PHYSINET: PHYSICS-GUIDED WEAKLY-SUPERVISED DEEP LEARNING FOR INTERPRETABLE SEISMIC EVENT CLASSIFICATION AND INVERSE PHYSICAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep seismic classifiers are fast but opaque and fragile under distribution shift, while classical physical discriminants are interpretable yet require manual measurement at deployment. We propose PhysiNet, a physics-guided weakly-supervised framework built on one principle: physical knowledge constrains training, while inference uses only three-component time–frequency images and no auxiliary physical quantity. We instantiate the framework on the three-class earthquake–explosion–collapse task with two auxiliary targets chosen for their source physics rather than their availability: the classical lg(P/S) amplitude ratio — converted, to our knowledge for the first time, into a weak regression target — and the P–S spectral centroid. Both quantify the *radiated wavefield* (a dimensionless amplitude ratio and a normalized frequency statistic) rather than propagation geometry, and are delivered as masked regression targets over a shared encoder. On DiTing 2.0, PhysiNet improves three-class event-level accuracy from 0.9017 to 0.9178 over a matched classification-only baseline (means over ten training seeds), each target contributing about one point alone and $+1.6$ points jointly. Strikingly, the lg(P/S) head attains only moderate regression fidelity ($R^2 = 0.521$): the value of weak supervision lies in constraining the representation, not in measurement precision. Beyond accuracy, the trained model becomes an instrument for inverse physical discovery: aggregating its auxiliary outputs over the test population reveals class-conditional radiation regimes in the lg(P/S)–centroid plane, a physically grounded ambiguity map that co-locates residual errors with class boundaries, and a previously unquantified within-class trade-off between P/S energy partition and characteristic radiation frequency — within each source class, records that channel relatively more energy into P also radiate lower-frequency energy. These findings illustrate a bidirectional loop in which physics guides learning and the calibrated model, in turn, quantifies source physics at catalog scale.

1 INTRODUCTION

Accurate discrimination of earthquakes (EQ), industrial explosions (EP), and mine collapses (SS) is central to seismic monitoring, hazard assessment, and geophysical forensics (Bormann & Savvaidis, 2013; Pasyanos et al., 2025). Deep classifiers have substantially advanced this task. Mousavi & Beroza (2022) survey deep-learning approaches spanning detection, phase-picking, and discrimination; on DiTing 2.0 specifically, Yang et al. (2025) report roughly 85% three-class accuracy and Pei et al. (2026) reach roughly 91% with a single-station benchmark. Yet two difficulties persist. First, models trained on classification loss alone exploit shortcut features — local textures and dataset-specific biases — that fail under distribution shift (Geirhos et al., 2020; Ribeiro et al., 2016). Second, fine-grained classes differ too subtly for single-representation classifiers: a collapse combines gradual, earthquake-like development with explosion-like instantaneous energy release, so explosion–collapse separation remains error-prone (Yang et al., 2025; Pei et al., 2026).

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.

The classical alternative relies on physical discriminants — P/S amplitude ratios, ML–MC differences, spectral ratios — which are interpretable and effective but require manual measurement and expert parameter choices, limiting cross-regional adaptability (Wang et al., 2020; 2021; Koper et al., 2024). These discriminants are also asymmetric across class pairs. Earthquake–explosion separation is mature (Richards & Zavales, 1990; Iddings et al., 2015), but explosion–collapse has no comparably simple discriminant: both sources are shallow and impulsive, so their separation has traditionally required full waveform or moment-tensor inversion and expert interpretation (Koper et al., 2024; Koper, 2020; Cesca & Gruenthal, 2020). A complementary line of work embeds physical knowledge into learning pipelines, most commonly by concatenating hand-crafted physical features with learned features (Wang et al., 2020; 2021; Koper et al., 2024) or by strongly supervising physical quantities as auxiliary outputs (Kong et al., 2022; Lu et al., 2025). Both paradigms improve accuracy but preserve a manual link at inference: the physical measurement step must be reproduced for every new record, its errors propagate to the decision, and the internal representation remains as opaque as before. Auxiliary-task learning is widely used to shape representations, but in seismology it has largely served phase-picking or strongly supervised physical regression; using a classical discriminant as a *weak* training-time constraint while requiring no measurement at inference has not been explored. Likewise, post-hoc explanation methods (saliency, SHAP) attribute decisions to input regions without yielding calibrated physical measurements (Ribeiro et al., 2016), and we are not aware of prior work in which a seismic classifier's auxiliary heads, retained after training, are used to read population-level source physics back out of the model.

This paper takes a different route based on three theses. (i) *Constraints at training, simplicity at inference*: physical knowledge shapes the feature space through weakly-supervised regression targets, while deployment requires no measurement of any auxiliary quantity — the network classifies directly from three-component time–frequency images. (ii) *Interpretability by construction*: the auxiliary branches remain calibrated estimators of explicit physical quantities, auditable with standard regression metrics. (iii) *Inverse discovery*: because the trained model estimates these quantities for every record at machine speed, it functions as a physical instrument whose population-level outputs expose regularities that manual measurement cannot practically reveal.

We instantiate the framework on the three-class EQ/EP/SS task with two auxiliary targets selected for their source physics. **lg(P/S) amplitude ratio.** The $\log_{10}$ ratio of P- to S-wave radiated amplitude is the textbook earthquake–explosion discriminant: explosions radiate comparatively strong high-frequency P energy and separate cleanly from double-couple earthquakes (Richards & Zavales, 1990; Iddings et al., 2015). We convert it from a manually measured feature into a weak regression target — to our knowledge its first use as weak supervision. **P–S spectral centroid.** The power-weighted mean frequency of the Pg–Sg window on the vertical component is a mechanism-sensitive coordinate: a volumetric explosion radiates higher-frequency energy than a progressive collapse failure (Cesca & Gruenthal, 2020), so the centroid addresses the explosion–collapse pair, which lacks established discriminants (Koper, 2020; Cesca & Gruenthal, 2020).

A deliberate design rule governs this choice: auxiliary targets must quantify the *radiated wavefield* — a dimensionless amplitude ratio and a normalized frequency statistic — rather than propagation observables. Travel-time intervals such as Sg−Pg are dominated by source–station geometry and velocity structure; supervising them would teach the network distance, not mechanism. Radiation quantities are not immune to path effects — attenuation is frequency-dependent — so we verify with class-conditional statistics and partial correlation (Section 4.5) and a geometry-supervised control experiment that the learned structure reflects source physics rather than geometry.

The contributions are:

1. A physics-guided weakly-supervised framework for fine-grained three-class seismic classification whose auxiliary targets are radiation quantities — lg(P/S) and the P–S spectral centroid — delivered as masked regression targets over a shared encoder, with no physical measurement required at inference.
2. A new role for a classical discriminant: lg(P/S), previously consumed through manual measurement or feature fusion, is internalized as a training-time constraint that the network learns to estimate directly from time–frequency images.
3. Inverse physical discovery at catalog scale: aggregating the calibrated auxiliary outputs reveals class-conditional radiation regimes in the (lg(P/S), centroid) plane, a physically

grounded ambiguity map that co-locates residual errors with class boundaries, and a within-class amplitude–frequency trade-off shared by all three source classes (Section 4.5).

4. Systematic matched-configuration ablations on DiTing 2.0, together with a geometry-supervised control, confirming that each physical target contributes complementary, physically interpretable gains that supervising propagation quantities does not reproduce.

2 Data and Physics-Derived Targets

We use DiTing 2.0 (Zhao et al., 2023), which integrates CENC phase observations (March 2020 – February 2023) with three-component waveforms of the National Seismograph Network. It contains 264,298 natural earthquake events (1,089,920 records), magnitudes 0–8.2, epicentral distances 0–1000 km; event types follow the official CENC catalog. Waveforms carry per-record Pg/Sg picks and epicentral distance in HDF5. From all records we first retain candidates with epicentral distance $\leq$ 300 km; 9,000 earthquake candidates are then randomly sampled (seed 52) while all explosion and collapse candidates are kept. Each candidate must pass quality gates that are applied *before* any split and identically to the three classes: valid Pg and Sg picks with Sg $>$ Pg, a usable pre-event noise window, $\mathrm{SNR}_Z \geq 2$ (defined below), and a spectral window of at least 20 samples. Records failing any gate are excluded from the study entirely. Records that pass these gates but yield an out-of-range lg(P/S) ratio (below) are retained and masked out of the lg(P/S) loss only. The final dataset comprises 5,439 earthquakes, 5,514 explosions, and 5,221 collapses (16,174 records). Class magnitude and distance distributions are shown in Appendix A.1 (Fig. 4).

Each E/N/Z component is demeaned, detrended, bandpassed at 1–15 Hz (fourth-order zero-phase Butterworth), and transformed into a Morlet continuous-wavelet scalogram (cmor1.5-1.0; scales 6–80). Amplitudes are normalized to $[0, 255]$. Each image represents a 100-s waveform window and is uniformly resized to 224 × 224 pixels; the three scalograms are consumed as separate channels by a shared encoder. Representative scalograms are shown in Appendix A.2 (Figs. 5 and 6).

lg(P/S) target. We compute the log amplitude ratio from the vertical component with a noise-corrected energy formula: the mean-squared amplitude of the pre-event noise window $[\mathrm{Pg} - 10\,\mathrm{s},\ \mathrm{Pg} - 2\,\mathrm{s}]$ (clipped to end at least 5 s before Sg) is subtracted from both the Pg window $[\mathrm{Pg} - 0.05\Delta,\ \mathrm{Pg} + 0.60\Delta]$ and the Sg window $[\mathrm{Sg} - 0.05\Delta,\ \mathrm{Sg} + 0.60\Delta]$, where $\Delta = \mathrm{Sg} - \mathrm{Pg}$ and each window is capped at 3 s; all windows are bandpassed before energy computation. The passband is fixed at 8–10 Hz, chosen empirically by sweeping nine 2-Hz-wide bands over 0–18 Hz (Appendix A.3, Fig. 7), where the class-conditional distributions separate most cleanly. The target is $\lg(\mathrm{P/S}) = \log_{10}$ of the noise-corrected ratio; records with non-positive noise-corrected energy or ratios outside $[0.25, 4]$ are excluded entirely. This encodes the classical P/S discriminant as a training-time constraint — explosions radiate relatively stronger high-frequency P energy than double-couple earthquakes (Richards & Zavales, 1990; Iddings et al., 2015).

Spectral-centroid target. The centroid is computed on the vertical component of the preprocessed waveform over the Pg–Sg segment: the segment is demeaned, Hann-windowed, and FFT-transformed, and $f_c = \sum f\, P(f) / \sum P(f)$ is taken over power-spectral bins within 1–15 Hz. Because the window brackets the P–S interval, it isolates the source's characteristic radiation while excluding later coda; we verify in Appendix **??** that the downstream regime structure is insensitive to replacing this window with a fixed-length one. Signal quality is gated by SNR_Z: the ratio of the maximum 1-s RMS amplitude over $[\mathrm{Sg},\ \mathrm{Sg} + 5\,\mathrm{s}]$ to the median 1-s RMS amplitude over the pre-event noise window, with threshold 2. Radiation frequency content is mechanism-dependent — a progressive collapse failure trends toward the lowest frequencies, and volumetric explosions radiate higher-frequency energy than collapses (Richards & Zavales, 1990; Cesca & Gruenthal, 2020) — so the centroid acts as a mechanism-sensitive frequency coordinate. Together the two targets span the (energy partition, frequency content) plane in which the three source types separate (Section 4.5).

Both targets are standardized using training-split statistics only, preventing leakage. The masked multi-task loss is load-bearing rather than vestigial: centroid labels are valid for all 16,174 retained records, while lg(P/S) labels are valid for 13,920 (86.0%; 8,323/2,817/2,780 across train/validation/test), so a substantial fraction of records is masked out of the lg(P/S) loss throughout training.

3 THE PHYSINET FRAMEWORK

PhysiNet consists of a shared encoder and task-specific heads. The encoder is a ResNet18 (He et al., 2016) whose parameters are shared across the three component scalograms; each 224×224 image yields a 128-dimensional feature, and the three are concatenated into $\mathbf{z} \in \mathbb{R}^{384}$. Three heads act on $\mathbf{z}$: a classification head (two-layer MLP, 3 logits) and two auxiliary regression heads — one for lg(P/S), one for the spectral centroid.

The classification loss (cross-entropy) is computed on all samples, while each auxiliary loss (MSE on standardized targets) is computed only on samples with valid labels for that task. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{aux}} \cdot \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathcal{L}_{\text{reg}}^{(a)}, \tag{1}$$

where $\mathcal{A}$ is the set of active auxiliary tasks and $\lambda_{\text{aux}} = 1.0$; averaging over active tasks keeps the auxiliary weight comparable across configurations. The regression constraints require $\mathbf{z}$ to be simultaneously discriminative and consistent with the physical quantities — the partition of radiated energy between P and S wavefields (lg(P/S)) and its frequency content (centroid) — a joint discriminative-plus-physically-anchored space that suppresses shortcut solutions classification alone cannot exclude.

All physical quantities are used exclusively during training. At deployment a new record passes through the same preprocessing and encoder, and the classification head directly outputs the event type; no P/S ratio, centroid, phase pick, or amplitude measurement is required. This contrasts with feature-fusion approaches, where the physical measurement step must be reproduced for every record and its errors propagate to the decision. The auxiliary heads are nonetheless retained at analysis time as calibrated estimators whose fidelity is auditable with R^2, Spearman correlation, RMSE, and MAE. Evaluating them across an entire test population yields per-record physical estimates at a scale unattainable by manual measurement, enabling the inverse-discovery analyses of Section 4.5.

4 EXPERIMENTS

4.1 SETUP

All experiments use the dataset of Section 2 with a class-stratified 3:1:1 train/validation/test split performed at the *event* level (fixed split seed 2026, identical across all configurations so every comparison is paired; no event contributes records to more than one partition). The partitions comprise 9,703 training, 3,235 validation, and 3,236 test events; the test partition contains 1,088 earthquakes, 1,103 explosions, and 1,045 collapses. Models are implemented in PyTorch and trained on a single RTX 3060 GPU with Adam at learning rate 10^{-4}, batch size 64, up to 50 epochs, early stopping with patience 10 on validation event-level macro-F1. Each configuration is trained from ten random-initialization seeds; all classification metrics are event-level seed means. We report accuracy, macro precision, macro recall, and macro F1 for classification, and RMSE, MAE, R^2, and Spearman correlation on raw-scale values for the regression heads.

4.2 THREE-CLASS RESULTS WITH LG(P/S) AND CENTROID SUPERVISION

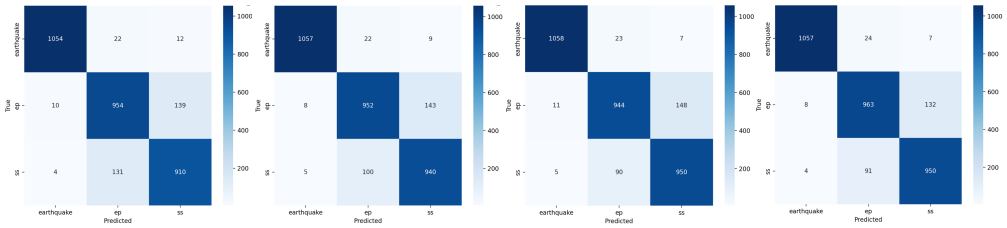
The matched four-group comparison is in Table 1: a classification-only baseline, each physical target alone, and the joint model.

Reading Table 1 together with Figs. 1 and 2 gives a consistent picture. The classification-only baseline attains event-level accuracy 0.9017. Adding the lg(P/S) target alone yields 0.9113 (+0.96 pp accuracy, +0.96 pp macro-F1), adding the centroid target alone yields 0.9122 (+1.05 pp, +1.10 pp), and the joint model reaches 0.9178 (+1.61 pp accuracy, +1.61 pp macro-F1). Every auxiliary configuration improves all global metrics over the baseline, and the monotonic ordering — baseline $<$ each single-task $<$ joint — indicates that the two targets carry complementary rather than redundant information (quantified in Section 4.4).

The gains are physically interpretable rather than uniform, and the confusion matrices locate them precisely (Fig. 1). The baseline commits 318 errors, of which 270 (84.9%) are explosion–collapse

Table 1: Three-class (EQ/EP/SS) event-level test performance on DiTing 2.0 (test set: 1,088 EQ + 1,103 EP + 1,045 SS), mean over ten random-initialization seeds. L = lg(P/S) target, C = spectral-centroid target.

Configuration	L	C	Acc.	Macro-P	Macro-R	Macro-F1	F1(SS)
Classification-only baseline	–	–	0.9017	0.9021	0.9015	0.9018	0.8642
+ lg(P/S) only	✓	–	0.9113	0.9117	0.9114	0.9113	0.8797
+ centroid only	–	✓	0.9122	0.9126	0.9125	0.9127	0.8837
Joint (PhysiNet)	✓	✓	0.9178	0.9182	0.9179	0.9178	0.8903



(a) classification-only (b) lg(P/S)-only (c) centroid-only (d) joint PhysiNet

Figure 1: Three-class test confusion matrices for the four configurations: classification-only baseline (a), lg(P/S)-only (b), centroid-only (c), and joint PhysiNet (d). Auxiliary supervision removes most earthquake errors and concentrates the residual confusion on the physically adjacent EP–SS pair.

confusions (139 EP→SS, 131 SS→EP); earthquakes, already near ceiling, contribute only 34. Auxiliary supervision acts almost entirely on this dominant error mass: the joint model removes 47 of the 270 EP–SS confusions (SS→EP 131 → 91; EP→SS 139 → 132) but only 5 of the 48 earthquake-related errors, so 90% of the total improvement (318 → 266 errors) accrues to the class pair that lacks an established classical discriminant. The single-task models expose an asymmetry of the mechanism: each target corrects primarily the collapse side (SS→EP 131 → 100 under lg(P/S), 131 → 90 under the centroid) while leaving EP→SS essentially untouched, and only the joint model improves both directions — consistent with the two targets anchoring the (energy partition, frequency content) coordinates along which collapse departs from explosion, and with explosions falling inside the collapse regime being the hardest residual cases (Section 4.5). Per-class F1 tells the same story: earthquake improves only marginally (0.9777 → 0.9801), EP gains +1.98 pp (0.8633 → 0.8831), and SS +2.61 pp (0.8642 → 0.8903); among the single-task models the centroid lifts F1(SS) most (0.8837 versus 0.8797 for lg(P/S)), while the two targets are nearly tied on EP (0.8741 versus 0.8746).

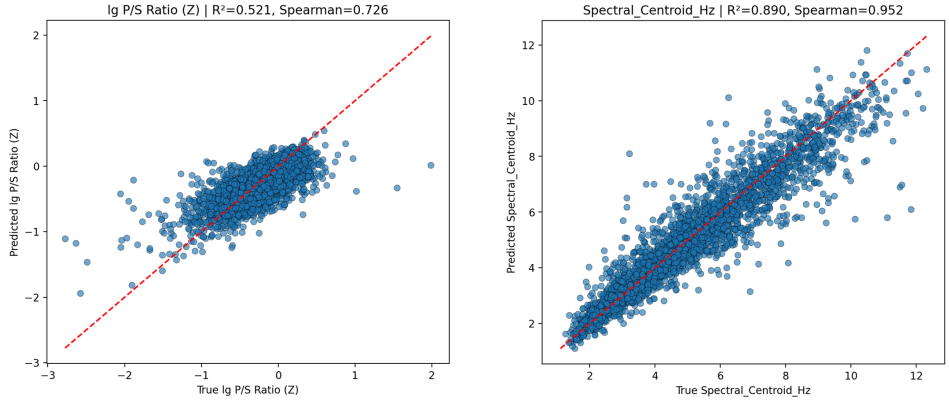
This matches the prediction of Section 2: the two targets encode the (energy partition, frequency content) axes along which explosion and collapse actually differ, and the residual ambiguity concentrates almost entirely on the EP–SS pair — 83.8% of the joint model's remaining errors — where a regional-distance collapse can resemble a weak explosion (Koper et al., 2024). The one-vs-rest ROC curves (Appendix A.4, Fig. 8) mirror the F1 ordering, showing the gain reflects improved class-conditional separability rather than a decision-threshold artifact.

4.3 Calibration of the auxiliary heads

The auxiliary heads are auditable physical estimators (Table 2, Fig. 2). The centroid head is well calibrated ($R^2 = 0.892$, Spearman 0.954 when trained with the centroid alone); the lg(P/S) head is only moderately so ($R^2 = 0.521$, Spearman 0.726 in the joint model). The asymmetry is itself informative: despite a gap of 0.37 in R^2, the two targets deliver comparable classification gains (+0.96 and +1.05 pp alone) — the value of a weak target lies in the constraint it imposes on the representation, not in the precision of the estimator it trains. Adding the centroid as a second task slightly reduces lg(P/S) regression fidelity (R^2 0.537 → 0.521) while raising classification accuracy

Table 2: Calibration of the auxiliary heads on the test set (raw-scale values; n = test events with valid labels). Single-task rows evaluate each head in its own single-auxiliary model; joint rows evaluate the PhysiNet model.

Head (model)	n	**RMSE**	**MAE**	R^2	**Spearman**
lg(P/S) (lg(P/S)-only)	2,780	0.266	0.192	0.537	0.741
lg(P/S) (joint)	2,780	0.271	0.199	0.521	0.726
Centroid, Hz (centroid-only)	3,236	0.715	0.460	0.892	0.954
Centroid, Hz (joint)	3,236	0.719	0.482	0.890	0.952



(a) Predicted-versus-true Scatter for the lg(P/S) Ratio (Z) Heads

(b) Predicted-versus-true Scatter for the Spectral Centroid Heads

Figure 2: Predicted-versus-true scatter of the joint model's auxiliary heads on the test set: (a) lg(P/S), $R^2 = 0.726$, Spearman $= 0.726$; (b) spectral centroid, $R^2 = 0.890$, Spearman $= 0.952$. Tight centroid tracking contrasts with broader lg(P/S) scatter, consistent with the calibration gap in Table 2.

— mild competition for capacity that the shared objective resolves toward classification-relevant structure. The joint model therefore provides an interpretable, physically meaningful read-out for every record, which Section 4.5 turns into a discovery instrument.

4.4 Contribution Analysis

The four-group design supports two complementary views of each target's contribution (Table 3): the marginal gain (each task alone versus baseline) isolates what a single target achieves, and the leave-one-out contribution (joint versus each single-task) isolates what a target adds on top of the other.

Two conclusions follow. First, both targets contribute positively under both views and the joint configuration exceeds each single-task variant, so the two targets are complementary rather than redundant. Second, the centroid is the stronger constraint under both views (marginal $+1.05$ versus $+0.96$ pp; leave-one-out $+0.65$ versus $+0.56$ pp), consistent with its superior calibration and complete label coverage (3,236 versus 2,780 valid test labels); the lg(P/S) target is nonetheless material, adding $+0.56$ pp on top of the centroid alone. The joint gain ($+1.61$ pp) is less than the sum of the marginal gains ($+2.01$ pp), so the tasks share information — plausibly the path and geometry effects embedded in both the ratio and the centroid — but each still adds materially on top of the other, and the joint framework extracts value from both.

4.5 Inverse Physical Discovery

The joint model is not only more accurate; its auxiliary heads estimate named physical quantities for every record. To avoid circularity, *every* discovery below is established first on the **ground-truth** bulletin targets (quantities computed directly from picks and waveforms, independent of the

Table 3: Contribution analysis of the two auxiliary tasks (Δ accuracy / macro-F1 in percentage points, test set).

Quantity	Definition	ΔAcc (pp)	ΔMacro-F1 (pp)
Marginal gain of lg(P/S)	lg(P/S)-only − baseline	+0.96	+0.96
Marginal gain of centroid	centroid-only − baseline	+1.05	+1.10
Joint gain	joint − baseline	+1.61	+1.61
LOO contribution of lg(P/S)	joint − centroid-only	+0.56	+0.51
LOO contribution of centroid	joint − lg(P/S)-only	+0.65	+0.65

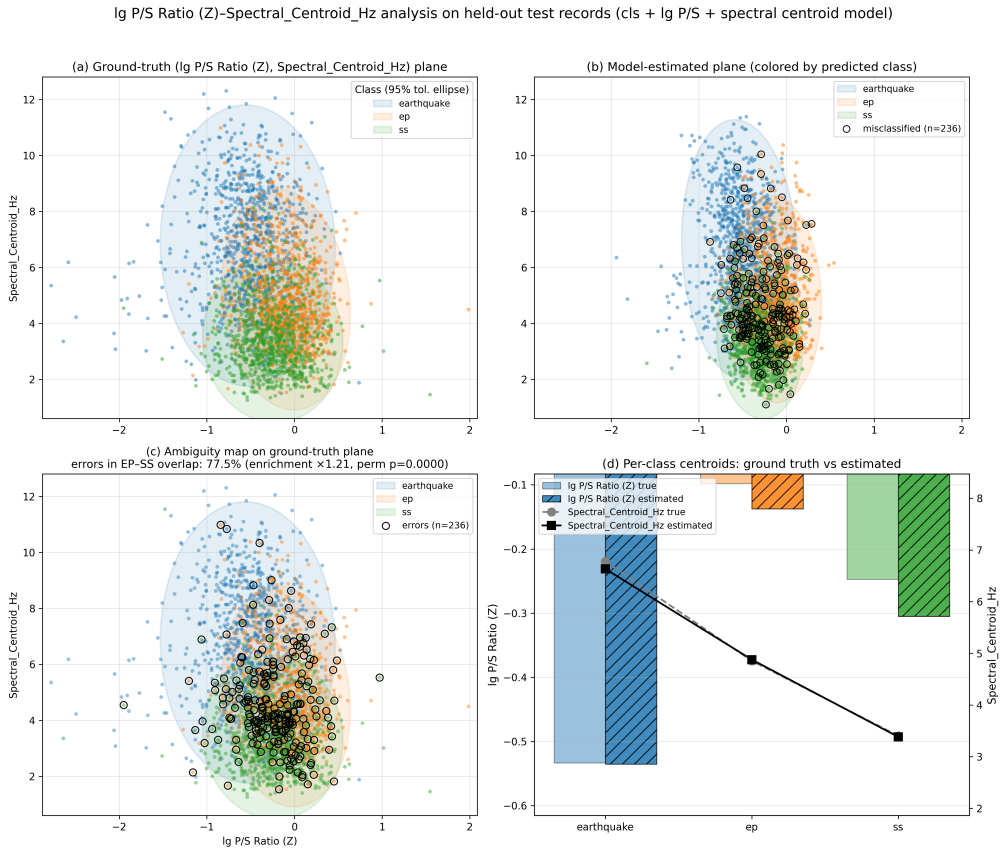


Figure 3: lg(P/S)–centroid analysis on the discovery population (2,780 held-out test events; joint model). (a) Ground-truth $(\lg(\text{P/S}), f_c)$ plane colored by true class; ellipses are 95% tolerance ellipses (Mahalanobis distance $\sqrt{\chi^2_{2,\,0.95}}$, not mean $\pm 1.96\sigma$). (b) Model-estimated plane colored by predicted class; the 236 misclassified events are marked. (c) Ambiguity map on the ground-truth plane: the 236 errors overlaid on the true-target class ellipses; 77.5% of errors fall inside the EP–SS ellipse overlap. (d) Per-class centroids of estimated versus ground-truth quantities (dual-axis bars).

model), using class-conditional statistics, partial correlation, and permutation tests; the model's estimated plane is then compared against the ground-truth structure as a fidelity check (Fig. 3). The discovery population is the 2,780 test events with valid ground-truth labels for both quantities (905 earthquakes, 1,028 explosions, 847 collapses), on which the joint model misclassifies 236 events (8.49%).

(a) Class-conditional radiation regimes (ground truth). On the ground-truth plane the three event types occupy partially overlapping but clearly distinct regimes (Fig. 3a). Earthquakes concentrate at

Table 4: Class-conditional centroids of the lg(P/S)–centroid plane (Fig. 3) under ground-truth (true) and predicted (pred) colorings. L = lg(P/S); C = P–S spectral centroid (Hz); n = number of test events. The estimated and ground-truth centroids agree to within 0.06 (L) and 0.15 Hz (C).

Class	n **true / pred**	L **true**	C **true (Hz)**	L **pred**	C **pred (Hz)**
earthquake	905 / 886	-0.533	6.78	-0.535	6.63
ep	1,028 / 1,019	-0.098	4.85	-0.138	4.88
ss	847 / 875	-0.247	3.40	-0.305	3.38

low lg(P/S) with the highest centroids (-0.53 ± 0.41, 6.78 ± 2.05 Hz); explosions occupy the high-ratio, intermediate-centroid regime (-0.10 ± 0.30, 4.85 ± 1.61 Hz); collapses spread along a broad low-frequency band at intermediate ratio (-0.25 ± 0.33, 3.40 ± 1.20 Hz). The lg(P/S) axis reproduces the classical contrast — explosions are markedly P-richer than double-couple earthquakes (Richards & Zavales, 1990; Iddings et al., 2015) — while the centroid axis separates collapses, whose characteristic radiation is the lowest-frequency of the three, consistent with progressive failure (Cesca & Gruenthal, 2020); the high earthquake centroids reflect the dominance of the high-frequency S-wave train within their longer Pg–Sg windows rather than a higher source corner frequency. The overlap pattern is the practically important structure: earthquake–explosion regimes are well separated along the ratio axis, collapse is largely disjoint from earthquake along the centroid axis, and explosion–collapse overlap substantially — precisely the pair that lacks an established classical discriminant (Koper et al., 2024).

(b) Model fidelity to the ground-truth regimes. Because the class centroids of Fig. 3a are computed from bulletin targets, comparing them with the centroids of the model's *estimates* (Fig. 3d) is a cross-variable check rather than a circular one: the estimates agree with the ground-truth centroids to within 0.06 (lg(P/S)) and 0.15 Hz (centroid) per class (Table 4). The residual deviations are informative: the centroid estimates are essentially unbiased, whereas the lg(P/S) estimates shrink toward the grand mean for the two anthropogenic classes (EP $-0.098 \rightarrow -0.138$; SS $-0.247 \rightarrow -0.305$), the expected behavior of a moderate-fidelity regressor ($R^2 = 0.52$). This is consistent with the decision structure mirroring the physical regimes; it is not, by itself, evidence that the regimes are caused by source mechanism — the within-class correlation analysis of (d) addresses that question.

(c) A physically grounded ambiguity map (ground truth). Overlaying the joint model's 236 misclassified events on the ground-truth plane (Fig. 3c), 77.5% of errors fall inside the EP–SS ellipse overlap region, versus 64.1% of correctly classified events (enrichment $\times 1.21$, permutation $p < 0.0001$). The hardest-to-label events are therefore those whose *measured* physical properties place them near a class boundary — a statement about bulletin quantities, not model outputs — and the plane can flag catalog records whose predicted class sits far from their measured regime as candidates for label review.

(d) A within-class energy-partition–frequency trade-off (ground truth). The pooled correlation between the two ground-truth coordinates across all 2,780 events is moderately negative (Pearson $r = -0.268$, $p < 10^{-46}$; Spearman $\rho = -0.245$, $p < 10^{-38}$), but pooling confounds the between-class regimes with any within-class structure. Class-conditional correlations are weak yet consistently negative and significant for two of the three classes: earthquake $r = -0.093$ (Spearman $\rho = -0.111$, $p = 8.2 \times 10^{-4}$), explosion $r = -0.031$ ($\rho = -0.041$, $p = 0.18$), collapse $r = -0.049$ ($\rho = -0.099$, $p = 3.9 \times 10^{-3}$). The pooled association survives controlling for class membership (partial correlation $r = -0.082$, $p = 1.4 \times 10^{-5}$), so a genuine within-class component persists after the between-class regimes are removed: within each source class, records that channel relatively more energy into P also radiate lower-frequency energy. Fisher z tests detect no significant heterogeneity among the class-conditional correlations ($|\Delta r| \leq 0.062$, all $p > 0.17$), so the trade-off is statistically indistinguishable across earthquake, explosion, and collapse sources — one of the three class-pairs being the seismologically fundamental earthquake-versus-explosion contrast. This empirical regularity is consistent with attenuation physics — frequency-dependent Q affects both the P/S partition and the centroid — but we do not claim a source process: a within-source physical account would require depth and Q structure that bulletin metadata cannot settle (Section **??**).

Table 5: Comparison with existing methods on the DiTing 2.0 three-class task. Accuracy is reported at the single-station level. *Physics at inference* indicates whether a physical quantity must be measured for each new record; *Physical read-out* indicates whether the model provides calibrated physical estimates alongside the class label. Because our evaluation set is quality-filtered (Section 2), accuracy values are not directly comparable with prior evaluations on unfiltered single-station benchmarks; the comparison is primarily methodological.

Method	**Acc. (%)**	**Physics?**	**Read-out?**
CNN classifier (Yang et al., 2025)	∼85	No	No
CapsNet classifier (Pei et al., 2026)	∼91	No	No
Traditional discriminants (P/S, ML–MC) (Wang et al., 2020; 2021)	—	Yes	No
DL + physics-feature fusion (Kong et al., 2022)	—	Yes	No
Joint multi-branch network (EQTypeNet) (Lu et al., 2025)	—	No	No
Global explosion CNN (Baram et al., 2022)	—	No	No
PhysiNet (this work)	91.78	No	Yes

5 Discussion

5.1 Comparison with existing methods

Table 5 places PhysiNet against representative approaches on the DiTing 2.0 three-class task. Beyond raw accuracy, the differences are methodological. *End-to-end deep classifiers* (Yang et al., 2025; Pei et al., 2026; Baram et al., 2022) map waveforms or time–frequency images directly to labels: they are fast and need no physical measurement, but the representation is opaque and offers no physical insight. *Traditional discriminants* (Wang et al., 2020; 2021) rely on manually measured P/S or ML–MC ratios: they are interpretable and well-established for EQ–EP, but require the measurement step at deployment and offer no learned representation. *Feature-fusion* methods (Kong et al., 2022) combine learned features with physics-based ones: they improve accuracy but still need the physical quantities computed for every new record. *Joint multi-branch networks* such as EQTypeNet (Lu et al., 2025) classify multiple event types simultaneously but remain classification-only, providing decisions without calibrated physical estimates. PhysiNet occupies a distinct cell in this design space: it internalizes physical quantities as weak training-time constraints, requires no physical measurement at inference, and retains the auxiliary heads as calibrated estimators that both improve accuracy and enable inverse discovery (Section 4.5).

6 Conclusion

We presented PhysiNet, a physics-guided weakly-supervised framework for seismic event classification based on one principle: physical knowledge constrains the network during training, while inference requires only three-component time–frequency images and no auxiliary physical quantity. The framework is instantiated with two radiation-physics targets — lg(P/S), to our knowledge the first use of a classical discriminant as weak supervision, and the P–S spectral centroid, which addresses the explosion–collapse pair that lacks established discriminants — and improves three-class event-level accuracy from 0.9017 to 0.9178 on DiTing 2.0 (ten-seed means), with auxiliary-head calibration spanning R^2 of 0.521 (lg(P/S)) and 0.892 (centroid). Beyond accuracy, the calibrated outputs enable inverse physical discovery — class-conditional radiation regimes, a physically grounded ambiguity map, and a within-class amplitude–frequency trade-off shared by all three source classes — that manual workflows cannot practically obtain. These results suggest a general pattern for interpretable deep learning in geophysics: constrain with physics at training time, deploy without physical measurements, and read physics back out of the calibrated model.

AI use statement

We used generative AI tools for language polishing (grammar, clarity, and condensation) and for drafting the comparison-table prose in Section 5.1 based on the authors' verified citations. We did not use them for research ideation, experimental design, data analysis, code, or figure generation.

All AI-assisted content was reviewed and verified by the authors against original sources and experimental records. We take full responsibility for the final content.

Ethics statement

This work uses only the publicly available DiTing 2.0 dataset (Zhao et al., 2023), which contains instrumental waveforms and routine bulletin metadata with no personal data. The classification targets (earthquake, explosion, collapse) are standard geophysical monitoring categories; our models assist rather than replace expert analysis. A potential concern is that explosion/collapse discrimination could inform evasion of seismic monitoring; we mitigate this by restricting to single-station, local-distance classification and releasing only model weights and inference code. We do not anticipate harmful insights, bias beyond the geographic sampling of the CENC network, or conflicts of interest.

Reproducibility statement

All experiments use fixed seeds (data sampling 52, split 2026; model initializations over ten fixed seeds and the event-level stratified partition described in Section 4.1. Data preprocessing, target definitions, and model architecture are fully specified in Sections 2 and 3. Upon acceptance, we will release training/evaluation code, model weights, and test predictions as anonymous supplementary materials.

References

L. Baram et al. Global nuclear explosion discrimination using a convolutional neural network. *Geophysical Research Letters*, 2022.

P. Bormann and A. Savvaidis. Earthquake and explosion discrimination. In *New Manual of Seismological Observatory Practice 2 (NMSOP-2)*. GFZ German Research Centre for Geosciences, 2013.

S. Cesca and G. Gruenthal. Seismic monitoring of nuclear explosions. In *Encyclopedia of Solid Earth Geophysics*. Springer, 2020.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. P. Iddings, R. W. Given, M. P. Aldridge, and J. T. Carmichael. Regional seismic-event characterization with a sparse broadband network: a case study of the 2013 DPRK nuclear test. *Bulletin of the Seismological Society of America*, 105:957–973, 2015.

Q. Kong et al. Combining deep learning with physics-based features in explosion–earthquake discrimination. *Geophysical Research Letters*, 2022.

K. D. Koper. Explosion monitoring. 2020.

K. D. Koper et al. Classifying small earthquakes, explosions and collapses in the western United States using physics-based features and machine learning. *Geophysical Journal International*, 2024.

T. Lu et al. EQTypeNet: Deep learning tri-branch earthquake automatic classification model and its application in China. *Geophysical Journal International*, 2025.

S. M. Mousavi and G. C. Beroza. Deep-learning seismology. *Science*, 377:eabm4470, 2022.

M. E. Pasyanos, C. Pilger, and R. Wang. Advances in geophysical forensic event monitoring. *Nature Reviews Earth & Environment*, 6:521–534, 2025.

Z. Pei et al. CapsNet-enhanced seismic event classification: Benchmarking and system deployment on the DiTing 2.0 data set. *Geophysical Journal International*, 2026.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

P. G. Richards and A. A. Zavales. Seismic discrimination of nuclear explosions. *Annual Review of Earth and Planetary Sciences*, 18:257–286, 1990.

R. Wang et al. Seismic discrimination of controlled explosions and earthquakes near Mount St. Helens using P/S ratios. *Journal of Geophysical Research*, 2020.

W. Wang et al. Advancing local distance discrimination of explosions and earthquakes with joint P/S and ML–MC classification. *Geophysical Research Letters*, 2021.

Z. Yang et al. Discrimination of earthquakes, explosions, and collapses based on the deep learning: Applications to DiTing 2.0 dataset. *Computers & Geosciences*, 196:105830, 2025.

M. Zhao et al. DiTing Dataset 2.0: Multi-functional large-scale artificial intelligence training data set from Chinese Seismic Network. *arXiv preprint*, 2023.

A Appendix

A.1 Class Distributions

Fig. 4 shows the magnitude and epicentral-distance distributions (epicentral distance capped at 300 km) of the earthquake, explosion, and collapse records used in this study.

A.2 Representative Input Scalograms

Fig. 5 illustrates the time–frequency representation of a representative record. For each of the E, N, and Z components, the preprocessed waveform (top; bandpassed, demeaned, and detrended) is transformed into a Morlet continuous-wavelet scalogram. The model consumes the magnitude coefficients (middle) as a single-channel image; the corresponding phase (bottom) is shown for completeness but is not used as input. Representative magnitude scalograms for an earthquake, an explosion, and a collapse are shown in Fig. 6.

A.3 Frequency-Band Selection for lg(P/S)

Fig. 7 shows the class-conditional P/S (a) and lg(P/S) (b) distributions across the nine 2-Hz-wide frequency bands swept in Section 2. Earthquake and explosion separate most cleanly at 6–12 Hz, motivating the fixed 8–10 Hz measurement band.

A.4 One-vs-Rest ROC Curves

Fig. 8 shows the one-vs-rest ROC curves for the four three-class configurations. The baseline already separates EQ almost perfectly but reaches only moderate AUC on EP and SS; each auxiliary target lifts the anthropogenic-class AUCs, and the joint model attains AUCs of 0.999 (EQ), 0.967 (EP), and 0.974 (SS). The ordering of AUCs across configurations mirrors the F1 ordering in Table 1, confirming that the gain reflects improved class-conditional separability rather than a decision-threshold artifact.

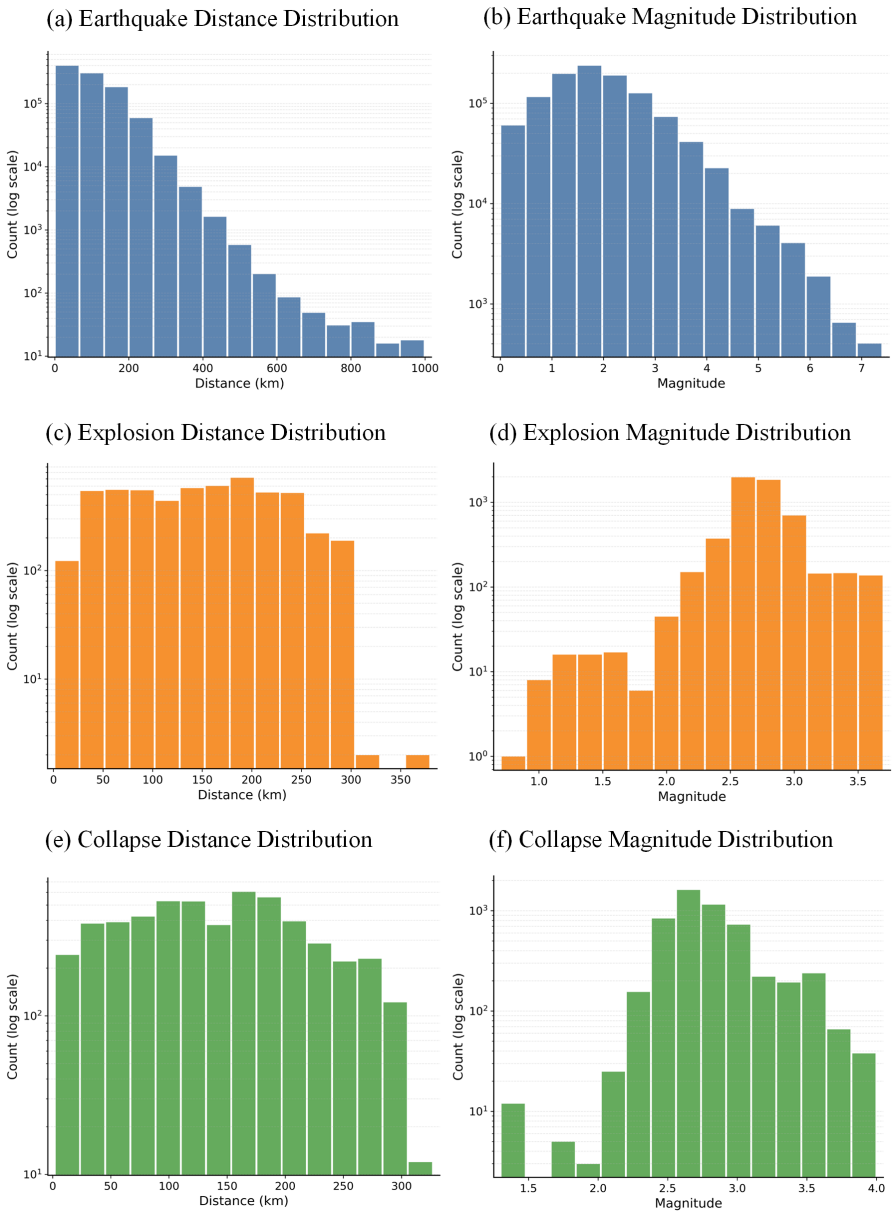


Figure 4: Magnitude and epicentral-distance distributions (distance $\leq$ 300 km) of the earthquake, explosion, and collapse records used in this study.

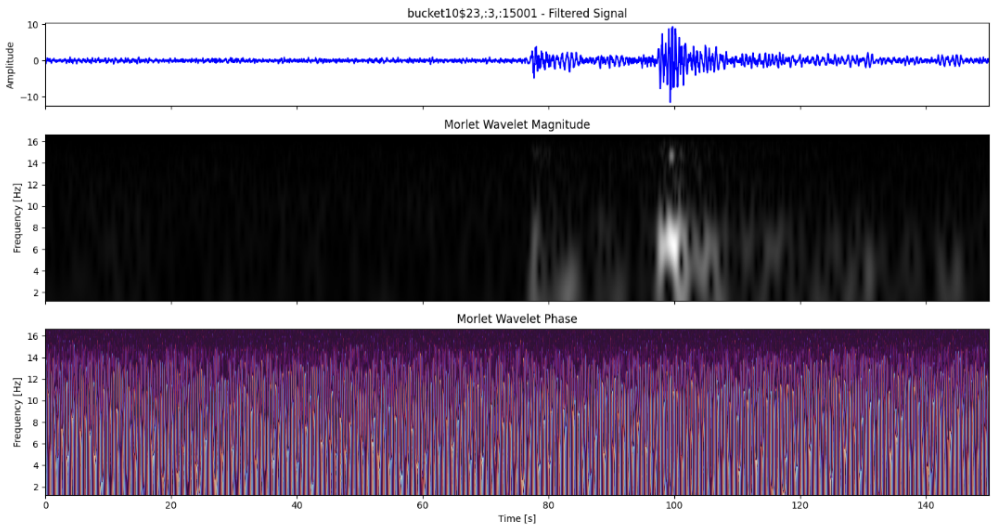


Figure 5: Time–frequency processing of a representative record. (Top) Filtered waveform (E component). (Middle) Morlet-CWT magnitude scalogram, used as model input. (Bottom) Morlet-CWT phase, shown for reference.

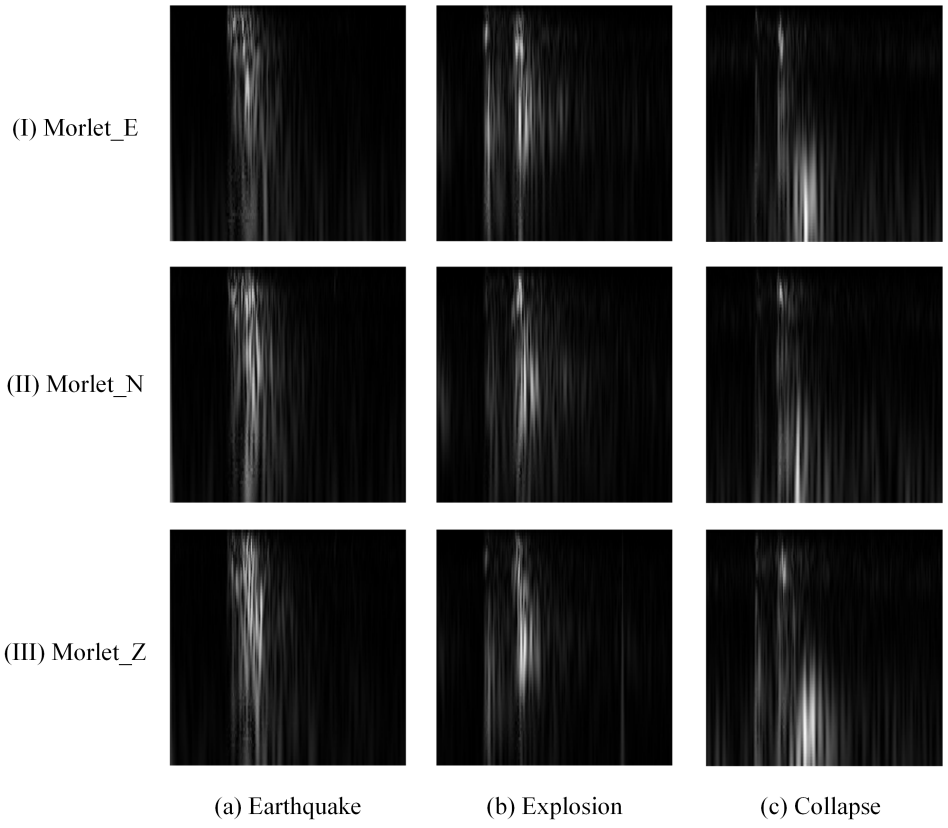
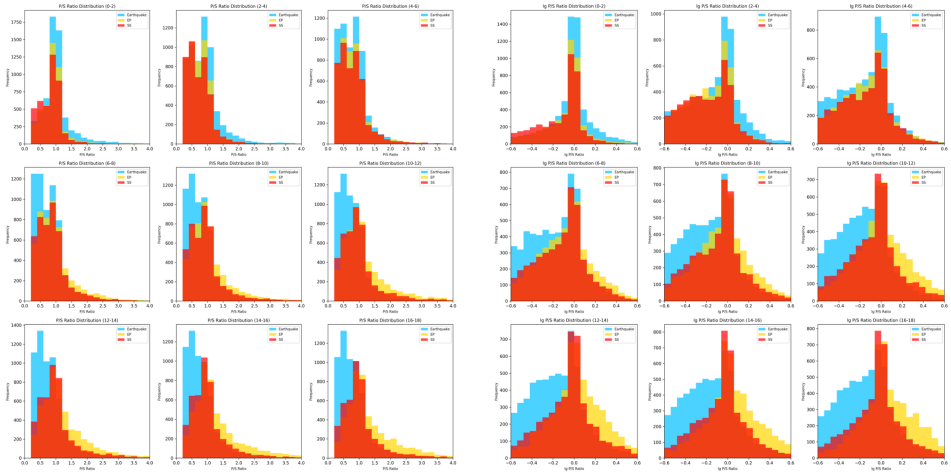
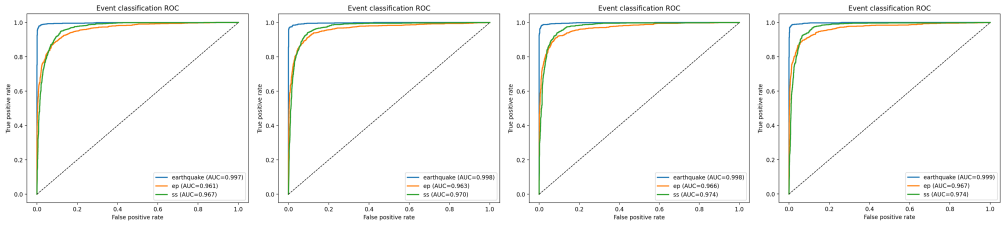


Figure 6: Representative Morlet-CWT magnitude scalograms of the input E, N, and Z components. (a) Earthquake, (b) explosion, (c) collapse.



(a) Class-conditional P/S Distributions (b) Class-conditional lg(P/S) Distributions

Figure 7: Class-conditional P/S (a) and lg(P/S) (b) distributions across nine 2-Hz-wide frequency bands for the three source classes. Earthquake and explosion separate most cleanly at 6–12 Hz.



(a) ROC Curves for Baseline (b) ROC Curves for lg(P/S)-only (c) ROC Curves for Centroid-only (d) ROC Curves for Joint PhysiNet

Figure 8: One-vs-rest ROC curves on the three-class test set for the four configurations: classification-only baseline (a), lg(P/S)-only (b), centroid-only (c), and joint PhysiNet (d). The joint model attains AUCs of 0.999 (EQ), 0.967 (EP), and 0.974 (SS).