WHAT DOES A NEURAL PDE SOLVER REALLY LEARN? A RESIDUAL-SPECTRUM DIAGNOSTIC

Ali Baheri
Department of Mechanical Engineering
Rochester Institute of Technology
Rochester, NY 14623, USA
akbeme@rit.edu

ABSTRACT

Neural PDE surrogates are typically evaluated using solution error (e.g., relative L_2), but low error does not guaranty that predictions satisfy the governing equations. We propose *Residual-Spectrum Diagnostics* (RSD), which evaluates physics compliance by analyzing the spatial frequency content of the PDE residual computed on model rollouts. RSD summarizes scale-dependent violations using two indices: *High-Frequency Violation* (HFV) for spurious small-scale artifacts and *Low-Frequency Violation* (LFV) for large-scale dynamical errors. On the 1D viscous Burgers equation, two models with similar relative L_2 error differ substantially in residual spectra: training with high frequency corrupted targets increases L_2 error by only 6% but increases HFV by 37% ($p < 10^{-3}$). These results show that residual-spectrum analysis reveals physics failures that aggregate error metrics can miss and provides an actionable complement to standard evaluation.

1 INTRODUCTION

Neural networks have emerged as powerful surrogates for solving partial differential equations, offering dramatic speedups over classical numerical methods. Architectures such as Fourier Neural Operators (FNO) and Physics-Informed Neural Networks (PINNs) can learn solution mappings from data, enabling rapid inference for applications ranging from climate modeling to engineering design optimization. However, as these models move toward deployment in safety-critical domains, a fundamental question arises: *how do we know when a neural PDE solver is actually learning the physics?* Current evaluation practice centers on solution accuracy metrics, primarily relative L_2 error between predicted and ground-truth solutions. While intuitive, this approach has a critical blind spot. A model can achieve low solution error by learning superficial patterns in the training data, interpolating between seen examples or exploiting statistical regularities without internalizing the governing equations. Such models may perform well on in-distribution test cases yet fail catastrophically when encountering novel initial conditions, longer time horizons, or different physical regimes. The solution error, by itself, cannot distinguish genuine physical understanding from sophisticated curve fitting.

This paper argues that **evaluating neural PDE solvers requires examining not just what they predict but also how well those predictions satisfy the underlying physics**. We introduce *Residual-Spectrum Diagnostics* (RSD), a framework that analyzes the frequency content of PDE residuals to characterize model failures. The key insight is that the physics residual—the degree to which a predicted solution violates the governing equation—carries rich diagnostic information when decomposed across spatial scales. A model producing spurious high-frequency oscillations will exhibit elevated residual energy at fine scales, even if its overall solution error remains low. Conversely, a model missing large-scale conservation properties will show residual energy concentrated at low frequencies. RSD provides two complementary indices: the *High-Frequency Violation* (HFV), measuring spurious small-scale artifacts, and the *Low-Frequency Violation* (LFV), capturing failures in bulk dynamics. These metrics reveal failure modes invisible to standard evaluation, enabling practitioners to diagnose problems and guide model improvement. Importantly, RSD requires only the ability to evaluate the PDE operator on model outputs; no ground-truth solutions are needed beyond an initial validation set. We validate RSD on the one-dimensional Burgers equation, demonstrat-

ing that models with nearly identical L_2 error can exhibit dramatically different HFV values. In controlled experiments, we show that a model trained on data corrupted with high-frequency noise achieves only 6% higher solution error than a cleanly-trained baseline, yet exhibits 37% higher HFV ($p < 0.001$). This confirms that spectral residual analysis captures physics violations that aggregate error metrics substantially underestimate.

Contributions. We make three contributions: (1) we formalize the distinction between solution accuracy and physics compliance, arguing that both are necessary for trustworthy neural PDE solvers; (2) we introduce the RSD framework with concrete diagnostic indices (HFV, LFV) that decompose physics violations by spatial scale; and (3) we provide empirical validation showing that RSD detects failure modes missed by standard metrics.

Organization. Section 2 presents the RSD methodology. Section 3 describes our experimental setup and reports numerical results. Section 4 discusses implications and limitations.

2 Related Work

Neural PDE Solvers. Deep learning approaches to solving partial differential equations have progressed rapidly in recent years. Physics-Informed Neural Networks (PINNs) embed governing equations directly into the training objective, enabling learning from sparse or unlabeled data (Raissi et al., 2019). Neural operators instead learn mappings between infinite-dimensional function spaces, with the Fourier Neural Operator (FNO) achieving strong empirical performance by parameterizing convolution kernels in the frequency domain (Li et al., 2021). DeepONet provides an alternative operator-learning framework grounded in universal approximation results for operators (Lu et al., 2021). Subsequent extensions include physics-informed neural operators (PINO) that combine data-driven and physics-based losses (Li et al., 2024), hierarchical and multiscale architectures (Liu & Cai, 2022), and transformer-based operator models (Li et al., 2023; Hao et al., 2023). While these approaches demonstrate impressive predictive accuracy, their evaluation is still dominated by aggregate solution error metrics.

Physics-Informed Learning. Incorporating physical structure into learning-based models has a long history. Prior work has explored Lagrangian and Hamiltonian formulations for learning dynamical systems (Greydanus et al., 2019; Cranmer et al., 2020), as well as enforcing conservation laws through architectural constraints (Beucler et al., 2021) or regularization terms in the loss function (Wang et al., 2021). Comprehensive surveys of physics-informed machine learning are provided by Karniadakis et al. (2021). These methods aim to improve generalization by embedding known physics into training; however, they largely focus on *how to train* physics-consistent models rather than *how to evaluate* whether a trained model respects the governing equations across scales.

Evaluation and Benchmarking of Neural PDE Models. Standard evaluation of neural PDE solvers relies on relative L_2 or L_∞ error against high-fidelity reference solutions (Takamoto et al., 2022). Benchmarks such as PDEBench (Takamoto et al., 2022) and PDEArena (Gupta & Brandstetter, 2022) provide standardized datasets and metrics spanning a wide range of equations and regimes. Recent studies have examined generalization across spatial resolutions (Li et al., 2021), physical parameters (Wang et al., 2023), and initial conditions (Brandstetter et al., 2022). However, these evaluations primarily assess predictive accuracy and do not directly quantify whether model predictions satisfy the underlying PDE. Krishnapriyan et al. (2021) analyzed failure modes of PINNs, highlighting optimization challenges that can prevent convergence despite sufficient model capacity. Our work complements these efforts by providing diagnostics that characterize *how* and *where* trained models violate the governing equations.

Spectral Methods and Frequency Analysis. Spectral analysis is a foundational tool in numerical PDEs (Canuto et al., 2007) and has recently influenced neural model design. The effectiveness of FNOs is partly attributed to the compact representation of many PDE dynamics in the frequency domain (Li et al., 2021). Neural networks are known to exhibit spectral bias, learning low-frequency components before higher frequencies (Rahaman et al., 2019), and Fourier feature mappings have been shown to facilitate the learning of high-frequency functions (Tancik et al., 2020). In the context of neural operators, Wang et al. (2022) analyzed how spectral properties affect approximation

quality. In contrast to prior work that applies spectral analysis to the *solution representation* or *model architecture*, our approach analyzes the *physics residual* in frequency space, using spectral decomposition as an evaluation and diagnostic tool rather than as a training mechanism.

Uncertainty, Reliability, and Multi-Scale Evaluation. Reliability assessment for neural PDE solvers has increasingly focused on uncertainty quantification. Bayesian neural operators and ensemble-based approaches provide predictive uncertainty estimates (Yang et al., 2021; Lakshminarayanan et al., 2017), and recent surveys summarize uncertainty quantification methods for physics-informed learning (Psaros et al., 2023). More recently, conformal prediction methods have been extended to function-valued outputs and neural operators, providing distribution-free uncertainty sets and diagnostics for autoregressive rollout degradation (Millard et al., 2025). Related work has also emphasized the importance of multi-scale structure in reliability assessment, developing conformal prediction schemes that guaranty finite-sample coverage across hierarchical resolutions (Baheri & Shahbazi, 2025). These approaches address *calibration and uncertainty* of predictions, but do not directly assess whether predicted solutions satisfy the governing equations. Our work is complementary: rather than quantifying uncertainty or coverage, we diagnose *physics noncompliance* by decomposing PDE residuals across spatial scales, revealing failure modes that may persist even when predictions are well-calibrated or accurate in aggregate.

3 METHODOLOGY

Standard evaluation of neural PDE solvers relies on solution error metrics such as relative L_2 error. While intuitive, these metrics can be misleading: a model may achieve low error by fitting training data without learning the underlying physics. We propose *Residual-Spectrum Diagnostics* (RSD), a complementary evaluation framework that directly measures physics compliance by analyzing the frequency content of PDE residuals.

3.1 PHYSICS RESIDUALS

Consider a time-dependent PDE $\partial_t u + \mathcal{L}[u] = 0$, where $\mathcal{L}$ is a spatial differential operator. Given a neural surrogate $\hat{u}(x,t)$, the *physics residual* quantifies the local violation of the governing equation:

$$r(x,t) = \frac{\partial \hat{u}}{\partial t} + \mathcal{L}[\hat{u}]. \tag{1}$$

For an exact solution, $r(x,t) = 0$ everywhere. Neural surrogates inevitably produce nonzero residuals, and the structure of these residuals—not just their magnitude—reveals how the model fails. The central observation motivating RSD is that solution accuracy and physics compliance are distinct properties. A model can approximate the solution well (low L_2 error) while generating residuals with systematic structure, indicating that it has learned a mapping that happens to match the data without respecting the PDE. Conversely, a model with moderate solution error might satisfy the physics constraints faithfully, suggesting better generalization potential.

3.2 SPECTRAL ANALYSIS OF RESIDUALS

To characterize the residual structure, we analyze its frequency content via the Fourier transform. The *residual power spectrum* is defined as

$$P(\omega) = \mathbb{E}_t \left[|\hat{r}(\omega, t)|^2 \right], \tag{2}$$

where $\hat{r}(\omega,t) = \mathcal{F}[r(\cdot,t)](\omega)$ is the spatial Fourier transform of the residual at time t, and the expectation averages over the temporal dimension. This spectrum reveals which spatial scales contribute most to physics violations. We partition the frequency domain into three bands: low frequencies ($1 \le |\omega| \le \omega_1$), mid frequencies ($\omega_1 < |\omega| \le \omega_2$), and high frequencies ($\omega_2 < |\omega| \le \omega_{\text{Nyq}}$), where $\omega_{\text{Nyq}} = N/2$ is the Nyquist frequency for a grid with N points. The cutoffs $\omega_1 = N/16$ and $\omega_2 = N/6$ provide reasonable defaults, though problem-specific tuning may be beneficial.

3.3 DIAGNOSTIC INDICES

From the power spectrum, we derive two scalar indices that summarize physics compliance.

Algorithm 1 Residual-Spectrum Diagnostics (RSD)

Require: Trained model f_θ, test initial conditions $\{u_0^{(i)}\}_{i=1}^M$, PDE operator $\mathcal{N}$
Ensure: HFV and LFV indices
1: **for** $i = 1$ to M **do**
2: Generate predicted trajectory: $\hat{u}^{(i)} \leftarrow \textsc{Rollout}(f_\theta, u_0^{(i)})$
3: Compute physics residual: $r^{(i)}(x,t) \leftarrow \mathcal{N}[\hat{u}^{(i)}]$
4: Compute power spectrum: $P^{(i)}(\omega) \leftarrow \mathbb{E}_t[|\mathcal{F}[r^{(i)}](\omega)|^2]$
5: **end for**
6: Aggregate spectra: $\bar{P}(\omega) \leftarrow \frac{1}{M}\sum_{i=1}^M P^{(i)}(\omega)$
7: Compute band energies: HFV $\leftarrow \sum_{\omega>\omega_2} \bar{P}(\omega) \,/\, \sum_{\omega\geq 1} \bar{P}(\omega)$
8: **return** HFV, LFV

The **High-Frequency Violation (HFV)** index measures the fraction of residual energy in high-frequency modes:

$$\text{HFV} = \frac{\sum_{\omega>\omega_2} P(\omega)}{\sum_{\omega\geq 1} P(\omega)}. \tag{3}$$

Elevated HFV indicates that the model generates spurious small-scale oscillations or fails to respect the dissipative nature of the PDE at fine scales. This is a common failure mode of neural surrogates trained purely on data, which may learn high-frequency artifacts present in the training set.

The **Low-Frequency Violation (LFV)** index captures residual energy at large spatial scales:

$$\text{LFV} = \frac{\sum_{1\leq\omega\leq\omega_1} P(\omega)}{\sum_{\omega\geq 1} P(\omega)}. \tag{4}$$

High LFV suggests that the model fails to capture bulk dynamics, such as conservation laws or large-scale transport. Together, HFV and LFV provide a coarse but informative decomposition of model failures across spatial scales.

3.4 Computational Procedure

Algorithm 1 summarizes the RSD evaluation pipeline. Given a trained model and a set of test initial conditions, we first generate predicted trajectories via autoregressive rollout (line 2). For each trajectory, we compute the physics residual by evaluating the PDE operator on the predicted solution (line 3), using central differences for time derivatives and standard finite difference stencils for spatial derivatives. We then compute the power spectrum of the residual via FFT (line 4), averaging over the temporal dimension. After processing all test cases, we aggregate the spectra (line 5) and compute the band energy fractions to obtain the final HFV and LFV indices (lines 6–7).

Computational Cost. Residual computation requires one forward pass through the model plus numerical differentiation, and the FFT scales as $O(N \log N)$ per snapshot. In practice, RSD evaluation adds negligible overhead compared to standard model inference. RSD assumes that the governing PDE is known and that the solution admits a meaningful Fourier representation. For problems on non-periodic domains, windowed Fourier transforms or alternative spectral decompositions (e.g., wavelets) provide natural extensions. The core principle, decomposing residuals by spatial scale—remains applicable across diverse settings.

3.5 Relationship to Existing Approaches

RSD complements existing evaluation paradigms rather than replacing them. Solution error metrics (L_2, L_∞) quantify predictive accuracy but cannot distinguish a model that learns physics from one that memorizes data. Physics-informed training losses penalize residual magnitude but provide only a scalar summary that obscures scale-dependent violations. Conservation diagnostics track specific invariants but miss violations orthogonal to conserved quantities. RSD fills this gap by revealing *where* in frequency space physics violations occur, enabling targeted diagnosis and model improvement.

4 NUMERICAL RESULTS

We validate the RSD methodology through controlled experiments on the one-dimensional viscous Burgers equation. Our experiments demonstrate that the High-Frequency Violation (HFV) index captures physics violations that standard error metrics fail to detect.

4.1 EXPERIMENTAL SETUP

Governing Equation. We consider the one-dimensional viscous Burgers equation:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, \quad x \in [0,1], \quad t \in [0,T], \tag{5}$$

with periodic boundary conditions and viscosity $\nu = 0.05$. Ground truth solutions are computed using a pseudo-spectral method with fourth-order Runge-Kutta time integration (relative tolerance 10^{-6}, absolute tolerance 10^{-8}).

Data Generation. Initial conditions are generated as random Fourier series:

$$u_0(x) = \sum_{m=1}^{3} \frac{1}{m}\left(a_m \cos(2\pi m x) + b_m \sin(2\pi m x)\right), \tag{6}$$

where $a_m, b_m \sim \mathcal{N}(0, 0.16)$. Each trajectory spans $T = 0.3$ time units with 15 uniformly spaced snapshots. We generate 8 training trajectories and 6 test trajectories per experimental seed.

Model Architecture. To isolate the effect of training data quality on physics compliance, we employ a simple linear time-stepping model:

$$\hat{u}^{n+1} = Au^n + b, \tag{7}$$

where $A \in \mathbb{R}^{N\times N}$ and $b \in \mathbb{R}^N$ are learned parameters, and $N = 32$ is the spatial grid resolution. Models are trained via gradient descent with learning rate 0.1 for 150 iterations, minimizing mean squared error on single-step predictions.

Experimental Design. We compare two training conditions:

- **Clean-trained**: Model trained on ground truth solutions $\{(u^n, u^{n+1})\}$.
- **Noisy-trained**: Model trained on targets corrupted with high-frequency noise:

$$\tilde{u}^{n+1} = u^{n+1} + \sum_{k=8}^{10} \frac{\eta_k}{\sqrt{k}} \sin(2\pi k x + \phi_k), \tag{8}$$

  where $\eta_k \sim \mathcal{N}(0, 0.04^2)$ and $\phi_k \sim \text{Uniform}(0, 2\pi)$.

This design creates models with similar data-fitting capacity but different physics compliance characteristics. All experiments are repeated across 10 random seeds for statistical reliability.

4.2 EVALUATION METRICS

Solution Error. We measure prediction accuracy using the relative L_2 error over full rollout trajectories:

$$\mathcal{E}_{L_2} = \frac{\|\hat{u} - u\|_2}{\|u\|_2}, \tag{9}$$

where $\hat{u}$ denotes the model prediction and u the ground truth, both evaluated over all spatial points and time steps.

High-Frequency Violation (HFV) Index. We compute the physics residual at each time step:

$$r(x,t) = \frac{\partial \hat{u}}{\partial t} + \hat{u}\frac{\partial \hat{u}}{\partial x} - \nu\frac{\partial^2 \hat{u}}{\partial x^2}, \tag{10}$$

using central differences for temporal derivatives and second-order finite differences for spatial derivatives. The HFV index quantifies the fraction of residual energy in high-frequency modes:

$$\text{HFV} = \frac{\sum_{\omega > \omega_{\text{mid}}} |\hat{r}(\omega)|^2}{\sum_{\omega \geq 1} |\hat{r}(\omega)|^2}, \tag{11}$$

where $\hat{r}(\omega)$ is the Fourier transform of the residual and $\omega_{\text{mid}} = N/6 \approx 5$ defines the boundary between mid and high-frequency bands.

4.3 Results

Main Findings. Table 1 presents the primary experimental results aggregated across 10 random seeds. Both models achieve comparable solution accuracy, with the noisy-trained model showing only 6.2% higher L_2 error. However, the HFV index reveals a striking difference: the noisy-trained model exhibits 37.1% higher high-frequency violations despite the modest difference in solution error.

Table 1: Comparison of clean-trained and noisy-trained models. Values are mean $\pm$ standard error across 10 seeds. Statistical significance assessed via paired t-test.

Metric	**Clean-trained**	**Noisy-trained**	**Δ(%)**	**p-value**
Relative L_2 error	0.478 ± 0.039	0.507 ± 0.038	$+6.2$	0.028
HFV index	0.405 ± 0.030	0.555 ± 0.028	$+37.1$	1.0×10^{-4}

The statistical analysis confirms that the HFV difference is highly significant ($p = 1.0 \times 10^{-4}$), while the L_2 error difference, though statistically detectable ($p = 0.028$), is much smaller in magnitude. This demonstrates that **HFV captures physics violations that L_2 error substantially underestimates**.

Spectral Analysis. Figure 1(a) shows the power spectrum of the physics residual averaged across all test cases. Both models exhibit similar residual power at low frequencies ($\omega < 5$), indicating comparable accuracy in capturing large-scale dynamics. However, the noisy-trained model shows consistently elevated power in the high-frequency band ($\omega > 5$), confirming that it generates spurious small-scale features that violate the governing physics.

Decoupling of L_2 Error and HFV. Figure 2 visualizes the relationship between solution error and physics compliance across individual experimental runs. The results reveal a critical insight: **models with nearly identical L_2 error can exhibit substantially different HFV values**. The clean-trained models (blue circles) cluster at lower HFV values, while noisy-trained models (orange squares) occupy a distinct region with higher HFV. The mean shift (indicated by diamond markers) shows that transitioning from clean to noisy training increases HFV by 37% while only increasing L_2 error by 6%.

Distribution Analysis. Figure 3 presents the full distribution of HFV values across seeds via violin plots. The clean-trained model exhibits a wider distribution with values ranging from 0.29 to 0.61, while the noisy-trained model shows a tighter distribution shifted upward (range: 0.38 to 0.73). Notably, there is minimal overlap between the distributions, indicating that HFV reliably distinguishes between the two training conditions across different random initializations.

4.4 Discussion

Our experiments validate the core hypothesis of the RSD methodology: **spectral analysis of physics residuals reveals failure modes that standard error metrics miss**. Several key observations emerge:

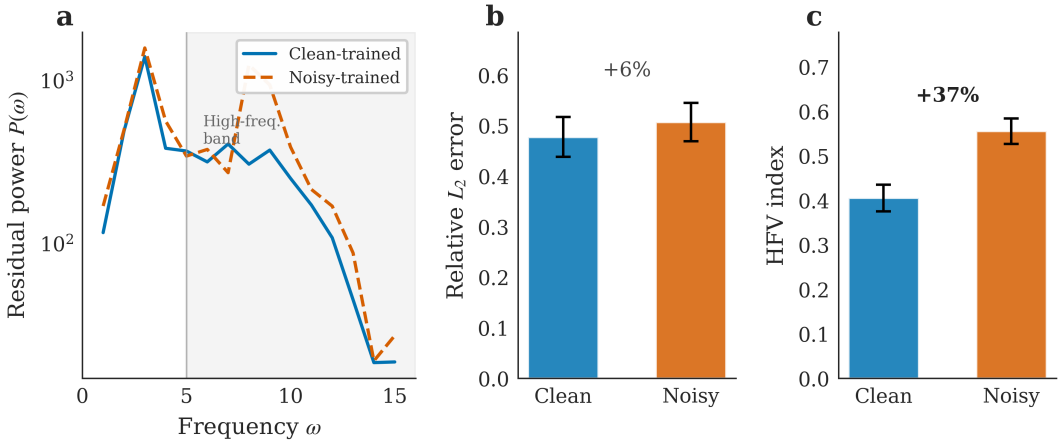

Figure 1: Experimental results comparing clean-trained and noisy-trained models. **(a)** Residual power spectrum showing elevated high-frequency content for the noisy-trained model. The shaded region indicates the high-frequency band ($\omega > \omega_{\text{mid}}$). **(b)** Relative L_2 error comparison showing modest 6% difference. **(c)** HFV index comparison revealing 37% higher physics violations for the noisy-trained model.

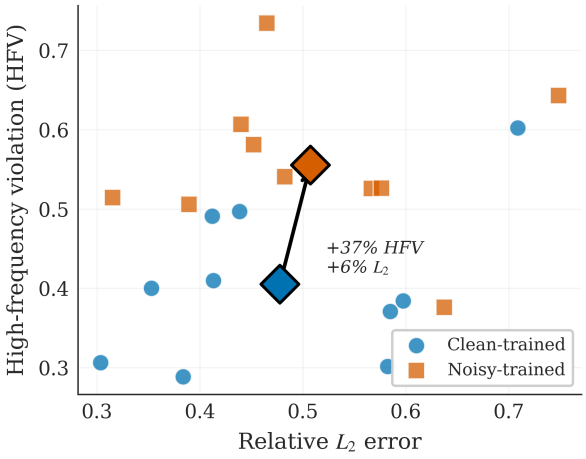

Figure 2: Scatter plot of relative L_2 error versus HFV index for all experimental runs. Each point represents one seed; diamond markers indicate group means. The arrow highlights the disproportionate increase in HFV (+37%) relative to L_2 error (+6%) when training on noisy data.

1. L_2 **error is insufficient for physics compliance assessment.** Two models with 6% difference in L_2 error showed 37% difference in HFV, demonstrating that solution accuracy does not guarantee physics compliance.
2. **HFV detects learned artifacts.** The noisy-trained model learned to reproduce high-frequency artifacts present in its training data. These artifacts manifest as elevated residual power at frequencies where the true Burgers dynamics should be damped by viscosity.
3. **RSD provides actionable diagnostics.** High HFV values serve as an early warning for potential downstream failures, including poor generalization to new initial conditions and instability in long-horizon rollouts.

Limitations. Our validation uses simplified linear models rather than state-of-the-art neural operators (e.g., FNO, PINO). While this design isolates the effect of training data quality, future work should validate RSD on production-scale architectures. Additionally, our experiments focus on a

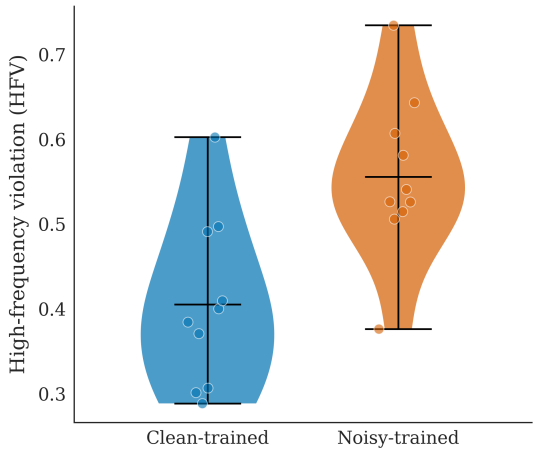

Figure 3: Violin plots showing the distribution of HFV values across 10 random seeds. Individual data points are overlaid. The noisy-trained model consistently exhibits higher HFV with minimal distributional overlap.

single PDE; broader validation across diverse physical systems would strengthen the methodology's generality.

Implications for Neural PDE Solvers. These results suggest that practitioners should supplement traditional error metrics with spectral residual analysis when evaluating neural PDE solvers. A model achieving low L_2 error may still generate non-physical solutions that violate the governing equations at specific frequency scales. The HFV index provides a complementary diagnostic that directly measures physics compliance.

5 Conclusions

We argued that evaluating neural PDE solvers requires measuring not only predictive accuracy but also physics compliance. We introduced *Residual-Spectrum Diagnostics* (RSD), which analyzes the frequency content of PDE residuals and summarizes violations with HFV and LFV. Experiments on viscous Burgers show that models with similar L_2 error can differ strongly in high-frequency physics violations, indicating failure modes that standard metrics underreport. RSD is a lightweight complement to error-based evaluation and can guide model selection and debugging when the governing PDE is known. Future work will validate the approach on modern neural operators, additional PDEs, and non-periodic domains.

REFERENCES

Ali Baheri and Marzieh Amiri Shahbazi. Conformal prediction across scales: Finite-sample coverage with hierarchical efficiency. *Results in Applied Mathematics*, 26:100589, 2025.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Claudio Canuto, M Yousuff Hussaini, Alfio Quarteroni, and Thomas A Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer Science & Business Media, 2007.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *arXiv preprint arXiv:2209.15616*, 2022.

Zhongkai Hao, Chengyang Ying, Zhengyi Wang, Hang Su, Yinpeng Dong, Songming Liu, Ze Cheng, Jun Zhu, and Jian Song. GNOT: A general neural operator transformer for operator learning. *arXiv preprint arXiv:2302.14376*, 2023.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2021.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/JMS Journal of Data Science*, 1(3):1–27, 2024.

Lingkai Liu and Wei Cai. HT-Net: Hierarchical transformer based operator learning model for multiscale PDEs. In *NeurIPS 2022 AI for Science Workshop*, 2022.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

David Millard, Lars Lindemann, and Ali Baheri. Split conformal prediction in the function space with neural operators. *arXiv preprint arXiv:2509.04623*, 2025.

Apostolos F Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, 477:111902, 2023.

Nasim Rahaman, Aristide Barber, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Maziar Raissi, Paris Perdikaris, and George E 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.

Makoto Takamoto, Timothy Praditia, Raphael Leber, Holger Ber, Alvaro Sanchez-Gonzalez, et al. PDEBench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks. *arXiv preprint arXiv:2203.07404*, 2022.

Liu Yang, Xuhui Meng, and George Em Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics*, 425:109913, 2021.

A Ablation Studies

We conduct ablation studies to examine the sensitivity of RSD metrics to experimental parameters. All experiments are repeated over 7 random seeds, and we report mean ± standard error. Figure 4 summarizes the key findings.

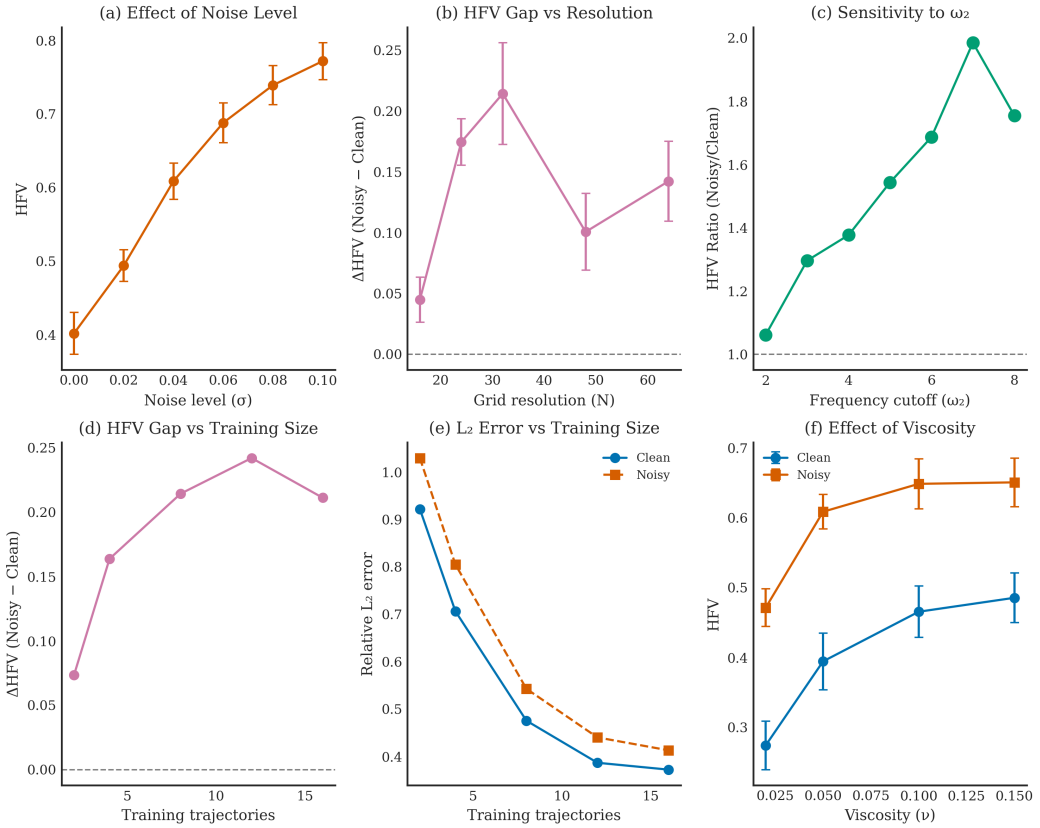

Figure 4: Ablation studies examining RSD sensitivity to (a) training noise level, (b) grid resolution, (c) frequency cutoff ω_2, (d-e) training data size, and (f) viscosity. The HFV gap between noisy-trained and clean-trained models (ΔHFV) remains positive across all experimental conditions, confirming the robustness of RSD diagnostics.

A.1 Effect of Noise Level

We first examine how the magnitude of high-frequency noise in training data affects the HFV index. Table 2 reports HFV and L_2 error for models trained with noise levels ranging from $\sigma = 0$ (clean) to $\sigma = 0.10$.

Table 2: Effect of training noise level on HFV and solution error (mean ± SEM, $n = 7$ seeds).

Noise Level (σ)	HFV	Relative L_2 Error
0.00	0.402 ± 0.028	0.496
0.02	0.494 ± 0.021	0.508
0.04	0.609 ± 0.025	0.543
0.06	0.688 ± 0.027	0.594
0.08	0.739 ± 0.027	0.658
0.10	0.772 ± 0.025	0.731

HFV increases monotonically with noise level, rising 92% from $\sigma = 0$ to $\sigma = 0.10$ (0.402 $\rightarrow$ 0.772), while L_2 error increases 47% over the same range (0.496 $\rightarrow$ 0.731). Critically, HFV is nearly twice as sensitive to training noise as the solution error metric, validating its use as a diagnostic for detecting high-frequency artifacts that corrupt physics compliance.

A.2 Effect of Grid Resolution

We evaluate whether RSD metrics generalize across spatial discretizations by varying the grid resolution from $N = 16$ to $N = 64$. To ensure fair comparison, we scale the noise injection frequencies proportionally to each grid's Nyquist limit (noise injected at modes $N/4$ to $N/2$).

Table 3: Effect of grid resolution on HFV gap (mean $\pm$ SEM, $n = 7$ seeds). Noise level $\sigma = 0.04$.

Resolution (N)	Clean HFV	Noisy HFV	ΔHFV
16	0.721	0.765	0.045 ± 0.019
24	0.255	0.430	0.175 ± 0.019
32	0.394	0.609	0.214 ± 0.042
48	0.602	0.702	0.101 ± 0.032
64	0.710	0.852	0.142 ± 0.033

The HFV gap (ΔHFV) remains strictly positive across all resolutions, confirming that RSD reliably distinguishes physics-compliant from non-compliant models regardless of discretization. The gap is largest at $N = 32$, which provides a balance between frequency resolution and model capacity. At coarse grids ($N = 16$), the limited frequency resolution constrains both clean and noisy models, reducing the gap. At fine grids ($N = 64$), both models have more capacity to fit high-frequency content, but the gap persists.

A.3 Sensitivity to Frequency Cutoff

The HFV index depends on the choice of cutoff frequency ω_2 that defines the high-frequency band boundary. We examine sensitivity to this parameter in Table 4, focusing on the meaningful range where the cutoff lies below the noise injection frequencies.

Table 4: Sensitivity of HFV to the high-frequency cutoff ω_2 (grid size $N = 32$, noise at modes 8–16).

ω_2	Clean HFV	Noisy HFV	Ratio (Noisy/Clean)
2	0.850	0.902	1.06$\times$
3	0.551	0.714	1.30$\times$
4	0.478	0.658	1.38$\times$
5	0.394	0.609	1.54$\times$
6	0.338	0.570	1.69$\times$
7	0.260	0.516	1.99$\times$
8	0.183	0.321	1.75$\times$

The HFV ratio between noisy and clean models exceeds 1.0 for all tested cutoffs, confirming robust discrimination. The ratio peaks at $\omega_2 = 7$ (ratio = 1.99$\times$), just below the noise injection band starting at $\omega = 8$. This validates our default choice of $\omega_2 = N/6 \approx 5$, which provides strong discrimination (1.54$\times$) while remaining robust across problem settings. Very low cutoffs ($\omega_2 = 2$) include too much mid-frequency content where both models behave similarly, reducing sensitivity.

A.4 Effect of Training Data Size

We investigate whether the HFV gap between clean and noisy models persists across different training set sizes (Table 5).

Table 5: Effect of training data size on HFV gap and solution error.

N_{train}	Clean L_2	Noisy L_2	ΔHFV
2	0.922	1.030	0.073
4	0.707	0.805	0.164
8	0.475	0.543	0.214
12	0.387	0.440	0.242
16	0.372	0.413	0.211

While L_2 error decreases substantially with more training data (from 0.92 to 0.37 for clean models), the HFV gap remains consistently positive across all training set sizes (ΔHFV = 0.07–0.24). Interestingly, the gap increases with training data up to $N_{\text{train}} = 12$, suggesting that better-trained models more faithfully reproduce the artifacts in their training targets. This demonstrates that the HFV difference reflects genuine differences in physics compliance rather than artifacts of limited data.

A.5 Effect of Viscosity

Finally, we examine how the physical viscosity parameter ν affects HFV (Table 6).

Table 6: Effect of viscosity on HFV (mean $\pm$ SEM, $n = 7$ seeds).

ν	Clean HFV	Noisy HFV	ΔHFV
0.02	0.274 ± 0.038	0.471 ± 0.045	0.197
0.05	0.394 ± 0.046	0.609 ± 0.066	0.214
0.10	0.466 ± 0.045	0.649 ± 0.057	0.183
0.15	0.485 ± 0.044	0.651 ± 0.050	0.165

The HFV gap persists across viscosities ranging from $\nu = 0.02$ to $\nu = 0.15$, with ΔHFV between 0.165 and 0.214. Absolute HFV values increase with viscosity for both models because higher viscosity damps physical high-frequency content, making the HF band relatively more dominated by numerical artifacts and noise. The consistent gap across viscosities confirms that RSD generalizes across different physical regimes of the Burgers equation.

A.6 Summary

The ablation studies support the following conclusions:

1. **Noise sensitivity:** HFV increases 92% across the tested noise range while L_2 error increases only 47%, confirming that HFV is more sensitive to high-frequency artifacts.
2. **Resolution robustness:** The HFV gap (ΔHFV > 0) persists across all tested grid resolutions from $N = 16$ to $N = 64$.
3. **Cutoff selection:** The default cutoff $\omega_2 = N/6$ provides robust discrimination (ratio $> 1.5\times$), with peak sensitivity near $\omega_2 = 7$ for $N = 32$.
4. **Data efficiency:** The HFV gap is detectable even with only 2 training trajectories and remains consistent as training data increases.
5. **Physical generalization:** Results hold across viscosities spanning an order of magnitude ($\nu = 0.02$ to 0.15).

These results confirm that RSD metrics are robust diagnostic tools that reliably detect physics violations across a wide range of experimental conditions.