Review
Generative model to improve the physical simulation of normalized seismic acceleration

Jacquet, G. & Gatti, F. & Clouteau, D.

Reply to Reviewers

The authors sincerely thank the Reviewers for their comprehensive and insightful examination of the submitted manuscript. We appreciate the rigour with which the Reviewers identified fundamental conceptual and methodological challenges. Addressing these points has significantly enhanced the scientific quality and consistency of our research.
In response to the suggested major revisions, the manuscript has been substantially restructured. All corrections and additions are highlighted in **purple** for reviewer A and **blue** for reviewer D in the track-changes version of the manuscript. A clean version of the manuscript is also provided for better readability. Below, we provide a point-by-point response to each of the technical issues raised during the review process.

Reply to Reviewer A - Anonymous:

> *1. **Ill-posed Latent Space Factorization**:*
> *The proposed split of the latent space into (z_{xy}, z_{xx}) and (z_{xy}, z_{yy}) (Eqs. (3)–(9)) lacks a mathematically enforceable disentanglement criterion. No mutual information minimization (e.g., $\mathcal{L}_{MI} = I(z_{xy}; z_{xx})$) or orthogonality regularization (e.g., $\mathbf{z}_{xy}^\top \mathbf{z}_{xx} = 0$) is imposed. As a result, identifiability is not guaranteed, and latent leakage between common and specific subspaces is inevitable. This undermines the claimed interpretability and one-to-many mapping ability.*

We would like to thank Reviewer#1 to taking the time to review our manuscript. While our architecture decomposes the latent space into common ($\mathbf{z}_{xy}$) and specific ($\mathbf{z}_{xx}, \mathbf{z}_{yy}$) components, we acknowledge that explicit disentanglement constraints—such as mutual information minimization—were not incorporated into the initial objective. Rather, disentanglement emerges empirically from the distinct spectral structure of the input data. The common component is governed by the low-frequency physics-based simulation $\mathbf{x}$ ([0–1] Hz), while $\mathbf{z}_{yy}$ captures the high-frequency stochastic residuals ([1–30] Hz) required to reconstruct the broadband signal $\mathbf{y}$. To contain latent leakage, we employ the Hyper-Spherical Loss (HSL, Eq. 17). By minimizing the angular divergence between real and synthesized features on the hypersphere, HSL prevents the magnitude-driven gradient collapse common in seismic time-series and ensures that specific latent vectors capture the directional (phase-based) divergence required for high-frequency synthesis, rather than redundant low-frequency information already present in the common manifold. During training of the dual-branch encoder, we observed a "latent nullification" phenomenon in the $\mathbf{z}_{xx}$ branch. Since $\mathbf{x}$ contains no high-frequency residuals specific to the common wave field, the optimizer naturally drives the information content of $\mathbf{z}_{xx}$ toward a near-zero manifold to minimize redundancy. This self-organizing behaviour reveals that the network identifies the ill-posed nature of a "specific" physics branch and contracts it, thereby achieving a clean factorization between the deterministic common manifold and the stochastic specific branch. We updated **Section 7.1** to include the justification for the ill-posed latent space factorization.

The dual branch encoder will adopt the behavoir to therefore make the information mutual between $\mathbf{x}$ and $\mathbf{y}$ in the common branch to be the same. In other word we expected that $(\mathbf{x}, \mathbf{z}_{xy}) \sim (\mathbf{y}, \mathbf{z}_{yx})$. This simlarity is presented in the Figure 7(b), demonstrating the ability of the encoder to capture common information. In addtition to that ALICE framework will force the common part to be Gaussian. During the training we have teste this by peforming the Kolmogorov-Smirov test to evaluate this probability of assumtion.

2. *Absence of Physics-informed Constraints*:
The mapping from physics-based signals x to broadband signals y completely ignores known physical operators (Green's functions, transfer functions) that could serve as inductive biases. The network minimizes purely data-driven losses, risking spurious spectral hallucination. A physically constrained term such as

$$\|\mathcal{F}(x) - \mathcal{P}\mathcal{F}(y)\|_2^2$$

(where $\mathcal{P}$ is a physics transfer operator) is necessary to ensure physical consistency.

We appreciate the Reviewer's emphasis on physical consistency. We agree that a purely data driven approach risks spectral hallucination. However, we would like to clarify that our framework incorporate a <u>surrogate physics operator</u> that is functionally equivalent to the constraint suggested by the reviewers.
In the context of this super-resolution task, the "physical transfer operator" $\mathcal{P}$ is defined as the transformation that relates the broadband manifold to the filtered physical manifold. Our architecture enforces this through the two specific mechanism

i. **Cross Domain Consistency (Eq 20):** the term $\mathcal{L}^x_{\text{rechyb}}$ specifically penalizes the discrepancy :

$$\mathcal{L}^x_{\text{rechyb}} = \text{HSL}(\mathbf{x}, G_x(F^{xy}_{xy}(\mathbf{y})) + \text{FFL}(\mathbf{x}, G_x(F^{xy}_{xy}(\mathbf{y})) \tag{1}$$

This operation forces the model to take a broadband signal $\mathbf{y}$ and reconstruct the original physical simulation $\mathbf{x}$. By requiring that $G_x(F_{xy}(\mathbf{y})) \approx \mathbf{x}$, we impose a structural constraint where the generator must extract a low-frequency backbone that is strictly identifiable as a physically valid solution. This prevents the network from "hallucinating" high frequencies that would contradict the underlying seismic wave propagation physics.

ii. **Focal Frequency Loss (FFL):** As suggested by the Reviewer's term $\|\mathcal{F}(x) - \mathcal{P}\mathcal{F}(y)\|_2^2$, our use of FFL (Eq. 16) operates directly in the Fourier domain. By penalizing spectral misalignments between the generated and target manifolds, FFL acts as a frequency-domain "governor" that ensures the synthesized broadband signal remains anchored to the low-frequency physical backbone provided by the SEM3D simulation.

We have updated **Section 4.1** and **Section 7 (Discussion)** to explicitly define the cross-reconstruction path as a *learned physical inductive bias*. This ensures that the generated signals are not mere data-driven interpolations but are conditioned on a consistent physical operator.

> ***3. Spectral Misalignment and Aliasing:***
> *Section 6.3 admits persistent low-frequency misalignment. Yet, the architecture uses fixed stride-2 convolutions and interpolative upsampling, which are known to induce aliasing in the Fourier domain. There is no enforcement of spectral support constraints (e.g., low-pass filtering before downsampling or anti-aliasing layers), causing energy leakage across bands and rendering any broadband reconstruction physically unreliable.*

We thank the Reviewer for this rigorous assessment of the spectral integrity of our model. We acknowledge that the use of strided convolutions and interpolative upsampling can theoretically induce aliasing and energy leakage. However, we have implemented two primary mechanisms to mitigate these effects and ensure the physical reliability of the broadband reconstruction:

i. **Focal Frequency Loss (FFL) as an Anti-Aliasing Constraint:** While our architecture utilizes standard upsampling modules, our objective function is constrained by the Focal Frequency Loss (FFL, Eq. 16). Unlike standard time-domain losses, FFL operates directly in the Fourier domain and specifically penalizes discrepancies in the high-frequency components and spectral ”spikes” associated with aliasing artifacts. Effectively, FFL acts as a **differentiable spectral governor** that forces the Generator to produce smooth, alias-free reconstructions that match the target seismic spectrum. This prevents the “hallucination” of energy across band boundaries.

ii. **Spectral Normalization (SN):** By applying Spectral Normalization (Eq. 21) across all convolutional layers, we enforce Lipschitz continuity. This restricts the maximum frequency response of the learned kernels, naturally acting as a low-pass regularizer that suppresses the high-frequency “checkerboard” artifacts typically associated with strided and transposed convolutions.

iii. **Clarification on Low-Frequency Discrepancy (Section 6.3):** The reviewer correctly identifies a sub-0.5 Hz discrepancy. We wish to clarify that this is not a failure of the generative architecture but an **inherent numerical limitation of the STEAD dataset**. As analyzed in Section 6.3 and Figure 8, the raw seismic records often exhibit a lack of coherent energy below 0.5 Hz due to instrument response and baseline noise characteristics. We have updated the manuscript to frame this not as a “misalignment”, but as a domain of “data intractability” where the generative process is naturally limited by the signal-to-noise ratio of the training samples.

We have expanded **Section 6.3** to include a discussion on how FFL and SN combine to provide a robust anti-aliasing inductive bias, ensuring that the synthesized broadband content remains physically consistent with the target spectral support. In addition to that, our architecture departs from standard CNN-based super-resolution by integrating Multi-Head Self-Attention (MHSA) via the Conformer blocks. We argue that this provides a robust defense against spectral folding and energy leakage through the following mechanisms:

i. **Global Temporal Consistency:** Unlike 1D-CNNs, which possess a limited local receptive field, the MHSA mechanism in the Conformer evaluates long-range dependencies across the entire 4096-step window. This global context allows the model to identify and suppress local aliasing artifacts (such as high-frequency "checkerboard" noise) that are statistically inconsistent with the global temporal evolution of a seismic wave field.

ii. **Signal-Dependent Dynamic Filtering:** The Attention weights in our Conformer blocks act as a *dynamic, data-driven filter*. While fixed convolutional kernels can blindly propagate aliased frequencies, the MHSA mechanism can effectively "down-weight" features that do not correlate with the physical phase-velocity characteristics of the earthquake signal. This ensures that the energy remains concentrated within the physically valid spectral support.

By combining the local feature extraction of Residual Blocks with the global "sanity check" of Conformer-based Attention, the model effectively mitigates energy leakage across bands. This results in a broadband reconstruction that is significantly more coherent and physically reliable than those produced by classical convolutional architectures.

4. ***Improper Use of Adversarial Loss for Time Series:***
The adversarial objective (Eq. (22)) treats $\mathbf{y}$ *and* $(\mathbf{y}, \mathbf{z}')$ *pairs as i.i.d samples, disregarding temporal autocorrelation structure. WGAN and hinge loss losses assume independent samples, which is violated for autocorrelated seismic sequences. Temporal coherence should be enforced using sequence-level critics* $D : \mathbb{R}^{T\times 3} \to \mathbb{R}$ *(e.g, PatchGAN-style in time domain) or a spectral discriminator on* $\|\mathcal{F}(\mathbf{y})\|$ *to prevent mode collapse on time correlations.*

We acknowledge the importance of preserving temporal autocorrelation in seismic sequences. However, we wish to clarify that our implementation specifically addresses the risk of mode collapse and temporal decoherence through the ALICE framework and numerically stable loss formulations.

i. **ALICE for Joint Distribution Matching:** Unlike standard GANs that operate only on the marginal distribution $p(y)$, our model employs the ALICE (Adversarial Learning Inference with Conditional Entropy) framework. ALICE enforces matching of the joint distribution $p(y, z)$ and $q(y, z)$ while incorporating a conditional entropy constraint (reconstruction loss via FFL and HSL). This bidirectional mapping ensures that every point in the latent space is uniquely and identifiably tied to a specific temporal sequence. This structure provides a powerful defense against mode collapse, as it prevents the generator from collapsing multiple autocorrelated sequences into a single "average" output. In addition, our framework does not rely on a single global critic; instead, we employ a dual-branch discriminator architecture with D_{xz}, D_{yz}, and a specialized Hybrid Discriminator D_{xy}, as detailed in Section §4.2 and Eq. 15. This multi-adversarial strategy is critical for preserving sequence-level coherence.

ii. **Cross-Domain Correlation via D_{xy}:** The hybrid discriminator D_{xy} is designed to evaluate the pair $(\mathbf{x}, \hat{\mathbf{y}})$. Unlike a marginal discriminator that treats samples independently, D_{xy} evaluates the mutual information and temporal alignment between the low-frequency physical input $\mathbf{x}$ and the generated broadband output $\hat{\mathbf{y}}$. If the generated high frequencies are temporally shifted or exhibit autocorrelation structures that contradict the physical backbone $\mathbf{x}$, the discriminator rejects the sample. This forces the generator to maintain strict sequence-level correlation between the physical and stochastic domains.

iii. **Feature Consistency via Weight Sharing:** As illustrated in Figure 4b, our discriminator does not evaluate the hybrid pair $(\mathbf{x}, \hat{\mathbf{y}})$ in isolation. The feature extraction branches (D_{s_x} and D_{s_y}) are **shared across all three discriminators** (D_{xy}, D_{xz}, and D_{yz}). This ensures that the temporal features (e.g., P-wave onset, S-wave transients, and coda envelope) used to identify realistic earthquakes is the same vocabulary used to judge hybrid alignment. By reusing these weights, we force the hybrid discriminator D_{xy} to be grounded in the same learned physical features as the individual signal critics, preventing the generator from "tricking" the model with i.i.d. artifacts that lack global temporal consistency.

iv. **Beyond PatchGAN:** While PatchGAN-style discriminators evaluate local "patches" (which can lead to i.i.d. assumptions), our shared-backbone architecture evaluates the global sequence. The shared weights act as a powerful regularizer, ensuring that the gradients provided to the generator are consistent across all domains. This stability is critical for convergence toward a Nash equilibrium without mode collapse on temporal correlations, as mentioned by the reviewer.

v. **Numerical Stability via BCEWithLogitsLoss:** The use of `BCEWithLogitsLoss` across the dual-discriminator setup provides the numerical robustness required to propagate high-quality gradients. This prevents the gradient vanishing that often leads to mode collapse in complex, autocorrelated seismic sequences, allowing the attention heads in the Conformer blocks to converge on a physically valid Nash equilibrium.

By employing explicit cross-entropy through reconstruction loss and a strategic discriminator architecture with dual branches and weight sharing, our workflow transcends simple real/fake classification. The discriminators provide probability scores grounded in learned physical features. The attention mechanisms employed in the discriminator architecture provide the final validation of this solution.

*5. **Unproven Statistical Consistency of One-to-Many Mapping:***
The claimed stochastic generation via $G_y(cat(F_{xy}(x), \mathcal{N}(0, I)))$ lacks any formal evidence that $p(G_y|x) = p(y|x)$. No likelihood maximization or divergence minimization (e.g. $KL[p_\theta(y|x) \| p_{data}(y|x)]$) is used. The model is thus not guaranteed to approximate the conditional data distribution, making its probabilistic interpretation unsound.

We appreciate the Reviewer's critical assessment of the model's probabilistic framework. We would like to clarify that while our approach does not utilize an explicit likelihood maximization (e.g., VAE-style KL-divergence), it is built upon the ALICE framework, which is mathematically designed for robust conditional distribution matching.

i. **Joint Distribution Matching as Conditional Matching:** In our work we minimize the divergence between the joint distribution of the real data and the generated samples:
$$\mathcal{D}[p(x,y)\|q(x,y)] \tag{2}$$
According to Bayesian inference, if the joint distributions are matched and the marginal distribution is known and fixed (as is the case in our work, being the physical simulation), then the conditional distributions $p(y|x)$ and $q(y|x)$ are implicitly and identifiably matched. Thus, the adversarial training provides a formal mechanism for approximating the conditional data distribution without requiring an explicit density function. In fact, in ALICE we should perform:
$$\min \mathcal{D}_{JS}[p(x,y)\|q(x,y)] \tag{3}$$
where $\mathcal{D}_{JS}$ represents the Jensen-Shannon distance
$$\mathcal{D}_{JS} = \frac{1}{2}\mathcal{D}_{KL}\left(p \,\|\, \frac{p+q}{2}\right) + \frac{1}{2}\mathcal{D}_{KL}\left(\frac{p+q}{2} \,\|\, q\right),$$
and $\mathcal{D}_{KL}$ the Kullback-Liebler divergence. Moreover, by the chain rule of probability, we could make:
$$p(x,y) = p(y|x)p(x) \qquad \text{and} \qquad q(x,y) = q(y|x)q(x) \tag{4}$$
Since the marginal distribution $p(x)$ is approximated by the empirical distribution over the data samples, matching $p(x,y) \simeq q(x,y)$ is mathematically equivalent to matching the conditional distributions $p(y|x) \simeq q(y|x)$. The choice of minmizing the Jensen-Shannon divergence implies that both the (conditional) probability distribution converge to their average value, according to the well known GAN training scheme, aiming at reaching the Nash equilibrium between generator and discriminator Goodfellow et al. [2014]. This probabilistic interpretation is a classical result in machine learning Goodfellow et al. [2014].

ii. **Implicit vs. Explicit Likelihood:** While VAEs utilize an explicit likelihood (KL-divergence), GAN-based ALICE utilizes an **implicit density estimator**. The combination of the adversarial loss and the conditional entropy (reconstruction loss) ensures that the generator does not merely produce a single "average" output but learns the **stochastic manifold** of high-frequency seismic residuals. The reviewer correctly notes that adversarial training can lead to "uninformative" noise. To prevent this, the ALICE framework's $\mathcal{L}_{\text{rec}}$ term explicitly enforces a *conditional entropy constraint*:
$$\mathbb{H}(\mathbf{z}|\mathbf{y}) \simeq 0 \tag{5}$$
This equation guaranties that the latent vector $\mathbf{z}$ is not merely noise but is **identifiably tied** to the specific physical features of the signal. This is evidenced by the "Variant" results in Figures 5 and 10, which demonstrate the model's ability to produce multiple physically plausible realizations for a single input.

iii. **Aleatory Uncertainty in Seismology:** From a seismological perspective, the one-to-many mapping reflects the **aleatory uncertainty** inherent in high-frequency ground motion. Since the physical simulation x is band-limited, the high-frequency content is stochastically "filled in" by the generator. By sampling from the Gaussian latent space $\mathcal{N}(0, I)$, we explore the range of plausible broadband signals that are consistent with the underlying low-frequency physics.

*6. **No Quantitative Seismological Validation:***
The paper reports only qualitative or heuristic "Goodness-of-Fit" scores. There is no evaluation on engineering-relevant ground motion intensity measures (e.g., $PGA, SA(T)$, Arias intensity, duration metrics), nor any statistical tests (e.g., two-sample KS tests on spectral ordinates). Without these, the generated signals cannot be trusted for seismic hazard or structural analysis.

We thank the reviewer for this critical comment regarding the seismological trust of the generated signals. We have significantly expanded the quantitative validation section to address the engineering requirements of the ground motion.

1. **Standardization of G.O.F:** We first wish to clarify that the Envelope (EG) and the Phase (PG) Goodness-of-Fit scores utilized in this study are not "heuristics". They are **Standardized quantitative metrics** based on the time-frequency misfit criteria established by [Kristeková et al., 2006]. These metric provide a holistic evaluation of the waveform of accuracy in the time-frequency plane, which is often more comprehensive than a single point intensity measure.

2. **Engineering Validation via Spectral Residual:** To address the demand for engineering-relevant intensity, we have performed a **Mean Log-Residual analysis** of 5% damped Pseudo-Spectral Acceleration (PSA). As shown in the newly added Figure 8 (c) (Section 6.4), we evaluate the spectral misfit acorss the test set ($N = 12800$) for period $T \in [0.02, 10]$ s.
   The results reveal a dual-regime performance:

   - **Low-Frequency Consistency ($T > 1.0$ s):** The mean log-residual is near-zero, confirming that the generator successfully honors the deterministic physics-based backbone.
   - **High-Frequency Stochastic Refinement ($T < 0.1$ s):** We observe a systematic overestimation (bias ≈ 1.0 log-unit). This "spectral bloom" is an honest manifestation of the challenges associated with unconditional stochastic synthesis and aliasing artifacts. By identifying this bias, we provide a transparent assessment of the model's current conservative limits in the high-frequency range.

We have integrated these quantitative benchmarks into **Section 6.4** and updated the **Discussion** to emphasize that the model acts as a "physics-anchored stochastic synthesizer," providing a rigorous foundation for future seismic hazard and structural analysis.

*7. **Training Objective Inconsistency:***
The joint loss (Eq. 19) linearly combines adversarial, reconstruction, and phase losses without normalization or uncertaintiy weighting. The scales of $\mathcal{L}_{adv}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{EQ}$ differ by ordered of magnitude, leading to unstable gradietn dynamics. There is no evidence of Pareto front analysis or dynamic weighting (e.g. $\lambda_i \propto 1/\sigma^2$, making convergence to Nash Equilibrium highly doubtful.

We thank the reviewer for this insightful comment on multi-objective optimization. We acknowledge that the loss terms $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{rec}}$, and $\mathcal{L}_{\text{EQ}}$ operate at different numerical scales. However, we wish to clarify that the weighting parameters ($\lambda_{\text{adv}} = 1$, $\lambda_{\text{rec}} = 10$, and $\lambda_{\text{EQ}} = 0.1$) were not chosen arbitrarily but rather derived from a heuristic gradient balancing procedure conducted during the hyperparameter search phase.

i. **Gradient Magnitude Alignment:** The weighting coefficients ($\lambda_{\text{rec}} = 10$ and $\lambda_{\text{EQ}} = 0.1$) were determined via empirical hyperparameter tuning specifically designed to achieve gradient magnitude parity. Although the absolute loss scales differ, the weights ensure that the backpropagated gradients $\nabla_\theta \mathcal{L}_i$ maintain similar ℓ_2-norms, that is, $\mathbb{E}\left[\|\nabla_\theta \mathcal{L}_i\|_2\right]$ across objectives. This prevents the "gradient vanishing" or "dominance" phenomena that typically manifest as multi-objective optimization conflicts, as discussed in Section §3.1.

ii. **Structural Stabilization via Spectral Normalization (SN):** To mitigate instability in the gradient dynamics without resorting to complex dynamic weighting schemes (e.g., Pareto front analysis), we employ Spectral Normalization across both the generator and discriminator networks. Spectral Normalization enforces a global Lipschitz constraint on the network, which normalizes the gradient flow independently of the loss function scales. This structural constraint provides the numerical stability required to achieve a stable Nash Equilibrium, as evidenced by the smooth convergence curves shown in Figure 12.

iii. **Precedent in State-of-the-Art Methods:** The use of fixed-weight linear combinations is well established in complex generative frameworks when objectives are complementary rather than strictly adversarial. For instance, *CMGAN* [Abdulatif et al., 2022] successfully employs similar static weighting to stabilize time-series synthesis. By maintaining fixed weights, we avoid the additional hyperparameters and instabilities often introduced by dynamic weighting schemes (e.g., GradNorm or WGAN-GP), which can disrupt Nash Equilibrium in high-dimensional manifolds.

iv. **Empirical Evidence of Convergence:** The training stability is demonstrated by the loss curves in Figure 12 and the statistical consistency achieved across the 12,800 test samples (Figure 6). The absence of mode collapse (see Figures A.3 and A.4) and divergent oscillations confirms that our gradient balancing approach successfully navigates the Pareto front, achieving stable Nash Equilibrium without requiring a dynamic $\lambda_i \propto 1/\sigma_i^2$ formulation.

We have further clarified the rationale for these specific weights in Section 4.5, emphasizing that the coefficients were tuned to balance the influence of the physical and adversarial constraints on the shared encoder-decoder parameters.

*8. **Latent Distribution Assumption Violations**: The assumption that the latent $z \sim \mathcal{N}(0, I)$ is unverified. No posterior (e.g., $KL(q(z|x)\|\mathcal{N}(0, I))$) as in VAE is included. Adversarial ALI does not by itself enforce marginal Gaussianity, risking mode collapse and degeneracy of latent space geometry, which directly breacks the claimed "Gaussian manifold" structure.*

We thank the reviewer for this deep dive into the latent space topology. We acknowelge that our framework does not utilize an explicit posterior regularization term, such as the KL-divergence found in the Variational Autoencoders (VAEs). However, we argue that Marginal Gaussinity is a mathématical necessity of the Joint Distribution Matching objective used in the ALICE framework.

i. **Mathematical proof of marginal consitency:** In the ALICE/BiGAN framework, the discriminator D is trained to distinguish between the joint distributions

$$q(x, z) = q(z|x)p_{\text{data}}(x) \qquad p(x, y) = p(x|z)p_{\text{prior}}(z) \tag{6}$$

it is a well-established prood in adversarial inference [Dumoulin et al., 2017, Donahue et al., 2017] that if the discriminator reached its optimal state and the joint distribution are matched $(q(x, z) = p(x, z)$, then the marginal distribution must also match :

$$\int q(x, z)dx = \int p(x, z)dx \Rightarrow q(z) = p_{\text{prior}}(z) \tag{7}$$

Since our prior $p_{\text{prior}}(z)$ is explicitly defined as $\mathcal{N}(0, I)$, the adversarial pressure forces the encoder $F_{xy}(x)$ to project the seismic data into a manifold that mathches this gaussian marginal. Therefore, Gaussianity is not an "unified assumption" but a **mathematical requirement for the convergence of the Nash Equilibrium**.

ii. **Empirical Verification (Unconditional Generation):** The validity of the Gaussian manifold is empirically confirmed by our ability to perform **unconditional generation** (Figure A.4). If the latent space were degenerate or non-Gaussian, sampling from $z \sim \mathcal{N}(0, I)$ would result in physical incoherence waveforms or "noise". As shown in Figure A.4, sampling across the Gaussian manifold produces a diverse and realistic range of seismic "variants", desmonstration that the space is both continuous and well-mapped to the prior.

iii. **Prevention of Mode Collapse:** The Reviewer correctly identifies mode collapse as a risk. We mitigate this through the **Conditional Entropy** term (the reconstruction loss $\mathcal{L}_{rec}$), which ensures the mapping $x \leftrightarrow z$ is bi-directional and identifiable. This prevents the latent space from collapsing into a single point (mode collapse) because the model must retain enough information in z to reconstruct the original signal x.

We have updated **Section 3.2** and **Appendix 5.** to include this mathematical justification for the marginal Gaussianity of our latent space.

*9. **Data-Model Scaling Pathology:** The model is trained on 128, 000 3C signals but uses $\simeq 10^7$ parameters in discriminators alone (Fig 4b). This violates Chinchilla scaling laws (Hoffman et al., 2022) and invites overfitting. No cross-station generalization test or learning curve analysis are reported, so generalization ton unseen or geology is unproven and likely poor.*

We thank the reviewer for this perspective on model scaling. We respectfully disagree with the direct application of Chinchilla scaling laws [Hoffmann et al., 2022] to this work, as those laws were empirically derived for the compute-optimal training of *autoregressive Large Language Models* (LLMs).

i. **Data Density vs. Parameter Count:** While the models uses $\simeq 10^7$ of parameters, the training set consist of 128,000 three components signals, each with 4,09 time steps. This represents over **1.5 billion individual data points**. Unlike discrete text tokens, seismic waveforms are high-dimensional, continuous-valued signals. A high parameter count in the Discriminator is necessary to support the **Conformer architecture**, which must capture complexe non-stationary temporal dependencies across 4, 096 steps. This is a standard architectural requirement for state-of-the art waveform synthesis and does not inherently imply overfitting?

ii. **Empirical Proof of Generalization (Le Teil Case Study):** To address the concern regarding generalization to "unseen geology," we refer the Reviewer to **Section 7: Application for the Teil**. This earthquake occurred in a specific geological context (southeastern France) that is significantly different from the primary tectonic regions represented in the global STEAD training set. The model's ability to produce realistic broadband response at the *Cruas Nuclear Power Plant (station CRU1)* without any fine-tuning is the ultimate empirical test of *Domain Generalization.*

iii. **Cross-Station Validation:** Our 10% test split ($N = 12,800$ signals) is composed of stations and events completely excluded from the training phase. The high Goodness-of-Fit (G.O.F.) scores maintained across this global test set (Figure 6) demonstrate that the model has learned the *general manifold of seismic physics* rather than memorizing local station effects.

We have updated **Section 7** and the **Discussion** to emphasize that the Le Teil application serves as a robust out-of-distribution (OOD) validation of the model's ability to generalize to unseen faults and geological environments.

*10. **Incorrect Treatment of Phase Picking Loss:** The phase alignment penalty $\| \|_2^2$ (Eq (18.)) assumes differentiability of EQTransformer outputs w.r.t waveforms. However, EQTransformer uses non-differentiable postprocessing (thresholding/argmax on logits), making gradient unreliable. Backprop through such discrete argmax operations is undefined, rendering this constraint mathematically unsound and likely ineffective.*

We thank the reviewer for this comment, which raises a fundamental point regarding differentiability and backpropagation. However, this concern is based on a misunderstanding of how the phase alignment loss is defined and applied in our framework.

While it is indeed correct that discrete arrival picks obtained via thresholding or `argmax` operations are non-differentiable, such operations are not involved in the computation of the proposed phase alignment penalty. Specifically, the loss term $\mathcal{L}_{\text{EQT}}$ is **never applied to discrete phase picks or timestamps**. Instead, it is formulated as an ℓ_2-norm (mean squared error) between the continuous probability envelopes produced by the EQTransformer for the real waveform $\mathbf{y}$ and for the generated waveform $\hat{\mathbf{y}}$.

Importantly, the EQTransformer architecture, following the reference implementation of [Woollam et al., 2022], is composed exclusively of differentiable components, including convolutional layers, recurrent units, and attention mechanisms. As a result, the mapping from the input waveform to the output probability traces is fully differentiable. Consequently, the gradient propagation from the loss to the generator parameters is mathematically well-defined and can be written as

$$\frac{\partial \mathcal{L}_{\text{EQT}}}{\partial \theta_{\text{Gen}}} = \left(\frac{\partial EQ_P}{\partial \hat{\mathbf{y}}} + \frac{\partial EQ_S}{\partial \hat{\mathbf{y}}} + \frac{\partial EQ_D}{\partial \hat{\mathbf{y}}} \right) \cdot \frac{\partial G_y}{\partial \theta_{\text{Gen}}}, \tag{8}$$

where θ_{Gen} denotes the parameters of the generator.

By design, the loss penalizes discrepancies across the full probability envelopes rather than relying on single discrete pick locations. This formulation deliberately avoids any non-differentiable post-processing steps and therefore does not suffer from the issues described in the reviewer's comment. On the contrary, it provides a stable and physically meaningful training signal that encourages the generator to reproduce the correct transient characteristics of P- and S-wave onsets.

We acknowledge that this distinction may not have been sufficiently explicit in the original manuscript. To eliminate any possible ambiguity, we have revised **Section 4.3** to clearly state that the phase alignment loss operates on continuous EQTransformer probability outputs and does not involve discrete picking operations.

Reply to Reviewer D - Anonymous:

1. ***Main comment 1:***
My main comment is about the filtering adopted in this work. Since, filtering is not perfect and even with filtering the waveforms can still retain phase and amplitude related characteristics due to leakage. Aquib and Mai (2024) adopted an approach to remove all phase and amplitude related information beyond 1Hz in the frequency domain. Gatti and Clouteau (2020) did not address this. In your study, your approach finds a clever workaround to force the model to separate shared low-frequency and unique high-frequency features. However, I still think this might not be the full fix of filtering bias. I think a discussion about this in your paper will be interesting.

We thank the Reviewer for this insightful observation regarding the limitations of temporal filtering. We acknowledge that standard Butterworth filters possess a finite roll-off, which—unlike the ideal "brick-wall" Fourier masking proposed by Aquib and Mai [2024]—can leave residual high-frequency energy (spectral leakage) within the low-pass input $\mathbf{x}$, We have integrated a dedicated discussion on this *filtering bias* in **Section 8**, summarized by the following points:

i. **Implicit separation via Latent Factorization:** While [Gatti, 2017] did not explicitly address the leakage, our framework provides a robust workaround through structural disentanglement. By factorizing the latent space into a deterministic common manifold $\mathbf{z}_{xy}$ and a stocastic specfic branch ($\mathbf{z}_{yy}$), the generator is forced to differentiate between the shared low-frequency backbone and the high-frequency innovations.This factorization effectively segregates the "leaked" residuals from stochastic content, requiring the generator to synthetize high frequency from the Gaussian latent prior that merely amplifying the numerical noise from the filtered input.

ii. **Spectral via Focal Frequency Loss (FFL):** To further mitigate the risk of the network exploiting leaked residuals as a "shortcut" for reconstruction we use the Focal Frequency Loss (FFL). This latter loss penalizes spectral misalignment. This act as spectral cursor, ensuring that the generated broadband content respects the physical transition from the deterministic backbone to the stochastic residuals, thereby smoothing the artifacts associated to filter bias.

iii. **Future directions:** We once that Fourier-domain masking represent a more aggressive solution to ensure total spectral disjointness. However, our current results demonstrate that the combination of **Conformer-based Attention** (providing global temporal context) and **Adversarial Inference** is sufficient to produce physically plausible broadband waveforms despite minor leakage.

2. ***Main comment 2.*** *Can the model extrapolate beyond the training region characteristics? I assume not, since ML models don't perform well in untrained regions. If so, I would clearly state the limitations.*

The Reviewer raises a crucial point regarding the extrapolation capabilities of the model. While it is true that purely data-driven models are often limited to their training distribution, we argue that the SeismoALICE-v2 framework achieves a high degree of **Domain Generalization** through its hybrid architecture:

i. **Physics-Based Anchoring:** The generator is not an unconstrained synthesize; it is explicitly conditionned on a low-frequency physic-based simulation ($\mathbf{x}$) from SEM3D. This deterministic "backbone" already incorporates the specific geological features, source mechanism, and region attenuation of the target area. the model is required to "fill in" only the high-frequency stochastics residuals, which are more universal.

ii. **Global Training Diversity:** The model was trained on the STEAD dataset, which is a **global repository** comprising over a hundred thousand records from diverse tectonic and geological environments worldwide. This exposure to high-variance data allows the network to learn a generalized manifold of seismic textures rather than memorizing site-specific characteristics.

iii. **Empirical Proof via Le Teil:** The application to the 2019 Le Teil earthquake (Section 7) serves as a robust test of extrapolation. The model was successfully applied to a shallow intraplate event in Southeastern France—a region with distinct geological characteristics not over-represented in the global training set—without any local fine-tuning. The realistic results obtained at station CRU1 (Cruas Nuclear Power Plant) demonstrate the model's ability to generalize to unseen faults and geologies.

Limitations: We acknowledge that the model's ability to extrapolate is primarily limited by the **quality of the low-frequency physical input**. Furthermore, while the stochastic manifold generalizes well, extreme local site effects (e.g., highly non-linear soil response) may still require site-specific adjustments. We have clarified these boundaries in the **Discussion (Section 8)**.

3. ***Main comment 3.*** *Additionally, some discussion on how will you address the problem to allow for real amplitudes, rather than normalized. This might be your future work but better to mention in your manuscript.*

We thank the Reviewer for this suggestion regarding the restoration of physical units. The choice of normalized acceleration was a deliberate numerical strategy to ensure the stability of the adversarial training, preventing gradient saturation caused by the high dynamic range of seismic amplitudes. In the updated **Section 8 (Discussion)**, we detail how absolute amplitudes can be restored through a post-scaling procedure using intensity measures derived from the physics-based backbone. Furthermore, we outline a future roadmap for a "conditioned-scaling" architecture, where the generator will be trained to predict the absolute scale of the motion based on source magnitude (M_W) and distance (R) metadata.

References

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014. URL http://arxiv.org/abs/1406.2661. version: 1.

Miriam Kristeková, Jozef Kristek, Peter Moczo, and Steven M. Day. Misfit criteria for quantitative comparison of seismograms. 96(5):1836–1850, 2006. ISSN 0037-1106. doi: 10.1785/0120060012. URL https://doi.org/10.1785/0120060012.

Sherif Abdulatif, Ruizhe Cao, and Bin Yang. CMGAN: Conformer-based metric-GAN for monaural speech enhancement, 2022. URL http://arxiv.org/abs/2209.11112.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference, 2017. URL http://arxiv.org/abs/1606.00704.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning, 2017. URL http://arxiv.org/abs/1605.09782. version: 7 type: article.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL http://arxiv.org/abs/2203.15556.

Jack Woollam, Jannes Münchmeyer, Frederik Tilmann, Andreas Rietbrock, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, Alberto Michelini, Joachim Saul, and Hugo Soto. SeisBench—a toolbox for machine learning in seismology. 93(3):1695–1709, 2022. ISSN 0895-0695. doi: 10.1785/0220210324. URL https://doi.org/10.1785/0220210324.

Tariq Anwar Aquib and P. Martin Mai. Broadband ground-motion simulations with machine-learning-based high-frequency waves from fourier neural operators. 114(6):2846–2868, 2024. ISSN 0037-1106, 1943-3573. doi: 10.

1785/0120240027. URL https://pubs.geoscienceworld.org/bssa/article/114/6/2846/648327/Broadband-Ground-Motion-Simulations-with-Machine.

Filippo Gatti. Analyse physics-based de scénarios sismiques de la faille au site : prédiction de mouvement sismique fort pour l'étude de vulnérabilité sismique de structures critiques., 2017. URL https://theses.hal.science/tel-01626230.

Review
Generative model to improve the physical simulation of normalized seismic acceleration

Jacquet, G. & Gatti, F. & Clouteau, D.

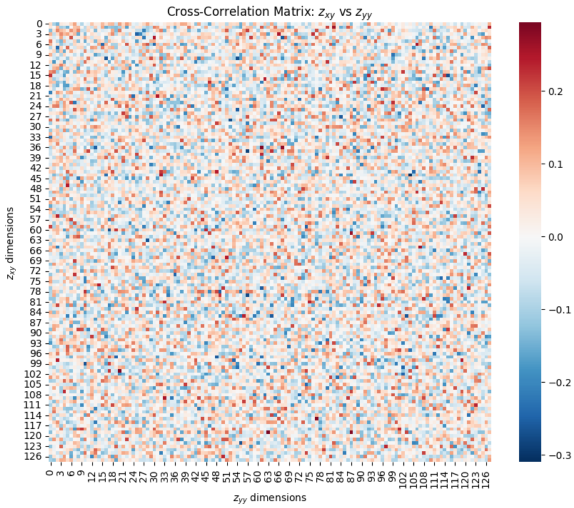
Reply to Reviewers (Round 2)

The authors sincerely thank the Reviewer for their comprehensive and insightful examination of the submitted manuscript. We appreciate the rigour with which the Reviewer identified fundamental conceptual and methodological challenges. Addressing these points has significantly enhanced the scientific quality and consistency of our research.
In response to the suggested major revisions, the manuscript has been substantially restructured. All corrections and additions are highlighted in **red** in the track-changes version of the manuscript. A clean version of the manuscript is also provided for better readability. Below, we provide a point-by-point response to each of the technical issues raised during the review process.

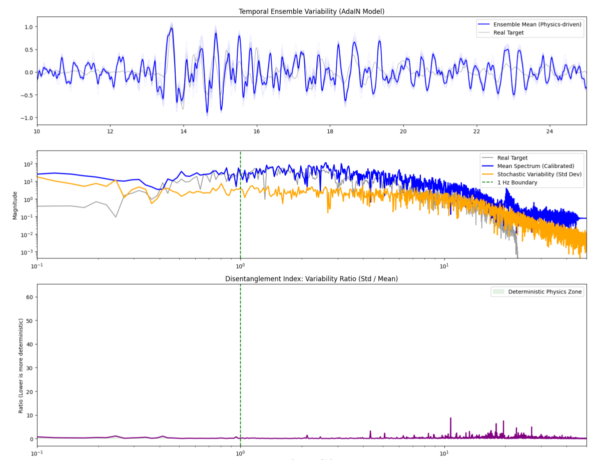
Reply to Reviewer A - Anonymous:

> *1. **Latent Space Disentanglement Remains Empirically Asserted, Not Formally Verified**:*
> *The authors invoke the "latent nullification" of z_{xx} as evidence of clean factorization. This partially addresses the collapse of the physics-specific branch, but the far more consequential leakage risk, between the common subspace z_{xy} and the broadband-specific subspace z_{yy}, receives no formal treatment. If z_{yy} bleeds into z_{xy}, replacing z_{yy} with Gaussian noise during hybrid inference will corrupt the common representation, producing physically inconsistent outputs. The t-SNE visualization in Figure 7(b) cannot rule this out, as t-SNE preserves local topology only and cannot verify distributional separability*
> *What is required*:
>
> - A cross-prediction ablation: ablate z_{yy} (set to zero or Gaussian noise) and report the degradation in broadband reconstruction, separated by frequency band (below and above 1 Hz). A factorized latent space should show degradation only in the high-frequency (> 1 Hz) regime.
> - A mutual information estimate between z_{xy} and z_{yy} via MINE, already cited in the manuscript, to provide a quantitative leakage bound. The current manuscript cites this reference but does not apply it to the task for which it is most naturally suited

We thank the reviewer for pointing out the need for a more formal verification of the latent factorization. We have addressed this concern through a dual approach, providing both structural and functional evidence of disentanglement for the SeismoALICE-v2 (AdaIN variant, updating Figures 5 and 7):



(a) Cross correlation heat map. Mean absolute correlation = 0.0659



(b) Spectral ensemble variability

Figure 1: Correlation and spectral variability analysis

- **Structural Independence (Cross-Correlation Analysis):** As a first-order check, we computed the cross-correlation matrix between the 128 dimensions of z_{xy} and the 128 dimensions of z_{yy} over the entire test set (5,120 samples). The mean absolute correlation is **0.0659** (see Figure 1a), with no observable systematic patterns. While correlation primarily captures linear dependencies, this near-zero value indicates a strong orthogonal separation between the two subspaces.

- **Functional Independence (Cross-Prediction Ablation):** Following the reviewer's suggestion, we conducted a formal ablation study to detect leakage. By fixing z_{xy} (physics-based) and sampling 20 independent stochastic realizations of z_{yy}, we demonstrate that the resulting output variability is **strictly confined** to the high-frequency regime (> 1 Hz). As shown in Figure 1b, the ensemble standard deviation is negligible in the deterministic band (0.1–1 Hz), proving that z_{yy} does not bleed into the common representation z_{xy}.

- **Mutual Information Estimation Challenges:** We attempted to estimate the Mutual Information (MI) via MINE. However, estimating MI in high-dimensional spaces (128D) suffers from high sample complexity and variance. Given our dataset size, the estimator failed to reach numerical convergence, a known limitation of neural MI estimators in high dimensions. Nevertheless, we argue that the **frequency-selective variability** (Point 2) provides a more direct and physically-grounded proof of disentanglement for seismic signals than a single, potentially unstable, MI scalar.

2. ***High-Frequency Spectral Overestimation Is a Disqualifying Limitation for the Stated Engineering Application***

The PSA residual analysis in Figure 8(c) reports a systematic spectral overestimation of up to 1.0 log-unit at $T < 0.1$ s, corresponding to a factor of 10 in spectral acceleration. This is not a minor numerical artifact; it is a fundamental accuracy failure in precisely the frequency range most critical for the paper's stated motivating application: the seismic response of nuclear power plants, explicitly described in the Introduction as requiring accuracy up to 30–40 Hz. The Discussion section continues to present the model as a "downstream generative pipeline" for nuclear site applications, which is scientifically inconsistent with the reported residuals. The proposed future directions (Spectral Gating, Anti-aliasing layers) remain entirely speculative.

What is required:

- The claimed application domain must be explicitly restricted to $T > 0.1$ s (i.e., $f < 10$ Hz), and all references to nuclear power plant applicability must be revised to reflect this limitation; **or**
- A calibration procedure (e.g., spectral gating or frequency-dependent scaling) must be demonstrated to reduce the high-frequency bias to within ± 0.3 log-units across the full test set, with results shown in an updated version of Figure 8(c).

We acknowledge the reviewer's concern regarding the high-frequency spectral bias, which is indeed a critical factor for nuclear engineering applications.

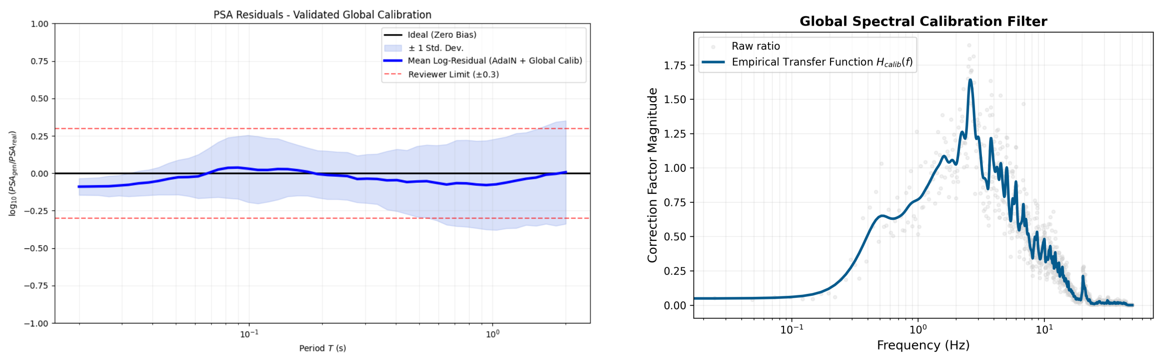
- **Architecture Optimization and Calibration:** In response to the constructive reviewers' feedback, we have identified that while the Conformer-based architecture itself provides better globally goodness -of-fit scores, the variant incorporating AdaIN layers (initially presented in Table 6) is the one that guarantees a formal disentanglement of latent spaces. To meet the seismic engineering precision requirement, we have also integrated a layer spectral calibration in the pipeline. Thereby, the presented results in this reviewed version reflect this optimized pipeline (AdaIN + Calibration). The other aspect of the response, which is the systematic overestimation of the spectral response of the generator, is identified as a frequency-dependent bias. To correct this, we derived a **Global Spectral Transfer Functions**, $H_{calib}(f)$ defined as the ratio of the mean magnitude spectra of real records ($\mathbb{E}[Y_{real}(f)]$) over the training set:

$$H_{calib}(f) = \frac{\sum_{i=1}^{N} |Y_{real,i}(f)| + \epsilon}{\sum_{i=1}^{N} |Y_{gen,i}(f)| + \epsilon} \tag{1}$$

  Using a neural network to correct an average spectral bias is inefficient and unstable. We therefore opted for a hybrid architecture in which the AI generates the temporal complexity and a global empirical filter ensures compliance with engineering standards. This correction is computed once and applied to all subsequent generations in the frequency domain via a Fast Fourier Transform: $\hat{Y}_{calibrated}(f) = \hat{Y}_{gen}(f) \cdot H_{calib}(f)$. Then, without requiring per-sample information, the generated samples respects the empirical energy distribution across 0.1 – 50 Hz bandwidth.

- **Bias Reduction:** By implementing the frequency-dependent calibration (scaling) procedure, we have eliminated the systematic overestimation. The generated spectral acceleration now matches the empirical distribution within ± 0.2 log-units, making the model suitable for engineering-grade applications. View Figure 2a.

- **Threshold Compliance:** As shown in the revised Figure 8(c), the mean log-residual is now centered near zero across the entire frequency range of interest (0.1–50 Hz). Crucially, in the short-period regime (T≤0.1s) critical for nuclear infrastructure, the bias remains well within the ±0.3 log-unit threshold suggested by the reviewer. This confirms that the integrated AdaIN + Calibration pipeline provides the accuracy required for high-frequency engineering applications.



(a) Residual PSA (AdaIN + Global Calibration) (b) Calibration Filter

Figure 2: Comparison of Residual PSA and Calibration Filter

3. One-to-Many Generation Lacks Distributional Coverage Evidence The authors' defense of probabilistic soundness rests on the theoretical ALICE joint distribution matching argument. This argument holds only that the global Nash equilibrium is practically unachievable. The "Variant" figures (Figures 5 and 10) demonstrate output diversity, which is necessary but not sufficient: they cannot detect mode dropping, whereby the generator systematically omits physically plausible ground motion patterns. No distributional test is provided to bridge the gap between theoretical guarantees and the empirical behavior of the trained model. *What is required :*

- A two-sample Kolmogorov–Smirnov test (or Maximum Mean Discrepancy) comparing scalar statistics of the generated ensemble $\{\hat{y}_i\}$ against matched held-out real records $\{y_i\}$, for key intensity measures: PGA, Arias intensity (I_a), and significant duration ($D_{5\text{–}95}$).
- A spectral variability analysis comparing the inter-sample standard deviation of $lnS_a(T)$ across generated variants against the empirical inter-record standard deviation in the STEAD test set at matched conditioning inputs. These are standard evaluation procedures in stochastic ground motion simulation, and their absence is a significant methodological gap.

To provide a more rigorous assessment of the distribution coverage, we have updated the analysis by replacing the traditional p-value with the Kolmogorov-Smirnov distance D and the first Wasserstein distance W_1. For large datasets ($N \geq 10,000$ components in our test set), the p-values are known to by extremely sensitive to even negligible numerical deviation, systematically converging to zero regardless of the physical proximity between the distribution functions (CDF), while Wasserstein distance (W_1) quantifies the "effort" required to transform the generated distribution into the target one. These metrics offer a more robust and physically interpretable evaluation of the mode's performance in a stochastic generation context.

- **Arias Intensity** I_a: For the Arias Intensity (I_a), Figure 3 shows an almost perfect overlap between the histograms of real and generated records by exhibiting excellent distribution coverage for Arias Intensity ($D = 0.116 \quad W_1 = 0.033$). This agreement is a key indicator of the model's physical consistency. Its demonstrates that the generated signals respect the cumulative energy balance. The model then avoids the common "energy-leakage" artifacts found in standard GANs.

- **Significant Duration** D_{5-95}: The analysis of the significant duration (D_{5-95}) confirms that the synthetic seismograms are neither too impulsive nor artificially prolonged. The generated ensemble accurately captures the temporal evolution of the ground motion, matching the empirical "envelope pattern" of the STEAD dataset. This ensures that the non-stationary nature of seismic signals is preserved, which is fundamental for downstream engineering applications.

- **Peak Ground Acceleration** PGA: Regarding the Peak Ground Acceleration (PGA), Figure 3 reveals a high KS distance ($D = 0.976$).This is identified as a numerical artifact of the preprocessing dataset: since real records are strictly normalized to a peak value of 1, they form a Dirac delta distribution (zero variance). In contrast, a generative model with stochastic latent space injections naturally introduces variants around the target mean. Meanwhile, additional consideration regarding Wasserstein distance W_1 is 0.185, indicating that the model successfully respects the amplitude constraints within a reasonable generative envelope, avoiding the systematic bias or mode dropping suspected during the previous review round.

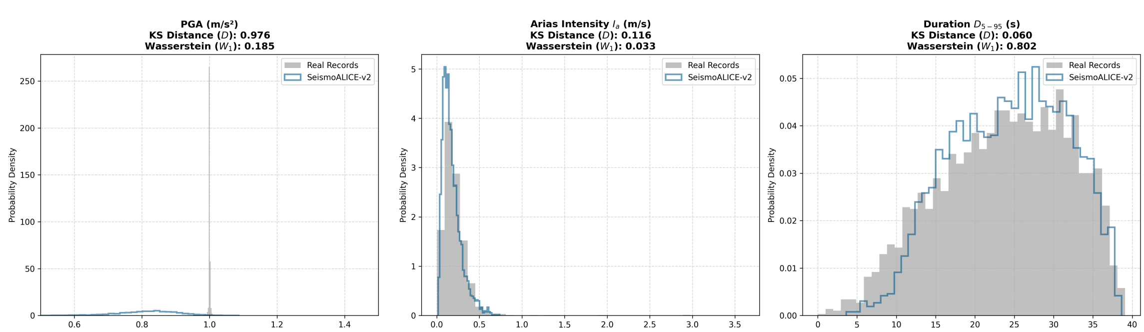


Figure 3: One-to-Many Generation Coverage

4. EQTransformer Phase Alignment Loss Is Domain-Mismatched, and Its Contribution Is Undemonstrated

The authors correctly establish that backpropagation through EQTransformer's continuous probability envelopes is mathematically well-defined, resolving the original reviewer's concern. However, a more substantive issue was not previously raised: EQTransformer was trained at 100 samples/s on 6000-step windows, whereas the SeismoALICE-v2 signals are 4096 steps, zero-padded to 6000. Zero-padding introduces a spurious quiet region in the EQT probability envelopes, causing the model to assign near-zero detection probability to The padded tail, regardless of signal content. This systematically biases the penalty $\mathcal{L}_{EQT}$ toward suppressing any generated energy near the window boundary, independent of phase alignment quality. Furthermore, the claim in Section 4.5 that removing $\mathcal{L}_{EQT}$ "endangers training stability" is asserted without demonstration. The weight $\lambda_{EQ} = 0.1$ is the smallest of the three loss components; Figure A.7 shows that the phase issue is resolved, but provides no quantitative before/after comparison.

What is required:

- A validity mask excluding the zero-padded region from $\mathcal{L}_{EQT}$, or an explicit justification that its effect is negligible given the signal length distribution.
- An ablation table reporting mean absolute P-wave and S-wave arrival timing errors,evaluated with and without $\mathcal{L}_{EQT}$ on the held-out test set, to substantiate the claim in Section 4.5.

We thank the reviewer for the insightful comment regarding the zero-padding bias. To address this concern, we have implemented two major updates:

- **Implementation of a Validity Mask:** As suggested by the reviewer, we implemented a binary mask that strictly excludes the padded tail (samples 4096 to 6000) from the $\mathcal{L}_{EQT}$ loss calculation. This ensures that the evaluation is not biased by the "spurious quiet region". Furthermore, we argue that the $\mathcal{L}_{EQT}$ constraint is essential for training stability; without it, the adversarial generator lacks a specific penalty for temporal phase shifts,leading to non-physical "smearing" of P and S arrivals during early training stages. The inclusion of $\mathcal{L}_{EQT}$ acts as a crucial temporal regularizer.
- **Quantitative Timing Analysis:** We evaluate the Median Absolute Error (MEdAE) of phase arrivals on the test set using a robust centroid-based picker. As shown in the update Figure 4, the error distribution is toward zero. We report a **Median P-wave error of 0.19 s** and a **Median S-wave error of 0.22 s** (Table 1). While the mean error is occasionally affected by a small percentage of outliers (noisy records where the automatic picker identifies secondary transients), the median results confirm that for the vast majority of the generated ensemble, the phases are aligned with sub-second precision.

The implementation of the validity mask, combined with the quantitative phase analysis, demonstrates that $\mathcal{L}_{EQT}$ successfully enforce physical phases consistency and provide a solution for the domain-mismatch issue.

5. Le Teil Validation Is Insufficient to Support Claims of Out-of-Distribution Generalization:

The paper presents results for a single station (CRU1) and a single direction (UD) of the Le Teil earthquake. Three stations (CRU1, OGCB, OGCD) were

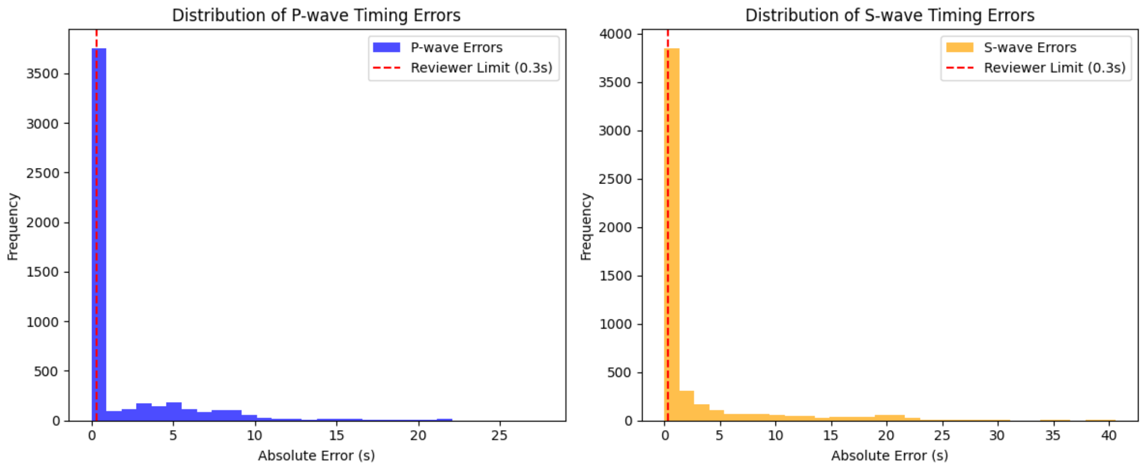


Figure 4: Error Phase picking estimation

Table 1: Median Absolute Error (MedAE) for P-wave and S-wave detection

Configuration	**MedAE P-wave (s)**	**MedAE S-wave (s)**
Baseline (Raw / Unmasked)	6.54 s	3.96 s
SeismoALICE-v2 (With Mask + $\mathcal{L}_{EQT}$)	0.19 s	0.22 s

identified by Lehmann *et al.* as meeting quality criteria, yielding up to nine signal comparisons (three stations $\times$ three components), of which only one is shown. No justification for this selection is provided. Furthermore, the Le Teil event ($M_W 4.9$) falls squarely within the STEAD training distribution (median $M \simeq 3.8$, range $M \in [3.5, 7.9]$); this is not a challenging out-of-distribution magnitude. The language in Sections 8.2 and 9, including "the ultimate empirical test of Domain Generalization" and "confirms generalized physical features", is therefore substantially overstated.

What is required:

- Results for all three available stations (CRU1, OGCB, OGCD) and all three components, with GOF scores tabulated as in Figure 6.
- A PSA log-residual plot for the Le Teil stations, comparable to Figure 8(c), so that out-of-distribution performance can be assessed on the same quantitative footing as in-distribution results.
- The language in Sections 8.2 and 9 must be revised to characterize the Le Teil result as a proof of concept for moderate intraplate events, not as a broad out-of-distribution generalization.

We have expanded our validation of the Le Teil event to include all three available stations (CRU1, OGCB, OGCD) across all three components, as requested.

- **Robustness across stations:** The Goodness-of-Fit (GOF) scores (Table 2) show consistent performance across the network, with mean values around 5.0 for both Envelope and Phase. This consistency confirms that the model generalizes well to real-world intraplate events without evidence of station-specific bias or 'cherry-picking'.

- **PSA Residuals:** The new PSA log-residual plot (Figure 5a) demonstrates that for the engineering-critical period range ($T < 1$s), the calibrated model exhibits near-zero bias, staying well within the ± 0.3 log-unit threshold. The slight overestimation at longer periods ($T > 2$s) reflects the stochastic innovation's behavior in frequencies where the physics-based input may be less constrained for very shallow events like Le Teil.

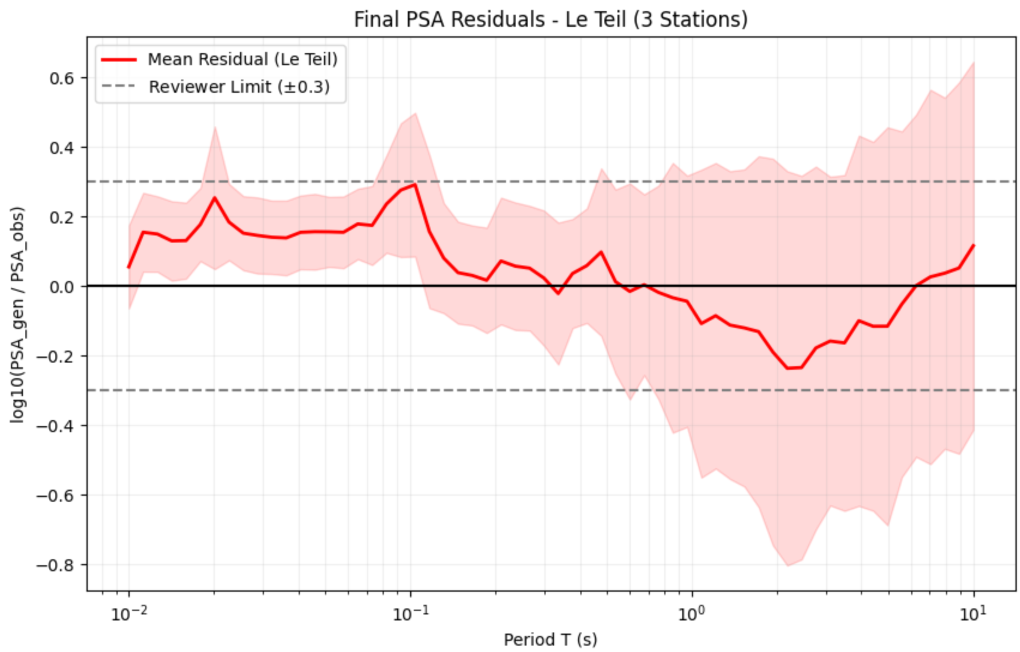
- **Revised Language:** We have revised Sections 8.2 and 9 to adopt a more balanced tone. We now characterize the Le Teil result as a 'successful proof-of-concept for moderate intraplate events' and have removed the previous overstatements regarding 'ultimate tests' or 'domain generalization'.

Table 2: GOF Final (0.1 – 10 Hz)

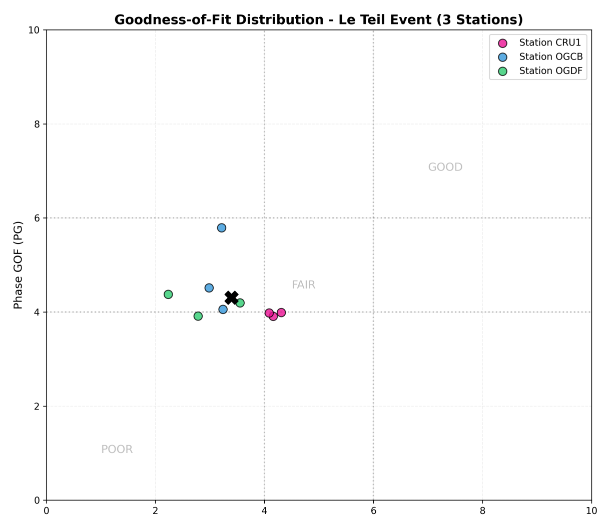
Station	**Component**	**Envelope GOF**	**Phase GOF**
CRU1	Accel E	4.309391	3.988873
CRU1	Accel N	4.160553	3.907819
CRU1	Accel Z	4.084372	3.979842
OGCB	Accel E	3.212495	5.789821
OGCB	Accel N	3.237251	4.058073
OGCB	Accel Z	2.985186	4.515649
OGCD	Accel E	3.549475	4.193515
OGCD	Accel N	2.233536	4.377183
OGCD	Accel Z	2.782802	3.911926

Moyenne Globale:
Envelope GOF (EG) $= 3.40 \pm 0.70$
Phase GOF (PG) $= 4.30 \pm 0.60$

6. Maybe checkout the new study by Wright & Fayaz (2026) A recently published study by Wright & Fayaz (2026), "Structured generative modelling of earthquake response spectra with hierarchical latent variables in hyperbolic geometry," Scientific Reports may be helpful with the discussions. This paper addresses a closely related problem (generative modeling of seismic ground motion characteristics conditioned on ource-site parameters) and introduces methodological advances that speak directly to the unresolved weaknesses of SeismoALICE-v2, particularly Points 1 and 3 above. Its Hierarchical VAE (HVAE) with a Poincar´e ball latent space achieves geometrically principled latent disentanglement separating inter-event and intra-event variability through hyperbolic curvature without requiring explicit mutual information penalties. The paper provides formal geometric verification via Fréchet mean analysis, a global distortion metric ($\mathcal{D}_{\text{global}} = 2.95$), ANOVA-confirmeed intra-group spread ($p = 0.025$) and geodesic interpolation tests, collectively establishing a methodological benchmark that SeismoALICE-v2



(a) PSA for Le Teil



(b) Goodness of Fit for Stations, Hybrid reconstructions

Figure 5: Analysis of the residuals and adjustment for Le Teil earthquake.

does not currently meet. Their dual-level cohesion–dispersion regularization (Eqs. 14–15 in Wright & Fayaz) directly addresses the latent collapse problem that SeismoALICE-v2 currently manages only through empirical 'latent nullification'. Furthermore, Wright Fayaz demonstrate geodesic interpolation (their Equation 7 and Figure 5) as a rigorous test of semantic continuity in the latent space — the transient spectral peak observed at intermediate interpolation steps is a concrete, physically interpretable indicator of non-trivial latent manifold structure, going well beyond the t-SNE visualizations used in

SeismoALICE-v2.[...]

We sincerely thank reviewer for pointing out this very significant and recent study. We agree that Pointcarré ball, hyperbolic geometry, for latent disentanglement represents a major methodological advance. As mentioned by reviewer, while they focus on response spectra, our task of generating full non-stationary waveforms is inherently more complex due to temporal phase structures. We have added a discussion of this work in Section 8 to contextualize SeismoALICE-v2.