Supplementary Material

GDEA: Global-Weight Deep Equilibrium Attention for Finite Element Systems

Junghun Lee[1*] and Conrad Tucker[1]

[1*]Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, 15232, PA, USA.

*Corresponding author(s). E-mail(s): junghunl@andrew.cmu.edu;
Contributing authors: conradt@andrew.cmu.edu;

S1 Linear Elastic Partial Differential Equation

To generate Deformed ABC, the Finite Element Method (FEM) is applied to meshes to solve linear-elastic partial differential equations presented in Eq. S1. Ω refers to the domain and $\sigma(u)$ is a Cauchy stress tensor presented in Eq. S2. C refers to the elastic tensor and $\varepsilon(u)$ is a linear strain tensor presented in Eq. S3, where u notates the displacement of the node.

$$\nabla \cdot \sigma(u) = 0 \quad \text{in } \Omega \tag{S1}$$

$$\sigma(u) = C : \varepsilon(u) \tag{S2}$$

$$\varepsilon(u) = \frac{1}{2}\left(\nabla u + (\nabla u)^T\right) \tag{S3}$$

Two different boundary conditions are applied. First, the Dirichlet boundary condition is applied following Eq. S4, where Γ_D is the boundary domain. Next, the Neumann boundary condition is applied, as in Eq. S5, where Γ_N is the boundary domain, n is the outward normal vector, and t_N is the traction vector.

$$u = 0 \quad \text{on } \Gamma_D \tag{S4}$$

$$\sigma(u) \cdot n = t_N \quad \text{on } \Gamma_N \tag{S5}$$

S2 Data Preprocessing

S2.1 Deformed ABC

We pre-process the Deformed ABC before training and evaluating the model. First, FEM-failed cases and geometries with multiple disconnected meshes are filtered out, as they result in rigid-body motion when the fixed and loaded surfaces are assigned to separate meshes. Of the 200,000 geometries selected from the ABC dataset [1], 115,194 completed finite element analyses and have a single mesh. Next, maximum displacements below 0.0001 mm and those exceeding 100 mm are removed as outliers. Fig. S1 (a) shows the distribution of maximum displacements.

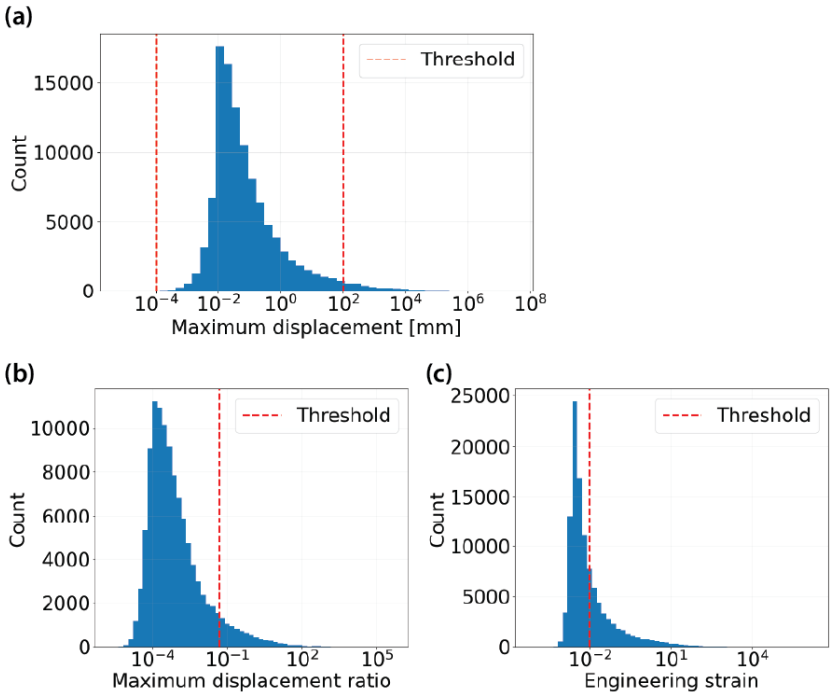

Fig. S1 This figure illustrates the distribution of the dataset according to (a) Maximum displacement, (b) Maximum displacement ratio, and (c) Engineering strain. Also, the red dashed line indicates the threshold used to filter the dataset.

9,953 geometries are filtered as outliers. Afterward, geometries exhibiting nonlinear deformation are excluded to ensure geometric linearity. A geometry is considered linear when both the maximum displacement ratio and the local engineering strain are below their respective thresholds, $R_{\text{th}} = 0.05$ and $\varepsilon_{\text{th}} = 0.01$, following Equation S6. Here, $\mathbf{u}$ denotes the displacement of the node, which is the solution of FEM, $\mathbf{x}$ is the coordinates of the nodes, and L_{char} represents the characteristic length, the diagonal of the bounding box enclosing the geometry. Fig. S1 (b) and (c) show the distribution of max displacement, engineering strain, and their threshold. 33,696 geometries are

filtered as nonlinear finite element solutions. Finally, the retained FEM dataset was converted into an .npz file for use in PyTorch. After excluding 37 files that failed during conversion, the final dataset contains 71,508 samples. The sensitivity test results in Fig. S3 show that the proposed model's performance plateaus with respect to data size, implying that the data is sufficient.

$$\frac{\max(\|\mathbf{u}\|)}{L_{\text{char}}} < R_{\text{th}} \quad \text{and} \quad \max\left(\frac{\|\Delta\mathbf{u}\|}{\|\Delta\mathbf{x}\|}\right) < \varepsilon_{\text{th}} \tag{S6}$$

Young's modulus, nodal loads, and displacement are logarithmically scaled to $\tilde{\mathbf{v}}$ using Equation S7, preserving vector directions while adjusting magnitudes. Here, each vector $\mathbf{v}$ is first converted to a unit vector, its magnitude is then scaled logarithmically, and the scaled vector is finally obtained by combining the direction with the scaled magnitude. ϵ is added to avoid taking the logarithm of zero. Young's modulus is log-scaled because it is logarithmically sampled. Loads and displacements, which depend on Young's modulus, are scaled in the same manner. Finally, the geometry is centered at the coordinate origin.

$$\tilde{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \cdot ((\log(\|\mathbf{v}\| + \epsilon) + \log(\|\mathbf{v}\|_{\min} + \epsilon)) \tag{S7}$$

To train and validate the model in Deformed ABC, 80% of the data is used for training and the remaining 20% for validation. For all experiments, a batch size of 8 with gradient accumulation of 8 is used for training and 1 for validation.

S2.2 Simulated Additive Manufacturing Displacement Fields Dataset

We pre-process the Simulated Additive Manufacturing Displacement Fields Dataset [2] before training and evaluating the model. We filter out geometries with floating meshes and geometries with 2D or 1D characteristics. If any dimension of the bounding box is shorter than 1% of the longest dimension, the geometry is considered non-3D and removed. Figure S2 shows examples of filtered data. Of the 24,880 geometries, 590 are filtered out, and 24,290 are retained as the final dataset. To train and validate the model, 80% of the data is used for training and the remaining 20% for validation. For all experiments, a batch size of 4 is used for both training and validation, following the original paper [2].

S2.3 Simulated Flag Waving in the Wind

We did not preprocess the simulated flag-waving-in-the-wind dataset [3] for model training and validation. Of the 1,200 scenarios, 1,000 are used to train the model and 200 to validate it, following the labeling of the original paper [3]. All datasets are converted from .tfrecord to .npz for the PyTorch implementation. For all experiments, a batch size of 8 is used for training and 1 for validation.

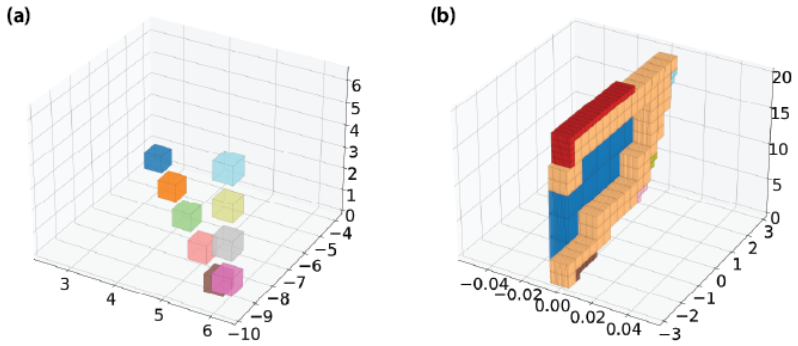

Fig. S2 This figure shows examples of filtered data. (a) is filtered as a floating mesh, and (b) is filtered as a non-3D mesh.

S3 Model Optimization

Hyperparameters of all models are optimized using Optuna [4]. Hyperparameters include the Multi-Layer Perceptron (MLP) architecture, the attention mechanism architecture, the graph representation, the learning rate, the weight decay, and parameters particular to each model. Optimization is conducted using 10% of each training and validation data for all datasets. Hyperparameter sets are evaluated over 10 training epochs with 100 trials. The median pruner is applied after the first 2 epochs. Hyperparameter configurations that result in out-of-memory errors are discarded. The hyperparameter ranges and optimized results for each dataset are presented below.

S3.1 GINO

Python library NeuralOperator [5] is used to implement GINO [6]. For GINO [6], hyperparameters are optimized over the following ranges: number of neural operator layers $\in [2, 8]$, hidden dimension $\in \{64, 128, 256\}$, lifting channels $\in \{64, 128, 256\}$, projection channels $\in \{64, 128, 256\}$, input MLP layers $\in [2, 8]$, output MLP layers $\in [2, 8]$, graph neural operator radius $\in [0.01, 0.1]$, mode values $\in [2, 16]$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. The optimized hyperparameter set for Deformed ABC is: number of neural operator layers=3, hidden dimension=256, lifting channels=256, projection channels=256, input MLP layers=3, output MLP layers=5, graph neural operator radius=9.566×10^{-2}, mode values=2, learning rate = 3.210×10^{-4}, and weight decay = 4.888×10^{-6}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: number of neural operator layers=6, hidden dimension=128, lifting channels=256, projection channels=64, input MLP layers=3, output MLP layers=3, graph neural operator radius=5.499×10^{-3}, mode values=16, learning rate = 2.124×10^{-4}, and weight decay = 9.8368×10^{-6}. The optimized hyperparameter set the simulated flag waving in the wind dataset [3] is: number of neural operator layers=2, hidden dimension=256, lifting channels=256, projection channels=256, input MLP layers=4, output MLP layers=3, graph neural operator radius=7.190×10^{-3}, mode values=14, learning rate = 8.076×10^{-4}, and weight decay = 3.336×10^{-6}.

S3.2 GNOT

For GNOT [7], hyperparameters are optimized over the following ranges: number of layers $\in [2, 8]$, hidden dimension $\in \{64, 128, 256\}$, number of attention heads $\in \{1, 2, 4, 8\}$, number of MLP layers $\in [2, 8]$, feed-forward network dropout $\in [0, 0.2]$, attention dropout $\in [0, 0.2]$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. The optimized hyperparameter set for Deformed ABC is: number of layers = 6, hidden dimension = 128, number of attention heads = 4, number of MLP layers = 2, feed-forward network dropout = 3.687×10^{-2}, attention dropout = 1.687×10^{-1}, learning rate = 5.434×10^{-4}, and weight decay = 2.795×10^{-5}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: number of layers = 2, hidden dimension = 128, number of attention heads = 4, number of MLP layers = 2, feed-forward network dropout = 4.888×10^{-2}, attention dropout = 9.467×10^{-2}, learning rate = 9.871×10^{-4}, and weight decay = 1.839×10^{-6}. The optimized hyperparameter set the simulated flag waving in the wind dataset [3] is: number of layers = 4, hidden dimension = 256, number of attention heads = 2, number of MLP layers = 8, feed-forward network dropout = 1.986×10^{-2}, attention dropout = 1.212×10^{-2}, learning rate = 1.605×10^{-4}, and weight decay = 7.696×10^{-4}.

S3.3 Transolver

For Transolver [8], hyperparameters are optimized over the following ranges: number of layers $\in [2, 8]$, hidden dimension $\in \{64, 128, 256\}$, number of attention heads $\in \{1, 2, 4, 8\}$, mlp ratio $\in \{1, 2, 3, 4\}$, slice number $\in \{16, 32, 64\}$, reference number $\in \{4, 8, 16, 32\}$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. The optimized hyperparameter set for Deformed ABC is: number of layers = 6, hidden dimension = 128, number of attention heads = 1, mlp ratio = 3, slice number = 32, reference number = 4, learning rate = 4.221×10^{-4}, and weight decay = 4.309×10^{-4}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: number of layers = 5, hidden dimension = 64, number of attention heads = 2, mlp ratio = 4, slice number = 64, reference number = 4, learning rate = 1.330×10^{-4}, and weight decay = 7.544×10^{-4}. The optimized hyperparameter set for the simulated flag waving in the wind dataset [3] is: number of layers = 6, hidden dimension = 128, number of attention heads = 4, mlp ratio = 3, slice number = 16, reference number = 4, learning rate = 1.060×10^{-4}, and weight decay = 4.956×10^{-4}.

S3.4 TAG-Unet

For TAG-Unet [2], hyperparameters are optimized over the following ranges: coarsen level $\in [2, 4]$, coarsen factor $\in \{4, 8, 16\}$, hidden dimension of MLP layers $\in \{64, 128, 256\}$, MLP layer depth $\in [2, 8]$, hidden dimension of convolution layers $\in \{64, 128, 256\}$, convolutional layer depth $\in [2, 8]$, channel number $\in \{64, 128, 256\}$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. In TAG-Unet, the coarse level corresponds to the number of layers since the model is a U-net structure. The optimized hyperparameter set for Deformed ABC is: coarsen level = 3, coarsen factor

= 16, hidden dimension of MLP layers = 128, MLP layer depth = 7, hidden dimension of convolution layers = 256, convolutional layer depth = 4, channel number = 64, learning rate = 5.929×10^{-4}, and weight decay = 7.045×10^{-5}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: coarsen level = 3, coarsen factor = 16, hidden dimension of MLP layers = 256, MLP layer depth = 4, hidden dimension of convolution layers = 128, convolutional layer depth = 5, channel number = 256, learning rate = 6.966×10^{-4}, and weight decay = 2.885×10^{-4}. The optimized hyperparameter set for the simulated flag waving in the wind dataset [3] is: coarsen level = 3, coarsen factor = 4, hidden dimension of MLP layers = 256, MLP layer depth = 3, hidden dimension of convolution layers = 128, convolutional layer depth = 7, channel number = 128, learning rate = 1.209×10^{-3}, and weight decay = 9.614×10^{-4}.

S3.5 AMG

For AMG [9], hyperparameters are optimized over the following ranges: number of layers $\in [2, 8]$, hidden dimension $\in \{64, 128, 256\}$, number of attention heads $\in 1, 2, 4, 8$, global ratio $\in [0.1, 0.4]$, global k $\in [4, 32]$, local ratio $\in [0.1, 0.4]$, local k $\in [4, 32]$, local nodes $\in [16, 128]$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. The optimized hyperparameter set for Deformed ABC is: number of layers = 2, hidden dimension = 256, number of attention heads = 1, global ratio 1.641×10^{-1}, global k = 13, local ratio = 2.618×10^{-1}, local k = 12, local nodes = 99, learning rate = 2.563×10^{-4}, and weight decay = 2.233×10^{-4}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: number of layers = 2, hidden dimension = 256, number of attention heads = 4, global ratio 1.350×10^{-1}, global k = 16, local ratio = 3.440×10^{-1}, local k = 17, local nodes = 128, learning rate = 3.517×10^{-4}, and weight decay = 2.677×10^{-6}. The optimized hyperparameter set for the simulated flag waving in the wind dataset [3] is: number of layers = 2, hidden dimension = 256, number of attention heads = 1, global ratio 3.437×10^{-1}, global k = 18, local ratio = 3.437×10^{-1}, local k = 21, local nodes = 52, learning rate = 2.343×10^{-4}, and weight decay = 8.110×10^{-6}.

S3.6 GDEA

For our model Global-Weight Deep Equilibrium Attention (GDEA), hyperparameters are optimized over the following ranges: number of layers $\in [2, 8]$, hidden dimension $\in 64, 128, 256$, number of attention heads $\in 1, 2, 4, 8$, number of MLP layers $\in [2, 8]$, feed-forward network dropout $\in [0, 0.2]$, attention dropout $\in [0, 0.2]$, convergence threshold $\in [0.01, 0.1]$, history size $\in [2, 8]$, learning rate $\in [10^{-5}, 10^{-3}]$, and weight decay $\in [10^{-6}, 10^{-3}]$. The optimized hyperparameter set for Deformed ABC is: number of layers = 2, hidden dimension = 128, number of attention heads = 2, number of MLP layers = 4, feed-forward network dropout = 1.257×10^{-1}, attention dropout = 3.333×10^{-3}, convergence threshold = 1.989×10^{-2}, history size = 5, learning rate = 9.097×10^{-4}, and weight decay = 1.321×10^{-4}. The optimized hyperparameter set for Simulated Additive Manufacturing Displacement Fields Dataset [2] is: number of layers = 4, hidden dimension = 128, number of attention heads = 2, number of MLP layers =

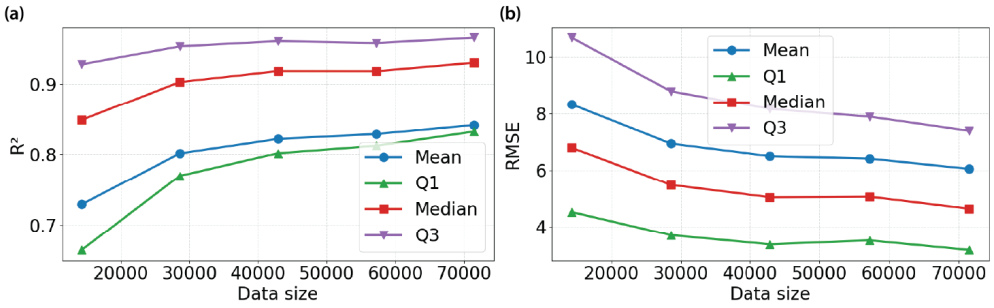

Fig. S3 This figure illustrates how (a) R^2 and (b) RMSE change according to the data size.

4, feed-forward network dropout = 1.624×10^{-1}, attention dropout = 1.436×10^{-1}, convergence threshold = 8.236×10^{-2}, history size = 5, learning rate = 6.725×10^{-4}, and weight decay = 9.891×10^{-4}. The optimized hyperparameter set for the simulated flag waving in the wind dataset [3] is: number of layers = 2, hidden dimension = 256, number of attention heads = 1, number of MLP layers = 6, feed-forward network dropout = 1.019×10^{-1}, attention dropout = 1.768×10^{-1}, convergence threshold = 2.065×10^{-6}, history size = 2, learning rate = 1.285×10^{-4}, and weight decay = 2.065×10^{-6}.

S4 Sensitivity Test

Table S1 Sensitivity study of the global weight matrix deep equilibrium attention on a Deformed ABC dataset.

Data size	R^2					RMSE				
	Mean (↑)	SD (↓)	Q1 (↑)	Median (↑)	Q3 (↑)	Mean (↓)	SD (↓)	Q1 (↓)	Median (↓)	Q3 (↓)
14,301	0.73	0.33	0.66	0.85	0.93	8.32	5.01	4.51	6.78	10.7
28,602	0.80	0.29	0.77	0.90	0.90	6.93	4.48	3.70	5.47	8.79
42,904	0.82	0.27	0.80	0.92	0.96	6.49	4.36	3.37	5.04	8.16
57,205	0.83	0.27	0.81	0.92	0.96	6.40	4.20	3.51	5.05	7.89
71,508	0.84	0.27	0.83	0.93	0.97	6.04	4.22	3.17	4.63	7.38

The performance of our model GDEA across different data sizes is shown in Table S1. The trend shown in Figure S3 indicates that performance improves with larger data sizes but eventually plateaus, suggesting that the current size of the Deformed ABC dataset is adequate.

Declarations

The authors declare no competing interests.

Data Availability

The new dataset used in this study is available on Hugging Face at huggingface.co/datasets/Junghunl/Deformed-ABC. Details of data generation are present in the paper and/or the Supplementary Information.

Code Availability

The code used in this study is available at github.com/AiPEX-Lab. Details of code implementation are present in the paper and/or the Supplementary Information.

References

[1] Koch, S. *et al. Abc: A big cad model dataset for geometric deep learning*, 9601–9611 (2019).

[2] Ferguson, K. *et al.* Topology-agnostic graph u-nets for scalar field prediction on unstructured meshes. *Journal of Mechanical Design* **147**, 041701 (2025).

[3] Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. *Learning mesh-based simulation with graph networks* (2020).

[4] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A next-generation hyperparameter optimization framework* (2019).

[5] Kossaifi, J. *et al.* A library for learning neural operators. *arXiv preprint arXiv:2412.10354* (2025).

[6] Li, Z. *et al.* Geometry-informed neural operator for large-scale 3d pdes. *Advances in Neural Information Processing Systems* **36**, 35836–35854 (2023).

[7] Hao, Z. *et al. Gnot: A general neural operator transformer for operator learning*, 12556–12569 (PMLR, 2023).

[8] Wu, H., Luo, H., Wang, H., Wang, J. & Long, M. Salakhutdinov, R. *et al.* (eds) *Transolver: A fast transformer solver for PDEs on general geometries.* (eds Salakhutdinov, R. *et al.*) *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, 53681–53705 (PMLR, 2024). URL https://proceedings.mlr.press/v235/wu24r.html.

[9] Li, Z., Song, H., Xiao, D., Lai, Z. & Wang, W. *Harnessing scale and physics: A multi-graph neural operator framework for pdes on arbitrary geometries*, 729–740 (2025).