TOWARD A PHYSICS-INFORMED TOKAMIND: LEARNING MAGNETOHYDRODYNAMIC EQUILIBRIUM LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although foundation models for scientific domains promise scalable prediction across diverse modalities, they often lack explicit constraints to governing physical laws. In tokamak dynamics, this is critical since surrogates must predict magnetic equilibrium features essential for safe operation, but predictions from standard data-driven learning approaches are not guaranteed to respect the governing laws of plasma confinement (which is worsened by the scarcity of experimental fusion data compared to other domains). We present a methodology to augment the multi-modal foundation model TokaMind with physics-informed losses derived from the Grad-Shafranov equation, the cornerstone constraint on tokamak equilibrium. We introduce two loss formulations based on strong-form and weak-form residuals, and develop new evaluation metrics that additionally target physical consistency and operator-space accuracy. Fine-tuning experiments based on the TokaMark benchmark reveal that physics-informed regularization substantially reduces mutual consistency errors between predicted flux and current density, improves flux accuracy in operator space, and softens small-scale structural disruptions. These improvements come at a modest cost in signal-space normalized RMSE, indicating a fundamental trade-off between data fidelity and physical consistency. Our work connects foundation models and physical reasoning, providing both a reusable template for physics-informed training in scientific machine learning, and evidence that governing laws improve the reliability of equilibrium predictions for high-stakes plasma physics applications. Our open-sourced implementations will be made available upon acceptance.

1 INTRODUCTION

Physics-Aware Machine Learning, and tokamak plasma dynamics. The umbrella term *physics-aware learning* describes a variety of machine learning (ML) methods that blend data-driven learning with governing physical laws, most commonly expressed as Partial Differential Equations (PDEs). We refer to a general review by Faroughi et al. (2024) for the physics-aware learning categories: *physics-guided* (e.g., Daw et al., 2022; Jia et al., 2021), *physics-informed* (e.g., Raissi et al., 2019), and *physics-encoded* methods (e.g., Rao et al., 2023; Chen et al., 2018). From these, physics-informed regularization has been reported to improve surrogate fidelity across multiple domains.

In fusion plasma research, tokamak dynamics are governed by fundamental magnetohydrodynamic (MHD) equations, yet neural surrogate models often learn these purely from data without explicit constraint from the underlying physical laws. The behavior of magnetically confined plasma in tokamaks is captured by a hierarchy of PDEs describing the coupling between the plasma, electromagnetic fields, and energy transport. Understanding these governing equations is essential for developing ML models to predict plasma behavior while respecting fundamental physical laws. Among these, the Grad-Shafranov (GS) equation (Grad & Rubin, 1958; Shafranov, 1958) is a fundamental PDE for tokamak equilibrium geometry, stability, and confinement: it encodes the balance between plasma pressure gradient and magnetic force, and is of particular relevance since its solution determines essential equilibrium quantities such as the flux function ψ, the last closed-flux surface, the toroidal current density J_ϕ, and the safety factor.

ML models that aim to predict tokamak dynamics must implicitly capture all the coupled PDE constraints from observational data. Physics-informed regularization could directly address this chal-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

lenge by suitably incorporating GS residuals into the learning objective. Physics-informed learning has been applied across fusion modeling, including active learning for turbulent transport surrogates (e.g., Zanisi et al., 2024), conservation-law-preserving operators for gyrokinetic transport (Ashourvan, 2026), neural surrogates for equilibrium reconstruction (e.g., Sun et al., 2024; Rutigliano et al., 2025; Wan et al., 2023; Liao et al., 2026), and solving the GS equation (Ding et al., 2025; Rizqan et al., 2025). The physics constraint is typically the modeling target of a bespoke, single-purpose network, rather than an auxiliary objective added to a general multi-modal foundation model.

TokaMark, TokaMind, and proposed TokaMind extension. TokaMind, a multi-modal transformer foundation model for tokamak plasma dynamics (Boschi et al., 2026), achieves strong performance on the TokaMark benchmark (Rousseau et al., 2026) through architecture innovation and scale. However, its tasks on predicting complex plasma equilibrium and stability features would benefit from explicit physical constraints that encode fundamental MHD principles. Augmenting TokaMind's multi-modal learning objective with GS constraints can enhance its robustness of equilibrium prediction, particularly for tasks where the model must capture fine details of magnetic configuration and stability margins (see also Liao et al., 2026). In this work, we augment the standard TokaMind's training objective with physics-informed losses based on the GS equation (Shafranov, 1958), as illustrated in Fig. 1. By computing and backpropagating residuals of the GS equation on TokaMind's training pipeline, we aim at rewarding predictions physically consistent with MHD equilibria. The physics-informed loss, via a combination of ground-truth and self-consistency terms, steers the training towards the manifold of valid GS solutions.

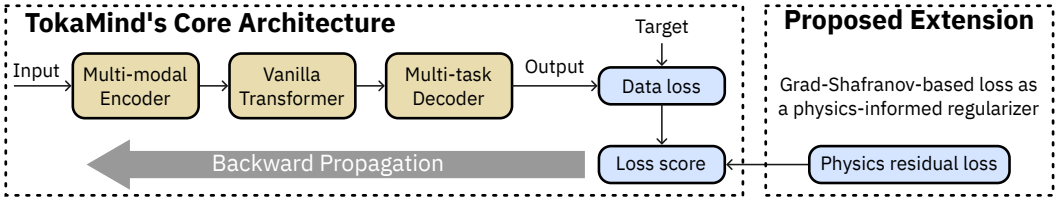


Figure 1: TokaMind's core architecture together with our proposed physics-informed extension.

Our contributions are summarized as follows: **(1)** A general methodology to define physics-informed losses, with applications beyond the GS equation and TokaMind; **(2)** two formulations of GS-based losses for their use with TokaMind; **(3)** new metrics to evaluate the proposed physics-informed TokaMind extension; and **(4)** numerical evaluation illustrating the benefits of this approach.

Paper outline. Section 2 provides a generic formulation of GS-based physics residuals and losses in terms of available model outputs, and two specific formulations are explored. Section 3 presents the proposed methodology to integrate the derived GS losses into TokaMind's multi-modal objective (including task/output specifications) and newly developed evaluation metrics. In Section 4 we validate our proposed TokaMind extension through tuning experiments on selected tasks, comparing predictions to both **(i)** a baseline TokaMind model tuned only with data loss, and **(ii)** reference MHD solutions computed with a conventional method, isolating the benefit of physics regularization.

2 Grad-Shafranov-based Losses for MHD-informed Training

2.1 General Conventions

Since the GS equation can be formulated in different ways (e.g., classical, variational, geometric), and each formulation can have different term arrangements, a common convention for the derivation of Grad-Shafranov losses is needed. For this purpose, we build the GS-based residuals and losses in terms of non-homogeneous (or inhomogeneous) formulations of the GS equation, i.e., of the form

$$T\psi = g\left(J_\phi\right), \tag{1}$$

where T in the Left-Hand Side (LHS) is a suitable elliptic operator acting on the poloidal magnetic flux function ψ, and g in the Right-Hand Side (RHS) is a fixed function usually dependent on the toroidal current density J_ϕ (and on the flux function itself), as shown in Appendix A.

2.1.1 GRAD-SHAFRANOV-BASED RESIDUALS

A well-defined practical formulation of Eq. (1) must include a LHS cleaning mask M of the form:

$$M \odot (T\psi) = g\left(J_\phi\right), \tag{2}$$

where M nullifies all the values of $T\psi$ outside the plasma boundary[1]. With the shorthands

$$\mathrm{GS}_{LHS}\left(\psi\right) \equiv M \odot (T\psi), \qquad \mathrm{GS}^{true}_{LHS} \equiv \mathrm{GS}_{LHS}\left(\psi^{true}\right), \qquad \mathrm{GS}^{pred}_{LHS} \equiv \mathrm{GS}_{LHS}\left(\psi^{pred}\right),$$
$$\mathrm{GS}_{RHS}\left(J_\phi\right) \equiv g(J_\phi), \qquad \mathrm{GS}^{true}_{RHS} \equiv \mathrm{GS}_{RHS}\left(J^{true}_\phi\right), \qquad \mathrm{GS}^{pred}_{RHS} \equiv \mathrm{GS}_{RHS}\left(J^{pred}_\phi\right),$$

we can formulate a generic GS residual as

$$\mathrm{GS}_{res}\left(\psi, J_\phi\right) = \mathrm{GS}_{LHS}\left(\psi\right) - \mathrm{GS}_{RHS}\left(J_\phi\right), \tag{3}$$

and different variants depending on ground-truth or model-prediction data, as summarized in Fig. 2:

(i) $\boxed{\mathrm{GS}^{\text{null}}_{res} = \mathrm{GS}^{true}_{LHS} - \mathrm{GS}^{true}_{RHS} = 0}$: This only involves the ground-truth, and via Eq. (2) we get:

$$\mathrm{GS}_{res}\left(\psi^{true}, J^{true}_\phi\right) = \mathrm{GS}^{true}_{LHS} - \mathrm{GS}^{true}_{RHS} = 0. \tag{4}$$

This null residual is a useful reference to, e.g., confirm the integrity of ground-truth data.

(ii) $\boxed{\mathrm{GS}^{\text{no-anchor}}_{res} = \mathrm{GS}^{pred}_{LHS} - \mathrm{GS}^{pred}_{RHS}}$: This is purely a function of ψ^{pred} and J^{pred}_ϕ values in Eq. (2), and, in a loss term, it guides the training towards self-consistent outputs by binding predicted values under a common objective. However, that does not necessarily guarantee the correspondence of the predicted equilibrium with respect to ground-truth data.

(iii) $\boxed{\mathrm{GS}^{\text{R-anchor}}_{res} = \mathrm{GS}^{true}_{LHS} - \mathrm{GS}^{pred}_{RHS} = \mathrm{GS}^{true}_{RHS} - \mathrm{GS}^{pred}_{RHS}}$: In a loss term, this residual penalizes discrepancies between RHS predictions and the corresponding ground-truth distributions.

(iv) $\boxed{\mathrm{GS}^{\text{L-anchor}}_{res} = \mathrm{GS}^{pred}_{LHS} - \mathrm{GS}^{true}_{RHS} = \mathrm{GS}^{pred}_{LHS} - \mathrm{GS}^{true}_{LHS}}$: In a loss term, this residual penalizes LHS predictions that do not correspond to ground-truth distributions. Additionally, if T in Eq. (2) is a linear operator, then $\mathrm{GS}^{\text{L-anchor}}_{res}$ yields a smoothing loss term on ψ^{pred} in the operator space (i.e., via $T(\psi^{true} - \psi^{pred})$).

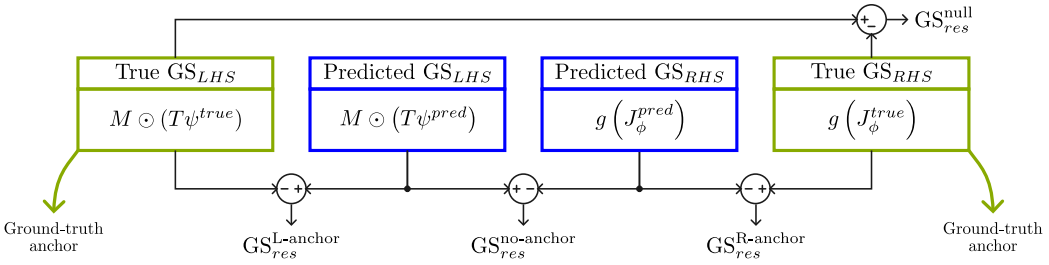


Figure 2: Visual representation of the generic Grad-Shafranov residuals as defined in Section 2.1.1.

2.1.2 GRAD-SHAFRANOV-BASED LOSSES

A generic GS-based loss function has the form

$$\mathcal{L}_{GS} = \mathcal{M}\left(\mathrm{GS}_{res}\right), \tag{5}$$

with $\mathcal{M}\left(\cdot\right)$ corresponding to a suitable loss function metric[2]. From that, a general formulation of a GS loss is given as the convex combination of non-null residuals as follows

$$\mathcal{L}^{Total}_{GS} = w_{LA} \cdot \mathcal{L}^{\text{L-anchor}}_{GS} + w_{RA} \cdot \mathcal{L}^{\text{R-anchor}}_{GS} + w_{NA} \cdot \mathcal{L}^{\text{no-anchor}}_{GS}, \tag{6}$$

with

$$\mathcal{L}^{\text{L-anchor}}_{GS} = \mathcal{M}\left(\mathrm{GS}^{\text{L-anchor}}_{res}\right), \quad \mathcal{L}^{\text{R-anchor}}_{GS} = \mathcal{M}\left(\mathrm{GS}^{\text{R-anchor}}_{res}\right), \quad \mathcal{L}^{\text{no-anchor}}_{GS} = \mathcal{M}\left(\mathrm{GS}^{\text{no-anchor}}_{res}\right). \tag{7}$$

[1] Special considerations for the cleaning mask are provided in Appendix B.

[2] Special considerations for the loss function metrics are provided in Appendix B.

2.2 Physics loss from the Grad-Shafranov Equation in Strong Form

The most common variant of the GS equation in strong form involves the flux function, the pressure-driven term, and the current-driven term (as shown in Eq. (29) of Appendix A), and is defined as

$$-\Delta^*\psi \equiv -R\,\nabla\cdot\left(\frac{1}{R}\,\nabla\psi\right) \;=\; \mu_0 R\left(R\frac{dP}{d\psi}+\frac{F}{\mu_0 R}\frac{dF}{d\psi}\right) \;\equiv \mu_0 R J_\phi\,. \tag{8}$$

For fixed finite-element geometry (as in the case of FAIR-MAST data (Jackson et al., 2025; 2024)), the Grad-Shafranov operator Δ^* is constant and can be precomputed (see, e.g., Dudson, 2024; Amorisco et al., 2024). This yields the following Strong Grad-Shafranov (SGS) residuals:

$$\overbrace{-M\odot\left(\Delta^*\psi^{pred}\right)-\mu_0 R J_\phi^{pred}}^{\mathrm{SGS}_{res}^{\text{no-anchor}}},\quad \overbrace{\underbrace{\mu_0 R J_\phi^{true}}_{\text{RHS anchor}}-\mu_0 R J_\phi^{pred}}^{\mathrm{SGS}_{res}^{\text{R-anchor}}},\quad \overbrace{\underbrace{M\odot\left(\Delta^*\psi^{true}\right)}_{\text{LHS anchor}}-M\odot\left(\Delta^*\psi^{pred}\right)}^{\mathrm{SGS}_{res}^{\text{L-anchor}}}. \tag{9}$$

From these residuals, we can finally define the SGS loss as

$$\mathcal{L}_{SGS}^{Total}\left(\psi^{pred},\psi^{true},J_\phi^{pred},J_\phi^{true}\right)=w_{LA}\cdot\mathcal{L}_{SGS}^{\text{L-anchor}}+w_{RA}\cdot\mathcal{L}_{SGS}^{\text{R-anchor}}+w_{NA}\cdot\mathcal{L}_{SGS}^{\text{no-anchor}}. \tag{10}$$

If J_ϕ is not a specified model output (as in the case of TokaMark's Tasks 1-3 and 2-3[3]), then we get

$$\mathcal{L}_{SGS}^{Total}\left(\psi^{pred},\psi^{true}\right)=w_{LA}\cdot\mathcal{L}_{SGS}^{\text{L-anchor}}. \tag{11}$$

2.3 Physics loss from the Grad-Shafranov Equation in Weak Form

The residuals of Subsection 2.2 are evaluated pointwise and therefore require the second derivatives appearing in Δ^*. Using standard Green identities, the weak (variational) formulation of the GS problem, satisfied by any solution $\{\psi, J_\phi\}$ of Eq. (8), is given by

$$\int_\Omega \frac{1}{R}\,\nabla\psi\cdot\nabla v\,\mathrm{d}\Omega \;-\; \oint_{\partial\Omega}\frac{1}{R}\,\frac{\partial\psi}{\partial n}\,v\,\mathrm{d}s \;=\; \mu_0\int_\Omega J_\phi\,v\,\mathrm{d}\Omega, \tag{12}$$

for test functions v on the poloidal domain Ω (Ciarlet, 2002; Brenner & Scott, 2008; Faugeras, 2020). We discretize Eq. (12) by a standard Galerkin construction with continuous piecewise-linear basis functions, one per grid node. On a structured $n_R\times n_Z$ grid with spacings h_R, h_Z, midpoint quadrature of $1/R$, and row-sum mass lumping, the discretization is reduced to an edge-flux operator

$$(W\psi)_{ij}=\sum_{(i',j')\in\mathcal{N}(i,j)}\frac{1}{R_e}\,\frac{\psi_{ij}-\psi_{i'j'}}{d^2},\qquad R_e=\tfrac{1}{2}\left(R_{ij}+R_{i'j'}\right), \tag{13}$$

with $d\in\{h_R,h_Z\}$ and $\mathcal{N}(i,j)$ the four axis neighbours. Appendix C gives the derivation, the treatment of the boundary term, and the properties of W. We implement Eq. (13) directly rather than assembling element matrices at training time. The adopted GS equation in weak form is then

$$W\psi=\mu_0 J_\phi. \tag{14}$$

Three properties make this formulation attractive for training purposes: **(i)** Eq. (13) involves only first differences of ψ, a common motivation for weak-form residuals in physics-informed learning (Kharazmi et al., 2019; 2021; Zang et al., 2020); **(ii)** W is symmetric positive semidefinite, so the squared residual is a non-negative quadratic form; and **(iii)** the construction is locally conservative and consistent with a locally integrated GS equation (Eymard et al., 2000; Samarskii, 2001).

We can then define the following Weak Grad-Shafranov (WGS) residuals:

$$\overbrace{W\psi^{pred}-\mu_0 J_\phi^{pred}}^{\mathrm{WGS}_{res}^{\text{no-anchor}}},\quad \overbrace{\underbrace{\mu_0 J_\phi^{true}}_{\text{RHS anchor}}-\mu_0 J_\phi^{pred}}^{\mathrm{WGS}_{res}^{\text{R-anchor}}},\quad \overbrace{W\psi^{pred}-\underbrace{W\psi^{true}}_{\text{LHS anchor}}}^{\mathrm{WGS}_{res}^{\text{L-anchor}}}. \tag{15}$$

[3]For a formal definition, see Section 3.1.

As in the derivations for SGS residuals in Eq. (9), the cleaning mask of Eq. (2) is a point-wise factor on GS_{LHS}, allowing for choices of the test functions that nullify the boundary integral. From these residuals, we can finally define the WGS loss as

$$\mathcal{L}_{WGS}^{Total}\left(\psi^{pred}, \psi^{true}, J_\phi^{pred}, J_\phi^{true}\right) = w_{LA}\cdot\mathcal{L}_{WGS}^{\text{L-anchor}} + w_{RA}\cdot\mathcal{L}_{WGS}^{\text{R-anchor}} + w_{NA}\cdot\mathcal{L}_{WGS}^{\text{no-anchor}}. \quad (16)$$

If J_ϕ is not a specified model output (as in the case of TokaMark's Tasks 1-3 and 2-3), then we get

$$\mathcal{L}_{WGS}^{Total}\left(\psi^{pred}, \psi^{true}\right) = w_{LA}\cdot\mathcal{L}_{WGS}^{\text{L-anchor}}. \quad (17)$$

2.4 Illustrative over-fitting experiments

To illustrate why signal-space accuracy alone may be insufficient, we compare data-only and GS-informed training in an over-fitting exercise described in Appendix D. Figure D.1 shows that, although the predicted ψ maps appear visually close to the reference, discrepancies become more apparent after applying the strong-form operator Δ^* or the weak-form operator W. Including the corresponding GS loss does improve agreement in operator space in these examples. This highlights how visually similar flux maps can differ in spatial structure relevant to the GS equation: small-amplitude, short-wavelength errors can be amplified by the second-order operator Δ^* and its weak-form discretization W, motivating supervision beyond pointwise data fidelity. This experiment does not assess generalization, which is evaluated separately in Section 4.2.

3 Methodology

3.1 Prediction tasks and output specifications

TokaMark (Rousseau et al., 2026) specifies the poloidal flux ψ as the sole gridded output of **Task 1-3** (flux-map reconstruction) and **Task 2-3** (flux-map reconstructive forecasting). These correspond to the *reduced cases* presented in Sections 2.2 and 2.3: with no J_ϕ^{pred} values available, the only computable residual is the L-anchored one, and both the strong- and weak-form losses collapse to a single operator-space term on ψ as seen in Eq. (11) and Eq. (17). To enable the use of the general formulations of GS-based losses (namely, Eq. (10) and Eq. (16)), and taking advantage of other signals available in FAIR-MAST data (Jackson et al., 2024; 2025), we define a collection of tasks that extend Tasks 1-3/2-3, as shown in Table 1. In the case of **Tasks 1-4/2-4**, the extension consists of adding J_ϕ as a gridded output alongside ψ, on the same 65×65 grid, using the same inputs, windows, and shot splits as Tasks 1-3/2-3. On the other hand, **Tasks 1-6/2-6** are variations of Tasks 1-4/2-4 where *Magnetics* signals are replaced with *Thomson scattering* signals as input diagnostics.

3.2 Evaluation metrics

Data fidelity is measured by the benchmark metric across all tasks in Table 1, and we report new companion diagnostics: **(i)** the GS mutual consistency of ψ and J_ϕ; **(ii)** the accuracy of ψ in the GS operator space; and **(iii)** the amount of high-frequency structure carried by the reconstruction. These diagnostics carry different requirements on the model: quantities built from ψ alone are available for all tasks in Table 1, while those that involve J_ϕ^{pred} are available only when J_ϕ is a specified output (i.e., not for Tasks 1-3/2-3). All these metrics are evaluation-only: they never enter the loss, the model inputs, or the benchmark aggregation, and none shares an operator with any training loss. Together, these newly introduced physics-aware metrics constitute a complementary evaluation framework for assessing aspects of model quality that standard signal-space errors cannot capture.

3.2.1 Data fidelity: NRMSE

We follow the TokaMark protocol without modification (Rousseau et al., 2026), defining $\mathrm{RMSE}_{k,i}$ and RMSE_k for output signal k and window i containing N_k flattened samples, where W_k is the number of windows for signal k in a given shot. The single-shot signal error is normalized by the empirical standard deviation σ_k of signal k over the evaluation shots. Shot aggregation is thus defined by NRMSE_k and $\mathrm{Signal}_{\mathrm{NRMSE}}$ in (Rousseau et al., 2026) with S being the total number of shots in the test set used for evaluation. As in Boschi et al. (2026), predictions are decoded to native

Table 1: Task families evaluated in our study, following the TokaMark definitions (Rousseau et al., 2026). New Tasks 1-4/2-4 and Tasks 1-6/2-6 extend Tasks 1-3/2-3 by adding a new output and replacing part of the original input diagnostics. Input and output windows are anchored at a reference time t_0: *Reconstruction* infers the outputs over the same interval $[t_0 - \delta_{in},\, t_0]$ spanned by the input diagnostics; *Reconstructive (RC) Forecasting* uses diagnostics over $[t_0 - \delta_{in},\, t_0]$ together with actuator trajectories over $[t_0 - \delta_{in},\, t_0 + \delta_{out}]$ to predict the outputs over $[t_0,\, t_0 + \delta_{out}]$, where δ_{in} and δ_{out} denote the input and output window lengths.

Task	Input Diagnostics	Input Actuators	Outputs	I/O Window	Type
1-3	Currents, Magnetics	–	ψ	5 / 5 ms	Reconstruction
1-4	Currents, Magnetics	–	ψ, J_ϕ	5 / 5 ms	Reconstruction
1-6	Currents, Thomson	–	ψ, J_ϕ	5 / 5 ms	Reconstruction
2-3	Currents, Magnetics	Voltages, NBI Power	ψ	5 / 25 ms	RC Forecasting
2-4	Currents, Magnetics	Voltages, NBI Power	ψ, J_ϕ	5 / 25 ms	RC Forecasting
2-6	Currents, Thomson	Voltages, NBI Power	ψ, J_ϕ	5 / 25 ms	RC Forecasting

signal space and scored against the original non-imputed targets at observed positions only: values below 1 indicate predictive skill, with lower values corresponding to better predictions.

3.2.2 Evaluation region

The grid-reduced GS diagnostics of Section 3.2.3 (namely E_{self}, E_ψ, and the roughness deviation) are averaged over the same per-slice region

$$\mathcal{C} = \left\{(i,j) : J_\phi^{true} \text{ finite},\ J_\phi^{true} > 10^{-6}\right\} \cap \mathcal{E}, \tag{18}$$

where $\mathcal{E}$ is the static MAST limiter mask eroded by one cell. The current density threshold reproduces the plasma mask of the physics loss; the intersection with $\mathcal{E}$ is an additional restriction, present only at evaluation, which guarantees that no five-point stencil evaluated on $\mathcal{C}$ reaches the domain boundary. Unlike the cleaning mask M of Section 2.1.1, $\mathcal{E}$ additionally erodes the limiter by one cell so the finite-difference (Δ^*_{FD} in Eq. (19) below) never involves the grid boundary. The region is derived only from J_ϕ^{true} (i.e., the *reference* current density) to prevent a model shrinking it to lower its own score. As $\mathcal{C}$ depends only on reference data, it is well defined for Tasks 1-3/2-3 even without J_ϕ being a specified model output, as $J_\phi^{true} = -\Delta^\star \psi^{true}/\mu_0 R$ is available via ψ^{true}.

3.2.3 Grad-Shafranov diagnostics

All quantities in this subsection are built from a centred finite-difference discretization of Δ^* (LeVeque, 2007) that no loss is optimized against during any model training:

$$\Delta^*_{\mathrm{FD}}\psi\big|_{ij} = \frac{\psi_{i,j+1} - 2\psi_{ij} + \psi_{i,j-1}}{h_Z^2} + \frac{\psi_{i+1,j} - 2\psi_{ij} + \psi_{i-1,j}}{h_R^2} - \frac{1}{R_{ij}}\frac{\psi_{i+1,j} - \psi_{i-1,j}}{2h_R}. \tag{19}$$

This finite-difference form is used deliberately, so that the same diagnostics will score models trained with either strong or weak GS forms on identical terms. With the GS shorthands

$$\mathrm{LHS}[\psi] = -\Delta^*_{\mathrm{FD}}\psi, \qquad \mathrm{RHS}[J_\phi] = \mu_0 R\, J_\phi, \tag{20}$$

we evaluate three discrepancy metrics as follows (further discussed in Appendix E.2):

$$\boxed{\text{GS self-consistency:}} \quad E_{\mathrm{self}} = |\mathcal{C}|^{-1} \textstyle\sum_{(i,j)\in\mathcal{C}} \left|\mathrm{LHS}[\psi^{pred}] - \mathrm{RHS}[J_\phi^{pred}]\right|_{ij}, \tag{21}$$

$$\boxed{\psi\text{-operator error:}} \quad E_\psi = |\mathcal{C}|^{-1} \textstyle\sum_{(i,j)\in\mathcal{C}} \left|\mathrm{LHS}[\psi^{true}] - \mathrm{LHS}[\psi^{pred}]\right|_{ij}, \tag{22}$$

$$\boxed{\text{RHS current density error:}} \quad E_J = |\mathcal{C}|^{-1} \textstyle\sum_{(i,j)\in\mathcal{C}} \left|\mathrm{RHS}[J_\phi^{true}] - \mathrm{RHS}[J_\phi^{pred}]\right|_{ij}. \tag{23}$$

We report E_{self} and E_ψ in our experiments, as E_J resolves to a weighted data loss.

3.2.4 ROUGHNESS DEVIATION

We quantify the roughness of ground-truth data and model predictions over $\mathcal{C}$ in Eq. (18) as

$$S[f] = \sqrt{|\mathcal{C}|^{-1} \sum_{(i,j)\in\mathcal{C}} \left(f - G_\sigma * f\right)_{ij}^2}, \tag{24}$$

where $G_\sigma * f$ denotes convolution of f with the Gaussian filter of width $\sigma = 2$ grid cells, which isolates structure at the stencil scale that Δ^* amplifies, and define

$$\epsilon_{\text{rough}} = \left(S\left[\psi^{\text{pred}}\right] - S\left[\psi^{\text{true}}\right]\right) / \, S\left[\psi^{\text{true}}\right], \tag{25}$$

as the relative deviation, where ψ^{true} is the corresponding EFIT reconstruction (Lao et al., 1985). This is a signed and dimensionless quantity, useful to compare profile across tasks. We evaluate Eq. (25) for ψ alone and report it for all tasks in Table 1.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

Data, windows, and training regime. All experiments use the TokaMark task definitions and splits (Rousseau et al., 2026; Boschi et al., 2026). Inputs and actuators are chunked into non-overlapping 5 ms chunks with a 5 ms window stride, retaining at most 50 chunks per role; outputs are embedded at window level. Non-finite values are handled by temporal followed by spatial interpolation with a zero fallback, and native targets are preserved for sparse supervision and evaluation. Input signals and chunks are randomly dropped with probability 0.05 during collation.

We train the TokaMind architecture from random initialization for 50 epochs, and every model in a comparison block sees the same data, schedule, and initialization policy, differing only in its loss. The backbone uses $d = 192$, 4 layers, 6 heads, feed-forward dimension 768, and dropout 0.05, giving approximately 2.0M trainable parameters. Optimization uses AdamW with per-block learning rates (token encoder 5×10^{-3}, backbone 10^{-3}, modality heads 5×10^{-3}, output adapters 5×10^{-3}), weight decay 0.01, a linear warmup over the first 2% of steps, batch size 512, bfloat16 mixed precision, and early stopping with patience 10. Appendix E.1 provides details on output embedding dimensions. All our experiments were run on an internal compute cluster, a collection of x86 compute nodes (100 GB / node) with NVIDIA A100 80gb GPUs. The implementation of the GS residuals, losses, and diagnostics is open-sourced and will be available upon acceptance.

Training objective and loss weights. The total objective combines the native-space sparse data loss of Boschi et al. (2026) with one GS physics term in a two-level convex combination, namely

$$\mathcal{L}_{\text{total}} = w_{GS} \cdot \mathcal{L}_{GS} + (1 - w_{GS}) \cdot \mathcal{L}_{\text{data}}, \tag{26}$$

with: $\mathcal{L}_{GS} \in \{\mathcal{L}_{SGS}^{Total}, \mathcal{L}_{WGS}^{Total}\}$; $w_{GS} = 0.1$, matched with $(w_{LA}, w_{RA}, w_{NA}) = (1.0,\ 0.0,\ 0.0)$ for Tasks 1-3/2-3, and $(w_{LA}, w_{RA}, w_{NA}) = (0.5,\ 0.0,\ 0.5)$ for all other tasks in all reported runs; and $\mathcal{M}(\cdot) = \mathcal{R}(\cdot)$, the masked mean of squares defined in Eq. (31) of Appendix B.2. The same weights are used for the strong-form (Eqs. 10 and 11) and weak-form (Eqs. 16 and 17) physics terms, so the two differ only in the discretization of the residual. Convexity at both levels means a single number controls the physics contribution. The choice $w_{RA} = 0$ follows from Section 3.2.3 since the resulting R-anchored term in our derivations acts similarly as the data term on J_ϕ, while the remaining two terms cover complementary failure modes: the no-anchor term enforces mutual consistency of ψ^{pred} and J_ϕ^{pred}, while the L-anchored term ties ψ^{pred} to the reference in operator space. The value $w_{GS} = 0.1$ is empirical, mainly to subsume different physical units and native orders of magnitude in $\mathcal{L}_{\text{data}}$ and $\mathcal{L}_{GS}$ onto a common loss, and not a statement of relative importance.

4.2 RESULTS

We compare three objectives on each task: **(i)** data loss alone, **(ii)** data loss with the strong-form GS term (SGS), and **(iii)** data loss with the weak-form GS term (WGS). Within each task block the shots, splits, seed, codecs, epoch budget, and optimizer settings are identical, so the three rows differ only in the physics term of Eq. (26). The availability of specified model outputs enables the calculation of a corresponding evaluation metric: NRMSE_{J_ϕ} and E_{self} require J_ϕ values and are reported only for those tasks, while E_ψ and ϵ_{rough} are built from ψ alone and are reported throughout.

Table 2: Test results. NRMSE follows the TokaMark protocol (Section 3.2); E_{self} and E_ψ are the GS diagnostics of Eqs. (21) and (22) over the region $\mathcal{C}$ of Eq. (18); lower is better for every column except ϵ_{rough} (Eq. (25)), for which values closest to zero are best. Within each task block, **bold** and underline mark the best and second-best values, respectively. Loss notation: $\mathcal{L}_{d+S}{=}0.9 * \mathcal{L}_{data} + 0.1 * \mathcal{L}_{SGS}$, and $\mathcal{L}_{d+W}{=}0.9 * \mathcal{L}_{data} + 0.1 * \mathcal{L}_{WGS}$.

Task, Loss	NRMSE_ψ	NRMSE_{J_ϕ}	E_{self}	E_ψ	ϵ_{rough}
1-3, $\mathcal{L}_{data}$	**0.1581**	–	–	0.3452	+0.1911
1-3, $\mathcal{L}_{d+S}$	0.1641 (↑3.8%)	–	–	0.1816	+0.0295
1-3, $\mathcal{L}_{d+W}$	0.1671 (↑5.7%)	–	–	**0.1670**	**+0.0164**
1-4, $\mathcal{L}_{data}$	**0.0929**	**0.1345**	0.3303	0.3310	+0.1989
1-4, $\mathcal{L}_{d+S}$	0.1045 (↑12.5%)	0.1348 (↑0.2%)	0.1711 (↓48.2%)	0.1746 (↓47.3%)	+0.0677
1-4, $\mathcal{L}_{d+W}$	0.1099 (↑18.3%)	0.1372 (↑2.0%)	**0.1624** (↓50.8%)	**0.1662** (↓49.8%)	**+0.0526**
1-6, $\mathcal{L}_{data}$	**0.0978**	**0.1563**	0.3291	0.3303	+0.1971
1-6, $\mathcal{L}_{d+S}$	0.1090 (↑11.5%)	0.1580 (↑1.1%)	0.1702 (↓48.3%)	0.1756 (↓46.8%)	+0.0665
1-6, $\mathcal{L}_{d+W}$	0.1142 (↑16.8%)	0.1608 (↑2.9%)	**0.1607** (↓51.7%)	**0.1667** (↓49.5%)	**+0.0494**
2-3, $\mathcal{L}_{data}$	**0.1289**	–	–	0.3380	+0.1806
2-3, $\mathcal{L}_{d+S}$	0.1364 (↑5.8%)	–	–	0.1830 (↓45.9%)	+0.0129
2-3, $\mathcal{L}_{d+W}$	0.1419 (↑10.1%)	–	–	**0.1644** (↓51.4%)	**-0.0040**
2-4, $\mathcal{L}_{data}$	**0.1195**	0.1467	0.3315	0.3319	+0.1944
2-4, $\mathcal{L}_{d+S}$	0.1252 (↑4.8%)	**0.1296** (↓11.7%)	0.1765 (↓46.8%)	0.1793 (↓46.0%)	+0.0620
2-4, $\mathcal{L}_{d+W}$	0.1296 (↑8.5%)	0.1307 (↓10.9%)	**0.1667** (↓49.7%)	**0.1702** (↓48.7%)	**+0.0445**
2-6, $\mathcal{L}_{data}$	**0.1201**	0.1510	0.3347	0.3341	+0.1924
2-6, $\mathcal{L}_{d+S}$	0.1283 (↑6.8%)	**0.1436** (↓4.9%)	0.1750 (↓47.7%)	0.1789 (↓46.5%)	+0.0554
2-6, $\mathcal{L}_{d+W}$	0.1333 (↑11.0%)	0.1471 (↓2.6%)	**0.1651** (↓50.7%)	**0.1698** (↓49.2%)	**+0.0388**

Physical consistency and flux accuracy. Both physics terms reduce E_{self} by roughly half in every general-case task against data-only baselines, with consistent and comparable effect across input-diagnostic sets and reconstruction and forecasting tasks. E_ψ falls under both physics terms in all six tasks, with approximately 50% reduction. The data-only baseline attains the *highest* E_ψ in all six tasks while attaining the *lowest* NRMSE_ψ in most tasks, which directly evidences that signal-space accuracy on ψ does not imply accuracy in the operator through which the physics acts.

Data fidelity, small-scale structure, and strong versus weak form. The gains reported above come with a trade-off concentrated primarily in a single quantity. While NRMSE_ψ grows slightly in the physics-trained configurations, NRMSE_{J_ϕ} is essentially unchanged throughout (consistent with the choice of $w_{RA}{=}0$). The asymmetry is informative: the physics terms act on ψ through the GS operator, so they displace the flux prediction in signal space while leaving the current density supervised by the data term alone. Nearly all reconstructions carry more small-scale structure than the EFIT reference ($\epsilon_{\text{rough}} > 0$), which however is substantially suppressed by the physics term, and toward the reference roughness rather than past it: the only negative entry is Task 2-3 under WGS, a marginal over-smoothing that is also the closest any configuration comes to the reference. The excess structure removed here is precisely the content that Δ^* amplifies, which explains why the roughness and operator-space improvements move together while NRMSE_ψ does not. Comparing the two formulations, WGS attains lower E_{self} in all four general-case tasks and lower E_ψ in all six, at the cost of a larger NRMSE_ψ than SGS in every task. This is consistent across tasks and diagnostics, but confirming or dismissing either trend requires running each configuration with repeated seeds.

Exploratory study on downstream transfer. We assess whether GS-informed source-task training does benefit downstream prediction on Tasks 1-1, 1-2, 2-1, and 2-2. These tasks reconstruct (Task 1 family) or forecast (Task 2 family) scalar plasma quantities (Tasks 1-1 and 2-1) and last-closed-flux-surface geometry (Tasks 1-2 and 2-2), which characterize the plasma state and boundary associated with the magnetic equilibrium represented by the source tasks' predicted ψ maps. Full task definitions are provided in TokaMark (Rousseau et al., 2026). Models trained on Task 1-4 initialize Tasks 1-1/1-2, while models trained on Task 2-4 initialize Tasks 2-1/2-2. This shared physical

structure motivates testing whether representations learned under GS constraints transfer to related plasma-prediction tasks, even when downstream training uses only data supervision. Appendix Table E.3 compares random initialization with warm starts from data-only, strong-GS, and weak-GS source models, using a 20-epoch downstream training budget and data-only supervision throughout. Transfer benefits therein are task-dependent: all three warm starts outperform scratch on Task 2-1, with strong GS achieving the lowest average NRMSE, whereas scratch performs best on the other three tasks. Among warm starts, strong GS achieves the lowest average NRMSE on three tasks, although several differences are small. These exploratory results suggest a potential transfer benefit in the Task 2-1 forecasting setting, but do not establish a general advantage of GS-informed initialization across downstream tasks.

Limitations. The reported results establish the effects on physical consistency and operator-space flux accuracy under a fixed budget and with a single seed per configuration, at the codec setting of Appendix E.1; a companion configuration at a higher codec threshold is reported in Appendix E.3. Both reported diagnostics are targeted by the loss by construction: they do not by themselves demonstrate that either formulation improves the benchmark objective, and NRMSE_ψ indicates a cost rather than a gain in every configuration. The weights of Eq. (6) were not swept jointly, and no comparison is made against the published TokaMark numbers, which use a different output codec (Appendix E.1). Finally, given how $\epsilon_{\mathrm{rough}}$ is normalized in Eq. (24), a "collapsed" low-amplitude reconstruction may register as smooth; while this does not affect the vast majority of our results, the metric should be read alongside NRMSE_ψ.

5 Concluding Remarks

We have demonstrated that augmenting TokaMind with physics-informed regularization from the Grad-Shafranov equation substantially improves physical consistency and operator-space accuracy. Our key contributions include a general methodology for Grad-Shafranov-based losses extensible to partial output specifications, two alternative formulations for Grad-Shafranov-based losses (in strong and weak form), and metrics that decouple physical consistency from data fidelity. Our experiments revealed a fundamental trade-off, with physics constraints reducing operator-space error by 46–52% and improving consistency by 47–51%, but increasing NRMSE_ψ by 4–18%. The effect on NRMSE_{J_ϕ} is comparatively small and is not uniformly detrimental, ranging from a 12% reduction to a 3% increase across the evaluated configurations. This, however, is not a failure but a reflection of data-driven and physics-driven objectives being orthogonal. The NRMSE increase indicates the model is steered toward physically plausible solutions, which may not minimize signal-space error due to measurement noise or sparse supervision. For applications that prioritize physical reliability (e.g., real-time control), this trade-off may represent an acceptable compromise between signal fidelity and physical consistency. Although our work has clear limitations, such as the use of empirical weight choice without systematic sweep and single-seed evaluation, the consistency of improvements across the evaluated tasks and formulations suggests the core findings are robust.

Some of the future directions we foresee as natural extensions of our work include adaptive weighting schemes to balance data and physics loss terms on-the-fly, integration of the transport equation to capture temporal profile evolution beyond instantaneous equilibrium, closed-loop validation in tokamak control applications, and cross-device transfer to leverage universal MHD principles.

AI USE STATEMENT

Generative AI tools were not used as a component of the core methodology of this work. Specifically, we did not use generative AI tools to generate synthetic data, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement the methods, clean or reformat datasets, or interpret experimental results. Assistance with translation and qualitative or thematic data analysis was not applicable to this work.

Our use of generative AI tools was limited to recommended disclosure tasks, primarily to assist with literature discovery and retrieval and to improve the readability, clarity, and concision of the manuscript, and assist with code debugging and refinement. AI tools were used to help identify potentially relevant prior work and sources, which were subsequently checked by the authors through manual literature review. AI tools were also used to assist in drafting and editing portions of the manuscript and in debugging and refining code; all AI-assisted text and code was reviewed, revised, and verified by the authors against the underlying mathematical derivations, cited literature, and experimental results before inclusion.

The authors reviewed all AI-assisted content and take full responsibility for the accuracy, originality, and integrity of the final submission, including all text, claims, mathematical statements, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only openly published fusion measurements and equilibrium reconstructions derived from those measurements. No human subjects are involved, and the study does not use personally identifiable, private, or otherwise sensitive information. We do not collect new data or release a new dataset as part of this work. We are not aware of harmful applications of the methodology presented here beyond those already inherent to fusion equilibrium modeling.

REPRODUCIBILITY STATEMENT

All experiments use publicly available data and benchmark definitions: FAIR-MAST data (Jackson et al., 2024; 2025) and the TokaMark task definitions and splits (Rousseau et al., 2026), with the TokaMind architecture of Boschi et al. (2026). The extended tasks are specified in Section 3.1 and Table 1. The strong- and weak-form Grad–Shafranov losses are fully specified in Section 2, with the derivation and properties of the weak-form operator in Appendix C, and the evaluation metrics are defined in Section 3.2 and Appendix B.2. Architecture, optimizer settings, training schedule, loss weights, and hardware are reported in Section 4.1, and the codec configuration for every signal in Appendix E.1 and Table E.1. An anonymized snapshot of the source code implementing the losses, diagnostics, and task configurations is provided as supplementary material, and the full implementation will be released as open source upon acceptance.

REFERENCES

Nicola. C. Amorisco, Adriano Agnello, George Holt, Matthijs Mars, James Buchanan, and Stanislas Pamela. FreeGSNKE: A Python-based dynamic free-boundary toroidal plasma equilibrium solver. *Physics of Plasmas*, 31(4):042517, April 2024. doi: 10.1063/5.0188467.

Arash Ashourvan. GKFieldFlow: A spatio-temporal neural surrogate for nonlinear gyrokinetic turbulence. *Nuclear Fusion*, 2026. doi: 10.1088/1741-4326/ae7e08.

Tobia Boschi, Andrea Loreti, Nicola C. Amorisco, Rodrigo H. Ordonez-Hurtado, Cécile Rousseau, George K. Holt, Eszter Székely, Alexander Whittle, Samuel Jackson, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Sue Thorne, and Mykhaylo Zayats. TokaMind: A Multi-Modal Transformer Foundation Model for Tokamak Plasma Dynamics. *arXiv preprint arXiv:2602.15084*, 2026.

Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, 3 edition, 2008.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 6571–6583, 2018.

Philippe G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2002.

Arka Daw, Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-Guided Neural Networks (PGNN): An application in lake temperature modeling. In *Knowledge-guided machine learning: Accelerating discovery using scientific knowledge and data*, chapter 15, pp. 353–372. 2022. doi: 10.1201/9781003143376-15.

Siqi Ding, Zitong Zhang, Guoyang Shi, Xingyu Li, Xiang Gu, Yanan Xu, Huasheng Xie, Hanyue Zhao, Yuejiang Shi, and Tianyuan Liu. Physics-Informed Neural Operator Learning for Nonlinear Grad-Shafranov Equation. *arXiv preprint*, pp. 2511.19114, 2025. doi: 10.48550/arXiv.2511.19114. Submitted November 24, 2025.

Ben Dudson. FreeGS: Free Boundary Grad-Shafranov Solver. `https://github.com/freegs-plasma/freegs`, 2024. Accessed: 2026-08-18.

Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite Volume Methods. In P. G. Ciarlet and J. L. Lions (eds.), *Handbook of Numerical Analysis, Vol. VII*, pp. 713–1020. North-Holland, Amsterdam, 2000. doi: 10.1016/S1570-8659(00)07005-8.

Salah A Faroughi, Nikhil M Pawar, Celio Fernandes, Maziar Raissi, Subasish Das, Nima K Kalantari, and Seyed Kourosh Mahjour. Physics-Guided, Physics-Informed, and Physics-Encoded Neural Networks and Operators in Scientific Computing: Fluid and Solid Mechanics. *Journal of Computing and Information Science in Engineering*, 24(4):040802, 2024.

Blaise Faugeras. An overview of the numerical methods for tokamak plasma equilibrium computation implemented in the NICE code. *Fusion Engineering and Design*, 160:112020, 2020. doi: 10.1016/j.fusengdes.2020.112020.

Jeffrey P. Freidberg. *Ideal MHD*. Cambridge University Press, Cambridge, 2014.

Harold Grad and Hanan Rubin. Hydromagnetic equilibria and force-free fields. In *Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy*, volume 31 of *Theoretical and Experimental Aspects of Controlled Nuclear Fusion*, pp. 190–197, Geneva, Switzerland, 1958. United Nations.

Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun de Witt, Stanislas Pamela, Rob Akers, Jeyan Thiyagalingam, and The MAST Team. FAIR-MAST: A fusion device data management system. *SoftwareX*, 27:101869, 2024. doi: 10.1016/j.softx.2024.101869.

Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun de Witt, Stanislas Pamela, Rob Akers, and Jeyan Thiyagalingam. An Open Data Service for Supporting Research in Machine Learning on Tokamak Data. *IEEE Transactions on Plasma Science*, 53(9):2440–2449, 2025. doi: 10.1109/TPS.2025.3583419.

Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan S. Read, Jacob A. Zwart, Michael Steinbach, and Vipin Kumar. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3):1–26, 2021. doi: 10.1145/3447814.

Ehsan Kharazmi, Zhongqiang Zhang, and George Em Karniadakis. Variational Physics-Informed Neural Networks for Solving Partial Differential Equations. *arXiv preprint arXiv:1912.00873*, 2019.

Ehsan Kharazmi, Zhongqiang Zhang, and George Em Karniadakis. hp-VPINNs: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021. doi: 10.1016/j.cma.2020.113547.

Lang L. Lao, H. St. John, R.D. Stambaugh, A.G. Kellman, and W. Pfeiffer. Reconstruction of current profile parameters and plasma shapes in tokamaks. *Nuclear fusion*, 25(11):1611–1622, 1985. doi: 10.1088/0029-5515/25/11/007.

Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Society for Industrial and Applied Mathematics, Philadelphia, 2007.

Wenbin Liao, Zhengping Luo, Yao Huang, Yuehang Wang, Zijie Liu, Kai Wu, Dalong Chen, Shengze Cai, and Bingjia Xiao. EAST plasma equilibrium reconstruction based on weight-adaptive physics-informed neural network. *Nuclear Engineering and Technology*, pp. 104297, 2026.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045.

Chengping Rao, Pu Ren, Qi Wang, Oral Buyukozturk, Hao Sun, and Yang Liu. Encoding physics to learn reaction–diffusion processes. *Nature Machine Intelligence*, 5(7):765–779, 2023. doi: 10.1038/s42256-023-00685-7.

Fauzan Nazranda Rizqan, Matthew Hole, and Charles Gretton. Evaluation and verification of physics-informed neural models of the Grad-Shafranov equation. *arXiv preprint*, pp. 2504.21155, 2025. doi: 10.48550/arXiv.2504.21155.

Cécile Rousseau, Samuel Jackson, Rodrigo H. Ordonez-Hurtado, Nicola C. Amorisco, Tobia Boschi, George K. Holt, Andrea Loreti, Eszter Székely, Alexander Whittle, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Sue Thorne, and Mykhaylo Zayats. Tokamark: A comprehensive benchmark for mast tokamak plasma models. *arXiv preprint arXiv:2602.10132*, 2026.

Novella Rutigliano, Riccardo Rossi, Andrea Murari, Michela Gelfusa, Teddy Craciunescu, Didier Mazon, and Pasquale Gaudio. Physics-informed neural networks for the modelling of interferometer-polarimetry in tokamak multi-diagnostic equilibrium reconstructions. *Plasma Physics and Controlled Fusion*, 2025. doi: 10.1088/1361-6587/addde6.

Alexander A. Samarskii. *The Theory of Difference Schemes*, volume 240 of *Pure and Applied Mathematics*. Marcel Dekker, New York, 2001.

Vitaly D. Shafranov. On magnetohydrodynamical equilibrium configurations. *Soviet Physics JETP*, 6(3):545–554, 1958.

Xuan Sun, Cihan Akçay, Torrin Bechtel, Scott E. Kruger, Lang L. Lao, Yueqiang Liu, Sandeep Madireddy, Joseph McClenaghan, and EFIT-AI Team. Impact of various DIII-D diagnostics on the accuracy of neural network surrogates for kinetic EFIT reconstructions. *Nuclear Fusion*, 64 (8):086065, 2024. doi: 10.1088/1741-4326/ad5d7b.

Chenguang Wan, Zhi Yu, Alessandro Pau, Olivier Sauter, Xiaojuan Liu, Qiping Yuan, and Jiangang Li. A machine-learning-based tool for last closed-flux surface reconstruction on tokamaks. *Nuclear Fusion*, 63(5):056019, 2023. doi: 10.1088/1741-4326/acbfcc.

John Wesson and David J Campbell. *Tokamaks*, volume 149. Oxford university press, 2011.

Yaohua Zang, Gang Bao, Xiaojing Ye, and Haomin Zhou. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020. doi: 10.1016/j.jcp.2020.109409.

Lorenzo Zanisi, Aaron Ho, Jackson Barr, Thandi Madula, Jonathan Citrin, Stanislas Pamela, James Buchanan, Francis J. Casson, Vignesh Gopakumar, and JET Contributors. Efficient training sets for surrogate models of tokamak turbulence with active deep ensembles. *Nuclear Fusion*, 64(3): 036022, 2024. doi: 10.1088/1741-4326/ad240d.

A MHD CONTEXT FOR THE GRAD-SHAFRANOV EQUATION

Tokamak plasma dynamics are governed by a coupled nonlinear system combining conservation of mass, momentum and energy, Maxwell's equations, a generalized Ohm's law, and transport equations for heat and particle diffusion (Freidberg, 2014; Wesson & Campbell, 2011). This appendix records only the chain relevant to this work: from that system to quasi-static force balance, and from force balance to the Grad-Shafranov equation.

From momentum balance to force balance. The plasma momentum equation, the analogue of Newton's second law for a continuous medium, is

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -\nabla p + \mathbf{J}\times\mathbf{B} + \mathbf{f}_{\text{visc}} + \mathbf{f}_{\text{ext}}, \tag{27}$$

where ρ is the mass density, $\mathbf{v}$ the bulk flow velocity, p the scalar pressure, $\mathbf{J}$ the current density and $\mathbf{B}$ the magnetic field. The Lorentz force $\mathbf{J}\times\mathbf{B}$ is the coupling to the electromagnetic field and is what confines the plasma against its own pressure gradient. Magnetic equilibrium is established on the Alfvén timescale, which is short compared with the transport and resistive timescales on which p and the current profile evolve (Freidberg, 2014). On the latter timescales the inertial terms in Eq. (27) are therefore negligible and, neglecting viscous and external forces, it reduces to the force balance condition

$$\nabla p = \mathbf{J}\times\mathbf{B}, \tag{28}$$

which states that pressure gradients are balanced pointwise by Lorentz forces.

From force balance to Grad-Shafranov. In axisymmetric toroidal geometry with nested magnetic flux surfaces, Eq. (28) combined with Ampère's law reduces to a single second-order nonlinear PDE for the poloidal magnetic flux function $\psi(R,Z)$ in cylindrical coordinates (R,Z,ϕ), with ϕ the toroidal angle: the Grad-Shafranov (GS) equation (Grad & Rubin, 1958; Shafranov, 1958),

$$\Delta^*\psi \equiv R\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial\psi}{\partial R}\right) + \frac{\partial^2\psi}{\partial Z^2} = -\mu_0 R^2\frac{\mathrm{d}p}{\mathrm{d}\psi} - F\frac{\mathrm{d}F}{\mathrm{d}\psi}, \tag{29}$$

where μ_0 is the vacuum permeability, $p(\psi)$ is the plasma pressure, and $F(\psi) = RB_\phi$ is the poloidal current function. Both source terms are flux functions: the pressure-driven term $-\mu_0 R^2\,\mathrm{d}p/\mathrm{d}\psi$ and the poloidal-current (diamagnetic) term $-F\,\mathrm{d}F/\mathrm{d}\psi$. Dividing by $-\mu_0 R$ identifies their sum with the toroidal current density J_ϕ, which is the form used in Eq. (8). The equation is solved on the poloidal cross-section subject to boundary conditions specified either at the plasma-vacuum interface (fixed-boundary problem) or determined self-consistently with external coil currents (free-boundary problem) (Lao et al., 1985).

B SPECIAL CONSIDERATIONS FOR GRAD-SHAFRANOV-BASED RESIDUAL AND LOSSES

B.1 CONSIDERATIONS FOR THE CLEANING MASK M

When TokaMind is trained following TokaMark's protocols (Rousseau et al., 2026), ψ is a specified model output in TokaMark's Tasks 1-3 and 2-3, and ψ must be reconstructed or predicted from ground truth data defined at every point in a 65x65 grid. With this, both ground truth and predicted values of ψ will have non-zero values outside the plasma boundary (e.g., around the toroidal magnetic coils), and thus the term $T\psi$ in Eq. (1) will behave similarly. Therefore, a suitable cleaning mask is needed to restrict the application of GS-based loss function metrics where the GS equation holds in practice, that is, inside the time-varying plasma region.

There are different ways to construct such cleaning mask (e.g., taking it directly from ground truth data), but in our work we select two methods:

- If J_ϕ is not a specified model output (as in TokaMark's Task 1-3 and 2-3), we use a fixed, geometry-based mask.

- If J_ϕ is also a specified model output (e.g., by extending TokaMark's Task 1-3 and 2-3 to also include it), we use J_ϕ^{true} as a proxy to build the mask as

$$M(i,j) = \begin{cases} 1, & \text{if } J_\phi^{true}(i,j) > 10^{-6}, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

B.2 Considerations for the loss function metric $\mathcal{M}(\mathrm{GS}_{res})$

To illustrate the special considerations for candidate loss function metrics, we select two options:

(i) the Mean-Squared Error (MSE), denoted by $\mathcal{R}(\cdot)$ and defined as

$$\mathcal{R}(\mathrm{GS}_{res}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{GS}_{res}(i)\right)^2, \tag{31}$$

(ii) the L2 norm, denoted by $\|\cdot\|_2$ and defined as

$$\|\mathrm{GS}_{res}\|_2 = \sum_{i=1}^{n}\left(\mathrm{GS}_{res}(i)\right)^2. \tag{32}$$

As an important remark, when $\mathcal{M}(\cdot)$ is set as $\mathcal{R}(\cdot)$, it is necessary to take the mean of squares over the active cells of a target field, which requires an additional step to identify the number of valid entries n. Alternatively, when $\mathcal{M}(\cdot) = \|\cdot\|_2$ is selected, the averaging action is not included, and therefore a step to calculate the number of valid entries can be avoided. With this, it is clear that the selection of $\mathcal{M}(\cdot)$ has a direct impact on the hyper-parameter tuning process for w_{GS} in Eq. (26).

C Weak Form of the Grad-Shafranov Equation and its Discretization

Throughout, $\nabla = (\partial_R, \partial_Z)$ is the gradient in the poloidal plane, h_R and h_Z are the grid spacings, and each grid rectangle is split into two right triangles by a fixed diagonal.

C.1 Variational identity

Dividing Eq. (8) by R and using $\Delta^*\psi = R\,\nabla\cdot\left(R^{-1}\nabla\psi\right)$ gives the divergence (conservation) form

$$-\nabla\cdot\left(\frac{1}{R}\,\nabla\psi\right) \;=\; \mu_0\, J_\phi, \tag{33}$$

in which both sides carry units of $\mathrm{Wb\,m^{-3}}$; the factor R multiplying J_ϕ in Eq. (8) is absorbed by the division. Multiplying by a test function and applying Green's first identity yields Eq. (12), the standard variational formulation used by finite-element equilibrium codes (e.g. Faugeras, 2020), the only difference being that we evaluate its residual rather than solve for ψ. The two formulations carry different units, which is why w_{GS} in Eq. (26) is fixed empirically: with $\mathcal{M}(\cdot) = \mathcal{R}(\cdot)$ the squared strong- and weak-form residuals have units $\mathrm{Wb^2\,m^{-4}}$ and $\mathrm{Wb^2\,m^{-6}}$ respectively, and neither is commensurable with $\mathcal{L}_{\text{data}}$.

The weak loss is evaluated only at nodes inside the cleaning masked region, so the contributing test functions are the nodal basis functions centred at those nodes. Since that region does not reach the grid boundary, every contributing test function is supported in the interior of Ω and the boundary integral in Eq. (12) vanishes identically, without any Dirichlet or Neumann assumption on ψ: outside the masked region ψ^{pred} is constrained only by $\mathcal{L}_{\text{data}}$. Note that the mask selects test functions; it is not a restriction of the computational domain to the plasma region, which would instead introduce a boundary term on the plasma boundary.

C.2 Reduction to the edge-flux operator

Expanding $\psi_h = \sum_j \psi_j \phi_j$ in the continuous piecewise-linear nodal basis and testing with $v = \phi_i$ gives the Galerkin system $A\psi = f$ with $A_{ij} = \int_\Omega R^{-1}\nabla\phi_i\cdot\nabla\phi_j\,\mathrm{d}A$ and $f_i = \mu_0\int_\Omega J_\phi\phi_i\,\mathrm{d}A$ (Ciarlet, 2002; Brenner & Scott, 2008). On this triangulation the gradients of two diagonal-neighbour basis functions are orthogonal on both shared triangles, so diagonal couplings vanish for any integrable

coefficient; and the basis is a partition of unity, so row sums vanish. Hence A is determined by its four axis off-diagonal entries and acts as a weighted graph Laplacian,$(A\psi)_i = \sum_{j\in\mathcal{N}(i)} \tau_{ij}(\psi_i - \psi_j)$ with $\tau_{ij} = -A_{ij}$.

For an R-direction edge the two adjacent triangles give $\nabla\phi_i \cdot \nabla\phi_j = -h_R^{-2}$ over a union of area $h_R h_Z$, so

$$\tau_{ij} = \frac{1}{h_R^2} \int_{T\cup T'} \frac{\mathrm{d}A}{R} = \frac{h_Z}{h_R} \cdot \frac{1}{L}, \tag{34}$$

with L the logarithmic mean of R_i and R_{i+1}. We replace this by midpoint quadrature, $1/L \to 1/R_e$ with $R_e = \frac{1}{2}(R_i + R_{i+1})$, a relative perturbation of order $(h_R/R)^2$. For Z-direction edges R is constant along the edge, so $R_e = R_i$, and the same quadrature incurs a perturbation of the same order relative to the exact triangle average.

Finally, the nodal basis satisfies $\int_\Omega \phi_i = h_R h_Z$ at interior nodes, so row-sum lumping gives $M_L = h_R h_Z I$ and $f_i \to \mu_0 J_{\phi,i} h_R h_Z$. Applying M_L^{-1} to the Galerkin residual returns exactly $W\psi - \mu_0 J_\phi$ with W as in Eq. (13). The rescaling is a global constant equal to the cell area and is absorbed into w_{GS}.

C.3 Properties of W

All edge weights are positive, so W is symmetric positive semidefinite with the constant fields in its kernel. Equation (13) is also the two-point-flux finite-volume discretization of Eq. (33) scaled by the cell area, and for the smooth positive coefficient $1/R$ on a uniform grid it is second-order consistent with Eq. (33) (Eymard et al., 2000; Samarskii, 2001). It is evaluated in practice by edge-wise flux accumulation, each edge visited once and contributing with opposite sign at its two endpoints, which makes the operator symmetric and locally conservative by construction. The discrete operator therefore plays the role of $-\Delta^*\psi/R$, which is what makes Eq. (14) the weak-form counterpart of Eq. (8), and never forms the second derivatives or the $R^{-1}\psi_R$ term of Δ^* explicitly. As in the strong form, the reference fields satisfy Eq. (33) inside the plasma boundary, so $J_\phi^{true} = \mu_0^{-1} W \psi^{true}$ up to discretization error; no explicit radial weight is needed because the factor R is already carried by the edge coefficients R_e^{-1}.

D Over-fitting exercise

We conducted over-fitting experiments on Task 1-4, which predicts the poloidal flux ψ and toroidal current density J_ϕ, using data from a single FAIR-MAST shot (Jackson et al., 2025; 2024) (shot ID: 30421). We trained three models for 10 epochs each, using data-only supervision, data supervision combined with the strong-form GS loss, and data supervision combined with the weak-form GS loss, respectively. Figure D.1 presents a qualitative comparison of the predictions at the end of the 10 training epochs. In these examples, GS-informed training improves agreement with the reference in operator space, even when differences between the flux maps themselves are visually subtle.

E Methodological considerations

E.1 Output embedding dimension

TokaMind represents each signal by a DCT3D codec whose retained coefficient set is tuned per signal by ranked energy selection: coefficients are ranked by aggregate energy over sampled training windows, and the smallest prefix reaching a target explained-energy threshold is retained, subject to a budget $K_{\max}$ (Boschi et al., 2026). Unlike a fixed low-pass truncation, this set may include high-frequency coefficients whenever they carry energy.

We tune codecs on 100 shots (at most 15,000 windows) with thresholds 0.999 for inputs and actuators and $K_{\max} = 4096$, and, for the gridded outputs ψ and J_ϕ, a threshold of 0.99 with $K_{\max} = 8192$. The output threshold is the one setting we deliberately lower relative to the published downstream configuration, which uses 0.999. The reason is specific to the diagnostics of Section 3.2.3: because ranked selection admits high-frequency coefficients whenever they carry energy, and because a fraction of that energy in experimental reconstructions is measurement noise

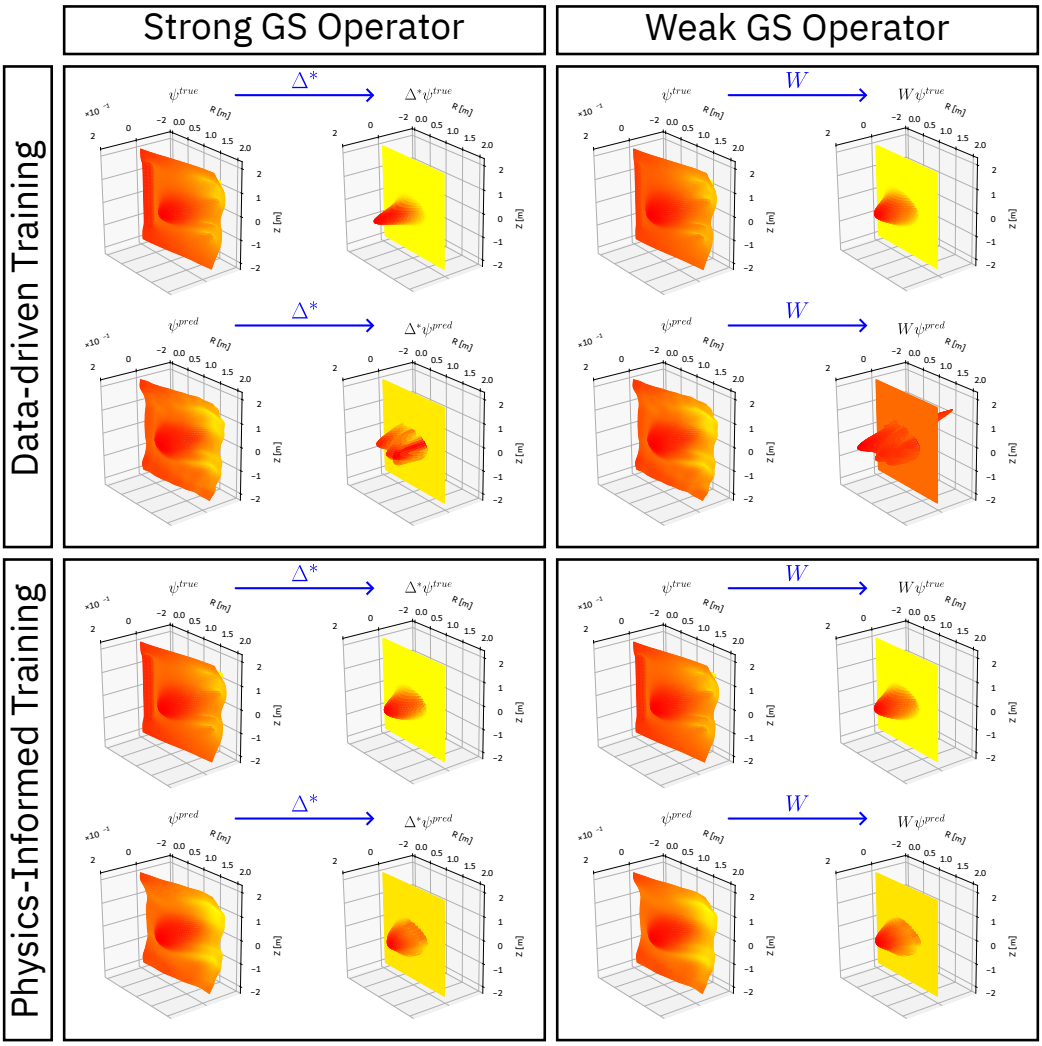


Figure D.1: Results from over-fitting experiments on Task 1-4 using TokaMind models trained via both data-driven and (GS-based) physics-informed learning approaches. Each block of plots shows two mappings for a given training approach and GS operator, namely, $\psi^{true} \rightarrow T\psi^{true}$ and $\psi^{pred} \rightarrow T\psi^{pred}$, where T is the corresponding GS operator (strong or weak), ψ^{true} is a sample of ground-truth data, and ψ^{pred} is a sample of predicted data at the end of the training process. Direct comparison between predicted results in operator space and the corresponding ground-truth values reveal a substantial improvement when using physics-informed training. [**REMARK:** for ease of visualization, data samples (2D arrays) were plotted via Matplotlib's *plot_surface* method.]

rather than equilibrium structure, a higher threshold admits oscillatory content into the decoded field. The operator Δ^* amplifies exactly that content, so oscillations that are minor in ψ are pronounced in $\Delta^*\psi$, which is the field both the physics term and the GS diagnostics act on. At 0.99, the decoded fields retained the equilibrium structure in our experiments while reducing this oscillation. Minimum-coverage guardrails (10 unique indices per spatial dimension) apply in addition, and for some signals it is the guardrail rather than the threshold that fixes the final dimension. Table E.1 reports the resulting settings.

Two consequences follow and are stated here so that the results are read correctly. First, the codec bounds achievable native accuracy: no model can reconstruct structure outside the span of its retained coefficients, so part of the reported NRMSE is a codec floor rather than a model error. Second, because the published TokaMind results for Tasks 1-3 and 2-3 use a threshold of 0.999, our

NRMSE values are not directly comparable with those numbers; comparisons in this work are internal, between models trained under identical codecs.

Table E.1: Tuned DCT3D output codecs at explained-energy threshold 0.99 and $K_{\max} = 8192$. K_g is the retained coefficient count and EE the achieved explained energy.

Task	**Signal**	(H, W, T)	K_g	**EE**
Task 1-3	equilibrium-psi	$(65, 65, 1)$	39	0.9918
Task 1-4	equilibrium-psi	$(65, 65, 1)$	39	0.9918
Task 1-4	equilibrium-j_tor	$(65, 65, 1)$	94	0.9901
Task 1-6	equilibrium-psi	$(65, 65, 1)$	39	0.9918
Task 1-6	equilibrium-j_tor	$(65, 65, 1)$	94	0.9901
Task 2-3	equilibrium-psi	$(65, 65, 5)$	48	0.9900
Task 2-4	equilibrium-psi	$(65, 65, 5)$	48	0.9900
Task 2-4	equilibrium-j_tor	$(65, 65, 5)$	356	0.9954
Task 2-6	equilibrium-psi	$(65, 65, 5)$	48	0.9900
Task 2-6	equilibrium-j_tor	$(65, 65, 5)$	356	0.9954

E.2 Discussion on GS evaluation metrics

What E_{self} measures. Equation (21) involves no ground truth. It asks whether the two fields the model emits are mutually compatible with force balance, that is, whether the pressure gradient implied by the predicted flux is in fact balanced by the predicted Lorentz force. A model can be accurate in NRMSE and still emit a (ψ, J_ϕ) pair that does not correspond to any MHD equilibrium, and E_{self} is the quantity that detects this; conversely it is exactly the quantity a no-anchor physics term acts on. Two cautions follow from its construction. First, low E_{self} certifies consistency, not correctness: the trivial pair $\psi = J_\phi = 0$ scores zero, and the GS equation admits infinitely many solutions, so the metric must be read alongside NRMSE rather than in place of it. Second, it requires a predicted current density and is therefore available only when ψ and J_ϕ are both specified model outputs.

What E_ψ measures. Because Δ^*_{FD} is a linear operator, Eq. (22) equals the mean absolute value of $\Delta^*_{\text{FD}}\big(\psi^{true} - \psi^{pred}\big)$: it is a data error on ψ measured in operator space rather than in signal space. This is worth reporting separately from NRMSE_ψ because the two are sensitive to different failure modes. Δ^* amplifies high-frequency content, so a small-amplitude, short-wavelength error that is invisible in NRMSE_ψ can dominate E_ψ, and two reconstructions with the same ψ-NRMSE can differ by orders of magnitude in E_ψ. The physical content is direct: every equilibrium quantity of interest is a derivative of ψ, since $B_R = -R^{-1}\partial_Z\psi$, $B_Z = R^{-1}\partial_R\psi$, and the current density follows from Ampère's law, and $\Delta^*\psi$ is precisely the combination of those derivatives that enters force balance. E_ψ therefore measures the accuracy of the predicted flux in the norm the physics actually uses, and a reconstruction that is accurate pointwise but rough at the grid scale will be penalized here even when NRMSE is satisfied. Crucially, E_ψ is built from ψ alone and is therefore available for all six tasks, including the ψ-only Tasks 1-3/2-3.

Reading E_{self} and E_ψ together. The two reported columns answer complementary questions: E_{self} asks whether the prediction is physically possible, and E_ψ asks whether it is the right one. They are related but not redundant. Since the reference fields satisfy $\text{LHS}[\psi^{true}] = \text{RHS}[J_\phi^{true}]$ on $\mathcal{C}$ up to discretization error, inserting this identity into Eq. (21) and applying the triangle inequality gives

$$E_{\text{self}} \;\le\; E_\psi + E_J, \tag{35}$$

so a model cannot achieve small errors on both sides while being self-inconsistent. The converse fails, which is the informative direction: E_{self} can be small while E_ψ and E_J are both large, corresponding to a self-consistent equilibrium that is not the observed one. Reporting the two components separately avoids conflating physical self-consistency with operator-space accuracy in a single aggregate metric.

Why E_J is not a headline metric. Equation (23) contains no differential operator. Since $\mathrm{RHS}[\cdot]$ is multiplication by the fixed positive field $\mu_0 R$, the quantity reduces to

$$E_J = \frac{1}{|\mathcal{C}|} \sum_{\mathcal{C}} \mu_0 R \, |J_\phi^{true} - J_\phi^{pred}|, \tag{36}$$

a radially weighted mean absolute error on the current density. It therefore measures the same thing as NRMSE_{J_ϕ} up to a smooth pointwise reweighting and the choice of absolute rather than squared deviation. This is the structural asymmetry between the two anchored quantities: E_ψ passes through Δ^*_{FD}, which amplifies high-frequency content and can therefore separate reconstructions that NRMSE_ψ ranks alike, whereas multiplication by $\mu_0 R$ shifts the relative weight of inboard and outboard cells but cannot make E_J sensitive to any spatial scale to which NRMSE_{J_ϕ} is blind.

Independence of the evaluation metrics from the training objective. Beyond redundancy with NRMSE_{J_ϕ}, an important consideration is whether a diagnostic can be evaluated independently of the training objective. Each of Eqs. (21)–(23) has the same algebraic form as one of the residual candidates derived in Sections 2.3 and 2.2; therefore, algebraic form alone does not distinguish a diagnostic from a training term. The distinction comes from the operator used to evaluate it. The strong-form residuals are assembled with the precomputed sparse operator of Eq. (8) and the weak-form residuals with the edge-flux operator W of Eq. (13), whereas E_{self} and E_ψ are built from Δ^*_{FD}, which neither loss optimizes. A model trained with either formulation is therefore scored on a discretization it never saw, which is the whole purpose of Eq. (19).

No such separation is available for E_J. The R-anchored residual involves no operator acting on a predicted field in either formulation: it is $\mu_0 R\,(J_\phi^{true} - J_\phi^{pred})$ in the strong form and $\mu_0 (J_\phi^{true} - J_\phi^{pred})$ in the weak form. The two differing only in whether the radial factor is applied pointwise or is already carried by the edge coefficients of W. Since there is no discretization in which the metric and the training term could differ, reporting E_J as a headline column would grade an R-anchored model on a quantity it was directly trained to minimize, reintroducing exactly the circularity that the finite-difference choice was made to avoid, and would in addition report the current data error a second time under a different name. We therefore compute and store E_J, where it serves as the third leg of the bound in Eq. (35) and as a check that the reference pair satisfies force balance on $\mathcal{C}$, and report the current density data error once, as NRMSE_{J_ϕ}.

This also fixes how the R-anchored term should be described on the training side, in both formulations. Carrying no operator, it acts as a supervised term on J_ϕ expressed in physics units rather than as an additional physics constraint, and we do not present it as the latter. For example, in the weak form its anchor is $\mu_0 J_\phi^{true}$ and the term therefore pulls J_ϕ^{pred} toward consistency with the flux the model is simultaneously being asked to reproduce, rather than toward J_ϕ^{true} in isolation. In either formulation, the no-anchor residual is the only component that constrains the predicted pair without reference to ground truth, so attributing a change in J_ϕ accuracy to the physics rather than to supervision requires the weights themselves: setting $w_{RA} = 0$ removes every direct supervision of J_ϕ^{pred} from within the physics loss and leaves the no-anchor residual as the sole term acting on the current density, while the native-space data loss continues to supervise it as usual. We treat this as the relevant ablation and make no claim at this stage about the relative contribution of the data and physics terms to J_ϕ accuracy.

E.3 Companion results at a higher codec threshold

The results of Table 2 use an output codec at explained-energy threshold 0.99 with $w_{GS} = 0.1$. We report here one companion configuration that uses a higher codec fidelity and a reduced w_{GS} value, as shown in Table E.2. Since a higher threshold retains more coefficients and the decoded fields therefore carry more of the high-frequency content that the GS residual acts on, holding w_{GS} fixed would change the realized contribution of the physics term rather than isolate the effect of the codec. Threshold and weight consequently vary together, so Table E.2 is indeed a companion configuration rather than a sweep of either setting on its own, and comparisons should be read within a table rather than across them. All other settings, including shots, splits, seed, epoch budget and optimizer, match Section 4.1.

Two points should be kept in mind when reading Table E.2. First, the codec bounds achievable native accuracy, so NRMSE values are comparable only among models sharing a codec; this applies between Table 2 and Table E.2 as much as it does against the published TokaMark numbers. Second, E_{self} and E_{ψ} pass through Δ^{*}_{FD}, which amplifies the high-frequency content that a higher threshold admits, so their absolute magnitudes are expected to differ between the two tables even where the effect of the physics term does not.

Table E.2: Companion results at explained-energy threshold 0.999 with $w_{GS} = 0.05$. Columns, region $\mathcal{C}$, and **bold**/underline conventions are as in Table 2; loss notation $\mathcal{L}_{d+S}$=$0.95 * \mathcal{L}_{data} + 0.05 * \mathcal{L}_{SGS}$ and $\mathcal{L}_{d+W}$=$0.95 * \mathcal{L}_{data} + 0.05 * \mathcal{L}_{WGS}$.

Task, Loss	NRMSE$_{\psi}$	NRMSE$_{J_\phi}$	E_{self}	E_{ψ}	ϵ_{rough}
1-3, $\mathcal{L}_{data}$	**0.1302**	–	–	0.3565	+0.2043
1-3, $\mathcal{L}_{d+S}$	0.1424 (↑9.4%)	–	–	0.2178 (↓38.9%)	**+0.1604**
1-3, $\mathcal{L}_{d+W}$	0.1462 (↑12.3%)	–	–	**0.2153** (↓39.6%)	+0.1667
1-4, $\mathcal{L}_{data}$	**0.0549**	**0.1078**	0.3455	0.3457	+0.2116
1-4, $\mathcal{L}_{d+S}$	0.0652 (↑18.8%)	0.1106 (↑2.6%)	**0.2238** (↓35.2%)	**0.2267** (↓34.4%)	**+0.2017**
1-4, $\mathcal{L}_{d+W}$	0.0729 (↑32.8%)	0.1142 (↑5.9%)	0.2335 (↓32.4%)	0.2365 (↓31.6%)	+0.2047
1-6, $\mathcal{L}_{data}$	**0.0636**	**0.1366**	0.3462	0.3468	+0.2059
1-6, $\mathcal{L}_{d+S}$	0.0709 (↑11.5%)	0.1375 (↑0.7%)	**0.2213** (↓36.1%)	**0.2263** (↓34.7%)	**+0.1951**
1-6, $\mathcal{L}_{d+W}$	0.0769 (↑20.9%)	0.1398 (↑2.3%)	0.2338 (↓32.5%)	0.2388 (↓31.1%)	+0.1952
2-3, $\mathcal{L}_{data}$	**0.0947**	–	–	0.3243	**+0.1389**
2-3, $\mathcal{L}_{d+S}$	0.1001 (↑5.7%)	–	–	0.2235 (↓31.1%)	+0.1448
2-3, $\mathcal{L}_{d+W}$	0.0987 (↑4.2%)	–	–	**0.2134** (↓34.2%)	+0.1494
2-4, $\mathcal{L}_{data}$	**0.0787**	0.1216	0.3189	0.3184	**+0.1404**
2-4, $\mathcal{L}_{d+S}$	0.0839 (↑6.6%)	**0.1206** (↓0.8%)	0.2153 (↓32.5%)	0.2174 (↓31.7%)	+0.1756
2-4, $\mathcal{L}_{d+W}$	0.0889 (↑13.0%)	0.1218 (↑0.2%)	**0.2120** (↓33.5%)	**0.2144** (↓32.7%)	+0.1746
2-6, $\mathcal{L}_{data}$	**0.0842**	**0.1388**	0.3170	0.3169	**+0.1339**
2-6, $\mathcal{L}_{d+S}$	0.0903 (↑7.2%)	0.1407 (↑1.4%)	0.2154 (↓32.1%)	0.2189 (↓30.9%)	+0.1674
2-6, $\mathcal{L}_{d+W}$	0.0963 (↑14.4%)	0.1431 (↑3.1%)	**0.2109** (↓33.5%)	**0.2150** (↓32.2%)	+0.1676

E.4 Exploratory study on downstream transfer

We include a companion table with results from an exploratory downstream-transfer study involving the fine-tuning of TokaMind models for Task 1-1, 1-2, 2-1, and 2,2 (see TokaMark (Rousseau et al., 2026) for task definitions) using our Task 1-4 and Task 2-4 trained models reported in Table 2 as warm-start checkpoints. All downstream runs use data-only supervision and a 20-epoch training budget. Warm-start runs freeze the token encoder and backbone for the first five epochs, then train the entire model for the remaining 15; scratch baselines train the entire model for all 20 epochs.

Table E.3: Downstream test NRMSE averaged across target signals (lower is better). Scratch denotes random initialization; *Data-only*, *Data with strong GS*, and *Data with weak GS* denote warm starts from the corresponding Task 1-4 or 2-4 source models. All downstream runs use data-only training with a 20-epoch budget. Percentages compare warm start values with the matching scratch value. Bold and underline mark the best and second-best results in each row.

Task	Average NRMSE			
	Scratch	Warm start		
		Data-only	Data with strong GS	Data with weak GS
1-1	**0.140618**	0.143842 (↑2.29%)	0.145213 (↑3.27%)	0.150255 (↑6.85%)
1-2	**0.042082**	0.043941 (↑4.42%)	0.043890 (↑4.30%)	0.044304 (↑5.28%)
2-1	0.180871	0.172324 (↓4.73%)	**0.166878** (↓7.74%)	0.169687 (↓6.18%)
2-2	**0.042383**	0.043481 (↑2.59%)	0.043027 (↑1.52%)	0.043213 (↑1.96%)