Supplementary Information for

Generative Inverse Design of Shell Metamaterials through Continuous-Latent Conditional Flow Matching

Sibo Zhang, Yunlan Zhang, and Krishna Kumar

Department of Civil, Architectural and Environmental Engineering,
The University of Texas at Austin, Austin, TX, USA

Correspondence: yunlan.zhang@austin.utexas.edu; krishnak@utexas.edu

Supplementary Contents

S1 Implementation details **2**
S1.1 Dataset split and evaluation targets . . . 2
S1.2 Function encoders . . . 2
S1.3 Conditional flow-matching models . . . 3
S1.4 3D-CNN surrogate . . . 3

S2 Finite-element validation details **4**

S3 Robustness and sensitivity analyses **4**
S3.1 Multi-seed robustness . . . 5
S3.2 Cross-surrogate assessment . . . 5
S3.3 Alternative validity criteria . . . 6
S3.4 Paired benchmark comparisons . . . 6
S3.5 Geometry coefficient dimension . . . 7

S4 Additional diagnostic experiments **7**

S1 Implementation details

S1.1 Dataset split and evaluation targets

The DiffuMeta dataset contains 23,534 periodic shell unit cells represented by symbolic implicit fields[1]:

$$\Psi_i(x, y, z) = \sum_{r=1}^{32} \alpha_{ir} T_r(x, y, z) + c_{0i}, \tag{S1}$$

where T_r denotes a periodic trigonometric term and $(\alpha_{i1}, \ldots, \alpha_{i32}, c_{0i})$ is the corresponding coefficient vector. Each symbolic field was evaluated on a periodic 64^3 grid to obtain the discrete input used by the geometry function encoder and the 3D-CNN surrogate. The Natural–Fourier baseline instead operates directly on the original 33-dimensional coefficient vector of the symbolic Fourier representation.

Using random seed 42, the dataset was divided into 21,181 training designs and 2,353 validation designs. Among the validation designs, 251 were also excluded from the released DiffuMeta training set. From this doubly held-out pool, 200 structures were sampled without replacement and used as the common evaluation targets for all five generative models.

S1.2 Function encoders

The geometry and stress-response function encoders used the differentiable ridge-regularized projection defined in Section 4.2, with $\lambda = 10^{-5}$. Both basis networks were trained using the batch-averaged squared relative reconstruction error

$$\mathcal{L}_{\mathrm{rec}}(f) = \frac{1}{B} \sum_{i=1}^{B} \frac{\left\| \widehat{f}_i - f_i \right\|_2^2}{\|f_i\|_2^2}, \tag{S2}$$

where $f_i = \Psi_i$ for geometry fields and $f_i = \boldsymbol{\sigma}_i$ for stress responses. The projected coefficients were standardized dimension-wise using statistics from the training set.

The geometry basis network maps coordinates $\mathbf{x} \in [0, 1]^3$ through fixed integer-frequency Fourier features followed by a multilayer perceptron (MLP). Spatial frequencies were selected from $\{-2, -1, 0, 1, 2\}^3$ using a canonical real-valued representation, yielding 125 Fourier features. Because all spatial frequencies are integers, every learned basis field and every decoded linear combination are periodic in all three coordinate directions. Each geometry-training update used 4,096 spatial points sampled from the 64^3 fields, while validation was performed every 200 steps using a larger sampled coordinate set. The checkpoint at step 19,800, which achieved the lowest sampled validation reconstruction loss, was retained. The frozen basis was then evaluated on the complete 64^3 grid for coefficient projection and all downstream geometry experiments.

The stress-response basis used polynomial features up to degree 4 and Fourier features up to order 4 over a strain range of 0.30. Its four frozen basis functions were evaluated at the prescribed strain coordinates before ridge projection, yielding a fixed-dimensional condition $\boldsymbol{\beta} \in \mathbb{R}^4$ for both fixed-grid and off-grid inputs. The stress-response basis was trained directly on the 11 stress values at 11 equally spaced strain levels from 0.03 to 0.30. Validation was performed every 200 optimization steps, and the checkpoint at step 5,800, which achieved the lowest reconstruction loss, was used for response-basis projection and reconstruction analysis. Reconstruction NRMSE was computed on the original 11 response values, and the mean, median, and p_{90} values are 0.0296, 0.0239, and 0.0560, respectively.

Table S1: Architectures and principal training settings of the function encoders and CFM models.

Component	Representation	Architecture	Training settings
Geometry function encoder	$\Psi(x, y, z) \to \mathbf{c} \in \mathbb{R}^{32}$ by basis projection	125 Fourier features; $125 \to 128 \to 128 \to 32$; SiLU	20,000 steps; batch size 64; 200-step warm-up
Stress-response function encoder	$\sigma(\varepsilon) \to \boldsymbol{\beta} \in \mathbb{R}^4$ by basis projection	Degree-4 polynomial and order-4 Fourier features; 3 hidden layers of width 64	8,000 steps; batch size 128; 100-step warm-up; cosine decay; gradient clipping at 1.0
CFM generator, 4×256	$\mathbf{q} \to \mathbf{c} \in \mathbb{R}^{32}$	4 residual blocks; width 256; 128-dimensional time/condition embeddings	30,000 steps; batch size 256; 500-step warm-up; gradient clipping at 1.0
CFM generator, 8×512	$\mathbf{q} \to \mathbf{c} \in \mathbb{R}^{32}$	8 residual blocks; width 512; 128-dimensional time/condition embeddings	30,000 steps; batch size 256; 500-step warm-up; gradient clipping at 1.0

All models used a learning rate of 3×10^{-4} and a weight decay of 10^{-5}. The condition $\mathbf{q}$ denotes either the raw 11-point response or the $L = 4$ stress-response coefficients. Natural–Fourier CFM used the same 4×256 velocity architecture and training settings.

S1.3 Conditional flow-matching models

The CFM velocity network was implemented as a residual MLP with adaptive layer normalization. Time was encoded using a 128-dimensional sinusoidal embedding, and the combined time and response embeddings modulated each residual block. During training, the response condition was replaced by a learned null embedding with probability $p_{\text{drop}} = 0.1$. At inference, conditional and unconditional velocity predictions were combined using classifier-free guidance:

$$\widetilde{v}_\theta\left(\mathbf{c}_t, t \mid \mathbf{q}\right) = v_\theta\left(\mathbf{c}_t, t \mid \varnothing\right) + w\left[v_\theta\left(\mathbf{c}_t, t \mid \mathbf{q}\right) - v_\theta\left(\mathbf{c}_t, t \mid \varnothing\right)\right], \tag{S3}$$

where $\varnothing$ denotes the learned null condition and $w = 2$ is the guidance scale. Sampling used a second-order Heun solver with 50 uniform time steps. Model architectures and principal optimization settings are summarized in Table S1.

S1.4 3D-CNN surrogate

The 3D-CNN surrogate mapped a standardized 64^3 implicit geometry field to the corresponding 11-point stress response. Geometry-field normalization statistics were estimated from 256 randomly selected training structures, while the stress responses were standardized component-wise using the complete training set.

The network contained four residual convolutional blocks with channel progression $1 \to 16 \to 32 \to 64 \to 128$ and spatial resolutions $64^3 \to 32^3 \to 16^3 \to 8^3 \to 4^3$. Each block used stride-2 downsampling, group normalization, SiLU activation, and a residual shortcut. Circular padding was applied to the $3 \times 3 \times 3$ convolutions. Global average pooling was followed by a $128 \to 256 \to 256 \to 11$ prediction head with SiLU activations and dropout of 0.1. The model contained 992,347 trainable parameters.

The surrogate was trained for 15,000 steps using AdamW with a batch size of 8, an initial learning rate of 10^{-3}, a weight decay of 10^{-4}, a 200-step linear warm-up followed by cosine decay, and gradient clipping at 1.0. The loss was the mean squared error in standardized stress space.

Validation was performed every 200 steps on 2,153 held-out designs remaining after exclusion of the 200 evaluation targets. The checkpoint at step 14,600 achieved the lowest mean validation NRMSE of 0.1092 and was frozen for the evaluation of all generated candidates and comparison methods.

S2 Finite-element validation details

Finite-element analysis (FEA) of the generated candidates followed the protocol used for the DiffuMeta reference dataset[1]. For each candidate, the periodic implicit field was reconstructed on a 64^3 grid, and its zero level set was extracted using marching cubes. The resulting surface was scaled to a cubic unit cell with side length $a = 10$ mm. The extracted mesh was cleaned by merging duplicate nodes and faces, removing near-zero-area and severely distorted triangles, and projecting boundary-adjacent nodes onto the corresponding unit-cell planes. The surface was then remeshed with a target edge length of approximately 0.20 mm and discretized using three-node reduced-integration shell elements (S3R). A uniform shell thickness was assigned according to

$$h^{(j)} = \frac{\rho a^3}{A_{\mathrm{mid}}^{(j)}}, \qquad \rho = 0.1, \tag{S4}$$

where $A_{\mathrm{mid}}^{(j)}$ is the midsurface area of candidate j.

The shell material followed the isotropic elastoplastic model used in the DiffuMeta reference simulations. Each unit cell was compressed between two rigid plates, with the lower plate fixed and the upper plate displaced in the compression direction. Linear constraint equations coupled the three translational degrees of freedom, U_1, U_2, and U_3, of matched nodes on opposing lateral faces to impose periodic boundary conditions. General contact was defined with a friction coefficient of 0.6. The analyses were performed in Abaqus/Standard using an implicit dynamic procedure with moderate numerical dissipation. The upper plate was displaced by 3 mm, corresponding to a maximum nominal compressive strain of 0.30. Quasi-static behavior was assessed by monitoring the ratio of total kinetic energy to total internal energy, with ALLKE/ALLIE < 0.01 used as the acceptance criterion.

The nominal compressive strain and stress were calculated from the vertical displacement and reaction force at the upper-plate reference point, where compression was reported as positive.

$$\varepsilon = \frac{\left|U_3^{\mathrm{top}}\right|}{a}, \qquad \sigma = \frac{\left|RF_3^{\mathrm{top}}\right|}{a^2}. \tag{S5}$$

To assess the numerical reproducibility of the FEA, 16 structures from the original dataset were simulated. The NRMSE between the recomputed responses and the corresponding dataset curves ranged from 0.016 to 0.063, with the largest discrepancies generally occurring near the peak and in the post-buckling regime. This range was interpreted as an empirical measure of numerical variability for the meshing and FEA workflow rather than as a strict lower bound on prediction error.

Across the 200 evaluation targets, 985 shortlisted candidates were submitted for FEA, of which 981 completed successfully. FEA-best was defined as the successfully completed candidate with the lowest FEA NRMSE within each predefined shortlist. At least three successful FEA responses were available for every target.

S3 Robustness and sensitivity analyses

This section examines the robustness of the principal conclusions to training seed, forward surrogate, and geometric validity criterion. It also reports target-level paired comparisons and the generation

sensitivity to geometry coefficient dimension.

S3.1 Multi-seed robustness

To assess training variability, five Raw-11D CFM (8×512) velocity networks were trained using different random seeds. All runs shared the same training data, geometry representation, network architecture, optimization settings, evaluation targets, candidate budget, and 3D-CNN surrogate. For each seed, 20 candidates were generated for each of the 200 target responses and evaluated using the periodic-wrap criterion.

Table S2: Multi-seed robustness of Raw-11D CFM (8×512).

Training seed	Mean NRMSE	Median NRMSE	p_{90} NRMSE	Validity
42	0.07388	0.06040	0.13283	73.00%
1	0.07214	0.06031	0.12556	75.13%
2	0.07171	0.05818	0.13263	73.63%
3	0.07000	0.05751	0.12396	74.20%
4	0.07348	0.06009	0.13308	75.85%
Mean ± s.d.	0.07224 ± 0.00155	0.05930 ± 0.00135	0.12961 ± 0.00447	$74.36 \pm 1.14\%$

Across the five training seeds, the mean best-of-20 NRMSE ranged from 0.07000 to 0.07388, and candidate validity ranged from 73.00% to 75.85%. The seed-42 network used in the main comparison was neither the lowest-error nor the highest-validity run, indicating that the reported result was not obtained using an unusually favorable training seed.

S3.2 Cross-surrogate assessment

To examine whether the benchmark conclusions depended on the forward surrogate model, the same candidate sets were evaluated using two independently trained models. The primary 3D-CNN operated directly on the reconstructed 64^3 implicit geometry field and achieved a mean validation NRMSE of 0.1092. The coefficient-space MLP mapped the standardized $K = 32$ geometry coefficients to the 11-point stress response and achieved a mean validation NRMSE of 0.1290.

For DiffuMeta and Natural–Fourier candidates, the reconstructed implicit fields were projected onto the learned $K = 32$ geometry basis before evaluation by the coefficient-space MLP. Candidates generated directly in the learned geometry representation were evaluated using their coefficients. Both surrogate evaluations used the same 200 target responses with 20 candidates per target.

Table S3: Cross-surrogate assessment of the five inverse-design methods.

Method	Validity	n_{eff}	3D-CNN mean	3D-CNN median	3D-CNN p_{90}	MLP mean	MLP median	MLP p_{90}
Raw-11D CFM (8×512)	**73.0%**	200	**0.0739**	**0.0604**	**0.1328**	**0.0596**	**0.0491**	**0.1013**
Raw-11D CFM (4×256)	61.8%	199	0.0891	0.0694	0.1482	0.0861	0.0697	0.1532
Stress-conditioned CFM ($L = 4$)	61.4%	200	0.0904	0.0746	0.1813	0.0862	0.0706	0.1647
Natural–Fourier CFM	51.6%	199	0.0972	0.0828	0.1723	0.0958	0.0805	0.1726
DiffuMeta	47.9%	186	0.1182	0.0864	0.2118	0.1182	0.0838	0.1970

Although the best-of-20 NRMSE values differed between the two surrogates, the ordering by mean NRMSE remained unchanged. Raw-11D CFM (8×512) achieved the lowest mean error under

both evaluations, while DiffuMeta achieved the highest. This shows that the comparative ordering of different inverse-design methods was stable to the choice of forward surrogate.

S3.3 Alternative validity criteria

To examine whether the validity results depended on the evaluation criterion, the same candidate sets were assessed using two procedures. Under the primary periodic-wrap criterion, each reconstructed 64^3 implicit field was periodically wrapped before zero-level surface extraction, and a candidate was classified as valid when the extracted surface formed a single connected component. Under the DiffuMeta-style criterion, each candidate was projected onto the Natural–Fourier representation, the resulting coefficients were rounded to increments of 0.05, and the implicit field was reevaluated on a 20^3 grid. The extracted surface was then required to form a single connected component and span all six faces of the unit cell.

Table S4: Candidate validity under the periodic-wrap and DiffuMeta-style evaluation procedures. Difference denotes DiffuMeta-style minus periodic-wrap in percentage points.

Method	Periodic-wrap	DiffuMeta-style	Difference (percentage points)
Raw-11D CFM (8×512)	**73.000%**	**73.725%**	+0.725
Raw-11D CFM (4×256)	61.800%	63.675%	+1.875
Stress-conditioned CFM ($L = 4$)	61.400%	61.575%	+0.175
Natural–Fourier CFM	51.575%	50.700%	−0.875
DiffuMeta	47.900%	46.900%	−1.000

Relative to the periodic-wrap criterion, the DiffuMeta-style procedure changed the absolute validity rates by between −1.000 and +1.875 percentage points, while preserving the ordering of all five methods. Overall, the validity results were stable across the two evaluation procedures.

S3.4 Paired benchmark comparisons

We compared the primary Raw-11D CFM (8×512) model with each of the four alternative methods on targets for which both methods produced at least one valid candidate. For target i, the paired difference was defined as

$$\Delta_i = e_i^{(\text{other})} - e_i^{(\text{primary})}, \tag{S6}$$

where $\Delta_i > 0$ indicates lower best-of-20 NRMSE for the primary model. Mean paired differences and 95% confidence intervals were estimated using 2,000 paired bootstrap resamples. A primary-model win was recorded when $e_i^{(\text{primary})} < e_i^{(\text{other})}$.

Table S5: Target-level paired comparisons between the primary Raw-11D CFM (8×512) model and the other benchmark methods.

Other method	Paired n	Mean Δ	95% CI	Primary wins	Win fraction
Raw-11D CFM (4×256)	199	0.0160	[0.0087, 0.0239]	140	0.704
Stress-conditioned CFM ($L = 4$)	200	0.0165	[0.0115, 0.0220]	136	0.680
Natural–Fourier CFM	199	0.0230	[0.0167, 0.0295]	147	0.739
DiffuMeta	186	0.0461	[0.0304, 0.0669]	143	0.769

The primary model achieved positive mean paired differences against all four comparison methods. The mean differences ranged from 0.0160 against Raw-11D CFM (4×256) to 0.0461 against DiffuMeta, and all four 95% confidence intervals remained above zero. The primary model achieved lower best-of-20 NRMSE on 68.0–76.9% of the paired targets, indicating that the improvement was observed across a majority of the common evaluable targets.

S3.5 Geometry coefficient dimension

We evaluated geometry coefficient dimensions $K \in \{8, 16, 32, 64\}$ using the same fixed set of the held-out designs. For each design, coefficients were obtained through ridge-regularized projection with $\lambda = 10^{-5}$, and reconstruction error was evaluated on the complete 64^3 implicit field as

$$E_i^{\mathrm{geo}} = \frac{\left\| \widehat{\Psi}_i - \Psi_i \right\|_2}{\left\| \Psi_i \right\|_2}. \tag{S7}$$

Generation was evaluated for $K = 32$ and $K = 64$ using matched Raw-11D CFM models with four residual blocks of width 256. Both models were evaluated using 20 candidates for each of the 200 common evaluation targets, the periodic-wrap validity criterion, and the frozen 3D-CNN surrogate.

Table S6: Geometry reconstruction and generation across coefficient dimensions. Dashes indicate that generation was not evaluated.

K	Recon. mean	Recon. median	Recon. p_{90}	Validity	Gen. mean	Gen. median
8	0.6159	0.6185	0.9641	–	–	–
16	0.1931	0.0152	0.6257	–	–	–
32	0.0012	0.0010	0.0016	61.8%	0.0891	0.0694
64	0.0012	0.0011	0.0016	34.4%	0.1381	0.1136

Reconstruction statistics summarize the per-design relative L_2 error, while generation statistics summarize the per-target best-of-20 NRMSE.

Increasing K from 8 to 16 substantially reduced the median reconstruction error, while the mean and p_{90} errors remained high, indicating that $K = 16$ accurately reconstructed only a subset of the geometry family. At $K = 32$, all three reconstruction statistics decreased to approximately 10^{-3}, and increasing the dimension further to $K = 64$ provided no additional reconstruction benefit.

The $K = 32$ and $K = 64$ representations had comparable reconstruction accuracy, but the larger representation produced weaker generation performance. Candidate validity decreased from 61.8% to 34.4%, while the mean best-of-20 NRMSE increased from 0.0891 to 0.1381. These results support $K = 32$ as the selected geometry coefficient dimension.

S4 Additional diagnostic experiments

The standard and Matryoshka geometry representations were evaluated using the same 4×256 CFM architecture, 200 target responses, 20 candidates per target, the frozen 3D-CNN surrogate, and the periodic-wrap criterion. The whitened and field-based PCA representations were compared only through their validation flow-matching losses.

Table S7: Additional geometry-representation diagnostics. B20 denotes best-of-20 NRMSE. The standard and Matryoshka representations both had $n_{\text{eff}} = 199$. Dashes indicate that the complete candidate-generation evaluation was not performed.

Representation	Dimension	Validation FM loss	Validity	Mean B20	p_{90} B20
Standard learned representation	32	0.6742	**61.80%**	**0.0891**	**0.1482**
Matryoshka representation	32	0.5791	48.75%	0.1033	0.1921
Whitened representation	32	0.7590	–	–	–
Field-based PCA	28	0.7774	–	–	–

Although the Matryoshka representation reduced the validation flow-matching loss from 0.6742 to 0.5791, its mean best-of-20 NRMSE increased from 0.0891 to 0.1033 and its candidate validity decreased from 61.80% to 48.75%. Among the 198 targets, the mean paired NRMSE difference, defined as Matryoshka minus standard, was 0.01328. A percentile 95% confidence interval of [0.00430, 0.02106] was obtained from 2,000 paired bootstrap resamples. The standard representation achieved lower error for 128 of the 198 paired targets. Thus, a lower validation flow-matching loss did not correspond to better candidate-generation performance. Whitening the learned geometry coefficients increased the validation flow-matching loss from 0.6742 to 0.7590, while the 28-dimensional field-based PCA representation produced a loss of 0.7774.

We also compared two forward-surrogate formulations. The direct coefficient-space MLP mapped the standardized $K = 32$ geometry coefficients to the raw 11-point response and achieved a validation mean NRMSE of 0.1290. The basis-to-basis MLP instead predicted four stress-basis coefficients before reconstructing the same 11-point response and achieved a mean NRMSE of 0.1444. The intermediate response representation therefore did not improve accuracy for the fixed 11-point forward-prediction task.

References

[1] Zheng, L., Kumar, S. & Kochmann, D. M. Algebraic language models for inverse design of metamaterials via diffusion transformers. *Nature Machine Intelligence* **8**, 628–640 (2026). ArXiv:2507.15753.