024), enabling activity to move continuously along a circular manifold, effectively maintaining a continuous variable over time (e.g. Drosophila's central complex (Kim et al., 2017)). In biological circuits, however, such idealized dynamics are necessarily approximate (Park et al., 2023; Ságodi et al., 2024). Variability in synaptic strengths, heterogeneity among neurons, and constraints from development all introduce deviations from the perfect symmetry assumed in theoretical models (Shimizu et al., 2021). Such variability highlights that neural circuits—and therefore neural dynamics—are not static. For example, synaptic

strengths can drift over time while the circuit's underlying computational principle remains intact (Chirimuuta, 2024; Ságodi et al., 2024).

These observations motivate a modeling approach that abstracts away spurious details and emphasizes simple, symmetric structures. By incorporating symmetry constraints derived from theory, models can become more **temporally robust**, capturing the **core computation** despite ongoing variability. Furthermore, symmetries constrain the network's activity patterns and the transformations it can implement, reducing complexity and improving interpretability: it becomes easier to understand what computation the circuit implements and to compare neural computations across individuals and species.

Exploiting Symmetry State-of-the-art methods for neural system identification, such as XFADS (Dowling et al., 2024), excel at inferring a low-dimensional nonlinear dynamical system from observations but do not incorporate inductive biases, such as symmetry, in the inferred vector field.

In Yang et al. (2024), the authors introduce a general framework for embedding Lie symmetries into equation discovery pipelines, including sparse regression and genetic programming approaches. They show that continuous symmetries of a differential equation correspond to equivariance of its flow map, providing a principled criterion for enforcing invariance during model learning. For groups acting linearly on the state space, constraints can be solved explicitly to reduce the search space, while for nonlinear or unknown symmetries they propose a regularization strategy that penalizes deviations from symmetry. Our approach is inspired by this perspective: rather than discovering governing equations, we apply symmetry regularization on a flexible nonlinear dynamical system within the variational state space model, similarly exploiting geometric priors to obtain interpretable and robust latent dynamics. We develop a framework that augments variational state space models with continuous symmetry constraints, minimizing deviations from perfect symmetry in the learned dynamics to recover interpretable neural representations.

Contributions 1. We develop a general formulation for symmetry-regularized learning in variational inference of neural dynamical systems. 2. Through comprehensive experiments on ring attractor dynamics, we demonstrate that our method recovers interpretable, symmetry-preserving vector fields from noisy high-dimensional observations.

2. Symmetry Regularization

This section summarizes our approach, a detailed treatment is found in App. A.

We propose a framework that augments variational state space models with symmetric regularization terms based on a chosen continuous invariance. Following Yang et al. (2024), we enforce that learned vector field $f_\theta \colon \mathbb{R}^n \to \mathbb{R}^n$ approximately commute with generators $v \in \mathfrak{g}$ of a predefined symmetry group G (e.g., rotations for ring attractors). The degree of non-commutativity is measured by the ℓ^2 norm of the Lie bracket:

$$L_{\text{lie}} = \|\mathcal{L}_{f_\theta}\, v\|_2^2 = \|f_\theta \cdot Dv - v \cdot Df_\theta\|_2^2 \,. \tag{1}$$

When this quantity vanishes, the flows generated by f_θ and v commute, indicating that f_θ is preserved by G. By adding this term to the standard evidence lower bound (ELBO), we create a trade-off between data fitting accuracy and symmetry satisfaction.

To impose a more global rotational symmetry, we use the following group equivariance loss term to penalize deviations from equivariance:

$$L_{\text{group}} = \|f(g \cdot x) - g \cdot f(x)\|_2^2 , \tag{2}$$

where $g \in SO(2)$ denotes an element of the group of all rotations in $\mathbb{R}^2$, sampled randomly at each training step to encourage generalization across the full group. See App. A.2 for XFADS implementation and training details and App. C for regularization strength optimization details.

3. Experiments

We validate our approach on a canonical neuroscience problem: recovering ring attractor dynamics from high-dimensional noisy observations. We perturb the perfect ring attractor vector field by adding a smooth, spatially correlated random field. Specifically, a Gaussian process is sampled on a grid of points and interpolated via radial basis functions to produce bounded perturbation vectors. The resulting dynamics are $f(x) = x\,(1 - \|x\|) + h_{\text{pert}}(x)$, where the first term defines the baseline ring attractor and $h_{\text{pert}}(x)$ is the RBF-interpolated perturbation of controlled magnitude, resulting in an approximate ring attractor (Fig. 2A). Further details on the data generation procedure are provided in App. A.1.

For regularization, we use a pure rotation vector field: $g_{\text{target}}(\mathrm{x}) = \frac{1}{2}\begin{bmatrix} -x_2 & x_1 \end{bmatrix}^\top$, derived from the generator of $SO(2)$. This encourages the learned dynamics to exhibit rotational symmetry characteristic of ring attractors. The model is trained by combining the standard ELBO loss (Eq. 12) with the rotational regularization (Eq. 2): $L_{\text{total}} = L_{\text{ELBO}} + \lambda_{\text{group}} L_{\text{group}}$.

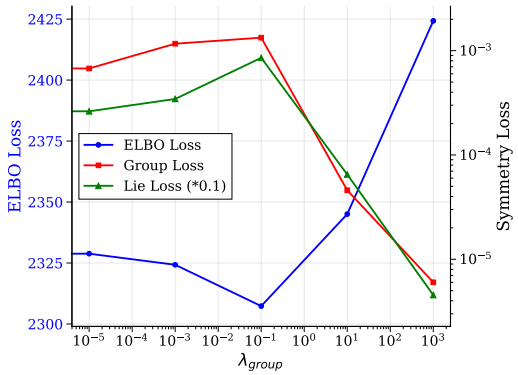

Figure 1: Trade-off between accuracy and symmetry regularization. Increasing the strength of the symmetry regularization improves symmetry but may eventually reduce accuracy. Local (L_{lie}, green) and global (L_{group}, red) symmetry are consistent, with values related by a simple scaling factor.

We evaluate model performance using multiple complementary metrics that capture both reconstruction quality and symmetry adherence. Data reconstruction is assessed via the ELBO loss, which quantifies how well the model reproduces observed trajectories. To evaluate the learned dynamics, we compare the inferred vector field to ground truth using vector field similarity, providing both visual and quantitative assessments. Symmetry preservation is measured in two ways: the Lie equivariance loss (L_{lie}) that calculates *infinitesimal* departures from symmetry using Lie derivatives, and the group equivariance loss (L_{group}) that quantifies semi-global departure from equivariance by *averaging* over G.

Symmetry regularization improves the model's adherence to rotational symmetry, as reflected in both the local L_{lie} and the global L_{group} (Fig. 1). Appropriate regularization preserves ring topology while enforcing rotational bias (Fig. 2), balancing reconstruction and symmetry (Fig. 1).

Training with local symmetry regularization, using the L_{lie} computed through the Jacobians (Eq. 14), did not result in symmetric vector fields for high regularization strengths (Fig. 4). Local symmetry regularization enforces only infinitesimal constraints via Jacobians, which appears insufficient for producing consistent global symmetry and can destabilize training, whereas global regularization directly enforces the full symmetry of the vector field.

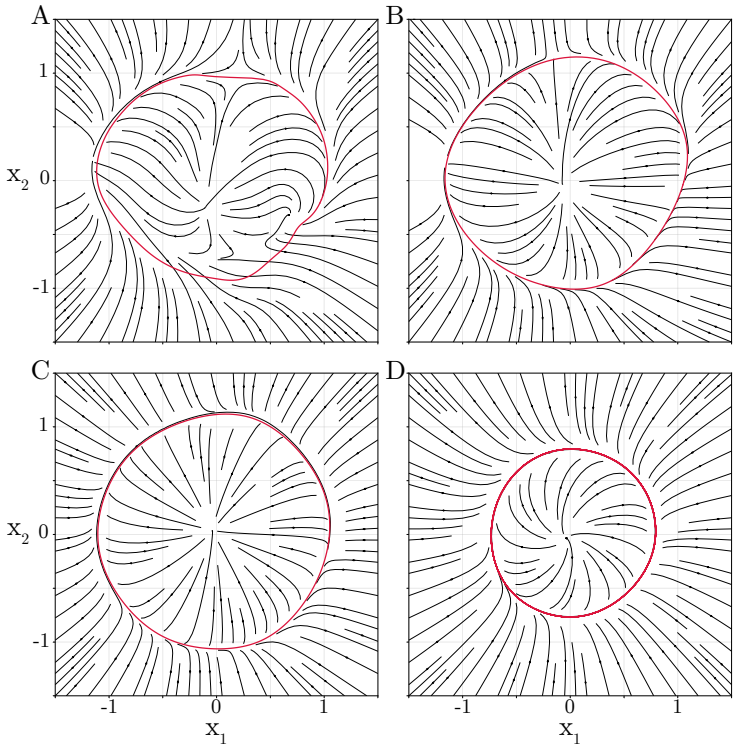

Figure 2: Flow fields for the target (perturbed ring attractor, A) versus the inferred vector fields with different symmetry regularization strengths (B-D) with the ring invariant manifolds (magenta, App. B.1). (B) Without symmetry regularization the inferred system inherits the asymmetries of the target system. (C) For a regularization strength that optimally trades-off some accuracy for more symmetry ($\lambda_{\text{group}} = 10^1$), the inferred system is close to a perfect ring attractor. (D) For the highly regularized inferred system ($\lambda_{\text{group}} = 10^3$), the vector field is symmetrical, but the manifold is rescaled.

4. Discussion

We address a fundamental challenge in computational neuroscience: extracting interpretable dynamical principles from complex, high-dimensional neural data. We present a framework for learning neural dynamics with geometric constraints, combining variational inference with symmetry regularization. Our approach successfully recovers interpretable, symmetry-preserving dynamics from high-dimensional observations and demonstrates that soft symmetry constraints via symmetry regularization provide an effective inductive bias for learning neural dynamics with discoverable symmetry. The Lie bracket formulation provides a foundation that could extend to other continuous symmetries beyond rotations, such as translational invariance in grid cell networks or scaling symmetries in sensory processing.

Recent theoretical work using persistent manifold theory shows that approximate continuous attractors with slow manifolds can perform analog memory tasks nearly as well as ideal attractors (Ságodi et al., 2024). Our regularization approach uses this insight by biasing learning towards symmetries, thereby smoothing out imperfections, instead of quantifying them (Ságodi and Park, 2025).

References

Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, 5(2):e1000291, 2009.

Mazviita Chirimuuta. *The Brain Abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press, 2024.

Matthew Dowling, Yuan Zhao, and Memming Park. eXponential FAmily Dynamical Systems (XFADS): Large-scale nonlinear Gaussian state-space modeling. *Advances in Neural Information Processing Systems*, 37:13458–13488, 2024.

CC Alan Fung, KY Michael Wong, and Si Wu. A moving bump in a continuous manifold: A comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Computation*, 22(3):752–792, 2010.

Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, 2022.

Benjamin K. Hulse and Vivek Jayaraman. Mechanisms underlying the neural computation of head direction. *Annual Review of Neuroscience*, 43:31–54, 2020.

Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853, 2017.

Sung Soo Kim, Ann M Hermundstad, Sandro Romani, LF Abbott, and Vivek Jayaraman. Generation of stable heading representations in diverse visual scenes. *Nature*, 576(7785): 126–131, 2019.

Marcella Noorman, Brad K. Hulse, Vivek Jayaraman, Sandro Romani, and Ann M. Hermundstad. Maintaining and updating accurate internal representations of continuous variables with a handful of neurons. *Nature Neuroscience*, Oct 2024. doi: 10.1038/s41593-024-01766-5. Epub ahead of print.

Il Memming Park, Ábel Ságodi, and Piotr Aleksander Sokół. Persistent learning signals and working memory without continuous attractors. August 2023.

Ábel Ságodi and Il Memming Park. Dynamical archetype analysis: Autonomous computation, 2025. URL https://arxiv.org/abs/2507.05505.

Ábel Ságodi, Guillermo Martín-Sánchez, Piotr A Sokol, and Il Memming Park. Back to the continuous attractor. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fvG6ZHrH0B.

Alexander Seeholzer, Moritz Deger, and Wulfram Gerstner. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS computational biology*, 15(4):e1006928, 2019.

H Sebastian Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, 1996.

Genki Shimizu, Kensuke Yoshida, Haruo Kasai, and Taro Toyoizumi. Computational roles of intrinsic synaptic dynamics. *Current opinion in neurobiology*, 70:34–42, 2021.

SM Stringer, TP Trappenberg, ET Rolls, and IETd Araujo. Self-organizing continuous attractor networks and path integration: One-dimensional models of head direction cells. *Network: Computation in Neural Systems*, 13(2):217–242, 2002.

Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431–439, 2014.

Si Wu, Kosuke Hamaguchi, and Shun-ichi Amari. Dynamics and computation of continuous attractors. *Neural computation*, 20(4):994–1025, 2008.

Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. *arXiv preprint arXiv:2310.00105*, 2023.

Jianke Yang, Wang Rao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry-informed governing equation discovery, 2024. URL https://arxiv.org/abs/2405.16756.

Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

Junfeng Zuo, Ying Nian Wu, Si Wu, and Wenhao Zhang. The motion planning neural circuit in goal-directed navigation as Lie group operator search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Qz7BfmWizk.

Appendix A. Implementation Details

A.1. Ring Attractor Data Generation

We generate synthetic data mimicking neural recordings from a perturbed ring attractor:

1. **Ideal dynamics**: Start with a perfect ring attractor:

$$f(x) = x\,(1 - \|x\|) \tag{3}$$

2. **Perturbation**: Add Gaussian process noise to simulate biological variability:

$$f_{\text{true}}(x) = f_{\text{ring}}(x) + \epsilon \cdot \delta f(x), \quad \delta f \sim \mathcal{GP}(0, k_{\text{RBF}}) \tag{4}$$

3. **Initial latent state**: We generate initial points for the latent state x_0 from a multivariate Gaussian distribution:

$$x_0 \sim \mathcal{N}(m,\, \sigma\,\mathrm{Id}), \tag{5}$$

where $m \in \mathbb{R}^n$ is the mean and σ is the variance.

4. **Latent state dynamics**: The system state evolves according to the perturbed vector field:

$$dx = f_{\text{true}}(x_t)dt + dB \tag{6}$$

or in discrete time:

$$x_{t+1} = x_t + f_{\text{true}}(z_t)\,\Delta t + \sqrt{\Delta t}\eta \tag{7}$$

for $\eta \sim \mathcal{N}(0, 1)$.

5. **Observations**: Linear readout with Gaussian noise:

$$y_t = Cz_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, R), \quad C \in \mathbb{R}^{100\times 2} \tag{8}$$

A.2. XFADS Framework

We build upon the XFADS variational state space model, which posits:

$$\text{Latent dynamics:} \quad z_t \sim \mathcal{N}(f_\theta(z_{t-1}), Q) \tag{9}$$

$$\text{Observations:} \quad y_t \sim p(y_t|z_t) \tag{10}$$

$$\text{Inference:} \quad q(z_{1:T}|y_{1:T}) \approx \prod_t \mathcal{N}(\mu_t, \Sigma_t) \tag{11}$$

The model is trained by maximizing the evidence lower bound (ELBO):

$$L_{\text{ELBO}} = \mathbb{E}_q[\log p(y_{1:T}|z_{1:T})] - \text{KL}[q(z_{1:T})||p(z_{1:T})] \tag{12}$$

where the expectation is approximated via Monte Carlo sampling and the KL divergence has a closed form for Gaussian distributions.

Our experiments use the following architecture:

- **Dynamics**: GRU with 32 hidden units
- **Local encoder**: MLP $[100 \to 64 \to 32 \to 4]$ (rank-2 output)
- **Backward encoder**: MLP $[100 \to 64 \to 32 \to 4]$ (rank-2 output)
- **Likelihood**: Gaussian with learned diagonal covariance

Training Hyperparameters

- Optimizer: Adam with learning rate 10^{-3}
- Batch size: 32 trials
- Maximum epochs: 25
- Early stopping: Patience 5 on validation ELBO
- Monte Carlo samples: 5 for training, 10 for evaluation
- Gradient clipping: Norm threshold 1.0

Appendix B. Analysis Methods

B.1. Approximation Of the Invariant Manifold

To approximate the invariant manifold, we focus on the convergence of the flow under the perturbed vector field. By simulating trajectories starting from various initial conditions, we identify regions where the dynamics quickly converge, revealing the low-dimensional structure that captures the long-term behavior of the system. This approach allows us to characterize the manifold without requiring an explicit analytical solution.

B.2. Lie Bracket Computation

For completeness, we provide the full derivation of the Lie bracket in coordinates. Given vector fields $f = (f^1, \ldots, f^n)$ and $g = (g^1, \ldots, g^n)$:

$$\mathcal{L}_f\, g^i = \sum_{j=1}^{n} \left(f^j \frac{\partial g^i}{\partial x^j} - g^j \frac{\partial f^i}{\partial x^j} \right) \tag{13}$$

For vector fields $f, g : \mathbb{R}^n \to \mathbb{R}^n$, the Lie bracket measures the failure of their (local) flows to commute:

$$\mathcal{L}_f\, g = f \cdot Dg - g \cdot Df, \tag{14}$$

where Df denotes the Jacobian of f. When this quantity vanishes, the flows generated by f_θ and v commute, indicating that f_θ is preserved by G.

Appendix C. Extended Results

Regularization Strength Sweep We conducted a comprehensive sweep across the following regularization strengths for L_{group} and L_{lie}:

$$\lambda_{\text{group}} \in \{0, 10^{-5}, 10^{-3}, 10^{-1}, 10^{1}, 10^{3}\}, \tag{15}$$

$$\lambda_{\text{lie}} \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\} \tag{16}$$

For each value, we trained models with:

- 1000 trials $\times$ 75 time bins $\times$ 100-dimensional observations
- Early stopping based on validation ELBO (patience = 5)
- Fixed random seed for reproducibility
- 5 Monte Carlo samples for variational inference

Our systematic parameter sweep reveals a tradeoff of ELBO loss and symmetry loss with increasing λ_{group} (Fig. 1). We plot the final flow fields for each run (Fig. 3, 4), noting the instability of inferred vector fields for the Lie algebra regularization sweep. This suggests that local symmetry constraints may be insufficient for capturing the global geometric structure of ring attractors. The choice of symmetry regularization may need to align with the spatial scale of the target dynamical system's geometric properties.

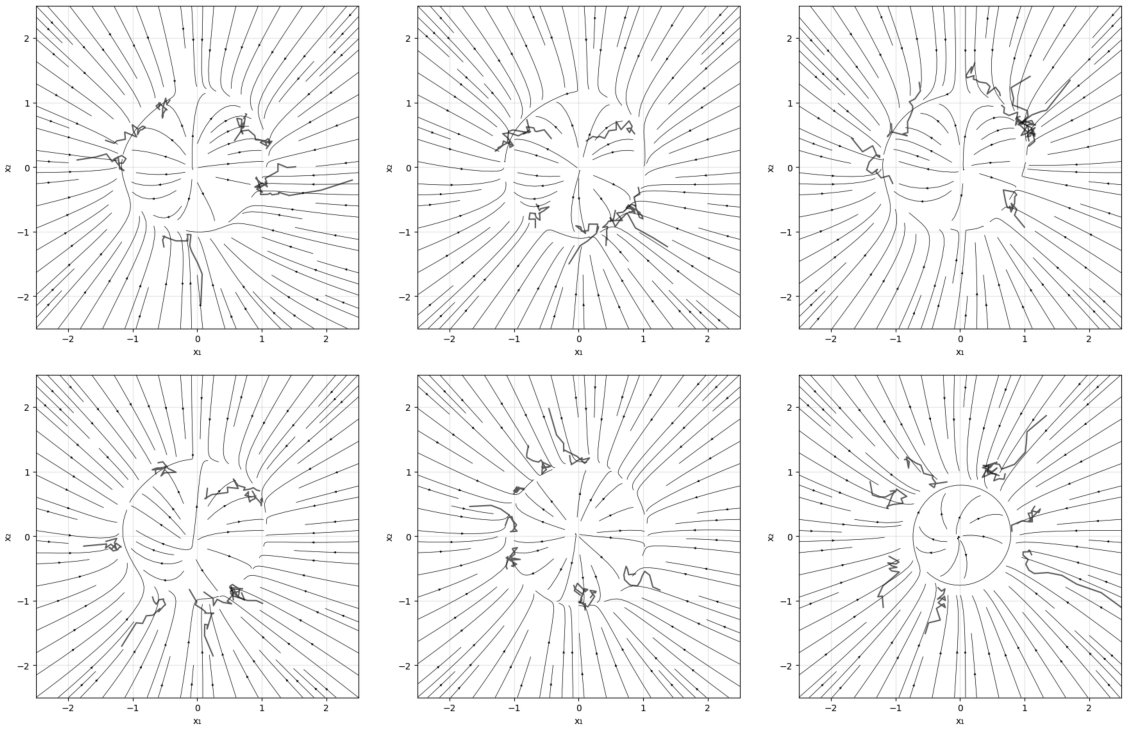

Figure 3: Inferred flow fields with different regularization strengths with short trajectories sampled from random initial conditions. From left to right: $\lambda_{\text{group}} \in \{0, 10^{-5}, 10^{-3}, 10^{-1}, 10^{1}, 10^{3}\}$

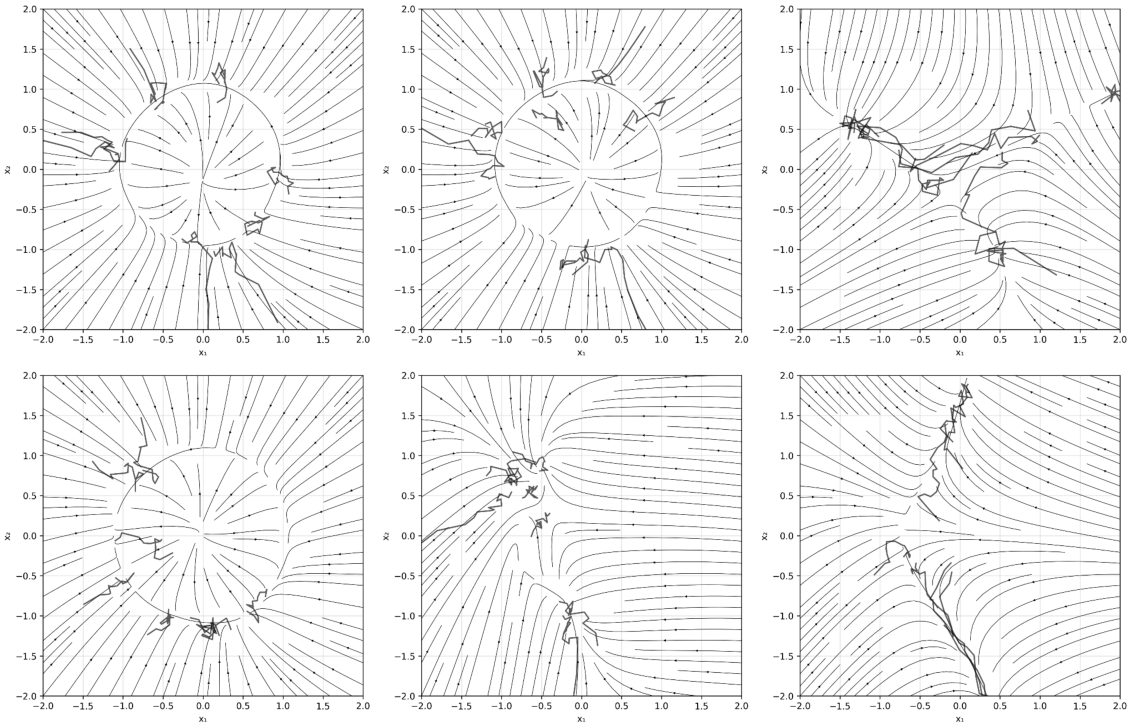

Figure 4: Inferred flow fields with different regularization strengths with short trajectories sampled from random initial conditions. From left to right: $\lambda_{\text{lie}} \in \{0, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}\}$

Appendix D. Limitations

Our current validation focuses on 2D ring attractors, and extending the framework to higher-dimensional manifolds, such as tori or spheres, will require addressing the associated computational complexity. Validation on experimental neural recordings introduces further challenges, including missing data, and unknown ground truth for the symmetry. Moreover, automatically identifying underlying symmetries remains an open challenge, which could be approached either by searching over the group space (Zuo et al., 2024) or by leveraging equivariant neural network architectures (Yang et al., 2023).