nature portfolio

Peer Review File

A Physics-Informed Foundation Model for Rapid High-Fidelity Structural Response Prediction

Corresponding Author: Professor Ying Zhou

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The Authors propose a quantitatively impressive approach for the obtainment of structural response predictions under the action of static loads and seismic excitation.

The proposed model is based on a very large amount of data, partially numerically generated through FE simulations and partially obtained from real structural responses. The data are processed by means of an innovative ML set of algorithms which core is an autoencoder.

The obtained results are noteworthy, mainly because they show the possibility to create physics-informed AI approaches for structural responses in realistic scenarios.

The originality of the work is related to the way in which the large amount of structural response data is generated and to the ambition of obtaining real-time high fidelity structural response predictions.

The claims are supported by the obtained results, nevertheless there are some points to be clarified, as listed below.

- In various parts of the article it is declared that the proposed model is able to cope with non-linear and even elasto-plastic structural responses. In the detailed description of the methodology the Authors declare to use standard non-linear models for concrete and steel, apparently no complete elasto-plastic models able to correctly describe the irreversible behaviour of elasto-plastic materials are used in the creation of structural responses.

- More generally, the proposed methodology was not trained and/or tested on damage scenarios, hence the claim in the abstract should be mitigated.

- With reference to the SDR module: how the torsional effects in a 3D full model are captured by the lumped one?

- The speed-up obtained with respect to FE simulations is not really exceptional, the claim of "real-time" predictions inserted in the Title should be mitigated.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors present SeisGPT, a physics-informed deep learning model for predicting structural dynamic responses. The model combines a graph neural network encoder with a Transformer decoder, trained on a large synthetic dataset of

~270,000 building models generated with AI and simulated via FEM, and fine-tuned on FEM simulations of 694 real structures. The authors claim ~40,000× speedup over finite element analysis with <5% normalized error.

In my honest view, the claimed novelty is overstated. The core components, namely physics-informed neural networks, graph neural networks for structural systems, transformer architectures for time-series prediction, and surrogate modeling via NNs for FEA, are all well-established. This leaves with only one potential contribution: one of scale and engineering integration rather than methodological innovation. As I will make clear in my review, even this aspect seems to face some challenges.

The authors acknowledge building directly on their prior Phy-Seisformer work (ref 35), and the architectural choices (GQA, RoPE, SwiGLU) are standard components from the contemporary Transformer literature, now widely adopted across sequence modeling applications. Calling this a "foundation model" invokes comparison to GPT-class systems, but the generalization demonstrated here (across building types within a constrained design space) is more modest than that framing suggests.

For a venue like Nature Communications, I would expect either a more fundamental methodological advance or a demonstration of capability that clearly opens new application domains. This paper does neither convincingly.

In addition to the issues of originality and impact, I have some technical concerns:

1. Insufficient sensitivity analysis for temporal resolution. The authors claim the model "robustly accommodates dynamic excitation inputs of varying durations" and test temporal resolution sensitivity across $\Delta t = 0.005$s to 0.04s (Extended Data Table 6). This represents only a factor of 8 range, spanning roughly 0.25× to 2× the training resolution of 0.02s. This is inadequate for establishing robustness. Standard practice for sensitivity analysis typically requires at least an order of magnitude variation in each direction from the nominal value. Structural dynamics applications may require much finer resolution for impact or blast loading ($\Delta t \sim 0.0001$s) or coarser resolution for legacy monitoring systems. The claim of robustness is not supported by the evidence presented.
2. Contradictory evidence regarding synthetic pretraining value. Figure 3a-b presents a puzzling result that the authors do not adequately address. Comparing configurations A1 (SeisGPT-Base, pretrained only), A2 (SeisGPT-Enhanced, pretrained + fine-tuned), and A3 (trained directly on real building data without pretraining) we see that for the displacement prediction, A3 achieves the highest R (0.9435 vs 0.9395), lowest FNMAE (0.0633 vs 0.0656), and lowest FNRMSE (0.0881 vs 0.0920). For acceleration, A3 similarly outperforms A2 across all metrics. If the model trained without the massive synthetic pretraining dataset performs better than the pretrained-then-fine-tuned model, this fundamentally undermines the paper's central premise. The 270,000 synthetic buildings and 10 billion time-steps of pretraining data—presented as a key contribution—appear to provide no benefit or even slight harm to real-world prediction accuracy. The authors note that A3 was "trained for additional epochs to ensure a fair comparison," but this explanation is insufficient. If extended training on real data alone outperforms the elaborate pretraining pipeline, the practical value proposition collapses.
3. Computational comparison methodology. The 40,000× speedup compares GPU-based neural network inference against CPU-based FEM. While speedup is expected, this comparison inflates the advantage. Modern FEM solvers with GPU acceleration would provide a more honest baseline. This is clearly not an apples to apples comparison.
4. Limited validation of synthetic data realism. The diffusion-model pipeline (ArchiFlux, StructFlux, BeamFlux) for generating training buildings is described but not validated. How do we know these synthetic structures are realistic? The claim of "code compliance" is asserted but not demonstrated with any verification study.
5. Exclusion of critical failure regimes. The training data explicitly excludes numerically unstable and collapse-prone structures. For a model positioned for "post-earthquake damage assessment," this is a serious limitation that restricts the model to precisely the scenarios where it is least needed.

Finally, there are also some minor Issues to note: the paper is lengthy and could be more concise, some notation inconsistencies between main text and methods, the "foundation model" terminology, while trendy, may not be appropriate for a domain-specific surrogate model

Recommendation

My final recommendation is to reject. The paper presents competent engineering work but does not meet the novelty threshold for Nature Communications. The contradictory results regarding synthetic pretraining (Figure 3) makes perhaps the main claimed contribution moot, and the sensitivity analyses must be substantially expanded to support the robustness claims. A more specialized venue focused on computational mechanics or structural engineering applications may be more appropriate.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Introduction
This reviewer is actively working on research about the integration of experimental data into digital models for Structural Health Monitoring (SHM) and is aware of the respective advantages and limitations of black-box, white-box, and grey-box approaches in this area, as well as the uncertainty propagated by experimental noise and model discrepancy. Conceptually,

the manuscript represents a significant contribution to the field of structural modelling applied to SHM, particularly in the area of grey-box approaches. However, I have several comments, primarily aimed at improving the clarity of the exposition, the generalizability of the article claims, and the scientific communication of the work. I recommend publication after a minor revision.
This is, in essence, a valuable piece of research, although I could have presented it with less narrative emphasis and in a more direct manner. Below, I report my comments, which I hope will be useful in strengthening the dissemination of this work.

Comment 1: It is unclear to me which topological variety was actually used to generate the synthetic data. The text seems to suggest the use of multi-story frame structures, but it is unclear whether unconventional typologies (e.g., churches, spatial structures, bell towers) are included. This is particularly relevant because the error indicators (FNMAE, FNRMSE) appear to be designed for multi-story structures only. It is therefore legitimate to ask: are the results also applicable with low errors to structures without defined floors (e.g., bell towers, churches)?
Furthermore, if the synthetic dataset does not include sufficient topological variety, the effectiveness of the framework in real-world scenarios could be compromised. This point is particularly important in light of what is stated in the text (e.g., line 107): […] “Additionally, a physics-informed graph neural network (GNN) is employed to encode complex topological and dynamic interactions within building structures, enabling detailed representation of structural variations under diverse external dynamic loading conditions.” […]
Please clarify.

Comment 2: The text is full of promising claims, but these are often presented without direct numerical support in the body of the text. Although the figures and tables are detailed in describing errors, it would be useful, from a scientific communication perspective, to include some key numerical value that support the claims in the main text, to facilitate the understanding even without consulting the figures.

Comment 3: Please clarify in the text which damage models were used to simulate masonry structures. Since this kind of systems have a wide response behaviour with respect to these models. Were elastoplastic models also used in this case? Or was a simplified model adopted? This clarification is important for assessing the validity of predictions in contexts with brittle nonlinearities. It is not the intention of the reviewer to criticize the use of one model over another, if correct, but rather to openly declare the families of damage models used in order to allow the future reader (e.g., expert in masonry modelling) to understand if SeisGPT is the tool that can be used for his purposes or not (e.g., brittle mechanisms).

Minor comments: The variables P_{t,f} (i.e., prediction) is referred in the text (e.g. above or below Eq. (2)), but no explicit definition is provided. It is suggested to clarify the meaning of these variable in the body of the text, to avoid interpretative ambiguities, as done for O_{t,f}. Finally, some typos in the text are highlighted: Lines 231, 236 and 237: the word “Figure” is repeated twice (“Figure Figure”) or the figures are called out without numbers (e.g. “see Figure” instead of “see Figure XX”).

(Remarks on code availability)
Not applicable

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all questions raised by the reviewers and have revised the manuscript accordingly.

This reviewer now believes the manuscript is publishable in Nature Communications.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I appreciate the effort the authors put in reviewing the article and addressing my concerns. Such a massive review constitutes almost a new manuscript. Unfortunately, I still think that the content of the manuscript, while it might be technically correct, does not posses enough novelty for a journal such as this one (for the reasons already stated in my original review). Thus, unfortunately, I cannot this for publication in this venue. I sincerely believe this is better suited for a specialized journal.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have responded satisfactorily to the comments of this reviewer, so the work is accepted for publication.

(Remarks on code availability)

Response to the Reviewers of the manuscript entitled "A Physics-Informed Foundation Model for Rapid High-Fidelity Structural Response Prediction"

We sincerely thank you for your careful and constructive assessment of our manuscript. In preparing this revision, we focused directly on the two principal concerns highlighted in the editorial decision letter. First, we substantially moderated our claims to align strictly with the evidential scope of the study. We now explicitly state that the transferability of structural response prediction across structures and systems is still limited, thereby rigorously defining our validated domain. Second, we resolved the concern regarding the exclusion of collapse-prone structures. To address this, we introduced a dedicated Incremental Dynamic Analysis based failure-adjacent evaluation. This rigorous new benchmark assesses the model on the converged nonlinear branch up to the onset of instability, fully incorporating near-collapse response points and drift-based damage-state intensities.

Beyond addressing these critical scope limitations, we materially strengthened the methodological and empirical foundation of our work in response to your rigorous feedback. Most notably, we introduce a fundamental algorithmic breakthrough in our core architecture. We developed the Spectral Duhamel-Green Mixer, designated as the SDG-Mixer, to serve as a new structural-dynamics-grounded decoder backbone. This mechanism replaces standard generic token interaction with a dynamics-governed propagation operator tailored specifically to structural response prediction. To substantiate this methodological innovation, we added a controlled ablation study that demonstrates its substantial improvements in both predictive accuracy and computational efficiency.

Furthermore, we conducted extensive supplementary experiments to resolve the remaining key technical concerns. We systematically broadened and reinterpreted the temporal-resolution sensitivity analysis. We provided a fully transparent breakdown of the path-dependent inelastic constitutive formulations used in our numerical simulations. Finally, we implemented a comprehensive verification framework that explicitly demonstrates the structural realism and code compliance of our synthetic

dataset.

Collectively, these extensive algorithmic enhancements and empirical additions present our study in a significantly more rigorous, precisely delimited, and strongly supported form. We also confirm full compliance with all Nature Communications editorial policies regarding data availability, code sharing, and reporting standards. All changes are highlighted in blue throughout the revised manuscript and Supplementary Information.

Reviewer #1 (Remarks to the Author):

The Authors propose a quantitatively impressive approach for the obtainment of structural response predictions under the action of static loads and seismic excitation. The proposed model is based on a very large amount of data, partially numerically generated through FE simulations and partially obtained from real structural responses. The data are processed by means of an innovative ML set of algorithms which core is an autoencoder. The obtained results are noteworthy, mainly because they show the possibility to create physics-informed AI approaches for structural responses in realistic scenarios. The originality of the work is related to the way in which the large amount of structural response data is generated and to the ambition of obtaining real-time high fidelity structural response predictions. The claims are supported by the obtained results, nevertheless there are some points to be clarified, as listed below.

1. In various parts of the article it is declared that the proposed model is able to cope with non-linear and even elasto-plastic structural responses. In the detailed description of the methodology the Authors declare to use standard non-linear models for concrete and steel, apparently no complete elasto-plastic models able to correctly describe the irreversible behaviour of elasto-plastic materials are used in the creation of structural responses.

Response 1:

We sincerely thank you for highlighting this important terminological distinction. We agree that our previous manuscript did not sufficiently clarify the specific level of elastoplastic constitutive modelling employed. To address your valuable feedback, we have revised the manuscript and Supplementary Information to make the representation of irreversible material behaviour fully explicit.

Elastoplastic modelling in structural dynamics can generally be categorized into three distinct levels. The first level comprises full continuum plasticity formulations defined through three-dimensional yield surfaces, flow rules, and hardening laws. The second level consists of member-level or section-level cyclic inelastic constitutive models utilized in macroscopic nonlinear structural analysis. The third level involves simplified nonlinear idealizations such as concentrated plastic hinges with idealized

backbone curves. The finite element database developed in this study specifically belongs to the second category. It was generated using path-dependent nonlinear inelastic constitutive formulations that explicitly admit irreversible cyclic structural responses, which are standard and rigorously validated in the nonlinear time-history analysis of building structures.

To correctly describe the irreversible behaviour of elastoplastic materials under seismic loads, we implemented advanced uniaxial material models within a fibre-section framework. For reinforced concrete frame members, the formulation employs the Concrete02 model for concrete and the Steel02 model for reinforcing steel. The Concrete02 model captures not only the nonlinear compressive yielding but also the post-peak softening, tension stiffening, and critical unloading and reloading stiffness degradation indicative of cumulative damage. Simultaneously, the Steel02 model accounts for the Bauschinger effect, kinematic hardening, and the permanent accumulation of irreversible plastic strains under cyclic loading. For reinforced concrete wall systems, the constitutive chain integrates these concrete and reinforcing steel cyclic laws within a macroscopic wall panel formulation. Collectively, these models comprehensively represent yielding, hysteresis, stiffness degradation, and residual inelastic deformation at both the component and system scales. Therefore, the structural responses in our database are fundamentally governed by irreversible elastoplastic and inelastic mechanisms, even though they are not derived from a single full three-dimensional continuum plasticity law for every component.

To make this constitutive scope completely transparent and directly address your concern, we have substantially strengthened the revised Supplementary Information. The revised text now states explicitly that the database relies on path-dependent nonlinear inelastic formulations, and it details the specific mechanisms through which the concrete, steel, and wall panel models generate irreversible response histories. We have also added supplementary constitutive illustrations and an updated component-to-model mapping table. This allows readers to directly verify the specific irreversible behaviours captured by each mathematical formulation.

Changes in the revised manuscript and Supplementary Information: We

revised the Methods subsection “Numerical module: Dataset acquisition and processing”, paragraph 9; and Supplementary Information Section 2.2, paragraphs 1 – 6. We also updated Supplementary Table 1 and the supplementary constitutive-behaviour figures associated with Section 2.2 (Supplementary Figures 4 and 5). All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

2. More generally, the proposed methodology was not trained and/or tested on damage scenarios, hence the claim in the abstract should be mitigated.

Response 2:

We thank you for raising this point. In the revised manuscript, we now state the scope of this claim more precisely.

SeisGPT was not trained as an explicit damage-classification model, and the revised text no longer suggests supervision by categorical damage labels. However, the training data were also not limited to undamaged or linear-elastic states. The model was pretrained on a large database of nonlinear finite-element response histories generated with path-dependent inelastic constitutive formulations. These trajectories already contain the mechanical consequences of structural deterioration under strong loading, including yielding, hysteresis, stiffness degradation or softening, and residual inelastic deformation where applicable. Accordingly, SeisGPT was trained on nonlinear structural responses that include damage-related inelastic evolution, even though it was not trained on manually annotated damage classes.

To address your concern more directly at the downstream engineering level, we added a dedicated IDA-based benchmark in the revised manuscript and Supplementary Information. In this benchmark, SeisGPT is evaluated not as a separate damage classifier, but as a surrogate for the converged nonlinear response branch used in standard earthquake-engineering assessment. We then derive drift-based limit-state intensities from the SeisGPT-predicted response envelopes using exactly the same response-to-MIDR-to-damage-state procedure applied to the FEM reference analyses.

Across the four Hazus-based damage states (LS1 – LS4), the correspondence between SeisGPT predictions and FEM results remains strong, with Pearson correlation coefficients of 0.967, 0.965, 0.958, and 0.960, respectively, and median absolute percentage errors of 8.872%, 12.035%, 12.952%, and 11.862%.

We therefore revised the manuscript in two complementary ways. First, we narrowed the abstract-level wording so that it no longer implies explicit damage-label supervision. Second, we now support the damage-related relevance of SeisGPT with an explicit IDA-based damage-state benchmark. The revised claim is therefore more precise: SeisGPT is a structural-response model trained on nonlinear inelastic response histories and validated for response-informed damage-state assessment, rather than a separately supervised end-to-end damage classifier.

Changes in the revised manuscript and Supplementary Information: We revised the Abstract; the Results subsection beginning "IDA-based evaluation in near-collapse and damage-state regimes", paragraphs 1 – 2; and Supplementary Information Sections 4.5.1, paragraphs 1 – 3, 4.5.5, paragraphs 3 – 4, and 4.5.6, paragraphs 1 – 9. We also added Extended Data Figure 6 and Extended Data Table 10. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

3. With reference to the SDR module: how the torsional effects in a 3D full model are captured by the lumped one?

Response 3:

We thank you for raising this important point. In the revised manuscript, we now clarify the role of the SDR module much more explicitly.

The SDR module is not intended to be a fully coupled reduced-order model that explicitly retains storey-level torsional degrees of freedom. In the present implementation, it is constructed separately for the X and Y directions and serves as a direction-specific translational prior. It therefore does not explicitly evolve rotational coordinates, torsional inertia, or cross-direction coupling blocks in the reduced state.

At the same time, the SDR prior is not extracted from an idealized planar shear-building model. It is derived from the parent three-dimensional FE system. Because the

directional operators are identified from the displacement field of the full 3D structure, their effective translational stiffness already reflects the influence of plan irregularity and torsion-related flexibility that are induced in the parent model under directional loading. In this sense, torsional effects are not explicitly resolved within the reduced coordinates, but they are not absent from the reduced operators either.

Most importantly, SeisGPT is not constrained to reproduce the SDR approximation itself. The SDR prior is used only as a coarse mechanics-informed input. The final model is trained against the full nonlinear time-history responses generated by the parent 3D OpenSees model, so the learned mapping is supervised by the complete three-dimensional reference behaviour. We have revised both the main text and the Supplementary Information to make this division of roles explicit: the SDR module provides a low-order directional prior, whereas the final network prediction is learned from the full 3D FE response.

Changes in the revised manuscript and Supplementary Information: We revised the Methods subsection “Simplified dynamic response (SDR) module”, paragraph 1; and Supplementary Information Section 3.3.1, paragraphs 1 – 7. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

4. The speed-up obtained with respect to FE simulations is not really exceptional, the claim of "real-time" predictions inserted in the Title should be mitigated.

Response 4:

We thank you for this careful comment. In the revised manuscript, we have moderated the wording accordingly.

We agree that the term real-time, especially at the title level, can be read too broadly. We therefore revised the title and related wording throughout the manuscript to use more precise formulations such as rapid or low-latency where appropriate. Our intention in the revision is to keep the efficiency claim strong, while aligning it more closely with the actual evidential scope of the benchmark.

At the same time, the revised benchmark still supports a substantial deployment-

level efficiency advantage. In the updated comparison, SeisGPT inference was measured on a single NVIDIA A800 using standard PyTorch bf16 at batch size = 1, without ONNX Runtime, TensorRT, or kernel-level optimization, whereas nonlinear FEA was benchmarked using the end-to-end OpenSees workflows used in this study. Under this conservative setting, SeisGPT requires 392.43 ms per 1,000 time steps, whereas nonlinear FEA requires 2,339.99 – 23,833.51 s on CPU and 2,166.65 – 14,621.79 s in the GPU-accelerated OpenSees/CuSP workflow. The revised manuscript therefore frames the central conclusion as a substantial practical latency reduction relative to nonlinear FEA, rather than as an unqualified universal real-time claim.

Changes in the revised manuscript and Supplementary Information: We revised the Title and the Abstract; the Results paragraph beginning “Computational performance and cost”, paragraph 1; and the Discussion, paragraph 1. We also updated Extended Data Table 5. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

Reviewer #2 (Remarks to the Author):

The authors present SeisGPT, a physics-informed deep learning model for predicting structural dynamic responses. The model combines a graph neural network encoder with a Transformer decoder, trained on a large synthetic dataset of ~270,000 building models generated with AI and simulated via FEM, and fine-tuned on FEM simulations of 694 real structures. The authors claim ~40,000× speedup over finite element analysis with <5% normalized error.

1. In my honest view, the claimed novelty is overstated. The core components, namely physics-informed neural networks, graph neural networks for structural systems, transformer architectures for time-series prediction, and surrogate modeling via NNs for FEA, are all well-established. This leaves with only one potential contribution: one of scale and engineering integration rather than methodological innovation. As I will make clear in my review, even this aspect seems to face some challenges.

The authors acknowledge building directly on their prior Phy-Seisformer work (ref 35), and the architectural choices (GQA, RoPE, SwiGLU) are standard components from the contemporary Transformer literature, now widely adopted across sequence modeling applications. Calling this a "foundation model" invokes comparison to GPT-class systems, but the generalization demonstrated here (across building types within a constrained design space) is more modest than that framing suggests.

For a venue like Nature Communications, I would expect either a more fundamental methodological advance or a demonstration of capability that clearly opens new application domains. This paper does neither convincingly.

Response 1:

We thank you for this important comment. We agree that, if the contribution were framed primarily as a combination of physics-informed learning, graph-based structural encoding, and contemporary sequence-modelling components, the methodological advance would not have been articulated with sufficient precision. In revising the manuscript, we therefore reconsidered the technical core of the model from the perspective raised in your comment, namely whether the principal contribution lies only in scale and engineering integration, or whether a genuinely problem-specific

computational mechanism is introduced.

Following this reconsideration, we revised the manuscript so that the claimed methodological contribution is no longer associated with standard architectural ingredients such as GQA, RoPE, or SwiGLU, which we agree are established components in the contemporary sequence-modelling literature. Instead, the revised manuscript identifies the principal advance more precisely as the introduction of a new decoder backbone, the Spectral Duhamel – Green (SDG) Mixer, whose central purpose is to replace generic token interaction with a dynamics-governed propagation operator tailored to structural response prediction. In this formulation, the relevant question is not whether individual high-level modules, viewed separately, have precedents in the literature, but whether the dominant internal mixing mechanism of the network has been reformulated in a manner that is specific to the governing physics of the target problem.

The motivation for this redesign is that long-range dependencies in structural response histories are not arbitrary sequence correlations. They are governed by mass – stiffness coupling, modal decomposition, damping evolution, and impulse-response propagation. For this reason, after considering the issue you raised, we introduced these structural-dynamics ingredients directly into the decoder computation. In the revised architecture, the decoder constructs a sample-specific mass-normalized stiffness operator, projects latent representations into a differentiable modal basis, and performs spectral propagation through a Green-function-inspired transfer process that captures phase evolution and damping decay. A strictly bounded rational spectral correction is further introduced to account for departures from idealized linear modal superposition under complex nonlinear response. The methodological point we now emphasize is therefore not the use of a generic Transformer-style decoder supplemented by engineering priors, but the replacement of the generic global mixing mechanism itself by a structure-dependent dynamics operator embedded within the network backbone.

To examine this point directly, we also added a controlled decoder-backbone ablation in which all other parts of the SeisGPT pipeline are held fixed and only the decoder is changed. In this comparison, the baseline decoder already incorporates strong contemporary design choices, including RMSNorm, SwiGLU, RoPE, and

grouped-/multi-query attention, so that these established components are treated as baseline engineering practice rather than as sources of claimed novelty. Under this matched protocol, the SDG-Mixer decoder improves the Pearson correlation coefficient from 0.9320 to 0.9423, reduces displacement FNMAE from 0.0811 to 0.0639, and reduces FNRMSE from 0.1114 to 0.0888 on the real-world building test set after fine-tuning. At the same time, total parameters decrease from 1764.08 M to 570.49 M, FLOPs per inference decrease from 761.07 GFLOPs to 185.12 GFLOPs, and throughput increases from 54.47 to 127.41 samples/s. We therefore revised the manuscript so that the methodological claim is anchored specifically to this decoder-level reformulation and its measured accuracy – efficiency benefit, rather than to standard components or to scale alone.

We also carefully revised the terminology surrounding the model's name and the term "foundation model". In the revised manuscript, SeisGPT is retained as the name of the overall pretrained model family for structural-response prediction, whereas the technical description now makes explicit that the present decoder backbone is the newly introduced SDG-Mixer. The retained family name is intended to preserve continuity across model versions and downstream variants, not to imply that the current architecture should be interpreted as a conventional GPT-style Transformer. For the same reason, we have narrowed the use of the term "foundation model" and now employ it only in a domain-bounded sense, namely as a reusable pretrained base model within the validated scope of floor-defined multistorey structural-response tasks. It is not intended to imply unrestricted GPT-class generalization across arbitrary scientific domains or structural morphologies. This revised framing, in our view, is more precise and more consistent with the actual evidential scope of the work.

Overall, your comment led us to revise both the technical presentation and the scope of the claim. The revised manuscript now distinguishes more clearly between established implementation choices and the methodological component that is specific to this work, namely the introduction of a structural-dynamics-grounded decoder backbone that embeds modal propagation and Green-function-inspired spectral evolution within the network computation. We are grateful for this comment, because

it prompted a more rigorous and more appropriately delimited presentation of the contribution.

Changes in the revised manuscript and Supplementary Information: We revised the Abstract and Introduction to refine the framing of SeisGPT and the domain-bounded use of the term foundation model; substantially revised the main-text section "Structural response prediction with SeisGPT" to present the SDG-Mixer as the new decoder backbone and attention replacement; and revised the Methods subsection "SeisGPT core deep learning model" to describe the mechanics-consistent spectral operator, modal propagation, and bounded rational spectral correction. We also updated Figure 1 and Figure 7 to reflect the revised architecture, and added a controlled decoder-backbone ablation in Supplementary Information Section 4.4 (including Supplementary Table 4) to isolate the contribution of the SDG-Mixer under a matched protocol. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

2. Insufficient sensitivity analysis for temporal resolution. The authors claim the model "robustly accommodates dynamic excitation inputs of varying durations" and test temporal resolution sensitivity across $\Delta t = 0.005$s to 0.04s (Extended Data Table 6). This represents only a factor of 8 range, spanning roughly 0.25× to 2× the training resolution of 0.02s. This is inadequate for establishing robustness. Standard practice for sensitivity analysis typically requires at least an order of magnitude variation in each direction from the nominal value. Structural dynamics applications may require much finer resolution for impact or blast loading ($\Delta t \sim 0.0001$s) or coarser resolution for legacy monitoring systems. The claim of robustness is not supported by the evidence presented.

Response 2:

We sincerely thank you for this insightful and critical comment. We fully agree with your assessment that our previous assertions regarding robustness were overstated and lacked sufficient empirical support. The temporal resolution experiment should not have been framed as evidence that the model can universally accommodate arbitrary

excitation inputs across vastly different sampling intervals.

Upon reflecting on your critique, we realize that the precise intent of this analysis was poorly articulated in our initial submission. This section was never intended to establish universal temporal applicability, nor to imply that SeisGPT would perform uniformly across all possible Δt regimes. Rather, its objective was strictly diagnostic. It was designed as a boundary characterization study to systematically identify the operational thresholds of the model when the inference resolution deviates from the training baseline.

Guided by your expert recommendations, we have comprehensively revised both the experimental scope and our interpretative claims. First, we significantly expanded the zero-shot temporal resolution sweep to encompass intervals of 0.001, 0.002, 0.005, 0.010, 0.020, 0.030, 0.040, 0.080, and 0.100 seconds. This expanded evaluation was conducted entirely without retraining or recalibration. As you rightfully pointed out, standard sensitivity analyses require a broader order of magnitude variation, and this comprehensive sweep now appropriately maps the operating envelope of our model, which was trained at a singular nominal resolution of 0.020 seconds.

Second, we have meticulously refined the manuscript narrative to align with this more rigorous and narrowed scope. We no longer present these findings as proof of generic robustness. Instead, we frame them as a zero-shot characterization of model behavior under controlled deviations from the training resolution. The results demonstrate that predictive accuracy peaks at the training resolution, maintains stability under moderate deviations, and progressively degrades as the sampling interval diverges, particularly at coarser resolutions. This revised presentation directly addresses your concern by explicitly defining the capability boundaries of the model and showing exactly where performance deterioration becomes substantial.

The revised manuscript clarifies that our current analysis does not imply suitability for such ultra fine regimes without dedicated retraining. When a target application necessitates a substantially finer temporal resolution, generating new data and retraining the model at that specific target resolution remains necessary. We have repositioned this entire section to delineate the effective operating range of the model

and to clarify the strict limits of zero-shot transfer across temporal resolutions. We appreciate your rigorous review, which has prompted a more precise and scientifically sound presentation of our work.

Changes in the revised manuscript and Supplementary Information: We revised the Results paragraph beginning "Sensitivity to temporal resolution (Δt)", paragraph 1, and updated Extended Data Table 6. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

3. Contradictory evidence regarding synthetic pretraining value. Figure 3a-b presents a puzzling result that the authors do not adequately address. Comparing configurations A1 (SeisGPT-Base, pretrained only), A2 (SeisGPT-Enhanced, pretrained + fine-tuned), and A3 (trained directly on real building data without pretraining) we see that for the displacement prediction, A3 achieves the highest R (0.9435 vs 0.9395), lowest FNMAE (0.0633 vs 0.0656), and lowest FNRMSE (0.0881 vs 0.0920). For acceleration, A3 similarly outperforms A2 across all metrics. If the model trained without the massive synthetic pretraining dataset performs better than the pretrained-then-fine-tuned model, this fundamentally undermines the paper's central premise. The 270,000 synthetic buildings and 10 billion time-steps of pretraining data—presented as a key contribution—appear to provide no benefit or even slight harm to real-world prediction accuracy. The authors note that A3 was "trained for additional epochs to ensure a fair comparison," but this explanation is insufficient. If extended training on real data alone outperforms the elaborate pretraining pipeline, the practical value proposition collapses.

Response 3:

We thank you for flagging this point. You identified a genuinely serious problem in the previous version. The apparent contradiction did not arise from the underlying experimental outcome, but from an erroneous labeling of the three compared configurations in the text and associated displays. In the previous manuscript, the experimental identities of A1/A2/A3 were mis-assigned in the presentation of Figure 3a – b and the corresponding summary table, which in turn led to an inverted interpretation of the comparison. We fully agree that, as presented previously, this error

undermined the clarity and credibility of the argument. In the revised manuscript, we have therefore re-checked and fully corrected the experiment definitions, figure labeling, table labeling, and the corresponding textual interpretation so that all components are now consistent.

The three configurations are now defined consistently throughout the revised manuscript as follows: A1, a model trained from scratch using only the limited real-world building response dataset, without synthetic pretraining; A2, SeisGPT-Base pretrained on the large-scale synthetic building dataset and evaluated on the real-world building dataset without further adaptation; and A3, SeisGPT-Enhanced, obtained by fine-tuning A2 on the real-world building dataset. With these corrected definitions, the trend is no longer contradictory. Instead, it is clear and monotonic: synthetic pretraining substantially improves performance relative to real-only training from scratch, and subsequent fine-tuning on real buildings provides a further gain. Specifically, displacement FNMAE decreases from 0.1189 in A1 to 0.0666 in A2 and further to 0.0639 in A3, while the corresponding Pearson correlation increases from 0.8406 to 0.9359 and then to 0.9423. For acceleration, FNMAE decreases from 0.0524 to 0.0227 and then to 0.0195, while Pearson correlation increases from 0.9079 to 0.9781 and then to 0.9829. The corrected results therefore support, rather than weaken, the central premise that large-scale synthetic pretraining is beneficial and that real-data fine-tuning yields an additional improvement.

We also revised the explanation of the phrase “trained for additional epochs to ensure a fair comparison”. This refers only to A1, the scratch-trained real-only baseline. Because the real-world dataset is much smaller than the synthetic pretraining corpus, using the same nominal epoch count for A1 would have resulted in far fewer parameter-update iterations than in the synthetic pretraining stage, thereby producing an artificially under-trained scratch baseline. In the revised manuscript, we now explain this point explicitly: the additional epochs for A1 were introduced only to bring the total number of optimization steps into a more comparable range, so that A1 would serve as a stronger and more informative baseline. This adjustment therefore makes the comparison more conservative, not less. We appreciate that this issue was important to

raise, because in the previous version the labeling error obscured the actual experimental conclusion. In the revised manuscript, that error has been fully corrected, and the corresponding text, figure caption, and table caption have been harmonized so that the contribution of synthetic pretraining and subsequent fine-tuning is now presented accurately and unambiguously.

Changes in the revised manuscript and Supplementary Information: We revised the Results section "Fine-tuning SeisGPT-Enhanced with real-world building response dataset", paragraph 1; the Figure 3 caption; and Extended Data Table 3 and its caption. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

4. Computational comparison methodology. The 40,000× speedup compares GPU-based neural network inference against CPU-based FEM. While speedup is expected, this comparison inflates the advantage. Modern FEM solvers with GPU acceleration would provide a more honest baseline. This is clearly not an apples to apples comparison.

Reponse 4:

We thank you for this important comment. We agree that a comparison between GPU-based neural-network inference and CPU-based finite-element analysis should not be presented as a hardware-neutral lower bound against all possible FEM implementations. In the revised manuscript, we therefore narrowed the claim and now frame the reported ratios explicitly as deployment-level end-to-end runtime comparisons under the specific software and hardware stacks used in this study.

We also believe it is important to clarify why the original FEM baseline was implemented in CPU-based OpenSees. This baseline was not chosen because it is artificially weak. Rather, it is the same nonlinear OpenSees workflow used throughout the study to generate the finite-element response database and to conduct all reference nonlinear time-history analyses. In addition, the CPU-side implementation already employed an accelerated sparse solver (UmfPack), so the original benchmark should be understood as a comparison against a streamlined and practically relevant nonlinear

structural-analysis workflow, rather than against a generic or intentionally slow FEM setup. For this reason, the original CPU-based comparison was intended as a realistic end-to-end deployment benchmark, not as a solver-level microbenchmark.

To address your concern more directly, we added an additional supplementary benchmark using the historical OpenSees/CuSP GPU sparse-solver pathway on NVIDIA A800 hardware. This additional benchmark is important, but it should also be interpreted carefully. In the OpenSees/CuSP workflow, the GPU acceleration applies primarily to the sparse linear-system solution stage within each nonlinear analysis step, whereas a substantial fraction of the total wall-clock time remains outside that stage, including stiffness-related assembly and update operations, element state updates, constitutive integration, convergence checks, adaptive step fallback, and recorder I/O. The supplementary GPU-FEM comparison is therefore not a fully GPU-resident nonlinear FEM pipeline, but a hybrid CPU+GPU workflow in which only part of the overall nonlinear analysis is materially accelerated. This is precisely why the end-to-end gain remains modest for the 1 – 30-storey nonlinear building analyses considered here. In fact, using GPUs for traditional FEM calculations is fundamentally a mismatched operation that rarely yields massive computational speedups. From their inception, GPUs were not designed for the highly sequential, sparse matrix solving inherent to finite element analysis; rather, their architecture is intended for the massive, dense parallel computations characteristic of deep learning models.

Under this more stringent benchmark, however, the central conclusion remains unchanged. As now reported in Extended Data Table 5, SeisGPT requires 392.43 ms per 1,000 time steps, whereas nonlinear FEA requires 2,339.99 – 23,833.51 s in the CPU OpenSees workflow and 2,166.65 – 14,621.79 s in the GPU-accelerated OpenSees/CuSP workflow. Accordingly, SeisGPT retains an orders-of-magnitude wall-clock advantage over the CPU baseline and remains substantially faster even relative to the supplemented GPU-accelerated OpenSees/CuSP benchmark. At the same time, we have revised the wording throughout the manuscript so that these ratios are no longer presented as universal lower bounds against all conceivable FEM solvers, but rather as realistic end-to-end comparisons under the exact computational settings used

in this work.

We therefore revised the manuscript in two complementary ways. First, we moderated the title-level and text-level wording by replacing broad uses of real-time with more precise expressions such as rapid or low-latency, where appropriate. Second, we expanded the computational benchmark to include the OpenSees/CuSP supplementary comparison and clarified why GPU acceleration in this setting does not remove the dominant CPU-bound costs of practical nonlinear time-history analysis. We believe the revised manuscript now presents a more precise and more rigorous efficiency claim: SeisGPT does not derive its practical advantage from comparison to an artificially weak FEM baseline, but from a genuine deployment-level reduction in end-to-end latency relative to the nonlinear structural-analysis workflows that are actually used in this study.

Changes in the revised manuscript and Supplementary Information: We revised the Results paragraph beginning “Computational performance and cost”, paragraph 1; the Discussion, paragraph 1; the Title and related efficiency wording in the Abstract and main text; and Extended Data Table 5. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

5. Limited validation of synthetic data realism. The diffusion-model pipeline (ArchiFlux, StructFlux, BeamFlux) for generating training buildings is described but not validated. How do we know these synthetic structures are realistic? The claim of "code compliance" is asserted but not demonstrated with any verification study.

Response 5:

We sincerely thank you for raising this highly insightful point. We completely agree that relying solely on procedural descriptions is insufficient to demonstrate the realism of the synthetic dataset. To address your valid concern, we have introduced a comprehensive validation framework in the revised manuscript to systematically substantiate both the structural plausibility and the code compliance of the generated models.

First, we provide direct qualitative evidence of structural realism. Structural

realism can be evaluated through multiple dimensions, such as material property assignment, component detailing, and overall structural coherence. In this study, we focus on demonstrating the structural coherence of the generated models. The revised Supplementary Information now includes representative examples of the accepted synthetic buildings. We detail their plan layouts alongside their corresponding three-dimensional finite element realizations. These examples were selected to capture significant variations in storey count, plan elongation, wall distribution, and lateral-force-resisting systems. They demonstrate that the algorithm generates structurally coherent three-dimensional assemblies composed of beams, columns, slabs, and walls. This results in realistic storey organization and plausible plan-level irregularities within the domain of floor-defined multistorey buildings.

Second, to address your specific concern regarding code compliance, we have implemented an explicit seismic verification study. Code compliance for earthquake resistance can generally be verified through various analytical approaches, including equivalent static analysis, nonlinear static pushover analysis, and nonlinear time-history analysis. To provide the most rigorous verification, we chose to conduct nonlinear time-history analysis using the three-dimensional OpenSees framework detailed in the revised Methods. The adequacy of a lateral-force-resisting system is fundamentally reflected in its deformation demands. Consequently, we adopted the governing inter-storey drift ratio as our principal engineering metric to ensure the structures perform within prescribed seismic limits. Following gravity load application, bidirectional horizontal seismic excitations representative of the target seismic hazard were independently imposed on each accepted synthetic building. We computed the governing inter-storey drift across all storeys and normalized it against the corresponding limit prescribed by the GB 50011 design code. As detailed in the revised Supplementary Information, this resulting normalized governing drift ratio is strictly less than 1.0 for all retained buildings. We hope this explicit response-level verification grounded in rigorous nonlinear dynamic analysis adequately addresses your concern regarding our previous algorithmic assumptions.

Third, we substantiate the realism of the dataset statistically. The statistical realism

of structural datasets can be assessed across various parameter domains, such as construction material volumes, economic indicators, or engineering descriptors. For our validation, we evaluate the accepted synthetic dataset against the empirical real-world building dataset across two complementary descriptor spaces. The first is a geometry-dynamics space defined by storey count and fundamental translational period T1. The second is a plan-topology space defined by plan aspect ratio and wall density. Utilizing both scalar descriptor coverage and joint-space highest-density-region metrics, we demonstrate that the synthetic dataset substantially aligns with the empirical distributions of real-world buildings. We believe this physically plausible, code-consistent, and statistically anchored coverage of the relevant design space provides strong evidence of overall dataset realism.

Finally, we have carefully bounded our claims in the revised text. We recognize that the transferability of structural response prediction across structures and systems is still limited. Therefore, we explicitly state that the synthetic dataset is validated specifically for floor-defined multistorey buildings. We do not claim an exhaustive representation of all real-world structures or applicability to typologies outside this intended domain. The realism of the dataset is now supported by a convergence of evidence encompassing representative structural topologies, explicit seismic drift verification under dynamic loading, and robust statistical overlap with real-world empirical data.

Changes in the revised manuscript and Supplementary Information: We revised the Methods subsection “Numerical module: Dataset acquisition and processing”, paragraphs 4 – 9; and Supplementary Information Sections 1.4. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

6. Exclusion of critical failure regimes. The training data explicitly excludes numerically unstable and collapse-prone structures. For a model positioned for "post-earthquake damage assessment," this is a serious limitation that restricts the model to precisely the scenarios where it is least needed.

Response 6:

We thank you for raising this important concern. In response, we have substantially revised the manuscript and Supplementary Information to clarify the engineering scope and provide direct validation in the failure-adjacent regime. We address this concern through two complementary approaches: (1) clarifying the definition of "post-earthquake damage assessment" within standard collapse-oriented seismic analysis, and (2) introducing a dedicated Incremental Dynamic Analysis (IDA)-based benchmark to rigorously evaluate SeisGPT in the most demanding regimes defined by conventional nonlinear finite element methods (FEM).

In conventional implicit continuum-based nonlinear FEM, approaching global instability causes the tangent stiffness matrix to become singular or severely ill-conditioned due to combined material deterioration, geometric nonlinearity, and P-Delta amplification effects [1]. Beyond this point, tracking true post-collapse kinematics (e.g., member separation, rigid-body motion, contact, or debris interaction) requires alternative simulation paradigms, such as explicit dynamics or discrete-element methods, which fall outside the standard nonlinear response-history workflows of routine PBEE [2,4]. Standard seismic assessment defines collapse via the failure-adjacent loss of stability on the converged nonlinear branch. This is the exact regime in which our added IDA validation is framed.

Consequently, excluding numerically unstable post-instability trajectories from the training data does not confine SeisGPT to mild or moderate response states. Rather, SeisGPT is trained specifically on converged nonlinear response histories — the physically interpretable data that conventional FEM reliably produces and from which critical engineering demand parameters are derived. In practical collapse analysis, the paramount requirement is accurately reproducing the converged nonlinear response branch up to the onset of instability to capture damage progression, near-collapse demand, and limit-state exceedance.

To directly address your concern, we have introduced a comprehensive IDA-based evaluation focusing on near-collapse and severe damage-state regimes (Supplementary Information Section 4.5). Utilizing 65 real-world buildings subjected to the

standardized suite of 22 FEMA P-695 far-field ground motions [3], we incrementally intensified nonlinear analyses until the FEM models lost convergence or entered the collapse bracket. This benchmark explicitly tests SeisGPT at the extreme edge of the converged IDA branch — the specific domain where practical collapse-oriented assessment occurs.

Within this benchmark, we executed two complementary evaluations. First, we isolated the final converged intensity level immediately preceding instability on each IDA trajectory, intentionally testing the single most demanding converged state before the FEM solution ceases to be well-posed. We then compared the FEM-derived and SeisGPT-derived maximum inter-storey drift ratios (MIDR) at this juncture (Extended Data Figure 6a; detailed extraction protocol in Supplementary Information Section 4.5.5). Second, to align with field-recognized metrics, we evaluated SeisGPT's capacity to reproduce drift-based damage states (LS1 – LS4: Slight, Moderate, Extensive, and Complete) defined by the FEMA Hazus framework [5]. We converted each IDA MIDR sequence into a monotone envelope and extracted the intensity at which each damage threshold was exceeded. This allowed for a rigorous comparison across the full sequence of escalating damage transitions used in fragility-oriented assessment [6,7] (Supplementary Information Section 4.5.6).

To demonstrate the practical implications, we provide both a representative case-level fragility comparison and a population-level concordance analysis (Supplementary Information Section 4.5.7). As illustrated in Extended Data Figure 6b, the SeisGPT-derived fragility curves for a representative building closely track the FEM reference curves across all four damage states. Crucially, Extended Data Figure 6c and Extended Data Table 10 summarize the benchmark-wide agreement for threshold-crossing intensities across the entire building population.

The statistical results confirm that excluding post-instability trajectories does not limit SeisGPT to low-damage conditions. Evaluated within the standard IDA workflow, SeisGPT maintains exceptional consistency with FEM in the highly demanding failure-adjacent regime. Pearson correlation coefficients between FEM- and SeisGPT-derived

limit-state intensities are 0.967, 0.965, 0.958, and 0.960 for LS1 – LS4, respectively. The corresponding median absolute percentage errors (MAPE) are tightly bounded (8.87%, 12.04%, 12.95%, and 11.86%), with the across-building median of building-median MAPE similarly constrained (8.78% – 12.93%). These metrics verify strong agreement not just for lower damage tiers, but for the most severe Hazus-consistent drift-based states considered in structural fragility assessment.

We have accordingly refined the manuscript to ensure our claims are precise, rigorous, and completely aligned with accepted earthquake engineering practice. We clearly state that SeisGPT's validated scope is IDA-based, failure-adjacent response prediction and drift-based damage-state assessment up to the onset of instability. Rather than a retreat from practically relevant domains, this rigorously defines the exact regime in which conventional collapse-oriented assessments are formulated and deployed [2 – 4].

Changes in the revised manuscript and Supplementary Information: We added and revised paragraphs 1 – 2 in the Results subsection "IDA-based evaluation in near-collapse and damage-state regimes"; revised paragraph 3 in the Discussion to clarify the engineering scope of failure-adjacent validation; and added Extended Data Figure 6 and Extended Data Table 10. Furthermore, we added the comprehensive Supplementary Information Section 4.5 (Sections 4.5.1 – 4.5.7), which thoroughly details the engineering rationale, IDA benchmark composition, near-collapse state extraction, Hazus-consistent limit-state definition, threshold-crossing protocol, representative fragility construction, and population-level concordance analysis. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

Reference:

[1] Ibarra, L. F., Medina, R. A. & Krawinkler, H. Hysteretic models that incorporate strength and stiffness deterioration. *Earthq. Eng. Struct. Dyn.* **34**, 1489–1511 (2005).

[2] Vamvatsikos, D. & Cornell, C. A. Incremental dynamic analysis. *Earthq. Eng. Struct. Dyn.* **31**, 491–514 (2002).

[3] Federal Emergency Management Agency. *Quantification of Building Seismic Performance Factors* (FEMA P-695, FEMA, 2009).

[4] Federal Emergency Management Agency. *Seismic Performance Assessment of Buildings, Volume 1 – Methodology*, 2nd edn (FEMA P-58-1, FEMA, 2018).

[5] Federal Emergency Management Agency. *Hazus 6.1 Earthquake Model User Guidance* (FEMA, 2024).

[6] Porter, K., Kennedy, R. & Bachman, R. Creating fragility functions for performance-based earthquake engineering. *Earthq. Spectra* **23**, 471–489 (2007).

[7] Jalayer, F. & Cornell, C. A. *A Technical Framework for Probability-Based Demand and Capacity Factor Design (DCFD) Seismic Formats* (PEER Report 2003-08, Pacific Earthquake Engineering Research Center, 2003).

7. Finally, there are also some minor Issues to note: the paper is lengthy and could be more concise, some notation inconsistencies between main text and methods, the "foundation model" terminology, while trendy, may not be appropriate for a domain-specific surrogate model.

Response 7:

We sincerely thank you for these thoughtful and constructive comments. We completely agree that addressing these presentation-level issues is crucial for enhancing the precision and overall readability of our manuscript. To address your valid concerns, we have revised the manuscript accordingly in three main aspects. First, we substantially streamlined the text to improve conciseness. Second, we fully harmonized the notation and terminology between the main text, Methods, and Supplementary Information. Third, we clarified and narrowed our use of the term foundation model to avoid any implication of unjustified universality.

First, regarding conciseness, we carefully re-edited the entire manuscript to remove repeated conceptual summaries and overly extended explanatory passages. In particular, we streamlined the Introduction, where the motivation, challenge framing, and model positioning had previously been stated with some redundancy. We also compressed the main text architecture overview by retaining only the core design logic

of SeisGPT in the main text, moving non-essential implementation details to the Methods and Supplementary Information. Furthermore, we reduced repetitive quantitative restatements in the Results and Discussion sections and shortened several application-oriented or forward-looking sentences that were not essential to our main claims. We believe the revised version is now significantly more concise while fully preserving the evidentiary basis of the work.

Second, we meticulously reviewed the manuscript and Supplementary Information for notation consistency and internal cross-referencing. Symbols, naming conventions, and related terminology have now been rigorously harmonized across the main text, Methods, and Supplementary Information to ensure a seamless reading experience.

Third, regarding the term foundation model, we highly appreciate your insight. We agree that without a precise definition, this term risks implying a degree of universality that is inappropriate for a model validated strictly within a specific engineering domain. Deep learning models in scientific research can generally be classified into three distinct categories based on their scope and transferability. The first category includes general-purpose foundation models designed for universal applications across entirely disparate scientific disciplines. The second category comprises domain-specific foundation models, which are large-scale, reusable, pre-trained base architectures designed to generalize across a spectrum of downstream applications strictly within a clearly defined scientific domain. The third category consists of narrow, task-specific surrogate models trained exclusively to replace a single, specific numerical simulator.

Our proposed SeisGPT strictly aligns with the second category. It is a large-scale pre-trained foundation model specifically developed for cross-structure seismic response prediction. This designation explicitly avoids any implication of universality across arbitrary structural topologies or unrelated disciplines. Instead, it reflects a technically justified definition within the clearly defined domain of floor-level multistorey structural response. Crucially, this pre-trained architecture can be transferred or adapted to diverse downstream tasks within this domain, including direct response prediction, fine-tuning on real building data, few-shot building-specific

adaptation, and sparse-sensor response reconstruction.

Our terminology aligns with an evolving consensus in high-impact scientific literature, where foundation models are no longer viewed exclusively as cross-domain, general-purpose systems. Increasingly, the designation applies to large pre-trained architectures demonstrating robust transferability within specific scientific disciplines. For instance, as detailed in Reference [1], Cui et al. introduced scGPT as a foundation model for single-cell biology, validating its utility across various downstream tasks via transfer learning. Similarly, as noted in Reference [2], Jiang et al. characterized DeepRETStroke as a domain-specific foundation model for eye and brain connectivity, facilitating applications ranging from silent brain infarction detection to stroke prediction. Grounded in these established precedents, our application of the term domain-specific foundation model is deliberate, precise, and well-supported by current literature.

To prevent any potential overstatement, we have systematically reduced non-essential uses of the term throughout the manuscript. We now retain the term domain-specific foundation model only where it serves a strict definitional purpose. Elsewhere, we employ more localized descriptors, such as pre-trained model, base pre-trained model, pre-trained architecture, or decoder backbone. We believe these revisions substantially enhance terminological rigor while preserving the core thesis of this work, which is that SeisGPT functions as a highly reusable, pre-trained model for advanced structural response prediction.

Changes in the revised manuscript and Supplementary Information: We revised the Abstract; the Introduction, paragraphs 3-4; and the Discussion, paragraphs 1-2. We also harmonized notation throughout the main text, Methods, and Supplementary Information. Furthermore, the entire manuscript has been rigorously edited for conciseness, streamlining the narrative and eliminating non-essential phrasing to improve overall readability. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

Reference:

[1] Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics

using generative AI. **Nat. Methods 21**, 1470–1480 (2024).

[2] Jiang, N. *et al.* A deep learning system for detecting silent brain infarction and predicting stroke risk. **Nat. Biomed. Eng.** (2025).

Reviewer #3 (Remarks to the Author):

This reviewer is actively working on research about the integration of experimental data into digital models for Structural Health Monitoring (SHM) and is aware of the respective advantages and limitations of black-box, white-box, and grey-box approaches in this area, as well as the uncertainty propagated by experimental noise and model discrepancy. Conceptually, the manuscript represents a significant contribution to the field of structural modelling applied to SHM, particularly in the area of grey-box approaches. However, I have several comments, primarily aimed at improving the clarity of the exposition, the generalizability of the article claims, and the scientific communication of the work. I recommend publication after a minor revision.

This is, in essence, a valuable piece of research, although I could have presented it with less narrative emphasis and in a more direct manner. Below, I report my comments, which I hope will be useful in strengthening the dissemination of this work.

1. It is unclear to me which topological variety was actually used to generate the synthetic data. The text seems to suggest the use of multi-story frame structures, but it is unclear whether unconventional typologies (e.g., churches, spatial structures, bell towers) are included. This is particularly relevant because the error indicators (FNMAE, FNRMSE) appear to be designed for multi-story structures only. It is therefore legitimate to ask: are the results also applicable with low errors to structures without defined floors (e.g., bell towers, churches)?

Furthermore, if the synthetic dataset does not include sufficient topological variety, the effectiveness of the framework in real-world scenarios could be compromised. This point is particularly important in light of what is stated in the text (e.g., line 107):

[…] “Additionally, a physics-informed graph neural network (GNN) is employed to encode complex topological and dynamic interactions within building structures, enabling detailed representation of structural variations under diverse external dynamic loading conditions.” […]

Please clarify.

Repsonse 1:

We sincerely thank you for raising this highly insightful point regarding the topological variety of our synthetic dataset. We completely agree that the scope of the training data directly dictates the applicability of the framework, and we deeply appreciate the opportunity to clarify these crucial boundaries.

Structural typologies in civil engineering can generally be categorized into various forms based on their spatial organization. These categories include floor-defined multistorey buildings, continuous spatial structures such as shells and domes, and specialized vertical structures like bell towers and churches. The synthetic dataset generated for this study, and consequently the SeisGPT model trained on it, is exclusively focused on the first category. By this, we mean buildings with explicit storey organization and floor-level structural response quantities, including the frame, frame-shear wall, and shear wall systems studied here. We recognize that the transferability of structural response prediction across structures and systems is still limited. Therefore, our model is not intended for, nor applicable to, structural forms that do not possess clear floor-wise organization.

Regarding the applicability of the results and error indicators, performance metrics in structural dynamics can be defined at different resolutions, such as the global structural level, the floor level, or the individual element level. The error indicators utilized in our study, specifically FNMAE and FNRMSE, are explicitly designed as floor-normalized metrics to evaluate responses at discrete storey levels. Consequently, it is perfectly legitimate for you to question their applicability elsewhere. We confirm that these metrics, along with the current predictive framework, cannot yield low errors or be directly applied to structures without defined floors. We have revised the manuscript to state this scope and limitation explicitly so as not to mislead readers.

Furthermore, to address your valid concern regarding whether the synthetic dataset includes sufficient topological variety to be effective in real-world scenarios, we must clarify the specific role of the Graph Neural Network in our study. Graph Neural Networks can be designed to encode interactions across arbitrary continuous spatial meshes or, alternatively, across structured topological grids. In our framework, the physics-informed Graph Neural Network is employed strictly in the latter capacity. It

is tailored to encode the complex topological variations and dynamic interactions solely within floor-defined systems, capturing features such as varying shear wall layouts, plan irregularities, and vertical stiffness changes across storeys. However, we fully acknowledge that our original phrasing implied an unrestricted generality across arbitrary structural morphologies.

We have therefore rigorously revised the wording describing the Graph Neural Network to ensure it accurately reflects the structural variation represented only within our validated domain. To further substantiate the effectiveness of the framework within this specific domain, we added a dedicated validation framework in the revised Supplementary Information Section 1.4. This includes representative topology visualizations and descriptor-level comparisons against empirical real-world building datasets to prove the topological richness within our intended scope.

Changes in the revised manuscript and Supplementary Information: We revised the Introduction, paragraph 4; the section "Structural response prediction with SeisGPT", paragraph 1; the Methods subsection "Numerical module: Dataset acquisition and processing", paragraphs 3 – 6; the Discussion, paragraph 3; and Supplementary Information Section 1.4. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

2. The text is full of promising claims, but these are often presented without direct numerical support in the body of the text. Although the figures and tables are detailed in describing errors, it would be useful, from a scientific communication perspective, to include some key numerical value that support the claims in the main text, to facilitate the understanding even without consulting the figures.

Response 2:

We sincerely thank you for this suggestion. In the revised manuscript, we systematically strengthened the main text by replacing broad qualitative statements with representative numerical results.

Specifically, we inserted explicit numerical values into the main narrative for the large-scale pretraining results, the real-building fine-tuning results, the zero-shot cross-

system evaluation, the computational benchmark, the temporal-resolution sensitivity analysis, and the pretraining-scale analysis. We made the same change in the reconstruction and discussion sections, so that the principal claims in the revised manuscript are now tied directly to quantitative evidence rather than being stated in mainly qualitative form. Our aim in the revision was not to overload the main text with every metric reported in the figures and tables, but to ensure that the major claims can be evaluated directly from the narrative itself. We believe this substantially improves the scientific communication of the manuscript.

Changes in the revised manuscript and Supplementary Information: We revised the Results sections "Pretraining SeisGPT-Base on the large-scale dataset for structural generalization", paragraph 2; "Fine-tuning SeisGPT-Enhanced with real-world building response dataset", paragraph 1; "Zero-shot prediction across different structural systems", paragraph 2; the Results paragraphs beginning "Computational performance and cost", "Sensitivity to temporal resolution (Δ t)", and "Pretraining data scale"; and the Discussion, paragraphs 1 – 2. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

3. Please clarify in the text which damage models were used to simulate masonry structures. Since this kind of systems have a wide response behaviour with respect to these models. Were elastoplastic models also used in this case? Or was a simplified model adopted? This clarification is important for assessing the validity of predictions in contexts with brittle nonlinearities. It is not the intention of the reviewer to criticize the use of one model over another, if correct, but rather to openly declare the families of damage models used in order to allow the future reader (e.g., expert in masonry modelling) to understand if SeisGPT is the tool that can be used for his purposes or not (e.g., brittle mechanisms).

Response 3:

We sincerely thank you for raising this point. In the revised Supplementary Information, we now define the masonry benchmark much more explicitly.

The masonry models used in this study are not based on a purely elastic

idealization, nor are they represented by a simple bilinear elastoplastic continuum law. Rather, they belong to the family of nonlinear discrete-homogenized macro-scale masonry models governed by a damage-plasticity constitutive formulation. In this framework, the structural response is represented through macro-regions connected by nonlinear deformable links, and the constitutive behaviour of these links admits both irreversible inelastic strain and progressive damage-induced stiffness degradation.

This modelling choice is important because it allows the benchmark to represent the dominant brittle nonlinear mechanisms relevant at building scale, including tensile cracking, compressive crushing, post-peak softening, and cyclic strength/stiffness degradation. To make this scope fully transparent for readers working specifically in masonry modelling, we substantially expanded the revised Supplementary Information and added a schematic supplementary figure illustrating the main constitutive branches of the equivalent nonlinear macro-links.

Changes in the revised manuscript and Supplementary Information: We revised Supplementary Information Section 2.7, paragraphs 2 – 9, and added Supplementary Figure 8. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.

4. Minor comments: The variables P_{t,f} (i.e., prediction) is referred in the text (e.g. above or below Eq. (2)), but no explicit definition is provided. It is suggested to clarify the meaning of these variable in the body of the text, to avoid interpretative ambiguities, as done for O_{t,f}. Finally, some typos in the text are highlighted: Lines 231, 236 and 237: the word "Figure" is repeated twice ("Figure Figure") or the figures are called out without numbers (e.g. "see Figure" instead of "see Figure XX").

Response 4:

We sincerely thank you for this careful reading. In the revised manuscript, we now define $P_{t,f}$ explicitly in parallel with $O_{t,f}$, and we further clarify the normalized predicted and observed quantities immediately around Eq. (2) to remove any possible ambiguity.

We also re-checked the manuscript and Supplementary Information for

typographical inconsistencies and incomplete or duplicated figure callouts. The specific issues you noted have been corrected, and the remaining figure references were standardized throughout the revised submission.

Changes in the revised manuscript and Supplementary Information: We revised the Results, paragraphs 2 – 3 (immediately following Eq. (2)), and corrected figure-callout and typographical inconsistencies throughout the main text and Supplementary Information. All revised text is highlighted in blue in the revised manuscript and Supplementary Information.