GEOMETRY-AWARE ERROR DIAGNOSIS FOR DATA ACQUISITION AND TRAINING STRATEGY SELECTION IN GRAPH NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

High-fidelity finite element analysis (FEA) provides reliable structural-response estimates, but each new surrogate-training specimen requires additional FE simulation effort. When a GNN stress surrogate fails in a specific geometry regime, a practitioner can either acquire more FE data or expose more informative geometry to the model. We study this choice for von Mises stress prediction on perforated plates using a controlled 2×2 experiment that crosses targeted FE data enrichment with explicit geometric conditioning. Across two seeds, node-level RMSE changes from 0.0889 MPa at baseline to 0.0862 MPa after adding 315 targeted training specimens, whereas conditioning the same base training set reduces RMSE to 0.0425 MPa. The observed conditioning effect is regime-dependent and is largest for narrow ligaments: there, the conditioned model trained on 1,605 specimens outperforms the unconditioned model trained on 1,920 specimens in both seeds. Ablation further shows that most of the gain is associated with node-wise relational geometry rather than specimen-level scalars. These results support checking representation sufficiency before committing additional FE budget.

1 INTRODUCTION

Surrogate models can be accurate on average yet fail systematically in localized regions of the input space. Such failures suggest two distinct interventions: acquire more examples from the failing region, or change the representation so that prediction-relevant information is easier to use. When labels are expensive, this becomes a resource-allocation problem rather than a purely predictive one.

This distinction is especially important for FEA-based structural surrogates. Every additional training specimen requires additional FE simulation effort. For a refined three-dimensional mesh this computational cost can be substantial, so enlarging the training set scales directly with simulation time. Geometric descriptors, in contrast, are available directly from CAD and mesh data. Thus, targeted data enrichment and geometric conditioning have very different costs, and adding data is not automatically the most efficient response to localized error.

The problem we consider constrains the choice of surrogate in two ways. Stress prediction requires a spatially resolved nodal field rather than a single scalar per design, because where the peak occurs matters as much as how large it is. The discretization itself is unstructured and changes with the geometry, so the model must accept a mesh of variable size and connectivity. Graph neural networks (GNNs) meet both requirements: they operate directly on irregular meshes and return node-wise physical fields (Pfaff et al., 2021; Fu et al., 2023; Patrignani & Pinho, 2026). What the baseline mesh representation does not necessarily expose, however, is the task-relevant geometric context. Structural stress depends not only on local connectivity but also on relations that span the domain, such as how close a feature lies to a loaded surface and how much material remains to carry the load along that path. If those relations are poorly exposed in the input, errors at high-stress nodes can persist even when aggregate accuracy is good.

We distinguish two mechanisms behind a localized failure. Under sampling insufficiency, the relevant information is present in the input but sparsely represented in training; targeted enrichment should then help. Under representation insufficiency, the relevant geometry is not sufficiently observable from the current input. If a relevant relation is not sufficiently observable from the sup-

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

plied input, adding further samples with the same representation may have limited value. What is needed instead is an input that exposes the missing relation. Our central question is therefore regime-specific: where does geometric conditioning matter most, and can its effect be substituted by additional FE data?

We answer this question for nodal von Mises stress on a region of interest in parametric perforated structures. A controlled 2×2 design independently varies geometric conditioning and targeted FE enrichment, separating representation effects at fixed data from data effects at fixed representation.

The main finding is that the value of conditioning depends on the geometry. The largest benefit appears for specimens in which a hole lies close to the loaded surface without cutting into it, leaving only a narrow band of material to carry the load across that section. Such configurations are both sparsely sampled under independent parameter sampling and poorly described by the baseline input, and our results indicate that representation is the more limiting factor under the investigated enrichment: for them the conditioned model trained on 1,605 specimens outperforms the unconditioned model trained on 1,920 specimens in both seeds. When the hole is far from the loaded surface, by contrast, the conditioning effect is small and changes sign across seeds. Representation quality and data quantity are therefore not uniformly interchangeable across the design space.

Our contributions are:

1. A controlled FE-surrogate framework that independently varies geometric representation and targeted data enrichment.
2. A regime-wise analysis showing that the benefit of geometric conditioning tracks the stress-response regime rather than a single global value.
3. A direct two-seed comparison showing that, for narrow ligaments, the investigated 315-specimen enrichment does not substitute for conditioning.

2 Related Work

2.1 GNN-Based Structural Surrogate Models

Graph-network simulators established message-passing learning for physical systems represented as particle graphs (Sanchez-Gonzalez et al., 2020), while MeshGraphNets extended this paradigm to simulation meshes and node-wise physical-field prediction (Pfaff et al., 2021). Structural FE applications have since predicted stress and strain fields while exploring boundary-aware embeddings, global graph attributes, alternative structural graph representations, and local-global attention (Fu et al., 2023; Jiang & Chen, 2023; Cai & Jelovica, 2024; Patrignani & Pinho, 2026). We keep architecture and loss fixed and instead study how representation and data allocation affect localized stress errors.

This emphasis differs from work that primarily improves network architecture, physical constraints, or loss formulation. Beyond GNNs, geometry-aware stress surrogates also show that explicitly encoded geometric structure can materially affect stress prediction; for example, EPINN-GF augments a physics-informed neural network with geometry-derived Laplacian eigenvectors for complex spinal stress prediction (Zhang et al., 2025).

2.2 Geometric Information in Graph-Based Physical Learning

Geometric learning can rely on coordinates and mesh connectivity or encode geometric relations and symmetries more explicitly (Bronstein et al., 2021; Satorras et al., 2021). For message-passing GNNs, long-range dependencies can be difficult to transmit because of bottlenecks and over-squashing (Alon & Yahav, 2021; Topping et al., 2022). These observations motivate representations that expose task-relevant geometric relations rather than requiring them to emerge only through repeated local propagation.

For stress prediction, those relations include hole position, ligament width, and distance to boundaries. We encode them as node-wise relational fields and structure-level descriptors, then ablate the two groups to identify where the observed gain originates.

2.3 Targeted Data Acquisition and Subgroup Failure

When expensive simulations are the source of labels, targeted sampling and active learning aim to spend the data budget where it is most informative. Prior work has examined sampling and representation jointly in data-efficient structure–property learning (Liu et al., 2024). Our question is narrower: within localized physical-response regimes, can additional FE training specimens substitute for a more informative geometric representation?

Related subgroup-robustness work shows that high average accuracy can coexist with poor performance on specific groups, and that reallocation or upweighting of data can improve worst-group behavior (Sagawa et al., 2020; Liu et al., 2021). Our setting differs because groups are defined by a continuous mechanical variable—load-face gap—and new labels require additional FE solves.

3 Method

3.1 Overall Framework

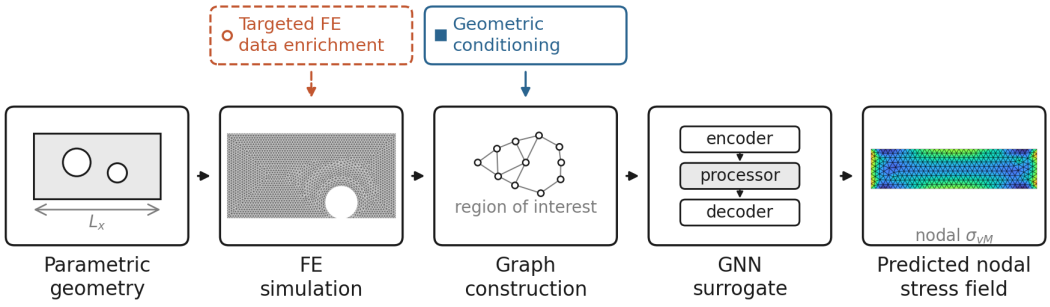


Figure 1: Overall framework. Parametric geometries are meshed and solved by finite element analysis, the region-of-interest nodes and their mesh connectivity are assembled into a graph, and a graph neural network regresses the nodal von Mises stress field. The two factors of the controlled 2×2 design act at different stages: targeted FE data enrichment changes which specimens are simulated and added to the training set, whereas geometric conditioning changes the node input assembled at graph construction.

The framework proceeds in five stages. A parametric description of the structure fixes the plate dimensions and the position and radius of each hole. Each sampled geometry is meshed and solved by finite element analysis, which yields the von Mises stress field. The nodes of the region of interest and their mesh connectivity are then assembled into a graph whose node and edge attributes carry the boundary conditions, the applied load, and, in the conditioned configurations, the geometric descriptors. A graph neural network is trained on these graphs to regress the nodal stress, and the trained model predicts the stress field on the region of interest for an unseen geometry. Across all experiments we vary only (i) explicit geometric conditioning and (ii) targeted FE enrichment; architecture, optimizer, target transform, and validation split remain fixed. Figure 1 summarizes the sequence and marks the two stages at which the experimental factors act.

3.2 Graph Representation and Baseline GNN

Each FE specimen is converted to a graph from its 10-node tetrahedral mesh. Physical quadratic-element edges are instantiated bidirectionally, with relative position and edge length as edge features. Only region-of-interest nodes and their connectivity are provided to the model; nodes outside the prediction region are excluded. Consequently, non-intersecting holes leave no direct baseline trace at the prediction nodes, which is central to the representation-insufficiency interpretation in Section 5.5.

The baseline contains 17 node features: three node coordinates, three prescribed-displacement indicators, three applied nodal-force components, six per-node local-stiffness indicators obtained from

the FE model, and two distances to the fixed and loaded faces. Zero-variance columns are removed on the training split, and the remaining features are standardized using training-set statistics. No feature in the baseline explicitly describes the hole geometry.

We use an encoder–processor–decoder mesh GNN with 10 message-passing blocks and a specimen-level global-context gate. Because architecture is not the contribution, detailed layer definitions are deferred to Appendix A.3.

Targets are normalized by total load magnitude using linear-elastic proportionality, then log1p-transformed and standardized. Predictions are inverse-transformed before computing MPa-scale errors. All conditions use the same plain-MSE objective.

3.3 Geometric Conditioning

Geometric conditioning appends 18 CAD- and mesh-derived features to the baseline node input; none uses FE stress labels. They divide into two groups according to whether they vary within a specimen.

Eight of the features describe node-wise relational geometry, that is, the position of each node relative to the holes and to the free surfaces. They comprise the signed distance from the node to the nearest hole and its reciprocal transform, four in-plane offsets from the node to the hole centers, and the distances to the nearest free faces in the width and thickness directions. Because the region of interest lies on the $x = L_x$ plane, six of these eight columns vary across the nodes of a specimen; the remaining two are constant within a specimen and vary only between specimens.

The other ten features are structure-level descriptors that characterize a specimen as a whole and are broadcast unchanged to all of its nodes. The first is the load-face gap

$$g = L_x - \max_i (x_i + r_i),$$

which measures how much material separates the nearest hole from the loaded face, together with its reciprocal transform $1/(\max(g, 0) + 1)$. The remaining descriptors are the intersection depth, an indicator of whether a hole cuts the loaded face, the corresponding gap at the fixed face, the number of holes, the two hole radii, the maximum radius, and the area fraction removed by the holes. Each descriptor is broadcast to all nodes of a specimen.

The reciprocal gap emphasizes the near-contact range, where stress changes sharply as ligament width decreases; intersecting geometries are handled separately by depth and indicator features.

3.4 Targeted FE Data Enrichment

The 2,000-specimen initial set was generated by independent geometry/load sampling, making high-stress cases relatively rare (median peak stress 12.6 MPa; 90th percentile 25.2 MPa under a 50 MPa ceiling). Baseline diagnostics identified thin-ligament and load-face-near/intersecting regimes as major failures, motivating 315 targeted training specimens in those regions.

Each targeted specimen uses a calibration FE solve to estimate load–peak-stress proportionality, followed by a production solve adjusted to place peak stress approximately in 38–48 MPa. Thus, 315 denotes added training specimens, not FE solves; throughout the paper it is the training-data budget unit.

Targeted regimes were selected using baseline errors on the fixed 395-specimen validation set and inspection of sparse regions in the training geometry distribution. The validation set is therefore not a fully held-out test set and is also used for checkpoint selection, so absolute errors should be read optimistically. This dependence favors enriched conditions B and D; the geometric features were fixed from CAD/physics rather than tuned to validation error, so the main B-versus-C contrast is conservative with respect to acquisition-design bias.

3.5 Controlled 2-by-2 Design

We treat the two interventions as independent factors and define the following four conditions.

Table 1: Controlled 2×2 experimental design.

	No enrichment	Targeted enrichment
No geometric conditioning	A (1,605 / 17 features)	B (1,920 / 17 features)
Geometric conditioning	C (1,605 / 35 features)	D (1,920 / 35 features)

Table 2: Training and validation configuration for the four experimental conditions.

	A	B	C	D
Targeted enrichment	No	Yes	No	Yes
Geometric conditioning	No	No	Yes	Yes
Training specimens	1,605	1,920	1,605	1,920
Node features	17	17	35	35
Validation specimens	395	395	395	395

A→B measures the effect of targeted data at fixed representation; A→C measures the effect of geometric conditioning at fixed data; C→D measures the effect of additional data after conditioning; and B↔C directly compares a more-data/unconditioned model with a less-data/conditioned model. The interaction is defined as

$$(C \to D) - (A \to B).$$

4 Experimental Setup

4.1 Parametric Structural Domain and FE Model

We consider a $200 \times 24 \times 100$ mm perforated steel plate with one or two through-holes with radii of approximately 5–20 mm. The material properties are $E = 200$ GPa and $\nu = 0.3$. The $x = 0$ face is fixed, and a distributed bending load is applied at $x = 200$ mm. SOLID187 10-node tetrahedral elements are used with a global element size of 3 mm and eight divisions through the thickness. Per-node load is sampled from 0.01–0.8 N and limited so that peak von Mises stress remains below 50 MPa.

The data pool contains 2,315 specimens: 2,000 initial specimens and 315 targeted specimens. A fixed split of the initial set provides 395 validation specimens; targeted specimens never enter validation.

4.2 Dataset and Training Configuration

All conditions use the same optimization schedule. Results are reported for seeds 42 and 43, controlling initialization, data order, and dropout. Checkpoints are selected by validation RMSE; hyperparameters are listed in Appendix A.3.

4.3 Geometry Regimes

We characterize each specimen by the load-face gap $g = L_x - \max_i(x_i + r_i)$ and express it relative to the specimen length as the normalized gap

$$\bar{g} = \frac{g}{L_x},$$

the signed load-face gap expressed as a fraction of the specimen length. Validation specimens are partitioned into Intersecting ($\bar{g} < 0$, $n = 72$), Narrow ligament ($0 \le \bar{g} < 0.10$, $n = 54$), and Far field ($\bar{g} \ge 0.10$, $n = 269$). The 0.10 boundary is motivated by a response scan over the full 2,000-specimen base set, in which the load-normalized peak stress rises sharply only below this value. The scan and a cutoff sensitivity check are reported in Appendix A.2. The normalization affects only how regimes are reported; the model input uses the unnormalized gap described in Section 3.3.

Table 3: Overall and regime-wise prediction performance. Values in brackets are two-seed $[\min, \max]$ ranges.

	A (base)	B (+data)	C (+geom)	D (+both)
Node RMSE (MPa)	0.0889 [0.0888,0.0890]	0.0862 [0.0861,0.0863]	0.0425 [0.0367,0.0484]	0.0363 [0.0356,0.0371]
Peak err.—Intersecting	2.332 [2.024,2.640]	1.683 [1.481,1.885]	1.531 [1.431,1.631]	1.991 [1.565,2.417]
Peak err.—Narrow	3.116 [2.769,3.464]	2.698 [2.275,3.120]	1.608 [1.606,1.611]	1.363 [1.089,1.637]
Peak err.—Far field	0.694 [0.585,0.802]	0.969 [0.667,1.271]	0.569 [0.437,0.701]	0.956 [0.377,1.534]
Peak err.—All	0.936 [0.736,1.137]	1.063 [0.811,1.316]	0.723 [0.575,0.872]	1.035 [0.474,1.596]

4.4 Evaluation Metrics and Contrasts

Field accuracy is measured by node-wise RMSE in MPa. Critical-response accuracy is measured by the median specimen-wise peak relative error,

$$e = \frac{|\max(\sigma_{\text{pred}}) - \max(\sigma_{\text{true}})|}{\max(\sigma_{\text{true}})} \times 100,$$

within each regime.

We also report a descriptive severe-underprediction count: nodes satisfying $\sigma_{\text{true}} > 3$ MPa and $\sigma_{\text{pred}} < 2.5$ MPa.

For regime r, define

$$\Delta E_{\text{geom},r} = E_r(A) - E_r(C),$$

the conditioning-induced peak-error reduction at fixed data, and

$$S_r = E_r(B) - E_r(C),$$

the substitutability contrast. $S_r > 0$ means that the conditioned model with fewer specimens outperforms the enriched unconditioned model. Brackets denote two-seed $[\min, \max]$ ranges; an asterisk denotes sign consistency across the two seeds.

5 Results

5.1 Overall Prediction Performance

Adding 315 targeted specimens changes unconditioned RMSE from 0.0889 $[0.0888, 0.0890]$ to 0.0862 $[0.0861, 0.0863]$ MPa, whereas conditioning the same base data reduces it to 0.0425 $[0.0367, 0.0484]$ MPa. Using both gives 0.0363 $[0.0356, 0.0371]$ MPa. The representation change therefore produces a much larger overall RMSE reduction than the investigated enrichment.

5.2 Geometry-Regime-Dependent Value of Conditioning

The largest observed conditioning effect occurs in the narrow-ligament regime. $\Delta E_{\text{geom},r}$ is $+1.858$ and $+1.158$ percentage points across the two seeds for Narrow, $+0.592$ and $+1.009$ for Intersecting, and $+0.365$ and -0.115 for Far field. The corresponding two-run ranges do not overlap, although two seeds are insufficient for a statistical ordering.

This observed regime-wise pattern is mechanically plausible. Far-field holes weakly perturb the load face; intersecting holes already alter baseline node distribution and local stiffness; narrow ligaments place stress-driving geometry immediately adjacent to, but not explicitly described by, the prediction input. We therefore interpret the main result as regime dependence rather than a single global value of conditioning. Figure 2 summarizes this regime-dependent behavior across the four experimental conditions.

5.3 Substitutability of Targeted FE Data and Geometric Conditioning

As shown in Figure 2, the most pronounced B-versus-C separation occurs in the narrow-ligament regime. B versus C is the direct substitutability test. In Narrow,

$$S_{\text{narrow}} = E_{\text{narrow}}(B) - E_{\text{narrow}}(C) = +1.089$$

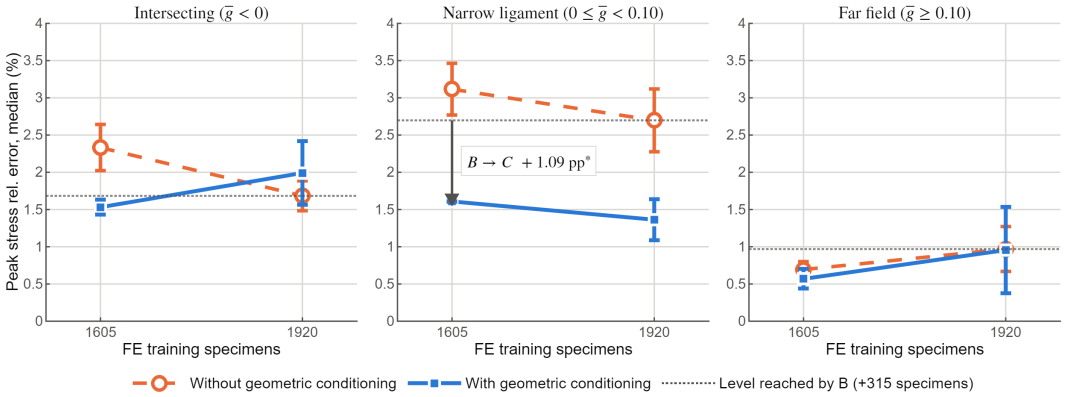


Figure 2: Regime-wise peak relative error across the controlled 2-by-2 design. Dashed curves with open circles denote models without geometric conditioning, whereas solid curves with filled squares denote models with geometric conditioning. The horizontal axis distinguishes the base 1,605-specimen set from the 1,920-specimen enriched set, and error bars show the range across the two evaluated seeds. In the Intersecting regime, the two curves cross because targeted enrichment improves the unconditioned model but worsens the conditioned model; this interaction is discussed in Section 5.4. The largest observed conditioning effect occurs in the narrow-ligament regime, where the conditioned model trained on 1,605 specimens (C) outperforms the enriched unconditioned model trained on 1,920 specimens (B). The horizontal reference line marks the error level reached by condition B after adding 315 targeted specimens.

Table 4: Regime-wise contrasts in median peak relative error. For pairwise contrasts, positive values indicate improvement by the second condition. The interaction is defined as $(C \to D) - (A \to B)$. An asterisk denotes the same sign across both seeds. Ranges are reported for the contrasts discussed in the text; the remaining entries give the two-seed mean.

Contrast	Intersecting	Narrow	Far field	All
A→B (data only)	$+0.649^{*}$ [0.138,1.159]	$+0.419^{*}$ [0.343,0.494]	-0.275	-0.127
A→C (geom only)	$+0.801^{*}$ [0.592,1.009]	$+1.508^{*}$ [1.158,1.858]	$+0.125$ [−0.115,0.365]	$+0.213$
B→C (geom vs. data)	$+0.152$ [−0.150,0.454]	$+1.089^{*}$ [0.664,1.515]	$+0.400^{*}$ [0.230,0.571]	$+0.340^{*}$ [0.236,0.444]
C→D (data after geom)	-0.460^{*} [−0.786,−0.134]	$+0.245$ [−0.026,0.516]	-0.387	-0.312
Interaction	-1.108^{*} [−1.944,−0.272]	-0.174	-0.112^{*} [−0.148,−0.075]	-0.185^{*} [−0.226,−0.143]

percentage points with a seed range of $[0.664, 1.515]$, positive in both seeds: 315 additional targeted specimens without conditioning do not match the conditioned 1,605-specimen model. This claim does not rely on monotonic learning curves or an FE-equivalent simulation count. Because acquisition design used validation error, any resulting bias favors B and D, making $S_{\text{narrow}} > 0$ conservative with respect to that dependence.

Far-field B→C is also positive in both runs, but A→C is small and one data-only run worsens; we therefore do not interpret Far field as a high-value conditioning regime. Intersecting B→C spans zero, so no substitutability claim is made there.

5.4 Interaction Between Data and Representation

The two quantities reported here answer different questions. C→D is the effect of the same 315-specimen enrichment when the model is already geometrically conditioned, whereas the interaction $(C \to D) - (A \to B)$ compares that effect against the effect of the identical enrichment on the unconditioned model. A negative interaction therefore states that enrichment buys less once the geometry is exposed, which is a claim about the two factors jointly rather than about either one alone.

Measured this way, the data effect changes after conditioning. C→D is negative in three of the four partitions and equals -0.460 $[-0.786, -0.134]$ in Intersecting, where both seeds worsen after enrichment. The interaction $(C \to D) - (A \to B)$ is negative in every partition and sign-consistent

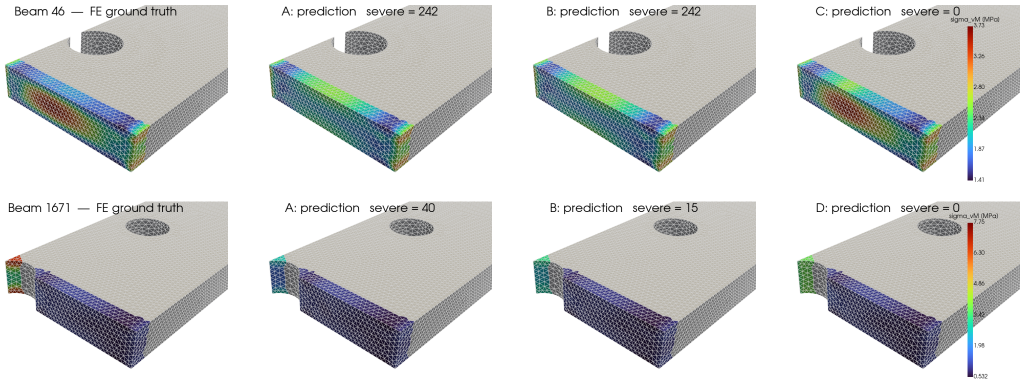


Figure 3: Representative sample-level failure modes for seed 42. Top: Beam 46, where a non-intersecting hole drives the stress concentration. Severe underpredictions remain unchanged from A to B (242 nodes) but disappear under geometric conditioning in C. Bottom: Beam 1671 ($\bar{g} = -0.11$, Intersecting), where the stress-driving geometry is locally reflected in the prediction-region input. Targeted enrichment reduces severe underpredictions from 40 nodes in A to 15 in B, and the combined condition D removes them. FE reference and predictions within each specimen use a common stress color range.

Table 5: Feature-level ablation at fixed training size. An asterisk denotes the same sign across both seeds.

	None	Node-wise	Structure-level	Both
Node RMSE (MPa)	0.0889 [0.0888,0.0890]	0.0381 [0.0371,0.0392]	0.0805 [0.0788,0.0822]	0.0425 [0.0367,0.0484]
Severe underpredicted nodes	320 [309,330]	16 [15,16]	324 [320,327]	40 [6,74]
Peak impr.—Intersecting	—	0.522	0.582*	0.801*
Peak impr.—Narrow	—	1.636*	1.396*	1.508*
Peak impr.—Far field	—	0.314*	0.013	0.125
Peak impr.—All	—	0.428*	0.147*	0.213

in Intersecting, Far field, and All, showing that the two interventions are not simply additive in this setting.

One explanation is distribution shift: targeted specimens concentrate at small/negative gaps and high stresses, and explicit gap features may make that shift easier to fit. We treat this as an interpretation rather than a demonstrated mechanism.

5.5 Sample-Level Failure Analysis

Representative specimens from seed 42 in Figure 3 illustrate the two mechanisms. In Beam 46, a non-intersecting hole drives the stress concentration but is not explicitly described at prediction nodes; 242 severe underpredictions remain after enrichment and disappear with conditioning, consistent with representation insufficiency.

In Beam 1671 ($\bar{g} = -0.11$, Intersecting), the hole cuts the prediction face, so the relevant geometry is reflected locally. Severe underpredictions fall from 40 to 15 after enrichment and vanish with enrichment plus conditioning, consistent with sampling insufficiency. The "thin ligament on the prediction face" describes the local mechanism, not the formal regime label.

5.6 Feature-Level Ablation: Node-Wise Versus Structure-Level Geometry

Figure 4 further shows that most of the observed improvement is associated with node-wise relational geometry. Node-wise relational features account for most of the field-level gain: RMSE is $[0.0371, 0.0392]$ versus $[0.0788, 0.0822]$ for structure-level scalars alone. Severe underpredictions fall from 320 to 16 with node-wise features but remain at 324 with structure-level scalars, indicating that specimen-level scalars alone are insufficient to localize severe errors in this setting.

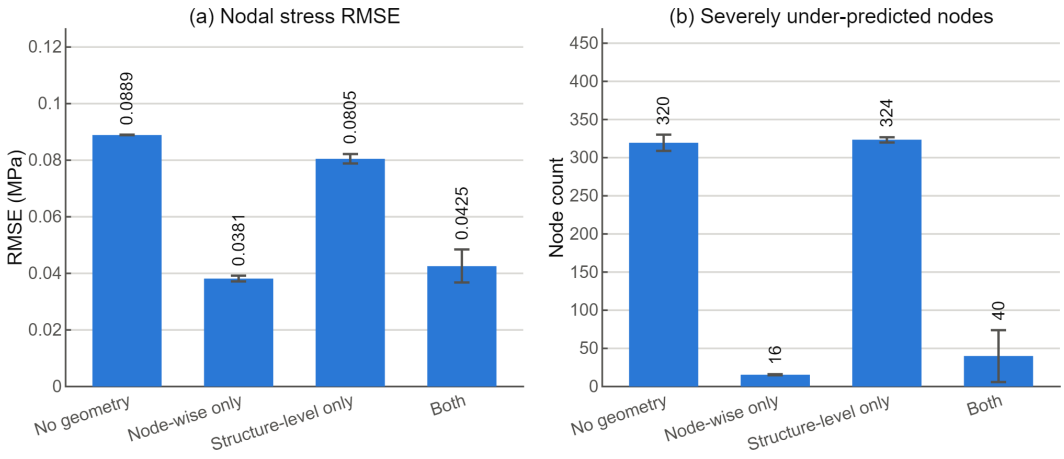


Figure 4: Feature-level ablation at the fixed base training size. (a) Node-wise RMSE and (b) the number of severely underpredicted nodes for no geometric conditioning, node-wise relational geometry only, structure-level descriptors only, and the full representation. Bars show the mean across two seeds and error bars show the corresponding seed range. Node-wise relational geometry accounts for most of the field-level and severe-underprediction improvements, whereas structure-level descriptors alone do not reproduce these gains.

The full 18-feature representation is not uniformly dominant and shows larger between-seed variation than node-wise only. We therefore do not claim node-wise-only superiority; rather, the current evidence attributes most of the observed gain to node-wise relational geometry, while the additional structure-level scalars do not consistently improve every metric.

A single-seed representation$\times$enrichment factorial with paired bootstrap is retained as supporting evidence in Appendix A.4 and is not used for the main two-seed claims.

6 Conclusion

We used a controlled 2×2 experiment to separate the effects of targeted FE enrichment and geometric conditioning on localized high-stress errors in a GNN stress surrogate. Overall RMSE changes little under the investigated 315-specimen enrichment but drops substantially when the same base data are geometrically conditioned.

The effect is regime-dependent. In Narrow, conditioning reduces peak error by 1.158–1.858 percentage points across the two seeds, and the conditioned 1,605-specimen model outperforms the unconditioned 1,920-specimen model in both runs. The investigated enrichment therefore does not substitute for conditioning in this regime. In Far field, the conditioning effect is small and changes direction across seeds, ruling out a single uniform representation value across the design space.

Ablation further attributes most of the gain to node-wise relational geometry rather than specimen-level scalars. Together with the non-additive data–representation interaction, this supports a practical ordering: assess whether response-driving geometry is sufficiently observable before allocating additional FE budget, then re-evaluate the marginal value of targeted data.

Limitations include one structural family, one architecture, two seeds, and a problem-specific regime boundary. The acquisition policy also used baseline errors from the same validation set used for checkpoint selection, so absolute errors are not independent-test estimates. Future work should validate the ordering across other topologies, loads, and materials and integrate representation-sufficiency diagnostics with adaptive acquisition.

AI USE STATEMENT

In this work, we used generative AI tools to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to help develop conceptual frameworks, to interpret results, to create or edit software code, and to assist with translation. We did not use generative AI tools to generate synthetic datasets, to clean or reformat the dataset, or to formulate mathematical claims; the proof-related and qualitative-analysis categories are not applicable to this work. We additionally used generative AI tools in the recommended-disclosure categories, including providing feedback on scientific figures, suggesting experimental parameters, drafting and editing the manuscript, organizing its structure, summarizing and identifying relevant literature, formatting references, searching for information, brainstorming, and proposing a title.

All finite element simulations, model-training runs, and primary result files were produced by the authors using their own computational workflow. No reported numerical result originated from generative AI output. The authors reviewed all AI-assisted code and cross-checked its outputs against the corresponding result files so that every value shown in a figure or table matches its source. Every AI-suggested reference was independently verified against the published source for existence, content, and relevance, and the surrounding text was revised where the source did not support the claim as drafted. Unsupported quantitative interpretations considered during drafting were removed rather than weakened; the resulting claim scope is stated in Appendix A.5. The authors made all final scientific and technical decisions and take full responsibility for the content of this work.

REFERENCES

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Yuecheng Cai and Jasmin Jelovica. Efficient graph representation in graph neural networks for stress predictions in stiffened panels. *Thin-Walled Structures*, 203:112157, 2024. doi: 10.1016/j.tws.2024.112157.

Xingyu Fu, Fengfeng Zhou, Dheeraj Peddireddy, Zhengyang Kang, Martin Byung-Guk Jun, and Vaneet Aggarwal. An finite element analysis surrogate model with boundary oriented graph embedding approach for rapid design. *Journal of Computational Design and Engineering*, 10(3): 1026–1046, 2023. doi: 10.1093/jcde/qwad025.

Chunhao Jiang and Nian-Zhong Chen. Graph neural networks (GNNs) based accelerated numerical simulation. *Engineering Applications of Artificial Intelligence*, 123:106370, 2023. doi: 10.1016/j.engappai.2023.106370.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792, 2021.

Hao Liu, Berkay Yucel, Baskar Ganapathysubramanian, Surya R. Kalidindi, Daniel Wheeler, and Olga Wodo. Active learning for regression of structure–property mapping: the importance of sampling and representation. *Digital Discovery*, 3:1997–2009, 2024. doi: 10.1039/D4DD00073K.

Luca Patrignani and Silvestre T. Pinho. Graph neural networks with hybrid local-global attention for effective prediction of mechanical response in structures. *Computer Methods in Applied Mechanics and Engineering*, 452:118753, 2026. doi: 10.1016/j.cma.2026.118753.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468, 2020.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332, 2021.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022.

Yue Zhang, Junhua Zhang, Hongjian Li, and Qiyang Wang. EPINN-GF: An enhanced physics-informed neural network integrating geometric features for stress prediction in complex spinal geometries. *Computer Methods and Programs in Biomedicine*, 271:109044, 2025. doi: 10.1016/j.cmpb.2025.109044.

A Appendix

A.1 Exploratory Learning-Curve Analysis

Conditioning ON/OFF learning curves were trained at base-set fractions 0.01, 0.02, 0.05, 0.10, 0.20, 0.35, 0.55, and 0.80 of the 1,605-specimen base training set, using nested subsets that preserve regime proportions and evaluating all models on the same 395 validation specimens.

Points below approximately $N = 160$ (fractions below 0.10) were excluded because the model collapses toward mean prediction and median peak error becomes misleading in the far-field-dominated validation set. The admissible range is therefore

$$N \in \{160, 321, 561, 882, 1284\},$$

and the sweep does not reach the full 1,605-specimen training set.

This sweep was performed with a single seed and exhibits non-monotonic behavior. We therefore treat it as exploratory only. We do not report a point estimate or lower bound of "FE-equivalent simulations" from these curves, and no main-text claim depends on a smoothed crossing.

A.2 Regime-Boundary Scan and Cutoff Sensitivity

The $\bar{g} = 0.10$ boundary is supported by a response scan over the full 2,000-specimen base set. For each specimen the load-normalized peak stress $\max(\sigma_{vM})/|F|$ is computed from the FE solution; because the problem is linear elastic, dividing by the applied load removes load variation and leaves only the geometric effect.

Table 6 shows that the median is nearly constant across all gap ranges (2.14–2.97), so the regimes are not distinguished by typical behavior. The difference is primarily in the upper tail: specimens with $0 \leq \bar{g} < 0.10$ reach 40.55, whereas no specimen with $\bar{g} \geq 0.10$ exceeds 3.44. Intersecting specimens ($\bar{g} < 0$) reach only 15.90, consistent with the reading that a hole which has removed part of the loaded face no longer concentrates load through a ligament.

A cutoff sensitivity check confirms that this characterization is not an artifact of the specific value 0.10. Moving the narrow/far boundary to 0.05, 0.075, 0.125 or 0.15 leaves the far-field maximum between 2.27 and 3.47 in every case, while the narrow-ligament maximum stays at 40.55. The separation therefore holds over a wide range of cutoffs; we use $\bar{g} = 0.10$ throughout. On the fixed validation set the resulting regime counts are 72 / 54 / 269, as stated in Section 4.3.

Table 6: Load-normalized peak stress as a function of the normalized load-face gap $\bar{g} = g/L_x$, over the 2,000-specimen base set.

$\bar{g}$ range	Specimens	Median peak/$\lvert F \rvert$	Max peak/$\lvert F \rvert$
$\bar{g} \geq 0.30$	878	2.14	2.15
$0.10 \leq \bar{g} < 0.30$	516	2.14	3.44
$0 \leq \bar{g} < 0.10$	260	2.17	40.55
$-0.05 \leq \bar{g} < 0$	141	2.97	15.90
$\bar{g} < -0.05$	205	2.20	12.25

A.3 Architecture and Optimization Details

Node and edge encoders are two-hidden-layer MLPs with width 64, layer normalization, ReLU activation, and dropout, mapping the inputs to a latent dimension of 64.

The processor contains 10 message-passing blocks with residual updates and a specimen-level global-context gate. The decoder maps the final latent representation to one scalar value per node.

Training uses Adam with an initial learning rate of 10^{-3}, weight decay of 10^{-4}, and cosine annealing to 10^{-6}. Each model is trained for 350 epochs with batch size 16 and dropout 0.05. Checkpoint selection is based on validation RMSE measured in MPa.

A.4 Supporting Single-Seed Representation–Enrichment Analysis

A single-seed factorial crossing node-wise, structure-level, and full geometric representations with targeted enrichment was evaluated using a paired graph-level bootstrap.

Within the narrow-ligament regime, the data effect was -0.233 percentage points with bootstrap interval $[-0.285, -0.157]$ under node-wise conditioning, $+0.325$ with interval $[+0.272, +0.422]$ under structure-level conditioning, and $+0.516$ with interval $[+0.063, +1.248]$ for the full feature set.

These intervals are bootstrap intervals over validation specimens at a single seed and are not directly comparable to the two-seed ranges reported in the main text. We therefore use this analysis only as supporting evidence for the non-additive interaction discussed in Section 5.4.

A.5 Claim Scope

To clarify the scope of the empirical claims, we adopt the following restrictions:

- We do not report a ratio such as "conditioning is worth k times more data," because the data effect changes sign across seeds and regimes.
- We do not report a point estimate or lower bound of FE-equivalent simulations from the single-seed learning curves.
- We do not claim that the node-wise-only representation is definitively more accurate than the full representation. Instead, we conclude that most of the observed gain is associated with node-wise relational features, while the full feature set shows larger between-seed variation in the two evaluated runs.
- The two evaluated seeds characterize the variation observed in these runs but do not estimate the full seed distribution. We therefore phrase the corresponding findings descriptively rather than as claims of statistical significance.