CHIP PROTECTION AREA DESIGN: DATASET AND EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The protection area surrounding a semiconductor chip guards its circuit against thermo-mechanical defects induced during wafer dicing. Its 2D designs are governed by manufacturing constraints and evaluated using test chips that require production processes with substantial costs. An ideal design should achieve optimal 3D mechanical states while meeting the hard constraints. This creates a highly nonlinear optimization landscape where data-driven approaches have shown strong promise in other domains. Given the lack of large-scale protection area datasets with 3D thermo-mechanical elasto-plastic evaluation, we propose the first 1.17 TB public dataset, containing 110,170 protection area layouts across 18 metal stacks, each paired with fields from thermo-mechanical elasto-plastic finite element simulations. We demonstrate that the dataset enables two key benchmarks: design-to-fields cross-dimensional forward prediction, and fields-to-design inverse generation. The baseline results establish the difficulty of these tasks and highlight opportunities for data-driven methods to address industrial design bottlenecks. In addition to the dataset, we provide an open-source end-to-end automated pipeline from design generation to simulation for reproducibility and potential extensions.

1 INTRODUCTION

Diminishing semiconductor chip size leads to increasing thermo-mechanical vulnerability, a critical contributor to chip assembly yield loss and chip lifespan. A prime example is mechanical dicing, where diamond grits separate the chips from the wafer and induce cracks propagating inwards from the protection area at the chip edge. Designing the protective structure is a nonlinear optimization problem: a design specifies layers of 2D layout, which are subject to design rules; the resulting mechanical states, which govern defects, depend on the full 3D composite geometry whose dimensions are fixed by the given metal-stack configuration (see Section 2 for more details). As a common alternative to the expensive hardware prototyping, 3D finite element analysis (FEA) can model these mechanical states as physical fields (e.g., strain energy density and stresses) based on meaningful assumptions. However, the poor efficiency of manual design generation and the expensive runtimes of 3D FEA simulation prohibit large-scale design space coverage.

Recent work in machine learning (Qiao et al., 2024; Park et al., 2026; Chen et al., 2026) develops thermo-mechanical simulation surrogates using either small-scale or private datasets. Existing public benchmarks for thermo-mechanical structural mechanics (Felten et al., 2025; Casenave et al., 2025) contain relatively few designs. Consequently, the lack of large-scale public datasets limits the development of data-driven approaches, a necessary complement to classical numerical methods in this domain. In this paper, we fill this gap with a large-scale dataset containing 110,170 designs across 18 metal-stack configurations paired with simulation results under thermo-mechanical loads.[1] To enable reproducing and extending the dataset, we also provide an open-source pipeline from design generation to thermo-mechanical simulation. Both the design rule checker and the simulator

[1]Due to hosting platform policies, it is not possible to properly anonymize the full dataset, which will therefore only be released upon acceptance. During review phase, the full metadata will be released with the code, and therefore one can re-generate the entire dataset given sufficient computation resources. For exploratory purposes, we prepare an anonymized small subset with 500 designs at `https://figshare.com/s/eed6f9b5abb5984fb519`.

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.

are validated separately against two industrial standard commercial software packages, namely the Cadence Virtuoso Layout Suite and Ansys Mechanical.

Most existing work learns mappings within a single modality, for example, predicting 3D stress fields from 3D meshes. However, in manufacturing settings, the primary design artifact is a 2D layout despite the robustness is tested in 3D. In view of this, we introduce two complementary benchmarks that model the relationship between the 2D design space and the 3D physical field space. The first task, design-to-fields, requires to learn mapping from 2D layout specifications to the 3D physical fields. The evaluations accept mesh and mesh-free representations by assessing the predictions on chosen spatial coordinates and regional field descriptors. The second task, fields-to-design, is a generative task where a model should produce valid 2D layouts based on field descriptors while fulfilling the hard constraints. A generated design will be evaluated for manufacturability using the design rule checker. If the design is valid, it will subsequently be fed into the simulator. Instead of layout reconstruction loss, this task assesses the accuracy of field descriptors derived from the simulation result. The baselines suggest much effort is needed to develop data-driven methods as effective approaches for industrial bottlenecks.

Our contributions can be summarized as follows: 1) We curate the first large-scale 1.17 TB public dataset comprising mostly manufacturable 2D semiconductor protection area designs and corresponding selected 3D fields from thermo-mechanical elasto-plastic finite element simulations that mimic manufacturing processes; 2) On this dataset, we propose two benchmarks with open-source evaluations to assess performance in forward prediction and inverse generation for industrial objectives; and 3) We provide an open-source data generation workflow implementation, including a design sampler and simulator. Both are validated against standard industrial commercial software for reproducibility and extensibility.

2 Preliminaries & Related Work

Due to the multi-disciplinary nature of this work, we provide a brief background of relevant ideas and terminology. We first introduce the terminology, concepts, and descriptions in the semiconductor industry, emphasizing the perspectives of engineering design and thermo-mechanical simulation; followed by the literature review in machine learning, consisting of three main areas: datasets and benchmarks, machine learning surrogates for thermo-mechanical elasto-plastic simulation, and inverse generation for engineering design.

Chip, Technology Node & Metal Stacks. A Complementary-Metal-Oxide-Semiconductor (CMOS) *chip* is a 3D stack of patterned metals embedded in dielectrics, manufactured layer by layer on a substrate through repeated cycles of material deposition, lithography, and etching to construct active devices and establish interconnectivity. Multiple chips are fabricated in parallel on a single wafer and split through separation processes such as mechanical dicing. A *technology node* defines a characteristic feature size along with a specific set of manufacturing processes and design constraints. Within a technology node, the interconnect between devices and I/O pads uses a predefined set of *metal stacks*, which are also used for the construction of the protection area construction. Each metal stack is a vertical stack of *meta-layers*, consisting of a metal layer and a via layer, each embedded in a corresponding dielectric layer. Metal stacks share the same xy footprint but differ in height (z) due to varying meta-layer combinations. Throughout this paper, we use *metals* to collectively refer to metal and via layers when discussing material properties, as both use ductile metallic materials (copper and aluminum) contrasting with brittle *dielectrics* (silicon oxide, silicon nitride, and silicon); we distinguish *metal layers* and *via layers* when discussing design constraints.

Protection Area. The protection area surrounding the chip circuit always contains a functional seal ring to protect the active area against ion contamination while acting as a moisture barrier. Additionally, the metals can be deposited in the remaining area, following the layout design, to enhance robustness against thermo-mechanical defects. Beyond the metal and via layers, there are two additional features for the protection area: the *optional* passivation groove and the *mandatory*[2] assessment zone with fixed size but variable position along the y-axis. By inducing a controlled fracture at a predetermined location, the passivation groove arrests cracks before they propagate into the active area. The assessment zone is a metal-free region that enables optical detection of

[2]Only in the context of this paper.

dicing-induced defects. A chip with any defect crossing the assessment zone is considered a failed product.

Design Constraints. The metals are the design variables. On a 2D design layout, an arbitrary number of axis-aligned rectangles can be inserted in each layer, subject to design constraints including: minimum and maximum width, minimum metal area, minimum spacing (a piecewise function of width), and minimum via enclosure. A via rectangle must be enclosed by metal rectangles in both the adjacent top and bottom layers. A passivation groove is continuous along the x-axis, spanning the full perimeter of the protection area, and has a fixed depth from the top of the chip to the top of the second-to-last metal layer. The structures of metal layer and via layer must not be coincident with the boundary of a passivation groove to prevent exposure. These design constraints define the valid design space within which metal arrangements must be optimized for thermo-mechanical performance.

Manufacturing Processes & Mechanical States. During material deposition, thermal cycling (e.g., cooling from 400°C to 22°C) induces intrinsic residual mechanical stresses due to mismatched thermal expansion coefficients between materials. These residual stresses undergo further nonlinear transformation when the protection area contacts the diamond grits on dicing blades during mechanical separation. Since strain energy density (W) and the maximum principal tensile stress (S_1) are well-established failure indicators for brittle materials (Li, 2001; Rankine, 1857), and the assessment zone contains only brittle materials, these two quantities are of interest in this context. A single metal feature can significantly alter the local state distribution due to its differing thermal expansion coefficient and stiffness relative to surrounding dielectrics. As a result, the metal arrangement is the primary design lever governing the mechanical states, making it both the input to the forward prediction problem and the output of the inverse design problem.

Datasets and Benchmarks. Recent attention on data-driven approaches for physical simulations has catalyzed the creation of several large-scale datasets with many in the domain of computational fluid dynamics (Hassan et al., 2023; Ashton et al., 2024; Ohana et al., 2024; Kanchi et al., 2025; Qiu et al., 2025) or atmospheric science (Ashkboos et al., 2022; Kaltenborn et al., 2023; Nguyen et al., 2023; Nathaniel et al., 2024; Aouni et al., 2025). Felten et al. (2025) and Casenave et al. (2025) include some subsets of thermo-mechanical simulations but fail to account for plasticity—which gives rise to material non-linearity—in ductile materials. Lejeune (2020) proposes Mechanical MNIST digits, which model the large mechanical deformation on MNIST using 2D FEM without considering any thermal effect. A few datasets (Kim et al., 2023; Zheng et al., 2023; He et al., 2025; Wang et al., 2025) for adjacent subdomains of the semiconductor industry have been proposed in recent years, underscoring data scarcity and the importance of data-driven approaches in the industry.

ML Surrogates for Thermo-Mechanical Simulations. Graph neural networks have demonstrated remarkable success as mesh-based physics surrogates for fluids and soft bodies; however, existing applications to thermo-mechanical problems typically neglect plasticity and consider only simple 3D geometries with homogeneous materials (Pfaff et al., 2020). Mozaffar et al. (2019) use recurrent networks to learn path-dependent plasticity from loading histories, while Mianroodi et al. (2021) adopt an image-based deep learning approach to capture elasto-plastic behavior. Both operate on simplified 2D geometry with strong emphasis on external loads. Regarding operator-learning approaches, Park et al. (2026) propose sequential DeepONets to capture thermal-viscoplastic behavior on one geometry; Chen et al. (2026) predict the thermo-mechanical distortion in 21 additive-manufactured parts using graph neural operators. Implicit neural fields are emerging as a promising representation for thermo-mechanical problems. Zong et al. (2023) model elastoplasticity and fracture using small scene-specific sets of training trajectories; Yamazaki et al. (2026) study thermoelastic operator learning using 5,000 generated coefficient 2D and 3D fields; and Thiamchaiboonthawee et al. (2026) predict post-cooling displacement from 164 one-way coupled thermo-elastoplastic simulations. In the semiconductor domain specifically, Qiao et al. (2024) adopts an image-partition approach to predict the vector of equivalent thermo-mechanical properties for 33k patch samples using a combination of transformer and convolutional neural networks.

Inverse Generation for Engineering Design. Generative models have emerged as powerful tools for engineering design. Several representative works (Mazé & Ahmed, 2023; Regenwetter et al., 2024; Etesam et al., 2025; Wang et al., 2026) propose different deep generative architectures but are restricted to low-dimensional parameterized domains or small 2D domains, with small or moderate data sizes. Several recent surveys (Regenwetter et al., 2022; Tanveer et al., 2026) report that scarcity

in engineering design and diversity in engineering data remain among the biggest bottlenecks, despite the rapid development of data-driven approaches. Regenwetter et al. (2022) emphasizes that feasibility, constraints, and manufacturability are central evaluation criteria for engineering design generation. Li et al. (2023) acknowledge this emphasis as the limitation in realism of large-scale conceptual designs (for example, sketches (Google Creative Lab, 2017) or CAD models (Wu et al., 2015)), while pointing out the importance of cross-modality datasets and tasks—a gap our benchmark directly addresses with its 2D layout to 3D stress field pairing. Picard et al. (2023) provides guidelines for generating synthetic datasets for engineering designs, focusing on data realism and test-set representativeness.

3 DATASET

We create a synthetic technology node derived from a confidential CMOS node to avoid intellectual-property restrictions. The design rules are sampled within the bounds of the real node, ensuring manufacturing feasibility. We define 18 industry-common metal-stack configurations (combo), as illustrated in Figure 1C (see Appendix A for more details). All configurations have four materials: Copper (Cu), Silicon Nitride (SiN), Silicon Oxide (SiO), and Silicon (Si), while configurations with topside metal have an additional material, Aluminium (Al). A design is always stacked bottom-up, following the order of meta-layers: fine pitch, double pitch, and, lastly, topside metal. By adopting the industrial assumption of design-pattern periodicity throughout the chip periphery, we focus on a single unit cell to reduce computational demand using only a fraction of the total geometry dimensions. Across 18 combos, the unit cells have the same width (x) and length (y), but different height (z) due to the number of layers. The per-configuration count (Figure 1C) is allocated proportionally to the number of layers, as more layers span a larger design space.

2D Layout. A design layout is a stack of 2D xy-plane describing the design patterns of metals, vias, trenches, and diffusion (see Figure 1A). The design patterns for each layer contain an arbitrary number of axis-aligned rectangles. For engineering design, a design that fulfills all rules is naturally deemed valid, feasible, and manufacturable. We develop a probabilistic approach for large-scale design generation. The algorithm (Appendix B) incorporates design rules as sampling bounds, ensuring a validity of 82.78% for $2\mu m$ width. The imperfect validity makes it feasible to use our dataset to extend the work of Regenwetter et al. (2024), which emphasizes the value of negative data in generative engineering design. The layout is stored as a binary file, namely `GDSII`, and can be accessed using various open source Python packages, such as KLayout (Koefferlein, 2019).

3D Mechanical States. The 3D geometry of a design is, globally, a cuboid with conformal interfaces, consisting of layers of dielectrics on top of a 10 μm silicon bar. The dielectric layers contain metal structures, which are the z-extrusions of the layout, following the pre-defined layer thicknesses. By running thermo-mechanical simulations on the geometry, we approximate the physical states of a design during manufacturing processes using a thermal load step (T), and subsequently three branched mechanical load steps ($\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3$). The first load step T simulates cooling from 400°C to 22°C during thermal cycling in the manufacturing process. Figure 1B (top right) shows the strain energy density under thermal load. Continuing from the thermally loaded state, the simulator applies mechanical loads of 20 GPa pressure *independently* on three representative $1\mu m^2$ areas on the outer side wall, approximating the mechanical interaction of the dicing process. As shown in Figure 1B (bottom) using red arrows, the load positions are z_{M_1} ($-1.5\mu m$, near the top of silicon bar), z_{M_2} ($0.5\mu m$, the bottommost of silicon oxide layer), and z_{M_3} (depends on stack configurations, the top of penultimate metal layer). The remaining simulation configuration for material models, boundary conditions, and solver can be found in Appendix C. The simulation results (stored as `NumPy NPZ` file) contain nodal displacements, Gauss-point plastic strains, and backstress tensors across four load steps (T, M_1, M_2, M_3). These fundamental quantities, together with the constitutive model and mesh geometry, are sufficient to reconstruct the full mechanical state, including Cauchy stress components, principal stresses, and strain energy densities.

Reproducibility & Extensibility. The entire pipeline—from design sampling to simulation—is open-source. For layout validation, we extend KLayout's design rule checker to cover the edge cases of our technology node, validated by domain experts against the Cadence Virtuoso Layout Suite. For simulation, we implement an elasto-plastic solver using `NumPy` (Harris et al., 2020), `SciPy` (Virtanen et al., 2020), and `scikit-sparse` (Cournapeau et al., 2008), validated against

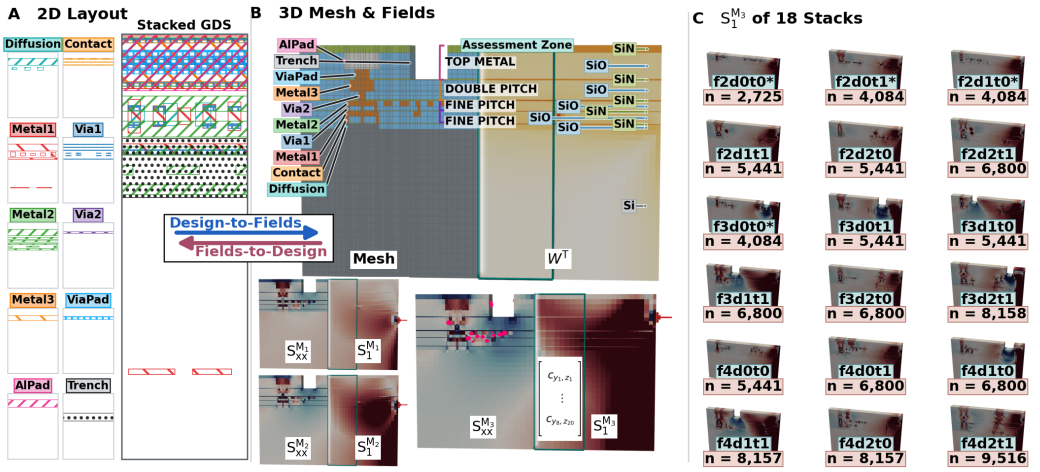

Figure 1: **A:** The miniature layouts on the left are per-layer pattern views, which are stacked together as shown on the right. **B:** The top view illustrates the 3D view of the design in A, in the form of mesh and strain energy density under thermal load (W^T). The annotation columns, from left to right, are patterned layers described in GDS, meta layers, and stack-specific dielectric layers. The assessment zone is bounded by a green box. The bottom views show the resulting $S_{xx}^{M_i}$ and $S_1^{M_i}$ fields at different mechanical load positions M_i (red arrows on the right edges). For W^T, the yellow intensiveness indicates the magnitude; while for $S_{xx}^{M_i}$ and $S_1^{M_i}$, the intensiveness of red and blue represents positive (tensile) and negative (compressive) magnitude. On the bottom right, the probe points and the coefficient vector in the assessment zone symbolize evaluations for the benchmarks, which are shown as bi-directional arrows between A and B. **C:** Each miniature 3D view is a sample from the named stack with annotated corresponding sample size. A total of 500 samples from stacks with asterisk (*) are included in the exploratory dataset.

Ansys Mechanical. Both tools are scoped to our use case rather than being general-purpose. A Croissant metadata file is provided for machine-readable dataset discovery and loading. The generation process takes 85 hours on 350 CPU cores, producing 110,170 designs (1.17 TB) with 15k–140k hexahedral elements each.

With hard manufacturing constraints, elasto-plastic simulation, and distinct metal-stack configurations, our dataset addresses the needs for engineering design—namely feasibility, realism, and representative test sets—as identified by Regenwetter et al. (2022), Picard et al. (2023), Li et al. (2023), and Tanveer et al. (2026), outlined in Section 2.

4 Benchmarks

The dataset enables the formulation of two machine learning problems, supporting design and evaluation cycles. The forward problem, namely design-to-fields, requests the mapping from the 2D design space to a set of 3D mechanical states. The returning route, fields-to-design, tasks a model to learn the underlying joint distribution of state space summary and the design space across metal-stack configurations, while adhering to the given technology specifications. We provide an open-source evaluation pipeline for both benchmarks, as well as the source code for all baselines.[3]

4.1 Design-to-Fields

Most existing work learns mappings within a single modality. For example, predicting 3D stress fields from 3D meshes. However, in manufacturing settings, the primary design artifact is a 2D layout that describes only ductile structures in 3D, while the technology configuration specifies the

[3] Available at `https://figshare.com/s/c155772e378c74c99549`

remaining bodies. One of the most common simulation approaches, finite element analysis, requires spatial discretization of the 3D geometry (a 3D mesh), posing a typically nontrivial trade-off between computational cost and simulation accuracy. Finding the balance could become a bottleneck, depending on domain expertise, geometric constraints, and prior assumptions about the thermo-mechanical response. A simulation result is not always sensible; a common quick assessment is to evaluate S_1 tensile stress hotspots near ductile-brittle material interfaces, resulting from differences in coefficients of thermal expansion. The quality-defining factor for a design, however, is defects in the assessment zone, which can be loosely approximated by the pattern-derived metrics based on W and S_1.

To maximize the advantage and compatibility of data-driven approaches, this task concerns field modeling *without* 3D spatial representation restriction based on the 2D layout, mechanical load positions, and layer specification, including thickness, elevation, and material of *all* brittle and ductile layers, according to the given combo configuration. Note that the thermal load (400°C to 22°C for the entire dataset) is not provided as input. Formally, let d denote the 2D design representation (e.g., a set of rectangle coordinates, a rasterized layout, or a sequence of design tokens), $\mathbf{h} \in \mathbb{R}^K$ the layer specifications, and $z_{\mathrm{M}_i} \in \mathbb{R}$ the load position of the mechanical steps on z-axis. The task is to learn a mapping

$$f : (d,\ \mathbf{h},\ z_{\mathrm{M}_i},\ p) \longmapsto \hat{\boldsymbol{\psi}}(p), \quad \forall\, p \in [0,1]^3, \tag{1}$$

where $\hat{\boldsymbol{\psi}} = \{\hat{\psi}_{\mathrm{W}}^{\mathrm{T}}, \hat{\psi}_{\mathrm{S_1}}^{\mathrm{M}_i}, \hat{\psi}_{\mathrm{S_{xx}}}^{\mathrm{M}_i}\}$ denotes the predicted physical quantity queried at arbitrary point p. We denote f as a point-wise mapping for simplicity. This notation is directly applicable to mesh-free approaches such as implicit neural fields; for mesh-based approaches, f is implemented as $\mathcal{Q}(\hat{f}, p)$, where $\mathcal{Q} : (\Omega_{\hat{\boldsymbol{\psi}}}, p) \longmapsto \hat{\boldsymbol{\psi}}(p)$ is a deterministic grid sampler that converts the given discretized fields in form of mesh $\Omega_{\hat{\boldsymbol{\psi}}}$ into a flow-field before probing p; and $\hat{f} : (d,\ \mathbf{h},\ z_{\mathrm{M}_i}) \longmapsto \Omega_{\hat{\boldsymbol{\psi}}}$ is the model that demonstrates spatial-physical understanding by generating a valid mesh containing the fields of interest. We particularly encourage the development of a mesh-based approach to generate an FEA-verifiable mesh for added reliability.

Point-wise coordinate query is the fundamental form of evaluation to embrace different representations: for every design d (after normalizing xyz-axis to [0,1]), we select a fixed set $\mathbf{P}_d$ of 256 3D coordinate points, where the points are narrowly separated by the brittle-ductile material interfaces. The normalized root mean squared error (NRMSE) is defined as

$$\mathrm{NRMSE} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sqrt{\frac{1}{|\mathbf{P}_d|} \sum_{p \in \mathbf{P}_d} \left(\frac{\psi(p) - \hat{\psi}(p)}{\sigma_{\psi}^{\mathrm{train}}} \right)^2}, \tag{2}$$

where $\sigma_{\psi}^{\mathrm{train}}$ is the selected quantitle (ψ) standard deviation of train set. The NRMSE of five physical quantities is evaluated: strain energy of thermal load (W^{T}), maximum principal stress of all three mechanical steps ($\mathrm{S}_1^{\mathrm{M}_1}$, $\mathrm{S}_1^{\mathrm{M}_2}$, $\mathrm{S}_1^{\mathrm{M}_3}$) and normal stress component in x-axis of M_3 ($\mathrm{S}_{\mathrm{xx}}^{\mathrm{M}_3}$). In contrast to the two failure indicators W and S_1, S_{xx} in our simulation setup reflects a sharp discontinuity at the material interfaces, making it a natural quantity for evaluating the model's capability to comprehend local material boundaries.

A **field descriptor** is our proposed approach to evaluate the spatial pattern of the quantity of interest in the assessment zone (az) of a design. Leveraging the point-wise coordinate query, we specify a set of anchor points $\{p\}_{\mathrm{az}}$ located in the assessment zone to approximate the Gaussian-smoothed fields using a volume-weighted least-square fitted B-spline with 160 coefficients, and three spline degrees. The worst fit across the entire dataset reports a volume-weighted Pearson correlation of 0.9963 and an NRMSE of 0.0863. We deem the vector of B-spline coefficients ($\mathbf{c}$) a reasonable representation of the smoothened fields; hence, the normalized mean absolute error (NMAE), defined as

$$\mathrm{NMAE} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{1}{|\mathbf{c}_d|} \sum_{c \in \mathbf{c}_d} \frac{|\hat{c} - c|}{\sigma_c^{\mathrm{train}}}, \tag{3}$$

between the coefficients fitted to the ground-truth field and to the model's predictions at the same anchor points is the performance metric.

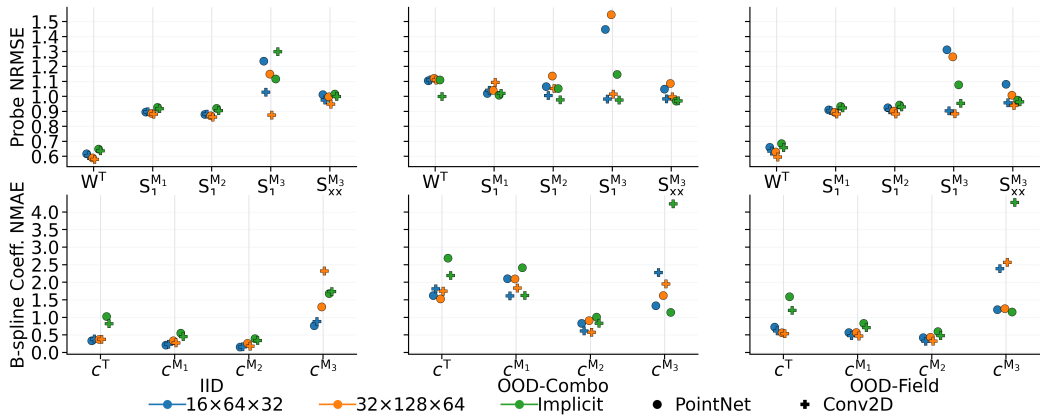

Figure 2: **Top:** The NRMSE 256 selected points of all test set designs over three splits, evaluated based on five physical quantities across four load steps. **Bottom:** The NMAE of coefficients fitted using least square over the fields of the assessment zone. **Color:** The $x \times y \times z$ legend entries (blue, orange) represent the grid sizes for the mesh-based approach, while implicit (green) indicates mesh-free models. **Metrics:** The lower the better.

We evaluate generalization along three axes: external loading, global geometry, and design layout. A common condition applies to all three splits: training uses only z_{M_1} and z_{M_2} load steps, while z_{M_3} is held out exclusively for testing. This probes the model's ability to capture the non-linear mechanical response signal, which is amplified under M_3 (see Figure 1B bottom)—particularly for taller stacks as illustrated in the Figure 1C. The first (**IID**) tests whether a model extrapolates to an unseen mechanical load position by partitioning the dataset by stratified random sampling across combos. The second (**OOD-Combo**) probes transferability across different metal-stack heights by using a test set of 6,000 designs from the three shortest combos and three tallest. The third (**OOD-Field**) assesses reliability on mechanically atypical designs using a test set comprising top 10% most distant Mahalanobis distance in terms of coefficients (field descriptors) per combo. We implement six baseline variants with **pointnet / conv2d encoders**, **mesh / mesh-free decoders**, and **coarse / medium resolutions** for the mesh decoder. The encoders examine the contribution of input intra-layer spatial dependencies, while the decoders provide insights into different spatial representations, with the impact of mesh discretization reflected in the resolutions. The meshes are regular voxel grids with xyz-dimension of $(16 \times 64 \times 32)$ and $(32 \times 128 \times 64)$. The baselines are trained with 20 epochs using an initial learning rate of 1e-3. 18 training variants take approximately 100 hours in total on an Nvidia A100. The split sizes, model architectures, training configurations and result tables are provided in the Appendix D.

Figure 2 summarizes the performance of the six baselines across three splits. **On the split level,** IID and OOD-Field show similar probe NRMSE, while OOD-Field generally yields larger coefficient NMAE, indicating that atypical regional fields may retain familiar local interface behavior. OOD-Combo is substantially harder across both metrics, identifying transfer across unseen combo configuration as the main generalization challenge. Thermal prediction is particularly sensitive to this shift, suggesting strong dependence on global geometry. **On the load-step level,** M_1 and M_2 have similar probe accuracy, although M_2 often has lower coefficient error, suggesting the plausibility of load influence masking distinct layout geometry. For M_3, several models perform well at interface probes but deteriorate substantially over the field descriptors. This confirms that local probe accuracy does not guarantee faithful field prediction. **On the model level,** conv2d generally performs well at probes, whereas pointnet often preserves M_3 field descriptors more reliably under distribution shift. Finer grid resolution tend to improve probe accuracy but do not consistently improve coefficient prediction, while implicit decoders provide no uniform advantage and depend strongly on the encoder, despite being free from explicit spatial discretization. The substantial normalized errors and mechanical MAEs of hundreds of MPa (see Probe MAE result table in Appendix D) demonstrate considerable headroom for better data-driven approaches.

Table 1: Inverse-design results on 5,000 targets per split. NMAE is computed only over the design rule check (DRC) valid designs.

Split	Baseline	DRC Valid (↑)	NMAE (↓)
IID	NN same-combo	**4,109**	**0.0772**
	NN cross-combo	40	0.2795
	ARG	704	0.2741
	ARG-Joint	609	0.2710
OOD-Combo	NN cross-combo	**114**	**0.6039**
	ARG	0	N/A
	ARG-Joint	2	1.1644
OOD-Field	NN same-combo	**4,171**	**0.1278**
	NN cross-combo	119	0.3563
	ARG	721	0.4386
	ARG-Joint	669	0.4808

4.2 FIELDS-TO-DESIGN

The inverse task reverses the direction: given desired states under the loads, produce a layout that realizes them. Unlike the forward problem, the inverse mapping is one-to-many—many distinct layouts could produce similar fields—and the output must satisfy hard manufacturing constraints (design rules). Formally, let $\mathbf{C} = \{\mathbf{c}_{\mathrm{T}}, \mathbf{c}_{\mathrm{M}_1}, \mathbf{c}_{\mathrm{M}_2}, \mathbf{c}_{\mathrm{M}_3}\} \in \mathbb{R}^{4\times 160}$ denote the target coefficient matrix, the task is to learn a conditional generator

$$G : (\mathbf{C}^*, \mathbf{h}) \longmapsto \hat{d}, \tag{4}$$

that produces a valid design $\hat{d}$ whose simulated field descriptors approximate the query $\mathbf{C}^*$. The generator must additionally pass the design rule check (DRC). This formulation provides a potential impact in quick robustness generalization to other configurations—assuming that $\mathbf{C}^*$ is a robust response against defects, enabling the possibility to minimize the costly hardware verification. In our context, geometry-wise comparison between generated and reference designs is not meaningful, as long as the generated designs fulfill the functional quality, which we evaluate through two metrics: the **DRC validity** and the **NMAE** of between target descriptors and the descriptors obtained from generated design.

Utilizing the three test sets specified in section 4.1, we sample 5000 designs each as targets. We provide four baselines using two approaches: retrieval and generative. The retrieval baselines search in the train set of **same-combo** (only for IID and OOD-Field) and **cross-combo** for designs with lowest descriptor NMAE against that of the targets. For cross-combo matches, geometry is transferred between identically named layers of source and target configurations; source-only layers invalidate the candidate, whereas target-only layers remain empty. The generative baselines (**ARG** and **ARG-Joint**) share an autoregressive setup by tokenizing the rectangles in layout. The target field descriptors and the target stack configurations are fed as input to decoder heads for predicting new designs and corresponding assessment zones. We enable ARG-Joint to be field-aware by adding a decoder head to predict field descriptors with a loss weight of 0.5. Across all splits, the retrieval baselines take around 47 seconds on 4 cores CPU while the generative baselines require approximately 97 hours for 100 epochs on an A100 GPU. All new designs are subject to DRC for validity, and the valid designs are simulated to obtain field descriptors, using same computational resources as detailed in Section 3. The model architectures, training configurations, per-combo breakdown tables and selected generated samples can be found in Appendix E.

Table 1 shows that, **on the split level**, this benchmark reflects a similar difficulty pattern to Design-to-Fields, in the order of IID, OOD-Field, OOD-Combo. The industrial bottleneck of generalizing proven designs across different configurations, as represented by OOD-Combo, seems to be a very hard task with critically poor performance from the baselines. **On the model level,** the retrieval baselines mostly outperform the generative baselines for both metrics. The same-combo variant, despite being trivial, sets a strong baseline for IID and OOD-Field. It is encouraging to see that ARG can already produce DRC valid designs without inducing the rules as explicit biases. With the addi-

tional feedback from field loss, ARG-Joint achieves slight improvement over IID NMAE at the two costs: fewer valid designs for IID and OOD-Field, and greater sensitivity to atypical fields. Overall, the baselines suggest generative approaches can produce some manufacturable designs under hard constraints, though reliability remains far from consistent.

5 LIMITATIONS

The proposed **dataset** contains the protection area design of 18 metal stacks from a synthetic technology node with 5 fixed materials (Cu, Al, SiN, SiO, Si). This is by no means a comprehensive coverage of semiconductor chip technology nodes and does not support the more common ML applications in semiconductors, such as floor planning, circuits, and routing, which emphasize electrical properties rather than mechanical defects. The sampled designs contain biases arising from choices of certain sampling distribution parameters and do not include complex patterns constructed from polygons. The simulation results assume fixed thermal loading from 400°C to 22°C and a constant pressure of 20 GPa across three positions. These assumptions can easily be violated with a technology node that uses different manufacturing processes, and in reality, both the pressure magnitude and positions should be modeled as distributions, which are expensive and difficult to establish empirically, rather than point estimates. The elastic assumption for brittle materials is often invoked to improve simulation stability and reduce computational costs. The proposed **benchmarks** do not yet showcase all information available in the dataset, including plastic states and negative DRC designs. Those remain useful for future development. The benchmark evaluations also do not enforce physics equilibrium, although we strongly encourage researchers to take this into account. The spatial domain is assumed to be regular and normalized, limiting the potential of transfer learning for irregular geometry. The simulation ground truths are conditioned on the mesh algorithm and chosen spatial resolution. The use of a smoothing kernel before B-spline fitting eliminates local field discontinuities, which are masked out in most aggregations, caused by very thin layers. The fields-to-design benchmark uses a relatively small test set containing only 5,000 samples to reduce the required computation resources for re-running large-scale finite element simulations during test-time evaluation.

6 CONCLUSION

The design and evaluation of chip protection area pose a unique perspective in which data-driven approaches are a natural fit for addressing cross-modality mapping governed by manufacturing constraints and thermo-mechanical physics. This paper fills a critical gap not only for semiconductor but also for general engineering design, with a large-scale dataset containing 1.17 TB of 2D design layouts and the corresponding 3D thermo-mechanical simulations with elasto-plastic properties. Additionally, the data generation pipeline is fully open-source, with key components such as a design rule checker and a simulator verified against the industrial counterpart within the scope of our context. The proposed benchmarks, namely design-to-fields and fields-to-design, mimic the design optimization cycle with an industrial objective in mind. The forward prediction task uses point-wise coordinate queries to support diverse 3D spatial representations while preserving regional field assessment in the form of B-spline coefficients fitted to a set of anchor probes. The baselines reveal preliminary insights about the spatial correlations of 2D layout input (encoders), explicit and implicit spatial representations (decoders), and explicit spatial resolution (grid sizes), and indicate that more sophisticated data-driven approaches are needed to reliably serve as good surrogates. It is especially interesting to observe that good performance on local probes does not guarantee similar results when modeling the physical fields by predicting B-spline coefficients. The inverse generative benchmarks prioritize functional equivalence over geometrical identity by evaluating the resulting field descriptors of the generated design layout instead of reconstruction discrepancy. A trivial nearest-neighbor retrieval sets a strong baseline for generative models to meet functional targets within manufacturing constraints. Future work includes modeling design layout as heterogeneous graphs, improving simulation realism, and expanding data coverage with diverse materials and stack configurations.

AI USE STATEMENT

In this work, we used generative AI tools for coding assistance, including implementing methods, refactoring codes, and writing documentations. We have not used generative AI tools to directly

generate synthetic data sets, help develop conceptual framework, design or provide feedback on research methodology and experiments, and interpret results. All tasks related to mathematical claims or proofs are not applicable to this work. Additionally, we used generative AI tools to identify relevant seed literature in different domains and polish writing. We have reviewed all AI-assisted work, including reviewing the codes, verifying outputs, reading the identified seed literature and manually search further literature, and revise the text multiple times. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

To ensure reproducibility, our main contributions including datasets, data generation pipeline, and the evaluations are open-source. Additionally, the baseline implementations and checkpoints are released on Figshare. The descriptions and configurations are written in the appendix, while the required computation resources are specified in Section 3 and Section 4. Moreover, it is possible to extend these assets for further use cases and other purposes to advance machine learning research.

References

Anass El Aouni, Quentin Gaudel, Juan Emmanuel Johnson, Regnier Charly, Julien Le Sommer, Van Gennip, Ronan Fablet, Marie Drevillon, Yann Drillet, and Pierre Yves Le Traon. OceanBench: A Benchmark for Data-Driven Global Ocean Forecasting systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

Saleh Ashkboos, Langwen Huang, Nikoli Dryden, Tal Ben-Nun, Peter Dominik Dueben, Lukas Gianinazzi, Luca Nicola Kummer, and Torsten Hoefler. ENS-10: A Dataset For Post-Processing Ensemble Weather Forecasts. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, June 2022.

Neil Ashton, Jordan B. Angel, Aditya Ghate, Gaetan Kenway, Man Long Wong, Cetin C. Kiris, Astrid Walle, Danielle C. Maddix, and Gary Page. WindsorML: High-Fidelity Computational Fluid Dynamics Dataset For Automotive Aerodynamics. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024.

Fabien Casenave, Xavier Roynard, Brian Staber, William Piat, Michele Alessandro Bucci, Nissrine Akkari, Abbas Kabalan, Xuan Minh Vuong Nguyen, Luca Saverio, Raphaël Carpintero Perez, Anthony Kalaydjian, Samy Fouché, Thierry Gonon, Ghassan Najjar, Emmanuel Menier, Matthieu Nastorg, Giovanni Catalani, and Christian Rey. Physics-Learning AI Datamodel (PLAID) datasets: A collection of physics simulations for machine learning, May 2025.

Zhonghao Chen, Haochen Mu, Zhao Zhang, Lei Yuan, Donghong Ding, Hongtao Zhu, Ninshu Ma, and Zengxi Pan. Graph neural operator: A DeepONet-based framework for learning thermo-mechanical distortion in metallic additive manufacturing. *Additive Manufacturing*, 115:105063, January 2026. ISSN 2214-8604. doi: 10.1016/j.addma.2025.105063.

David Cournapeau, Nathaniel Smith, Dag Sverre Seljebotn, Leon Barrett, Yuri, Antony Lee, Alex Grigorievskiy, Joscha Reimer, Justin Ellis, Aaron Johnson, and Bernard Roesler. Scikit-Sparse, 2008. URL `https://github.com/scikit-sparse/scikit-sparse`.

Yasaman Etesam, Hyunmin Cheong, Mohammadmehdi Ataei, and Pradeep Kumar Jayaraman. Deep Generative Model for Mechanical System Configuration Design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16496–16504, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i16.33812.

Florian Felten, Gabriel Apaza, Gerhard Bräunlich, Cashen Diniz, Xuliang Dong, Arthur Drake, Milad Habibi, Nathaniel James Hoffman, Matthew Keeler, Soheyl Massoudi, Francis VanGessel, and Mark Fuge. EngiBench: A Framework for Data-Driven Engineering Design Research. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

Google Creative Lab. Quickdraw-dataset, 2017. URL https://github.com/googlecreativelab/quickdraw-dataset.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

Sheikh Md Shakeel Hassan, Arthur Feeney, Akash Dhruv, Jihoon Kim, Youngjoon Suh, Jaiyoung Ryu, Yoonjin Won, and Aparna Chandramowlishwaran. BubbleML: A Multiphase Multiphysics Dataset and Benchmarks for Machine Learning. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Hongquan He, Zhen Wang, Jingya Wang, Tao Wu, Xuming He, Bei Yu, Jingyi Yu, and Hao Geng. LithoSim: A Large, Holistic Lithography Simulation Benchmark for AI-Driven Semiconductor Manufacturing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

Julia Kaltenborn, Charlotte Emilie Elektra Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Rohit Sunil Kanchi, Benjamin Melanson, Nithin Somasekharan, Shaowu Pan, and Sicheng He. UniFoil: A Universal Dataset of Airfoils in Transitional and Turbulent Regimes for Subsonic and Transonic Flows. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

Geonu Kim, Byunggook Na, Gunhee Kim, Hyuntae Cho, Seungjin Kang, Hee Sun Lee, Saerom Choi, Heejae Kim, Seungwon Lee, and Yongdeok Kim. Benchmark of Machine Learning Force Fields for Semiconductor Simulations: Datasets, Metrics, and Comparative Analysis. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Matthias Koefferlein. KLayout Layout Viewer And Editor, 2019. URL https://www.klayout.de.

Emma Lejeune. Mechanical MNIST: A benchmark dataset for mechanical metamodels. *Extreme Mechanics Letters*, 36:100659, April 2020. ISSN 2352-4316. doi: 10.1016/j.eml.2020.100659.

Q. M. Li. Strain energy density failure criterion. *International Journal of Solids and Structures*, 38 (38):6997–7013, September 2001. ISSN 0020-7683. doi: 10.1016/S0020-7683(01)00005-1.

Xingang Li, Ye Wang, and Zhenghui Sha. Deep Learning Methods of Cross-Modal Tasks for Conceptual Design of Product Shapes: A Review. *Journal of Mechanical Design*, 145(4), April 2023. doi: 10.1115/1.4056436.

François Mazé and Faez Ahmed. Diffusion Models Beat GANs on Topology Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9108–9116, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i8.26093.

Jaber Rezaei Mianroodi, Nima H. Siboni, and Dierk Raabe. Teaching solid mechanics to artificial intelligence—a fast solver for heterogeneous materials. *npj Computational Materials*, 7(1):99, July 2021. ISSN 2057-3960. doi: 10.1038/s41524-021-00571-z.

M. Mozaffar, R. Bostanabad, W. Chen, K. Ehmann, J. Cao, and M. A. Bessa. Deep learning predicts path-dependent plasticity. *Proceedings of the National Academy of Sciences*, 116(52):26414–26420, December 2019. doi: 10.1073/pnas.1911815116.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A Multi-Channel, Physics-Based Benchmark for Subseasonal-to-Seasonal Climate Prediction. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024.

Tung Nguyen, Jason Kyle Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. ClimateLearn: Benchmarking Machine Learning for Weather and Climate Modeling. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer, and Shirley Ho. The Well: A Large-Scale Collection of Diverse Physics Simulations for Machine Learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024.

Jaewan Park, Kazuma Kobayashi, Qibang Liu, Amar Alem Koric, Diab W. Abuiedda, Syed Bahauddin Alam, and Seid Koric. Sequential deep operator neural networks for plastic and thermo-viscoplastic transient material behavior. *International Journal of Plasticity*, 201:104689, June 2026. ISSN 0749-6419. doi: 10.1016/j.ijplas.2026.104689.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning Mesh-Based Simulation with Graph Networks. In *International Conference on Learning Representations*, October 2020.

Cyril Picard, Jürg Schiffmann, and Faez Ahmed. Dated: Guidelines for Creating Synthetic Datasets for Engineering Design Applications. In *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection, November 2023. doi: 10.1115/DETC2023-111609.

Wen Qiao, Zhiyan Zhao, Guoshun Wan, Hao Zheng, Wenqian Li, Mengxuan Cheng, Xiaohui Zhao, and Yuxi Jia. Efficient Prediction of Equivalent Thermo-Mechanical Properties for Embedded Unit Cells in Reliability Simulations of Complex Circuit Wiring Structures Using a Transformer-CNN Model. In *2024 25th International Conference on Electronic Packaging Technology (ICEPT)*, pp. 1–6, August 2024. doi: 10.1109/ICEPT63120.2024.10668535.

Jiyan Qiu, Lyulin Kuang, Guan Wang, Yichen Xu, Leiyao Cui, Shaotong Fu, Yixin Zhu, and Ruihua Zhang. DrivAerStar: An Industrial-Grade CFD Dataset for Vehicle Aerodynamic Optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

William John Macquorn Rankine. II. On the stability of loose earth. *Philosophical Transactions of the Royal Society of London*, 147(147):9–27, December 1857. ISSN 0261-0523. doi: 10.1098/rstl.1857.0003.

Lyle Regenwetter, Amin Heyrani Nobari, and Faez Ahmed. Deep Generative Models in Engineering Design: A Review. *Journal of Mechanical Design*, 144(071704), March 2022. ISSN 1050-0472. doi: 10.1115/1.4053859.

Lyle Regenwetter, Giorgio Giannone, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. Constraining Generative Models for Engineering Design with Negative Data. *Transactions on Machine Learning Research*, June 2024. ISSN 2835-8856.

Mohad Tanveer, Muhammad Muzammil Azad, Dohoon Kim, Salman Khalid, and Heung Soo Kim. Generative Design for Engineering Applications: A State-of-the-Art Review. *Archives of Computational Methods in Engineering*, 33(1):53–79, January 2026. ISSN 1886-1784. doi: 10.1007/s11831-025-10302-y.

Kittipong Thiamchaiboonthawee, Ghadi Nehme, Ram Mohan Telikicherla, Jiawei Tian, Balaji Jayaraman, Vikas Chandan, Dhanushkodi Mariappan, and Faez Ahmed. FLARE: A Data-Efficient Surrogate for Predicting Displacement Fields in Directed Energy Deposition, April 2026.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Ling-Qi Wang, Jing-Zhong Tong, Ken E. Evans, and Jiajia Shen. Performance-based inverse design of nonlinear mechanical metamaterials via enhanced conditional generative adversarial networks. *Computers & Structures*, 325:108183, April 2026. ISSN 00457949. doi: 10.1016/j.compstruc.2026.108183.

Zhihai Wang, Zijie Geng, Zhaojie Tu, Jie Wang, Yuxi Qian, Zhexuan Xu, Ziyan Liu, Siyuan Xu, Zhentao Tang, Shixiong Kai, Mingxuan Yuan, Jianye Hao, Bin Li, and Feng Wu. Benchmarking End-To-End Performance of AI-Based Chip Placement Algorithms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Yusuke Yamazaki, Reza Najian Asl, Markus Apel, Mayu Muramatsu, and Shahed Rezaei. Tackling multiphysics problems via finite element guided physics-informed operator learning. *Results in Engineering*, 32:112132, December 2026. ISSN 2590-1230. doi: 10.1016/j.rineng.2026.112132.

Su Zheng, Haoyu Yang, Binwu Zhu, Bei Yu, and Martin D. F. Wong. LithoBench: Benchmarking AI Computational Lithography for Semiconductor Manufacturing. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Zeshun Zong, Xuan Li, Minchen Li, Maurizio M. Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural Stress Fields for Reduced-order Elastoplasticity and Fracture. In *SIGGRAPH Asia 2023 Conference Papers*, SA '23, pp. 1–11, New York, NY, USA, December 2023. Association for Computing Machinery. ISBN 979-8-4007-0315-7. doi: 10.1145/3610548.3618207.

A Dataset Summary

Table 2 shows an overview of the dataset. The suffix of combo name **f***i***d***j***t***k*** indicates the meta-layer composition: *i* **f**ine pitch layers, *j* **d**ouble pitch layers, and *k* **t**op-metal layers. All combos by default contain a diffusion layer and a contact layer, in addition to the layers specified in naming, resulting in different number of layers and height, contributing to the difficulty on predicting M_3 for OOD-Combo split. The number of designs per combo is proportional to the number of layers, since the more the layers the larger the combo design space. The last three columns to the right of Table 2 summarize the design spatial distributions using first quantile, median, and third quantile of the number of rectangles, occupancy, and dispersion. Occupancy reflects the total areas of rectangles over the total layer areas excluding trench. The dispersion reveals the spread of rectangles over the layout length by computing the area-weighted entropy over 18 equal bins, normalized to $[0, 1]$. These distributions can be indirectly manipulated by changing the parameters in sampling metals and vias, or swapping the sampling distributions, as detailed in Appendix B.

Table 2: Dataset composition and geometric complexity across the 18 metal-stack configurations. Patterned layers count conductor, via, and trench layers. Stack height is measured above the silicon bar. Rectangle count, occupancy, and dispersion report in format of "median [Q1, Q3]". Configurations marked with † are held out from training in the OOD-Combo split, while those with $*$ are included in the 500-design exploratory dataset.

Combo	Layers	Height	Designs	Rectangles	Occupancy (%)	Dispersion
01_f2d0t0†∗	6	2.41	2,725	45 [34, 59]	10.5 [8.8, 12.4]	0.72 [0.65, 0.76]
02_f2d0t1∗	8	4.22	4,084	48 [37, 61]	12.5 [11.3, 13.9]	0.69 [0.64, 0.73]
03_f2d1t0∗	8	3.79	4,084	49 [37, 63]	9.5 [8.2, 11.1]	0.68 [0.62, 0.72]
04_f2d1t1	10	5.60	5,441	51 [39, 65]	12.3 [11.1, 13.5]	0.68 [0.63, 0.71]
05_f2d2t0	10	5.17	5,441	52 [40, 67]	10.0 [8.8, 11.4]	0.68 [0.63, 0.72]
06_f2d2t1†	12	6.98	6,800	55 [43, 69]	12.5 [11.6, 13.6]	0.67 [0.63, 0.70]
07_f3d0t0†∗	8	3.10	4,084	63 [48, 80]	10.6 [9.1, 12.1]	0.74 [0.69, 0.77]
08_f3d0t1	10	4.91	5,441	66 [52, 82]	12.6 [11.5, 13.9]	0.70 [0.66, 0.73]
09_f3d1t0	10	4.48	5,441	67 [52, 84]	9.8 [8.6, 11.1]	0.71 [0.67, 0.74]
10_f3d1t1	12	6.29	6,800	70 [56, 87]	12.7 [11.7, 13.8]	0.69 [0.65, 0.72]
11_f3d2t0	12	5.86	6,800	70 [55, 88]	10.5 [9.4, 11.6]	0.69 [0.65, 0.72]
12_f3d2t1†	14	7.67	8,158	74 [60, 90]	13.0 [12.1, 13.9]	0.68 [0.65, 0.71]
13_f4d0t0†	10	3.79	5,441	80 [63, 100]	10.6 [9.3, 12.1]	0.74 [0.70, 0.77]
14_f4d0t1	12	5.60	6,800	85 [68, 103]	12.8 [11.8, 13.9]	0.71 [0.67, 0.74]
15_f4d1t0	12	5.17	6,800	84 [67, 105]	10.1 [8.9, 11.2]	0.72 [0.68, 0.75]
16_f4d1t1	14	6.98	8,157	89 [73, 108]	13.0 [12.1, 14.0]	0.70 [0.67, 0.73]
17_f4d2t0	14	6.55	8,157	89 [71, 108]	10.9 [9.9, 11.9]	0.69 [0.66, 0.72]
18_f4d2t1†	16	8.36	9,516	94 [77, 112]	13.4 [12.6, 14.3]	0.69 [0.66, 0.71]
Overall			**110,170**	**71 [53, 92]**	**11.9 [10.4, 13.3]**	**0.69 [0.65, 0.73]**

B Design Sampling Algorithm

We represent each design as a periodic unit cell with planar dimensions $2\,\mu\mathrm{m} \times 18\,\mu\mathrm{m}$. Each cell contains a functional seal ring (FSR; the tower-like structure in the upper part of Figure 1B) and a $5\,\mu\mathrm{m}$ assessment zone (AZ). Algorithm 1 gives a high-level overview of the sampling process. Starting from an empty stack template, the sampler constructs each component within its rule-defined feasible domain.

Algorithm 1 Sampling a protection-area design

Require: Layer specification $\mathcal{H}$ and random seed s
Ensure: Design candidate d

```
Initialize the random generator with s
Initialize a periodic cell d with W = 2 μm and H = 18 μm
y_FSR ← ConstructFSR(d, H)
δ ← FSRClearance(H)
A ← SampleAZ(y_FSR, δ)
T ← SampleTrench(d, H, y_FSR, A)
for all metal layers m ∈ H do
    SampleMetal(d, m, y_FSR, A, T)
end for
for all via layers v ∈ H do
    SampleVia(d, v)
end for
return d
```

ConstructFSR. The FSR spans the entire x-axis and is placed at the inner end of the protection area using a deterministic procedure. It has a pyramidal structure, with the metal width generally increasing from the top to the bottom layers. The widths are selected to satisfy the minimum metal-width and via enclosure rules. For each pair of adjacent metal layers, the enclosed region determines how many FSR vias can be placed. These vias are distributed symmetrically within the available region.

SampleAZ. The AZ is a $5\,\mu$m-wide interval along the y-axis that spans the full x-extent and stack depth. Its position is sampled uniformly between the outer boundary of the design and the FSR, subject to the required clearance. No metal, via, or trench is placed inside the AZ.

SampleTrench. If the layer stack contains a trench, it is included with probability 0.5. The trench spans the entire x-axis and has the stack-defined extent along the z-axis. Its size along the y-axis is sampled uniformly between the minimum and maximum values specified by the design rules.

The AZ divides the available domain into at most two disjoint intervals. Among the intervals that accommodate the sampled trench size, one is randomly selected. The trench position is then sampled uniformly within that interval. If no feasible interval exists, the trench is omitted.

SampleMetal. For each metal layer, the FSR, AZ, applicable trench region, and their required spacing are removed from the sampling domain. This produces a set of disjoint intervals along the y-axis, which are processed independently.

For each interval, the maximum number of metal rows is estimated from the available space and the minimum metal dimensions and spacing. The requested number of rows is drawn from an exponential distribution with a mean of five and limited by this maximum. Rows are placed from top to bottom. When sampling each row, sufficient space is reserved for the remaining rows. The row height is drawn from a log-uniform distribution within bounds determined by the piecewise width and spacing rules.

Along the x-axis, a row is represented by a full-width bar with probability 0.32. Otherwise, a common pitch is sampled uniformly from its feasible range. This pitch determines how many rectangles fit across the cell. The width of each rectangle is sampled uniformly, and its height is sampled log-uniformly, subject to the minimum width, spacing, and area rules.

The exponential and log-uniform distributions are used to produce both sparse designs with large metal features and dense designs with smaller features. This also diversifies the enclosed regions available for via placement.

SampleVia. For each via layer, the sampler considers the geometries of its adjacent top and bottom metal layers. Their overlapping regions are reduced according to the corresponding enclosure requirements, producing the regions in which vias may be placed.

Within each region, the sampler first determines how many via rows can fit and samples their positions along the y-axis while maintaining the minimum spacing. If a region spans the full x-axis, each via row has equal probability of being a full-width bar or an equally spaced array of vias. Otherwise, the sampler draws the number of vias in the row and then samples their positions along the x-axis while reserving sufficient space for the remaining vias.

C SIMULATION CONFIGURATION

The simulation is a static structural analysis with one-way thermo-mechanical coupling. Each sampled design is converted into a three-dimensional material geometry and discretized using a conforming structured mesh of eight-node hexahedral elements. Material boundaries are included as mesh boundaries. The maximum element size is $0.5\mu m$, features separated by less than $0.05\mu m$ are merged during mesh construction, and element sizes grow with a grading factor of 1.3. A $3\mu m$ dielectric buffer is added at each end of the design along y-axis, and a $10\mu m$ silicon substrate spanning xy-axes is placed beneath the design.

Material models. Copper and aluminum are modeled using temperature-dependent elasticity and J_2 plasticity with Chaboche kinematic hardening. Silicon, silicon oxide, and silicon nitride are modeled as thermoelastic materials. Young's modulus and coefficient of thermal expansion are temperature-dependent, where applicable. Silicon nitride also exhibits anisotropic thermal expansion and intrinsic film strain.

Boundary conditions. Displacement normal to the $x_{\min}$, $y_{\max}$, and $z_{\min}$ surfaces is constrained. The x-displacements on the opposite x surface are coupled to represent the repeated cross-sectional response while preventing rigid-body motion.

Solver. We implement a small-strain, quasi-static, thermo-mechanical solver with eight-point integration and volumetric locking correction. Plastic equilibrium is iterated for at most 50 iterations with a relative displacement tolerance of 10^{-8}.

D DESIGN-TO-FIELDS DETAILS

Dataset splits. The first (**IID**) tests whether a model extrapolates to an unseen mechanical load position by partitioning the dataset by stratified random sampling across combos (train — val — test: 89,249 — 9,909 — 11,012); the second (**OOD-Combo**) probes transferability across different metal-stack heights by using a test set of 6,000 designs from the three shortest combos (f2d0t0, f3d0t0, f4d0t0) and three tallest (f2d2t1, f3d2t1, f4d2t1) with the split train — val — test: 66,104 — 7,342 — 6,000; and the third (**OOD-Field**) assesses reliability on mechanically atypical designs using a test set comprising top 10% most distant Mahalanobis distance in terms of coefficients (field descriptors) per combo (train — val — test: 89,236 — 9,909 — 11,025).

Model architectures. All baselines receive the 2D design layout, design layer specifications, and the normalized mechanical load z-position. The **pointnet encoder** represents each design using at most 512 rectangles (more than the dataset's maximum), each described by its normalized coordinates and its associated layer index. A shared MLP with widths 64, 128, 256, and 256 is applied to each rectangle, followed by max pooling. The **conv2d encoder** instead rasterizes the geometry into binary 32×32 layer masks. It applies four convolutional layers with 32, 64, 128, and 256 channels, followed by global average pooling. Both encoders produce a 256-dimensional geometry embedding.

The design layer specifications are encoded separately from normalized elevation, thickness, material, and layer type using two conv1d layers followed by max pooling. The resulting layer-stack embedding is added to the geometry embedding to form the thermal latent representation. A two-layer MLP encodes the normalized mechanical-load z-position, and its output is added to the thermal latent to form the mechanical latent representation. Thus, the mechanical branch is conditioned on both the complete thermal representation of the design and the applied load, whereas the thermal branch is independent of the mechanical load. The thermal branch predicts W, while the mechanical branch predicts S_1 and S_{xx}. Each branch has three hidden layers with a width of 128. We compare two mesh (**coarse/moderate**) decoders and one mesh-free (**implicit**) decoder. The mesh decoders use a fully-connected layer to project the latent vector onto a low-resolution 3D feature grid with 128 channels. Three transposed convolution layers then progressively upsample this grid to the target field resolution. The mesh-free decoder concatenates the latent representation with Fourier-encoded query coordinates using six frequencies. All outputs are standardized using train set statistics. Table 3 summarizes our six baseline models.

Table 3: Baseline models and trainable parameter counts.

Decoder	Field representation	PointNet	Conv2D
Coarse	$16 \times 64 \times 32$	3.591M	3.941M
Moderate	$32 \times 128 \times 64$	18.328M	18.679M
Implicit	Neural continuous field	0.311M	0.661M

B-spline field representation. To obtain a representation-agnostic regional representation, we project each field within the AZ onto a B-spline basis over normalized y and z coordinates. The basis contains 8 functions along y and 20 along z, yielding 160 coefficients per field. Before fitting, the three-dimensional element field is smoothed using a Gaussian kernel with $\sigma = 0.5\mu m$ and aggregated along the x-axis using elemental volume as weights. Coefficients are then obtained by volume-weighted least squares over the element centroids. We retain coefficients for the thermal strain energy density W^T and the maximum principal stress S_1 under M_1, M_2, and M_3, yielding a 640-dimensional response descriptor per design.

Training configuration. Training and validation use mechanical states M_1 and M_2. M_3 is excluded from both and introduced only at test time. For each design–load pair, the model is queried at 256 fixed interface probes and at 4,096 stratified samples of the reference-mesh centroids within the AZ. The centroid sampling probabilities are weighted by their physical volumes, so that the sampled B-spline fit approximates the full-field fit without bias toward sparsely populated regions.

The training objective combines two terms. First, an L_1 probe loss is averaged over standardized W, S_1, and S_{xx}. Second, the predicted W and S_1 values at the sampled centroids are projected onto the fixed B-spline basis by weighted least squares and compared with the reference coefficients using

an L_1 loss. This coefficient term is assigned a weight of 0.01. Validation uses the same objective, and the checkpoint with the lowest validation loss is retained.

Table 4: Forward-baseline training configuration.

Configuration	Value
Maximum epochs	20
Batch size	1024
Optimizer	AdamW
Initial learning rate	$1e^{-3}$
Weight decay	$1e^{-4}$
Learning-rate schedule	Cosine annealing
Gradient clipping	1.0
Probe/coefficient loss weights	1.0 / 0.01
Centroid samples per design	4,096
Maximum rectangles	512
Latent dimension	256
Precision	Automatic mixed precision
Early stopping	Patience 5 after epoch 15
Minimum improvement	$1e^{-3}$
Model selection	Lowest validation loss

Test-time evaluation. At test time, the selected checkpoint is evaluated on all loads. Probe accuracy is evaluated at the 256 design-dependent locations using NRMSE in normalized field space. Regional accuracy is evaluated by querying the prediction at *all* reference centroids, fitting the B-spline coefficients using volume-weighted least squares, and reporting coefficient NMAE normalized mode-wise by the train set standard deviations. The mesh-based baselines are interpolated at these evaluation coordinates, whereas the implicit models allow direct queries. For M_3, normalization uses mixed statistics from M_1 and M_2, ensuring no leakage of M_3 information.

Results. Table 5 reports probe NRMSE after standardizing each field using its train set moments. Table 6 reports B-spline coefficient NMAE, computed mode-wise using the corresponding train set standard deviation. Table 7 reports the MAE in MPa to allow physical interpretation. Note that the quantities have different physical scales and thus are not comparable across columns. Boldface marks the lowest error within each split and column.

Table 5: Probe NRMSE (↓).

Split	Decoder	Encoder	W^T	$S_1^{M_1}$	$S_1^{M_2}$	$S_1^{M_3}$	$S_{xx}^{M_3}$
IID	Coarse	PointNet	0.618	0.896	0.880	1.236	1.013
		Conv2D	0.604	0.897	0.881	1.028	0.976
	Moderate	PointNet	0.589	0.886	0.871	1.150	0.997
		Conv2D	**0.580**	**0.882**	**0.862**	**0.874**	**0.948**
	Implicit	PointNet	0.649	0.928	0.920	1.117	1.016
		Conv2D	0.637	0.918	0.905	1.299	0.999
OOD-Combo	Coarse	PointNet	1.105	1.020	1.066	1.447	1.048
		Conv2D	1.110	1.042	1.006	0.983	0.983
	Moderate	PointNet	1.120	1.041	1.136	1.546	1.087
		Conv2D	1.110	1.093	1.052	1.015	0.997
	Implicit	PointNet	1.110	**1.009**	1.052	1.146	**0.970**
		Conv2D	**0.999**	1.021	**0.977**	**0.976**	0.971
OOD-Field	Coarse	PointNet	0.660	0.910	0.925	1.131	1.081
		Conv2D	0.634	0.905	0.910	0.903	0.957
	Moderate	PointNet	0.630	0.894	0.902	1.265	1.009
		Conv2D	**0.594**	**0.882**	**0.883**	**0.884**	**0.939**
	Implicit	PointNet	0.685	0.933	0.943	1.078	0.975
		Conv2D	0.658	0.923	0.930	0.953	0.964

Table 6: B-spline coefficient NMAE(↓).

Split	Decoder	Encoder	W^T	$S_1^{M_1}$	$S_1^{M_2}$	$S_1^{M_3}$
IID	Coarse	PointNet	**0.331**	**0.209**	**0.153**	**0.762**
		Conv2D	0.386	0.236	0.159	0.881
	Moderate	PointNet	0.372	0.335	0.261	1.296
		Conv2D	0.373	0.271	0.180	2.319
	Implicit	PointNet	1.025	0.550	0.397	1.676
		Conv2D	0.821	0.449	0.339	1.736
OOD-Combo	Coarse	PointNet	1.620	2.104	0.829	1.334
		Conv2D	1.812	**1.609**	0.615	2.274
	Moderate	PointNet	**1.529**	2.098	0.906	1.622
		Conv2D	1.742	1.835	**0.571**	1.952
	Implicit	PointNet	2.687	2.412	1.008	**1.140**
		Conv2D	2.192	1.621	0.832	4.231
OOD-Field	Coarse	PointNet	0.725	0.569	0.422	1.222
		Conv2D	0.610	0.477	**0.315**	2.384
	Moderate	PointNet	0.564	0.566	0.437	1.251
		Conv2D	**0.535**	**0.469**	0.321	2.563
	Implicit	PointNet	1.589	0.832	0.593	**1.155**
		Conv2D	1.199	0.709	0.477	4.272

Table 7: Probe MAE (↓). Strain-energy density and the stresses are reported in MPa. These quantities have different physical scales and should be compared only within columns.

Split	Decoder	Encoder	W^T	$S_1^{M_1}$	$S_1^{M_2}$	$S_1^{M_3}$	$S_{xx}^{M_3}$
IID	Coarse	PointNet	0.565	192.8	173.5	237.0	271.6
		Conv2D	0.557	193.1	173.8	199.9	257.2
	Moderate	PointNet	0.547	188.2	169.5	231.1	267.4
		Conv2D	**0.539**	**185.9**	**167.0**	**176.6**	**250.1**
	Implicit	PointNet	0.597	203.1	184.9	219.4	271.6
		Conv2D	0.588	200.2	181.0	267.2	262.8
OOD-Combo	Coarse	PointNet	0.899	227.4	214.5	303.4	294.5
		Conv2D	0.929	232.1	206.6	**207.3**	273.1
	Moderate	PointNet	0.858	231.6	225.8	317.5	302.7
		Conv2D	1.065	243.8	215.8	211.2	277.2
	Implicit	PointNet	0.837	**224.0**	205.1	229.5	**257.6**
		Conv2D	**0.797**	227.5	**200.1**	208.6	273.3
OOD-Field	Coarse	PointNet	0.599	196.9	181.7	259.9	286.9
		Conv2D	0.584	194.7	177.8	182.0	254.1
	Moderate	PointNet	0.574	190.1	174.0	251.3	271.3
		Conv2D	**0.550**	**185.0**	**168.6**	**176.7**	**246.0**
	Implicit	PointNet	0.618	202.7	186.6	209.4	253.0
		Conv2D	0.606	200.0	183.4	194.6	255.6

E FIELDS-TO-DESIGN DETAILS

Since Nearest-Neighbour (NN) as a baseline is straightforward, in this section we only clarify relevant details for the generative baselines. The additional result tables include all baselines.

Model Architectures. The inverse baselines share the same architecture: an autoregressive generator (ARG) that maps a target field descriptor and the given metal-stack configurations to a layout. The condition is the 640-dimensional B-spline response descriptor defined above (160 coefficients for thermal strain energy density W^T followed by 160 S_1 coefficients for each of M_1, M_2, and M_3), standardized with train set statistics. The design layouts are preprocessed as sequences of rectangles on the $0.01\,\mu m$ design grid. Each rectangle is described by six typed tokens: layer, $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$, and a phase-shift flag which determines whether the rectangle has an cross-frame offset by half the width (left half on a given cell and right half on the left adjacent cell forms the whole rectangle). Coordinates are discretized to 200 steps in x and 2,400 steps in y over the buffered simulation domain. Sequences are delimited by start and end tokens. The condition encoder passes the target field descriptor through a two-layer MLP (widths 256, GeLU) and adds a 256-dimensional projection of the layer-stack embedding, computed with the same encoder as the forward baselines. The sum is layer-normalized, passed through GeLU, and expanded by a linear layer into four prefix tokens that are prepended to the layout token sequence. The generator is a decoder-only Transformer with causal attention. Each token is represented by a 256-dimensional vector, formed by summing a learned token embedding and a learned absolute position embedding. The vectors pass through six Transformer blocks with pre-layer normalization. Each block contains multi-head self-attention with eight heads and a position-wise feed-forward network that expands the 256-dimensional representation to 512 dimensions and projects it back with dropout 0.1. The final hidden state at each position is fed to task-specific output heads.

Training configuration. We train two models for each of the splits: IID, OOD-Combo, OOD-Field. The objective is token cross-entropy loss, averaged over all non-padding target tokens, plus a MSE loss on the predicted assessment zone (AZ) center and width. The ARG-Joint variant adds an auxiliary field loss: a linear head reconstructs the 640 standardized response coefficients from the pooled hidden states of the completed rectangles, using a forward pass with only the layout tokens and layer stack as the input. This loss is weighted by 0.5 and is used only during training. All other settings, the seed, and the data order are identical between ARG and ARG-Joint as specified in Table 8. The checkpoint with the lowest validation loss on the total training objective is retained.

Table 8: Generative baselines training configuration.

Configuration	Value
Epochs	100
Batch size	256 (2 accumulation steps, effective 512)
Optimizer	AdamW
Peak learning rate	3×10^{-4}
Weight decay	0.01
Learning-rate schedule	5 warmup epochs, cosine decay
Gradient clipping	1.0
Dropout	0.1
Model width / layers / heads	256 / 6 / 8
Feed-forward width	512
Maximum rectangles / tokens	256 / 1,538
Vocabulary size	2,623
Input field descriptor dimension	640
Field loss weight (ARG / ARG-Joint)	0 / 0.5
Sampling temperature	0.7
Precision	Automatic mixed precision
Model selection	Lowest validation loss

Test Time Generation. At test time, tokens are sampled with temperature 0.7. Sampling is constrained by masking layers that are not patterned in the target stack, forbidding an early stop, and enforcing $x_{\min} < x_{\max}$ and $y_{\min} < y_{\max}$. One layout is generated per target, and the predicted AZ

location is attached to it. Layouts are passed to the design rule checker for validity assessment. The valid candidates are then passed to the simulator for new simulations.

Results. Table 9 and Table 10 report the per-combo break down for both metrics separately. Figure 3, Figure 4, and Figure 5 demonstrate selected samples from all combos for IID. NN Cross-combo is left out due to poor combo coverage. Interestingly, the generative baselines can generate designs with deterministic functional seal ring without explicit enforcement during training. It is also encouraging to observe that the baselines can approximate the location of target assessment zones based solely on the field descriptors.

Table 9: Number of valid inverse designs out of 5,000 targets per configuration. NN-S and NN-X separately denote NN Same-combo and NN Cross-combo. "n/e" indicates configurations not included in the OOD-Combo test set and not evaluated.

	IID				OOD-Combo			OOD-Field			
Combo	NN-S	NN-X	ARG	ARG-Joint	NN	ARG	ARG-Joint	NN-S	NN-X	ARG	ARG-Joint
f2d0t0	111	0	31	21	0	0	2	115	0	29	33
f2d0t1	166	0	50	37	n/e	n/e	n/e	170	0	52	47
f2d1t0	146	0	52	32	n/e	n/e	n/e	159	0	35	33
f2d1t1	213	0	66	62	n/e	n/e	n/e	221	0	56	50
f2d2t0	214	0	54	43	n/e	n/e	n/e	216	0	41	34
f2d2t1	266	0	76	68	0	0	0	276	0	50	56
f3d0t0	151	18	36	22	0	0	0	160	21	35	32
f3d0t1	196	0	37	27	n/e	n/e	n/e	211	0	42	42
f3d1t0	208	0	33	32	n/e	n/e	n/e	205	0	34	31
f3d1t1	266	0	44	44	n/e	n/e	n/e	265	0	59	33
f3d2t0	246	0	31	36	n/e	n/e	n/e	251	0	43	30
f3d2t1	315	0	59	48	0	0	0	315	0	51	50
f4d0t0	197	22	23	23	0	0	0	210	98	33	35
f4d0t1	244	0	31	24	n/e	n/e	n/e	245	0	28	31
f4d1t0	235	0	26	18	n/e	n/e	n/e	239	0	18	27
f4d1t1	296	0	33	27	n/e	n/e	n/e	293	0	45	40
f4d2t0	276	0	28	19	n/e	n/e	n/e	284	0	26	31
f4d2t1	363	0	31	26	114	0	0	336	0	44	34
Overall	**4,109**	**40**	**741**	**609**	**114**	**0**	**2**	**4,171**	**119**	**721**	**669**

Table 10: Per-configuration normalized mean absolute error (NMAE) over all 640 standardized field descriptors. Values are means conditional on validity. NN-S and NN-X separately denote NN Same-combo and NN Cross-combo. "n/e" indicates configurations not included in the OOD-Combo test set, therefore not evaluated. NN-S uses the stored solved results while the remaining baselines require running new simulations for the valid design candidates. A dash denotes undefined NMAE because no candidate was valid.

	IID				OOD-Combo			OOD-Field			
Configuration	NN-S	NN-X	ARG	ARG-Joint	NN	ARG	ARG-Joint	NN-S	NN-X	ARG	ARG-Joint
f2d0t0	0.0810	–	0.2567	0.1998	–	–	1.1644	0.1760	–	0.5930	0.4106
f2d0t1	0.0694	–	0.2710	0.2655	n/e	n/e	n/e	0.1262	–	0.5486	0.4861
f2d1t0	0.0691	–	0.2655	0.2540	n/e	n/e	n/e	0.1395	–	0.3694	0.5862
f2d1t1	0.0650	–	0.2364	0.2689	n/e	n/e	n/e	0.1134	–	0.3863	0.5067
f2d2t0	0.0695	–	0.2854	0.3381	n/e	n/e	n/e	0.1127	–	0.4647	0.4577
f2d2t1	0.0685	–	0.2842	0.2791	–	–	–	0.1220	–	0.4079	0.4575
f3d0t0	0.0819	0.2658	0.2240	0.2454	–	–	–	0.1429	0.3307	0.4882	0.4514
f3d0t1	0.0758	–	0.2965	0.3520	n/e	n/e	n/e	0.1246	–	0.4247	0.4479
f3d1t0	0.0781	–	0.2023	0.2625	n/e	n/e	n/e	0.1238	–	0.4265	0.4812
f3d1t1	0.0751	–	0.2859	0.2717	n/e	n/e	n/e	0.1180	–	0.3966	0.4791
f3d2t0	0.0739	–	0.2367	0.2341	n/e	n/e	n/e	0.1215	–	0.3859	0.4735
f3d2t1	0.0762	–	0.2521	0.2501	–	–	–	0.1236	–	0.3900	0.4204
f4d0t0	0.0892	0.2906	0.3221	0.2579	–	–	–	0.1505	0.3618	0.4465	0.4612
f4d0t1	0.0815	–	0.4020	0.2693	n/e	n/e	n/e	0.1271	–	0.4663	0.6310
f4d1t0	0.0838	–	0.3052	0.2888	n/e	n/e	n/e	0.1347	–	0.4611	0.5160
f4d1t1	0.0786	–	0.3415	0.2607	n/e	n/e	n/e	0.1296	–	0.4382	0.4594
f4d2t0	0.0819	–	0.2760	0.2275	n/e	n/e	n/e	0.1211	–	0.4392	0.5230
f4d2t1	0.0853	–	0.2538	0.3104	0.6039	–	–	0.1294	–	0.4237	0.4693
Overall	**0.0772**	**0.2795**	**0.2741**	**0.2710**	**0.6039**	–	**1.1644**	**0.1278**	**0.3563**	**0.4368**	**0.4808**

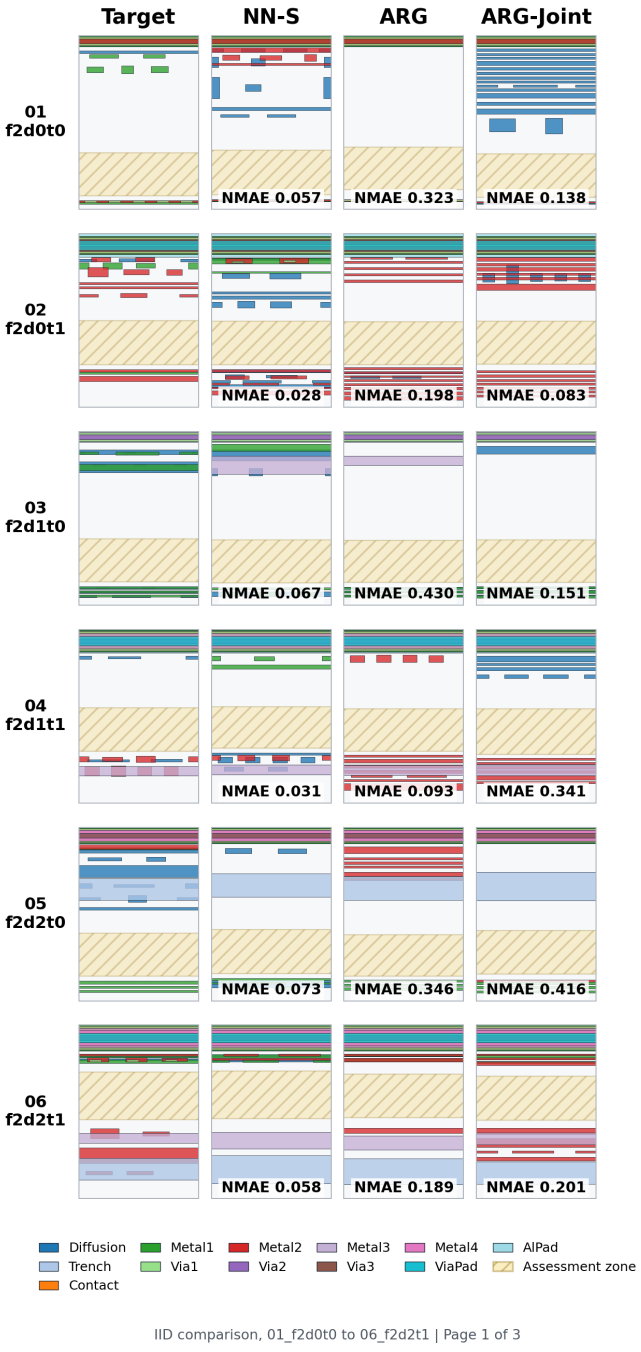

Figure 3: Selected samples of Combo 01 to Combo 06 from IID test set and the corresponding valid generated designs of NN Same-combo (NN-S), ARG, and ARG-Joint. The NMAE shows the discrepancy between the field descriptors of the assessment zone (marked with striped yellow) against that of the target design.

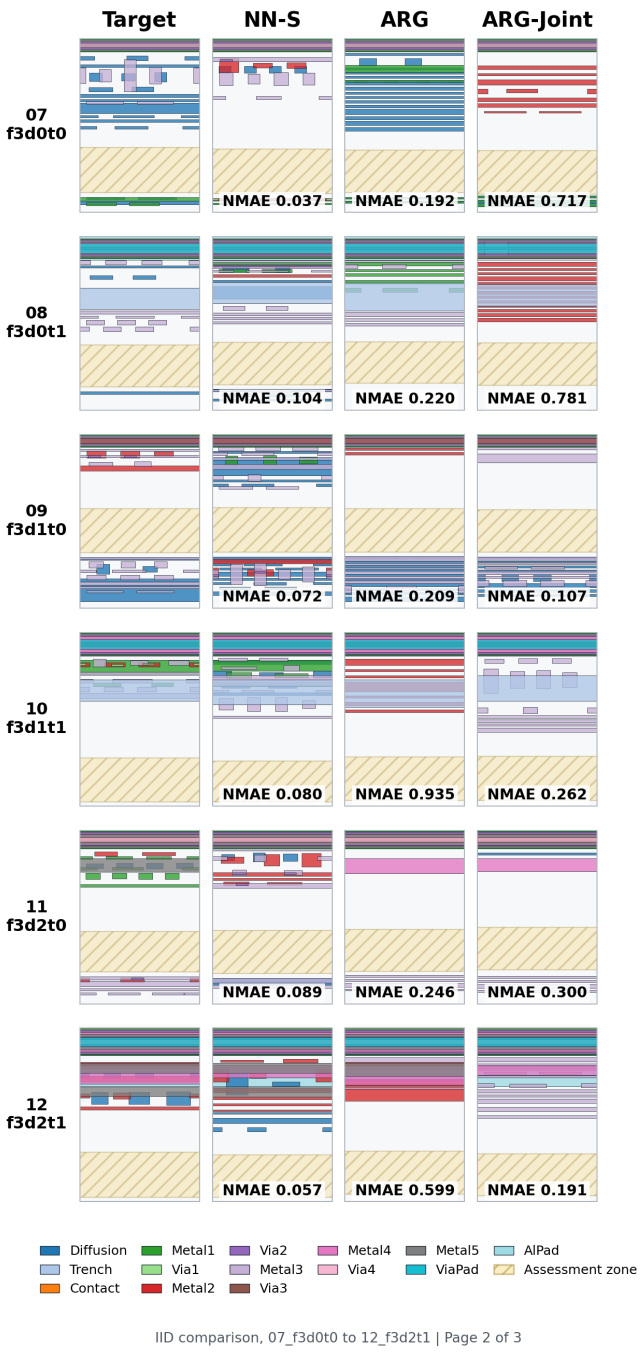

Figure 4: Selected samples of Combo 07 to Combo 12 from IID test set and the corresponding valid generated designs of NN Same-combo (NN-S), ARG, and ARG-Joint. The NMAE shows the discrepancy between the field descriptors of the assessment zone (marked with striped yellow) against that of the target design.

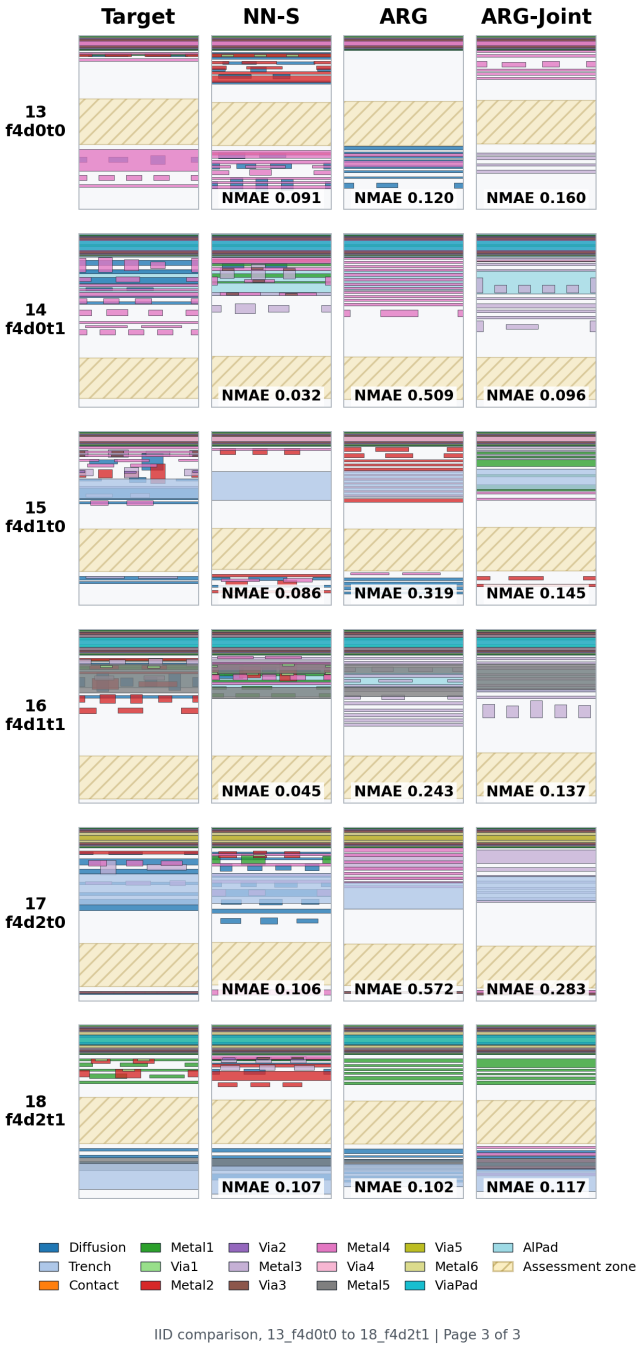

Figure 5: Selected samples of Combo 13 to Combo 18 from IID test set and the corresponding valid generated designs of NN Same-combo (NN-S), ARG, and ARG-Joint. The NMAE shows the discrepancy between the field descriptors of the assessment zone (marked with striped yellow) against that of the target design.