Fermilab

Digital twin framework for PIP-II linac: AI-driven multi-scale modeling from ion source to 800 MeV

FERMILAB-CONF-25-1035-AD-PIP2

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-WEPD097

DIGITAL TWIN FRAMEWORK FOR PIP-II LINAC: AI-DRIVEN MULTI-SCALE MODELING FROM ION SOURCE TO 800 MeV*

A. Pathak†, P. Hanlet, and T. Miceli
Fermi National Accelerator Laboratory, Batavia, IL, USA

Abstract

The PIP-II linac will enable > 1.2 MW beam power for DUNE, requiring unprecedented operational reliability across its warm front end (RFQ, MEBT) and five distinct SRF sections operating at 162.5/325/650 MHz. We present a comprehensive digital-twin framework that combines a fully differentiable fast beam-transport core with neural-network surrogates trained on high-fidelity PIC simulations, capturing space charge and nonlinear dynamics beyond traditional envelope codes while achieving $10^4\times$ speedup at < 1% accuracy. End-to-end differentiability enables gradient-based optimization across 500+ parameters simultaneously, previously impractical with conventional tools, while the model incorporates static/dynamic errors and serves as a virtual-commissioning platform for diverse hardware integration. The framework facilitates reinforcement learning for pulsed/CW mode transitions, predictive maintenance through anomaly detection, and autonomous tuning algorithm development with real-time execution capability. Validation against physics simulations shows excellent agreement for the front end, with initial results indicating potential for 30% commissioning-time reduction and proactive fault mitigation, providing a scalable blueprint for operating next-generation high-intensity accelerators.

INTRODUCTION

The PIP-II superconducting linac is being built to deliver proton beam power in excess of 1.2 MW to Fermilab experiments using a chopped bunch structure and a sequence of multi-frequency SRF cavities. Achieving and sustaining such operating points requires tight control of uncontrolled losses at the $O(1\,\mathrm{W/m})$ level, while mitigating SRF microphonics and LLRF amplitude/phase jitter that perturb longitudinal capture and transverse stability [1–3]. In this intensity regime, seemingly minor deviations in cavity phase/gradient or in magnetic lattice settings can seed emittance dilution and halo growth, and slow technical drifts rapidly erode the predictive value of static optics models.

Conventional online modeling tools used during commissioning sit at two ends of a spectrum. Fast linear or second–moment descriptions are responsive enough for control-room use but miss strong-coupling, space-charge, and field-map effects that dominate at high current. High-fidelity PIC or detailed field-map tracking captures the relevant physics, yet their turnaround time is typically incompatible with interactive setup and anomaly triage. In parallel, the facility continuously produces time-stamped diagnostics that are seldom fused with models in an uncertainty-aware fashion to enable reliable state estimation, anomaly detection, and safe optimization [4].

We address this gap with a controls-integrated digital twin (DT) that couples the physical machine (PT) to a fast, differentiable predictor synchronized to live signals via EPICS [5]. The DT advances a reduced machine state using a structured update $\Phi_\theta = M \circ K_\theta \circ E$, in which a symplectic transport core M is complemented by a learned corrective operator K_θ and a disturbance injector $E(\xi)$ representing measured or inferred perturbations. Figure 1 provides an overview, while Eqs. (1) and (2) formalize the state evolution and the tuning objective used for commissioning. Routine tasks are posed as constrained inverse problems that trade data agreement against actuator effort and machine-protection limits, enabling both advisory (digital-shadow) and guarded closed-loop operation. This approach aligns with efforts to bridge differentiable simulation and accelerator modeling for controls and optimization [6].

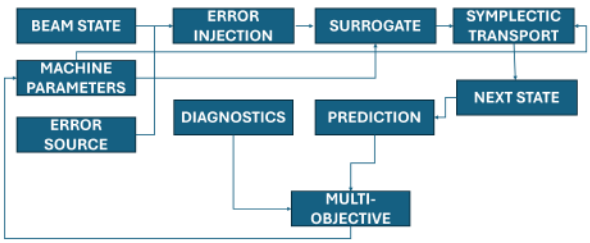

Figure 1: Schematic of the differentiable update $x_{k+1} = M(p_k) \circ K_\theta(x_k, p_k) \circ E(\xi_k)$ and of the objective $J(\mathbf{p})$ that compares virtual diagnostics $\hat{\mathbf{y}}(s;\mathbf{p})$ against measurements $\mathbf{y}_{\mathrm{obs}}(s)$ in Eq. (2).

A translator layer reconciles lattice descriptions with control-system PVs (naming, frames, units, signs), allowing PV snapshots to be round-tripped through the model for residual attribution and model-to-machine alignment. Central to the implementation is an accelerated engine, LINAC_GEN, which blends physics-based transport with compact, conditioned surrogates trained on field-map/TraceWin trajectories and curated machine data. Together, these components furnish a practical path to faster, safer commissioning and robust high-intensity running: proposed changes are evaluated in silico before application, uncertainties gate risky moves, and multi-objective strategies co-tune trajectory, optics, and RF within loss and power envelopes.

* This manuscript has been authored by Fermi Forward Discovery Group under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.
† abhishek@fnal.gov

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-WEPD097

DIGITAL TWIN FRAMEWORK

Problem Formulation

Let $\mathbf{p} \in \mathbb{R}^{N_p}$ denote the vector of machine settings (RF amplitudes/phases, magnet gradients, corrector setpoints, chopper timing), and let $\mathbf{x}$ represent the reduced beam state used online. The twin advances the state with a differentiable update

$$\mathbf{x}_{k+1} = \Phi_\theta(\mathbf{x}_k, \mathbf{p}_k, \xi_k) = M(\mathbf{p}_k) \circ K_\theta(\mathbf{x}_k, \mathbf{p}_k) \circ E(\xi_k), \quad (1)$$

where M is a symplectic transport core, K_θ is a learned corrective operator trained on high-fidelity trajectories, and $E(\xi_k)$ injects disturbance channels ξ_k (e.g., misalignments, RF amplitude/phase jitter, microphonics, supply ripple).

Tuning tasks are posed as constrained inverse problems that compare virtual diagnostics to measurements and penalize mismatch and loss proxies:

$$\min_{\mathbf{p}} J(\mathbf{p}) = \sum_{s \in \mathcal{S}} \Big[w_{\text{orb}} \big\| \mathbf{y}_{\text{obs}}(s) - \hat{\mathbf{y}}(s; \mathbf{p}) \big\|_2^2 + w_{\text{long}} \mathcal{L}_{\text{long}}(s) + w_{\text{loss}} \mathcal{R}_{\text{loss}}(s) \Big], \quad (2)$$

with $\mathcal{S}$ indexing BPM/TOF/toroid locations, $\hat{\mathbf{y}}$ denoting twin-predicted observables, $\mathbf{y}_{\text{obs}}$ the corresponding measurements, $\mathcal{L}_{\text{long}}$ a penalty on phase/energy errors, and $\mathcal{R}_{\text{loss}}$ a surrogate for uncontrolled losses (e.g., halo mass beyond $n\sigma$). Robust formulations evaluate J under sampled ξ (expectation or CVaR_α) without optimizing over ξ. *Figure 1 summarizes the mapping in Eqs. (1)–(2) and the associated objective.*

Learned Physics Operator

The operator K_θ provides a differentiable correction that captures collective and RF nonlinear effects not represented by M. In single-particle coordinates $(\mathbf{r}_i, \pi_i)$,

$$\Delta\pi_i = f_\theta(\mathbf{r}_i, \mathbf{p}, \mathbf{c}), \qquad \Delta\mathbf{r}_i = g_\theta(\mathbf{r}_i, \mathbf{p}, \mathbf{c}), \quad (3)$$

where $\mathbf{c}$ encodes local lattice/field descriptors and $\mathbf{p}$ are machine settings (not canonical momenta). The networks are conditioned and compact, share weights across element classes, and are regularized to preserve smooth parameter dependence and stable updates. Training minimizes

$$\mathcal{L}(\theta) = \lambda_1 \sum_t \big\| \hat{\mathbf{x}}_t - \mathbf{x}_t^{\text{PIC}} \big\|_2^2 + \lambda_2 \mathcal{L}_{\text{moments}} + \lambda_3 \mathcal{L}_{\text{stability}}, \quad (4)$$

which combines trajectory/state errors against PIC targets, moment-level consistency, and stability controls (e.g., bounded kicks, monotonic response versus detune). See Eq. (3) for the kick parameterization and Eq. (4) for the composite training objective; this design is consistent with current differentiable-simulation practices for accelerators [6]. The composition $M \circ K_\theta \circ E$ remains fully differentiable with respect to $\mathbf{p}$, enabling gradient-based tuning using Eq. (2).

MODES OF OPERATION AND USE CASES

The twin supports four operational modes that span early commissioning through routine operation (see Fig. 2). In this contribution we exercise read-only modes; closed-loop variants are reserved for future deployment.

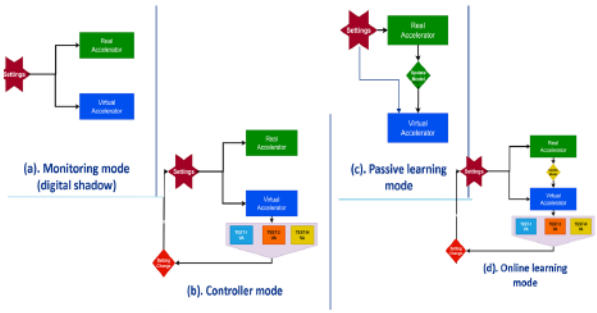

Figure 2: Operational modes of the twin: (a) Monitoring (digital shadow), (b) Replay / passive learning, (c) Controller (hybrid closed loop), and (d) Online learning (adaptive closed loop). Results reported here cover (a)–(b); modes (c)–(d) are slated for deployment.

(i) *Monitoring (digital shadow)*: a strictly read-only mirror that fuses live diagnostics into a running state estimate and performs what-if prediction. It produces virtual diagnostics $\hat{\mathbf{y}}(s; \mathbf{p})$ aligned with the measurement model so that residuals in $J(\mathbf{p})$ are directly comparable to $\mathbf{y}_{\text{obs}}(s)$ from BPM/TOF/toroid streams via Eqs. (1)–(2).

(ii) *Replay / passive learning*: time-synchronized replays of BPM/TOF/FFC/toroid/LLRF data for optics/phasing inference, residual attribution, anomaly ranking, and lightweight calibration refinement of gains/bias terms; no actuation is permitted.

(iii) *Controller (hybrid closed loop)*: supervisory optimization that evaluates $\nabla_{\mathbf{p}} J$ (Eqs. (1)–(2)) and issues feasibility-projected, rate-limited EPICS writes under machine-protection guard rails with automatic rollback.

(iv) *Online learning (adaptive closed loop)*: as in (iii) but with in-situ surrogate and calibration updates driven by residuals and curated replay batches, all under the same safety policy.

Scope here: (i)–(ii) only; (iii)–(iv) are planned for phased commissioning.

Use cases. *Pulse→CW transitions*: multi-parameter ramps of RF phases/voltages and chopper duty factor synthesized with bound, slew, and MPS constraints, validated in shadow/replay before live execution. *Autonomous tuning*: multi-objective fits to orbit/envelope/TOF incorporating loss proxies (e.g., halo mass beyond $n\sigma$) within $J(\mathbf{p})$; corrective actions are ranked in shadow mode with actuator-aware feasibility projection. *Predictive maintenance*: unsupervised detectors (parity-space, autoencoders) on LLRF/interlock streams gate controller updates and trigger replay-driven diagnosis [7, 8]. *HLA pretesting*: ORM analyzers and orbit-correction HLAs target the same APIs to ensure reproducible transition from virtual to live operation [9].

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-WEPD097

CONTROL SYSTEM INTEGRATION

For the results reported here the twin was run in *read-only* mode and connected to the control system through a soft IOC. Subscriptions covered BPM, TOF/FFC, toroid, and LLRF channels (amplitude, phase, detune, interlocks). A translator reconciled PV namespaces with model variables, mapping both into (i) the parameter vector $\mathbf{p}$ used by the state update in (1)–(2) and (ii) measurement operators that generate virtual observables $\hat{\mathbf{y}}(s;\mathbf{p})$ for like-for-like residuals inside $J(\mathbf{p})$. Incoming data were batched into time-aligned frames tagged with timestamps and configuration hashes for M, the K_θ checkpoint, and the disturbance model $E(\xi)$.

Each frame advanced through a deterministic processing chain: state estimation (optics/phasing reconstruction), residual computation, and objective/sensitivity evaluation via automatic differentiation through $M+K_\theta+E$. From these, the twin produced *rank-ordered* advisory corrections in actuator space (RF phases/voltages, magnet gradients, chopper timing) together with predicted changes to orbit/phase/energy and loss proxies. Actuator writes were disabled; recommendations were exported for operator review. Feasibility was enforced during ranking using on-line constraints (hardware bounds, slew limits, RF family couplings), while MPS/interlock PVs acted as gates. Even in read-only operation, all candidates were validated against the same policy through dry-run checks.

To ensure bitwise replay, we logged PV snapshots alongside software versions, dataset hashes, surrogate identifiers, and seeds for ξ. The EPICS-facing API was exercised by existing HLAs (ORM, analyzers, orbit correction) to verify compatibility and to rehearse commissioning workflows without enabling writes.

VALIDATION AND INITIAL RESULTS

Validation focused on the RFQ→MEBT interface and on HWR cells using two data sources: (i) PIC-derived hold-outs with domain-randomized phase/voltage settings and controlled mismatch, and (ii) time-synchronized replays of BPM, TOF/FFC, toroid, and LLRF measurements. The twin operated read-only; virtual diagnostics $\hat{\mathbf{y}}(s;\mathbf{p})$ were compared to $\mathbf{y}_{\mathrm{obs}}(s)$ inside $J(\mathbf{p})$.

Accuracy metrics included transverse state/moment trajectories ($x_{\mathrm{rms}}(s)$, $y_{\mathrm{rms}}(s)$), longitudinal projections (ϕ, W), transmission, and a halo indicator (mass beyond $n\sigma$). Across phase and voltage scans and under deliberate mismatch, envelope and longitudinal predictions remained within the target tolerances; transmission and halo proxies agreed with PIC references. Replay runs reproduced optics/phasing trends and reduced BPM/TOF residuals in line with the inferred corrections. Runtime was measured as mean wall-clock per meter of propagated lattice under identical sampling of diagnostics. A representative envelope comparison is shown in Fig. 3. The differentiable transport with the learned kick delivered the expected $\sim 10^4\times$ speedup relative to high-fidelity references while maintaining task-level accuracy suitable for shadow/replay decision support.

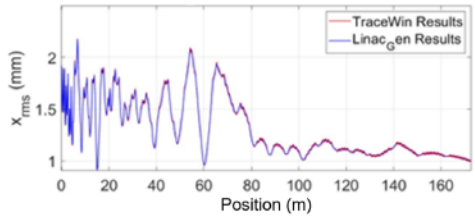

Figure 3: Representative rms-envelope comparison: $x_{\mathrm{rms}}(s)$ from PIC versus the twin prediction.

DEVELOPMENT PHASES AND REPRODUCIBILITY

Work proceeded in five tightly scoped stages: (i) *Data & specification* – operating envelopes, diagnostics, and error channels were defined and PIC campaigns executed; (ii) *Modeling* – the symplectic core M and the learned operator K_θ were implemented (element-wise for magnets/cavities or cell-wise blocks), and uncertainty channels $E(\xi)$ were instrumented; (iii) *Training & validation* – domain randomization with hold-out lattices/scans, cross-checks against PIC and measurement replays, and calibration procedures; (iv) *Controls binding* – PVs were mapped to $\mathbf{p}$ and to measurement operators, shadow mode was completed, and cycle-time determinism verified; (v) *Closed-loop trials* – candidate policies were exercised in safe mode under MPS constraints, with mature recipes promoted toward operations.

Reproducibility is ensured by configuration hashes for optics and objectives, versioned surrogate checkpoints for K_θ, fixed seeds for ξ, and archived PV snapshots. All software and datasets are under version control, and plotting scripts regenerate figures directly from logged artifacts.

OUTLOOK

Next steps extend surrogate coverage across the full SRF chain, strengthen uncertainty quantification in the estimator and objective, and generalize controllers to multi-beam user modes. The same stack supports hardware-in-the-loop trials and portable deployment to machines with similar front-end topologies, enabling a staged transition from shadow to guarded closed-loop operation.

REFERENCES

[1] M. A. Plum, "Beam loss in linacs," in *Beam Loss and Accelerator Protection*, CERN, Geneva, Switzerland, Rep. CERN-2016-002, pp. 39–62, Jan. 2016. doi:10.5170/CERN-2016-002.39

[2] L. Tchelidze and J. Stovall, "Beam loss limits in high power proton linear accelerators," in *Proc. IPAC'13*, Shanghai, China, May 2013, pp. 3930–3932.

[3] R. Stanek *et al.*, "PIP-II project overview and status," in *Proc. SRF'23*, Grand Rapids, MI, USA, Jun. 2023, pp. 19–24. doi:10.18429/JACoW-SRF2023-MOIXA02

[4] R. Roussel *et al.*, "Bayesian optimization algorithms for accelerator physics," *Phys. Rev. Accel. Beams*, vol. 27, p. 084801,

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-WEPD097

Aug. 2024.
doi:10.1103/PhysRevAccelBeams.27.084801

[5] S. A. Mnisi, "Systems modelling, AI/ML algorithms applied to control systems," in *Proc. ICALEPCS'23*, Cape Town, South Africa, Oct. 2023, pp. 257–261.
doi:10.18429/JACoW-ICALEPCS2023-TU1BCO03

[6] J. Kaiser *et al.*, "Bridging the gap between machine learning and particle accelerator physics with high speed, differentiable simulations," *Phys. Rev. Accel. Beams*, vol. 27, p. 054601, May 2024. doi:10.1103/PhysRevAccelBeams.27.054601

[7] A. Eichler, J. Branlard, and J. H. K. Timm, "Anomaly detection at the European XFEL using a parity space based method," *Phys. Rev. Accel. Beams*, vol. 26, p. 012801, Jan. 2023.
doi:10.1103/PhysRevAccelBeams.26.012801

[8] M. M. Rahman *et al.*, "Accelerating cavity fault prediction using deep learning at Jefferson Laboratory," *Mach. Learn.: Sci. Technol.*, vol. 5, no. 3, p. 035078, Sep. 2024.
doi:10.1088/2632-2153/ad7ad6

[9] P. Zhu *et al.*, "Multi user virtual accelerator at HEPS for high level application development and beam commissioning," in *Proc. ICALEPCS'23*, Cape Town, South Africa, Oct. 2023, pp. 1388–1390.
doi:10.18429/JACoW-ICALEPCS2023-THPDP033