KOOPMAN-STABILISED WORLD MODELS FOR OFFLINE REINFORCEMENT LEARNING IN ACCELERATOR CONTROL

S. Hirlaender*, O. Mironova, S. Trausner, University of Salzburg, Salzburg, Austria
L. Fischl, EBG MedAustron GmbH, Wiener Neustadt, Austria
L. Grech, University of Malta, Msida, Malta
A. Santamaria Garcia, University of Liverpool, Liverpool, UK

Abstract

Particle accelerators generate vast amounts of historical data, yet learning-based control still heavily relies on risky online optimisation. To better utilise this data and avoid online exploration, we present an offline reinforcement learning (RL) workflow. First, we use XSuite to generate high-fidelity steering trajectories across representative scenarios, yielding a synthetic dataset of expert and non-expert behaviour. Second, we learn an uncertainty-aware, Koopman-stabilised world model, lifting nonlinear beam dynamics into a latent space with approximately linear, spectrally constrained evolution. This provides stable long-horizon rollouts and estimates of epistemic uncertainty.

The resulting surrogate enables model-based offline RL, optimising policies entirely on pre-generated data while using epistemic uncertainty to penalise out-of-distribution behaviour. We benchmark these policies against an online Proximal Policy Optimisation (PPO) agent. Results show that policies trained purely offline on our Koopman world model drastically outperform standard online PPO without requiring any real-machine interaction, demonstrating a safe pathway for turning historical accelerator data into effective control policies.

INTRODUCTION

Accelerator facilities are "data-rich" [1]: decades of operational logs exist, yet reproducing nonlinear beam dynamics in simulation remains expensive. Here we use high-fidelity XSuite simulations as a proxy for such operational data. Consequently, RL agents are typically trained online [2, 3], risking beam loss and downtime.

Figure 1 illustrates the problem: standard multilayer perceptrons (MLPs) produce spurious energy gains, violating Liouville's theorem and making them unsafe for long-horizon Hamiltonian control. Hamiltonian Neural Networks (HNNs) [4] enforce symplectic structure to produce only physically plausible phase drift. While HNNs require explicit Hamiltonian decomposition, the *Koopman* framework [5] achieves similar spectral guarantees for arbitrary systems by lifting dynamics to a linear latent space, fitting high-dimensional, partially observed beam dynamics.

We propose **S^{2}K-MOPO** (Stable Spectral Koopman Model-based Offline Policy Optimisation): an offline RL framework that builds a high-fidelity surrogate from data via a *World Model* [6], encoding known accelerator physics as structural inductive biases. Unlike modern sequence models [7, 8] that lack spectral guarantees and suffer artificial damping, we use a **Stable Koopman State-Space Model** [9, 10] ensuring linear, spectrally bounded latent evolution.

Contributions: (1) Volume-preserving Koopman operator ($K \in SO(d)$) via Cayley transform; (2) Lie-algebra ensemble averaging; (3) Stein Variational Gradient Descent (SVGD) regularised epistemic uncertainty for S^{2}K-MOPO penalisation; (4) Offline validation on a Focusing–Defocusing (FODO) ring and AWAKE.

METHODOLOGY

The pipeline (Fig. 2) maps a **Nonlinear AutoRegressive eXogenous (NARX) history** s_t → **Non-linear Independent Components Estimation (NICE) encoder** ψ → **Cayley Koopman** K → **NICE decoder** ψ^{-1} → predicted orbit $\hat{s}_{t+1}$. An SVGD ensemble of N=3 particles provides epistemic uncertainty for the S^{2}K-MOPO penalty.

Physics-Informed Koopman Operator

A NARX stack of past Beam Position Monitor (BPM) readings and actions (s_t) is mapped to latent $z_t = \psi(s_t)$ via **NICE coupling layers** [11]. NICE layers have a unit Jacobian ($|\det J_\psi| \equiv 1$), ensuring exact volume preservation of the augmented state by construction—a structural proxy for **Liouville's theorem** preventing topological failure. Since NICE layers are invertible, the decoder is the exact analytical inverse $s = \psi^{-1}(z)$. Past actions causally separate beam momentum from corrector kicks. Latent evolution follows:

$$z_{t+1} = K\, z_t + B\, a_t,$$

mirroring the one-turn map: $K \in SO(d)$ encodes lattice optics and B linearly maps corrector kicks, reflecting the physical transfer-matrix structure of steering magnets. K is unconditioned on a_t; nonlinear actuator coupling is deferred to future work.

The conservative baseline constrains K via the Cayley transform of skew-symmetric A:

$$K = (I - A)(I + A)^{-1}, \quad K^T K = I.$$

For non-conservative effects (radiation damping, resistive-wall losses), a leaky extension $K_d = D \cdot K$, with learnable $D \in (0, 1]^d$, retains unconditional stability ($|\lambda_i| \leq 1$) while admitting physically motivated dissipation.

* simon.hirlaender@plus.ac.at

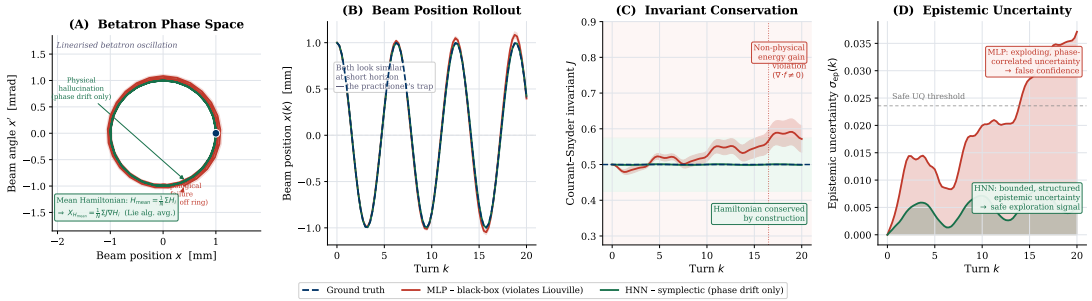

Figure 1: Standard MLP vs. HNN on a linearised betatron oscillator. *(A)* Phase space. *(B)* Rollout. *(C)* Invariant: MLP violates Liouville's theorem. *(D)* Uncertainty: MLP explodes; HNN remains bounded.

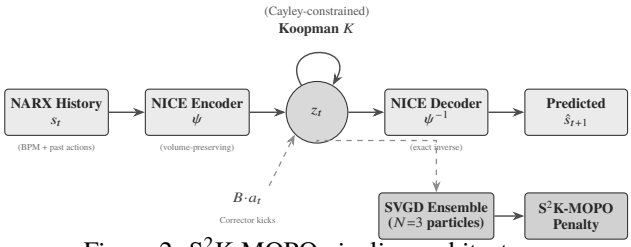

Figure 2: S^{2}K-MOPO pipeline architecture.

SVGD Ensemble & Lie-Algebra Averaging

We employ SVGD [12] to maintain N=3 particles approximating the model posterior. Unlike independent ensembles, SVGD's repulsive kernel maximally disperses particles in weight space, providing diversity disproportionate to the particle count. Direct averaging of K matrices destroys orthogonality (det $\bar{K}$<1). We instead average in the Lie algebra: $\bar{A} = \frac{1}{N}\sum_i A_i$, recovering

$$K_{\text{mean}} = (I - \bar{A})(I + \bar{A})^{-1} \in SO(d),$$

which serves as a first-order approximation of the Fréchet mean on $SO(d)$ [13]. Figure 3 illustrates this: standard averaging artificially damps to zero, while Lie-algebra averaging preserves oscillatory dynamics and zero volume drift.

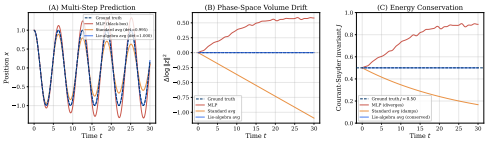

Figure 3: Symplectic vs. standard ensemble averaging. Standard damps to zero; Lie-algebra preserves zero volume drift.

Model-Based Offline RL (S^{2}K-MOPO)

We train policy π_ϕ via S^{2}K-MOPO [14], using ensemble disagreement as an epistemic safeguard against out-of-distribution actions. Let $f_\theta(z, a)$ denote the world-model next-state prediction; high disagreement across ensemble members triggers an automatic penalty:

$$r_{\text{S}^2\text{K}}(z, a) = r\big(s_{\text{dec}}\big) - \lambda \cdot \max_k \text{std}_\theta[f_\theta(z, a)_k],$$

confining the policy to historically validated strategies.

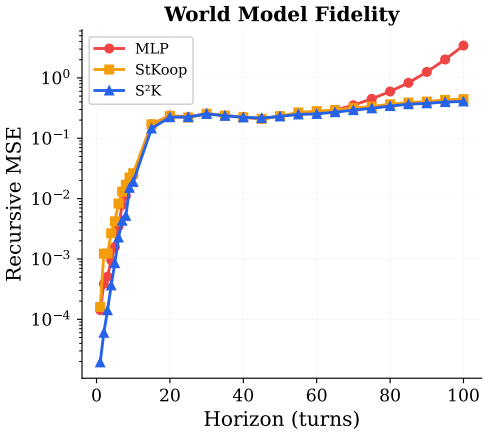

Figure 4: World model fidelity: recursive MSE over 100 turns. MLP diverges while S^{2}K variants stay bounded.

EXPERIMENTAL SETUP

Control task. At each step the agent reads 10 horizontal + 10 vertical BPM values and applies 10 corrector kicks (±50 µrad). A NARX window of W=3 past observations and $W-1$ past actions forms an 80-dim state ($s_t \in \mathbb{R}^{80}$), providing the history needed for non-Markovian dynamics. The reward (root mean square, RMS) is $r = -\text{RMS}(s_t) - 0.01\|a_t\|^2 - 0.01\|a_t - a_{t-1}\|^2 + 0.1 \cdot \text{I}_{\{\text{RMS}<0.1\,\text{mm}\}}$, with discount γ=0.95.

Benchmarks. *(I) XSuite FODO Ring* [15]: 10 FODO cells (k_1=0.08 m^{-2}), quadrupole alignment errors (200 µm RMS), BPM noise (50 µm RMS), optics jitter (±5% k_1); 15 000 offline transitions (50% expert SVD, 50% random exploration), following an analogous offline dataset protocol to D4RL [16]. *(II) AWAKE Electron Line* [17]: CERN TT43 single-pass line with 10 BPMs and 10 correctors; lower-triangular response matrix from MAD-X Twiss optics; quadrupole strengths randomised ±25% across episodes; only 1 000 transitions.

Baselines. Analytical SVD (oracle: $a = -R^+ s$); Latent Linear–Quadratic Regulator (LQR, infinite-horizon on learned K, B). Note that SVD requires a known, static, linear response matrix R and full observability—assumptions that break under model mismatch, nonlinearity, and partial sensor coverage. These are precisely the regimes where a learned policy becomes essential.

Key hyperparameters: latent dim d=48, NARX window W=3, N=3 SVGD particles, LR 3×10^{-4}, batch 256, 50 WM epochs, 50 k RL steps, MOPO penalty λ=1.0.

RESULTS

World Model Fidelity: At H=100, the MLP diverges (mean squared error, MSE=3.43) while both Koopman variants stay bounded: StKoop (standard Koopman without SVGD) reaches ≈0.1 and S^{2}K (full SVGD ensemble) ≈0.04 (Fig. 4). Naïve averaging causes volume drift of 456 over 200 turns; Lie-algebra averaging limits volume drift to 2.07. For the **Courant–Snyder invariant**, the MLP diverges, but Pure Cayley ($|\lambda_i|$=1) tracks ground truth perfectly. The leaky extension adds physical decoherence.

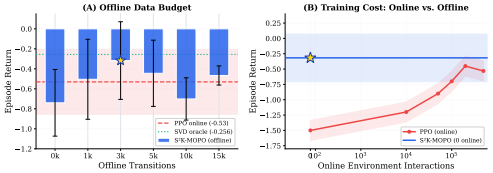

Figure 5: Sample efficiency vs. offline data budget. S^{2}K-MOPO requires zero online steps.

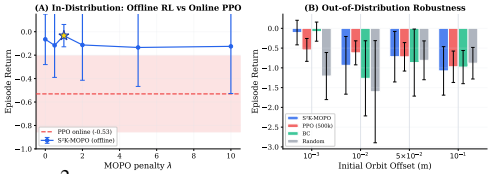

Figure 6: S^{2}K-MOPO (offline, zero online steps) vs. PPO (5×10^5 online steps) across initial orbit offsets.

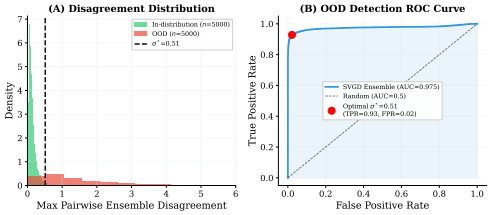

Figure 7: OOD detection via SVGD uncertainty (AUC=0.975).

Offline RL: Figure 5 shows 3,000 offline transitions suffice to surpass PPO [18]. S^{2}K-MOPO achieves **−0.033±0.095** from purely offline data vs. −0.531±0.324 for PPO with 5×10^5 online steps (Fig. 6). The offline dataset includes 50% expert SVD demonstrations, biasing the comparison; the key result is policy extraction from a fixed dataset without online interaction. The SVD oracle (−0.0002) remains superior in the nominal linear regime but requires a known response matrix. On **AWAKE**, 1,000 transitions yield world-model MSE 1.4×10^{-4}. Epistemic disagreement achieves **ROC-AUC**=0.975 for OOD detection (Fig. 7).

Physics Extraction: Filtering Koopman eigenvalue harmonics (Fig. 8) recovers physical decoherence ($|\lambda|$≈0.993) and **betatron tunes** (1.0% error vs. MAD-X).

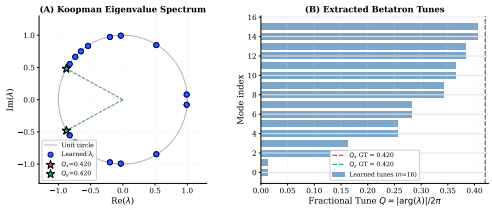

Figure 8: Learned Koopman eigenvalue spectrum recovers betatron tunes (1.0% error vs. MAD-X).

Table 1: Stress tests: control paradigms and sensor failure.

Experiment	Metric	Value
Control Paradigm Comparison		
S^{2}K-MOPO (offline RL)	Return	**−0.033 ± 0.095**
Online PPO (5×10^5 steps)	Return	−0.531 ± 0.324
Analytical SVD (oracle)	Return	−0.0002 ± 0.001
Latent LQR (learned K)	Return	−0.076 ± 0.274
Broken-BPM Robustness (50% broken)		
S^{2}K-MOPO (pristine)	Return	−0.468 ± 0.697
S^{2}K-MOPO (50% broken)	Return	**−0.540 ± 0.638**
SVD controller (pristine)†	Return	−0.942
SVD controller (50% broken)*†	Return	−0.288

*Masking zero-pads inputs, artificially inflating the reward metric.

†Single-seed deterministic controller (no stochastic variation).

Stress tests (Table 1): Larger initial offsets and hardware jitter explain the higher absolute returns. Unconstrained LQR violates actuator limits, performing worse than "No Control." S^{2}K-MOPO natively respects limits and degrades by only 15% under 50% sensor failure; the SVD controller's apparent gain is an artefact of zero-padded inputs lowering the RMS penalty.

CONCLUSION

S^{2}K-MOPO embeds accelerator physics constraints as structural priors into neural architectures via volume-preserving NICE encoders, Cayley-constrained Koopman operators, and SVGD ensembles for epistemic uncertainty. These physics-informed constraints serve as structural safeguards that improve reliability within the training distribution, though they do not formally certify correctness under arbitrary out-of-distribution conditions. While SVD remains superior on nominal benchmarks, S^{2}K-MOPO operates from offline data alone, degrades gracefully under model mismatch and sensor failure [19], outperforms online PPO, and extracts betatron tunes (1.0% error). Future work targets nonlinear extensions [5], offline RL baselines [20], and validation on real operational data.

ACKNOWLEDGEMENTS

Supported by WISS 2025 'IDA Lab Salzburg' (20102/F2300464-KZP, 20204-WISS/225/348/3-2023).

REFERENCES

[1] A. L. Edelen *et al.*, "Neural Networks for Modeling and Control of Particle Accelerators", *IEEE Trans. Nucl. Sci.*, vol. 63, pp. 878–897, 2016. doi:10.1109/TNS.2016.2543203

[2] S. Hirlaender and N. Bruchon, "Model-Free and Bayesian Ensembling Model-Based Deep RL for Particle Accelerator Control Demonstrated on the FERMI FEL", 2020. doi:10.48550/arXiv.2012.09737

[3] N. Bruchon *et al.*, "Basic Reinforcement Learning Techniques to Control the Intensity of a Seeded Free-Electron Laser", *Electronics*, vol. 9, p. 781, 2020. doi:10.3390/electronics9050781

[4] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian Neural Networks", *NeurIPS*, vol. 32, 2019.

[5] B. Lusch, J. N. Kutz, and S. L. Brunton, "Deep Learning for Universal Linear Embeddings of Nonlinear Dynamics", *Nat. Commun.*, vol. 9, p. 4950, 2018. doi:10.1038/s41467-018-07210-0

[6] D. Ha and J. Schmidhuber, "Recurrent World Models Facilitate Policy Evolution", *NeurIPS*, vol. 31, 2018. doi:10.48550/arXiv.1809.01999

[7] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces", *ICLR*, 2022. doi:10.48550/arXiv.2111.00396

[8] P. Kidger *et al.*, "Neural Controlled Differential Equations for Irregular Time Series", *NeurIPS*, 2020.

[9] B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space", *PNAS*, vol. 17, no. 5, pp. 315–318, 1931. doi:10.1073/pnas.17.5.315

[10] S. L. Brunton *et al.*, "Modern Koopman Theory for Dynamical Systems", *SIAM Rev.*, vol. 64, pp. 229–340, 2022. doi:10.1137/21M1401243

[11] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear Independent Components Estimation", *ICLR Workshop*, 2015. doi:10.48550/arXiv.1410.8516

[12] Q. Liu and D. Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm", *NeurIPS*, pp. 2370–2378, 2016.

[13] M. Moakher, "Means and averaging in the group of rotations", *SIAM J. Matrix Anal. Appl.*, vol. 24, pp. 1–16, 2002. doi:10.1137/S0895479801383877

[14] T. Yu *et al.*, "MOPO: Model-based Offline Policy Optimization", *NeurIPS*, vol. 33, pp. 14129–14142, 2020.

[15] G. Iadarola *et al.*, "Xsuite: An Integrated Beam Physics Simulation Framework", in *Proc. 68th Adv. Beam Dyn. Workshop High-Intensity High-Brightness Hadron Beams (HB'23)*, Geneva, Switzerland, Oct. 2023, pp. 73–80. doi:10.18429/JACoW-HB2023-TUA2I1

[16] J. Fu *et al.*, "D4RL: Datasets for Deep Data-Driven Reinforcement Learning", 2020. doi:10.48550/arXiv.2004.07219

[17] S. Hirlaender *et al.*, "Towards few-shot reinforcement learning in particle accelerator control", in *Proc. IPAC'24*, Nashville, TN, May 2024, pp. 1804–1807. doi:10.18429/JACoW-IPAC2024-TUPS60

[18] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms", 2017. doi:10.48550/arXiv.1707.06347

[19] J. Degrave *et al.*, "Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning", *Nature*, vol. 602, pp. 414–419, 2022. doi:10.1038/s41586-021-04301-9

[20] S. Levine, A. Kumar, G. Tucker, J. Fu, "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems", 2020. doi:10.48550/arXiv.2005.01643