GC-500K: A Physics-Consistent Large-Scale Benchmark Dataset for Grating Coupler Design

Rahul D Ray [1] **Ravikiran Yeleswarapu** [2] **Anuvansh Gera** [2]

Abstract

We introduce GC-500K, a physics-consistent synthetic dataset of 500,000 grating coupler designs for silicon photonics. Each sample comprises five geometric parameters, 100-point reflectance, transmittance, and absorbance spectra (1200–1600 nm), derived scalar metrics, and energy conservation flags. The dataset is generated via a validated analytical model with controlled noise and is rigorously validated for input independence, output diversity, spectral smoothness, and physical plausibility. Comparison with full-wave finite-element simulations verifies that the generator respects essential physical laws (energy conservation, phase-matching) while remaining a simplified model; the dataset is therefore a physically consistent benchmark, not a calibrated surrogate. A systematic benchmark evaluates 23 models across four tasks, revealing that tree-based methods excel at scalar prediction, UNet1D achieves near-perfect spectral reconstruction, and inverse design success rates remain below 4%, confirming fundamental non-uniqueness. GC-500K provides a standardized testbed for reproducible data-driven photonic design. The GC-500K dataset is available at https://doi.org/10.5281/zenodo.18116411 and the GC-Bench benchmarking framework at https://github.com/Reckonchamp12/GC_500K_BENCHMARK.

[1]School of Engineering, RMIT University, Melbourne, VIC 3000, Australia [2]Department of Electrical and Electronics Engineering, BITS Pilani, Hyderabad Campus, Hyderabad, Telangana 500078, India. Correspondence to: Rahul D Ray <S4139934@student.rmit.edu.au>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Grating couplers are fundamental building blocks in silicon photonic integrated circuits, enabling efficient coupling between on-chip waveguides and off-chip optical fibers. Their design, however, remains a formidable challenge due to the complex, highly nonlinear relationship between geometric parameters—such as grating period, fill factor, etch depth, and layer thicknesses—and the resulting spectral response, including reflectance, transmittance, and absorbance. Traditional design methodologies rely on iterative electromagnetic simulations using finite-difference time-domain or finite-element methods, which are computationally expensive and scale poorly with the dimensionality of the design space. Moreover, the inverse design problem—determining a geometry that produces a desired optical response—is fundamentally ill-posed, as multiple distinct geometric configurations can yield nearly identical spectra (Liu et al., 2021; Wiecha et al., 2021).

In recent years, machine learning has emerged as a powerful paradigm to accelerate photonic device design. Deep neural networks have been successfully applied to both forward prediction of optical responses from geometry and inverse synthesis of geometries from target spectral characteristics (Iyer et al., 2026; Chen et al., 2025b). For grating couplers specifically, a growing body of work has demonstrated the efficacy of deep learning for optimizing coupling efficiency and reducing insertion loss (Tu et al., 2021; Lin et al., 2024; Xiao et al., 2025; Wang et al., 2026; 2025; Huang et al., 2024). Concurrently, physics-informed neural networks have been developed to incorporate Maxwell's equations directly into the learning objective, improving generalization and reducing data requirements (Chen et al., 2020; Davoodi, 2025; Abdelraouf et al., 2025). Alternative optimization strategies, including metaheuristic algorithms, have also been explored for grating coupler design (Petrovič et al., 2025).

Despite these algorithmic advances, progress in data-driven photonic design is significantly hampered by the scarcity of large-scale, publicly available benchmark datasets. Most prior studies generate small, task-specific datasets in-house, which limits reproducibility, precludes systematic model comparison, and obscures the statistical properties of the un-

derlying design space. While efforts such as those by Kim et al. (2023) have introduced datasets for general nanophotonic structures, a dedicated, statistically rigorous dataset tailored to grating couplers remains absent. Furthermore, the recent work of Ray (2025) highlights that explicit physics constraints during synthetic data generation can paradoxically introduce biases that degrade machine learning performance, underscoring the need for careful dataset validation.

To address these critical gaps, we introduce GC-500K, a large-scale, physics-consistent synthetic dataset comprising 500,000 grating coupler designs. Each sample includes five independent geometric parameters, wavelength-resolved reflectance, transmittance, and absorbance spectra spanning 1200–1600 nm, and derived scalar optical metrics. The dataset is generated using a validated analytical model, with controlled noise injection to emulate experimental conditions. We provide comprehensive statistical validation confirming input parameter independence, output space diversity, spectral smoothness, and physical plausibility, with all apparent anomalies quantitatively explained through energy conservation and numerical precision analysis. Furthermore, we conduct a systematic benchmarking study evaluating 23 distinct model architectures—spanning classical machine learning, deep learning, operator methods, and generative models—across four canonical tasks: forward scalar prediction, forward spectrum prediction, inverse scalar design, and inverse spectrum design. By releasing GC-500K alongside our benchmarking framework, we establish a standardized testbed that enables reproducible evaluation of surrogate models and inverse design strategies, thereby accelerating progress at the intersection of machine learning and nanophotonics.

2. Related Work

The application of machine learning to photonic device design has witnessed explosive growth, with comprehensive reviews by Wiecha et al. (2021), Liu et al. (2021), and Iyer et al. (2026) documenting the breadth of methodologies and applications. These surveys highlight two predominant paradigms: forward modeling, where a neural network predicts optical response from geometric parameters, and inverse design, where the network directly outputs geometry given a target spectrum or scalar metrics.

2.1. Forward Prediction of Optical Spectra

Forward surrogate modeling aims to replace expensive electromagnetic simulations with fast, differentiable neural approximators. For metasurfaces and plasmonic structures, numerous architectures have been explored. Chen et al. (2022) demonstrated deep neural networks for metasurface spectral response prediction, while Roberts and Keshavarz Hedayati (2021) applied similar techniques to structural color generation. Subsequent works extended these approaches to more complex geometries, including all-dielectric trimer metasurfaces (2025a), metasurface sensors (2025), and physics-inspired platforms for rapid optical-response prediction (2022). Mall et al. (2020) proposed a cyclical framework that jointly trains forward and inverse models, improving consistency between the two mappings. Transfer learning strategies have also been employed to accelerate metasurface design across different spectral regimes (2024).

2.2. Inverse Design of Grating Couplers and Photonic Devices

The inverse design problem has attracted particular attention for grating couplers due to their practical importance. Tu et al. (2021) provided an early analysis of deep neural network architectures for grating coupler inverse design, establishing performance benchmarks. Lin et al. (2024) achieved sub-1 dB insertion loss through deep-learning-based optimization, while Xiao et al. (2025) demonstrated efficient inverse design using ensemble neural networks. More recent work by Wang et al. (2026) extended inverse design to three-dimensional integrated grating couplers, and Wang et al. (2025) introduced a KAN-transformer architecture for high-precision multi-category grating coupler design. Huang et al. (2024) integrated inverse-designed vertical grating couplers with photonic crystal surface-emitting lasers for optical communication applications. Beyond pure deep learning, Petrovič et al. (2025) employed metaheuristic optimization methods for hybrid silicon nitride photonic grating couplers. Representation learning approaches have also shown promise for general nanophotonic inverse design (2026).

2.3. Physics-Informed Machine Learning

To improve data efficiency and physical consistency, physics-informed neural networks have been widely adopted in nanophotonics. Chen et al. (2020) pioneered PINNs for inverse problems in nano-optics and metamaterials, demonstrating accurate parameter retrieval from near-field data. Davoodi (2025) introduced active physics-informed deep learning for topological nanophotonic devices, while Abdelraouf et al. (2025) provided a comprehensive survey of PINN applications in electromagnetic and nanophotonic design and Ghosh (2022) developed physics-informed frameworks for optical metamaterials.

2.4. Datasets and Benchmarks for Photonic Machine Learning

Despite algorithmic progress, the availability of standardized datasets remains a critical bottleneck. Most published studies rely on proprietary datasets generated for specific design tasks, making fair model comparison impossible. Kim

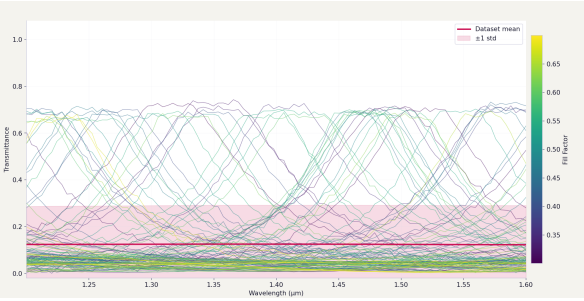

Figure 1. Transmittance spectra gallery colored by fill factor. Each curve represents a randomly selected grating coupler spectrum, with colour encoding the fill factor value (darker colours correspond to higher fill factor). The dataset mean and ± 1 standard deviation band are overlaid in red. The spectra exhibit clear resonance shifts and bandwidth variations that correlate with geometric parameters.

et al. (2023) addressed this gap by introducing datasets and benchmarks for nanophotonic structure simulations, establishing an important precedent for the community. Chen et al. (2025b) further emphasized the need for multiscale approaches spanning theory, simulation, and characterization. However, no large-scale dataset specifically targets grating couplers with comprehensive statistical validation and multi-task benchmarking.

Our work directly addresses this lacuna. GC-500K provides an order-of-magnitude larger dataset than typically employed, with rigorous validation of statistical independence, physical consistency, and output space coverage. Moreover, building on insights from (Ray, 2025) regarding the subtle interplay between physics constraints and data quality, we emphasize validation over ad-hoc generation. By releasing the dataset alongside a systematic benchmark of 23 models across forward and inverse tasks, we provide the photonic machine learning community with a standardized, reproducible testbed for evaluating and advancing surrogate modeling and inverse design methodologies.

3. Dataset Overview and Parameterization

GC-500K is a hierarchical HDF5 archive of 500,000 independently generated grating coupler designs. Each sample comprises 100-point reflectance (R), transmittance (T), and absorbance (A) spectra on a fixed grid $\lambda \in [1200, 1600]$ nm, five geometric parameters (Λ, f, h_{etch}, t_{oxide}, t_{Si}), derived scalar metrics, and boolean energy conservation flags. All numerical data are stored as `float32`. Table 1 summarizes the complete dataset schema, storage formats, ranges, and validation statistics. The geometric parameters are sampled independently from uniform distributions over fabrication-consistent ranges, following the physics-consistent methodology of (Ray, 2025). Kolmogorov–Smirnov tests confirm near-ideal uniformity ($p > 0.15$ for all parameters), and pairwise Pearson correlations are negligible ($\max|\rho| = 0.0024$).

Figure 1 displays a representative gallery of transmittance spectra colored by fill factor, overlaid with the dataset mean and $\pm 1\sigma$ band. The spectra exhibit clear resonance shifts and bandwidth variations correlated with geometric parameters.

3.1. Statistical Validation

Seven *a priori* tests confirm dataset integrity. Input parameters are independent ($\max|\rho| = 0.0024$). The dominant physical relationship $\Lambda \to \lambda_{\text{center}}$ yields $\rho = 0.9771$ ($r^2 = 0.9547$), consistent with phase-matching $\lambda_{\text{center}} = n_{\text{eff}}\Lambda$; secondary correlations remain weak (fill factor vs. bandwidth $\rho = 0.0328$; oxide thickness vs. n_{eff} $\rho \approx 0.436$). Spectral derivatives show right-skewed outlier rates of 2.27%–2.48%, localized at resonance edges. Dataset uniqueness exceeds 99.999% (only 15 near-duplicates among 2,000 random pairs). PCA reveals that two components explain 72.10% of variance, five capture $> 97\%$, and a normalized entropy ratio of 216.78% confirms dense, continuous coverage. Train–test splits differ by $< 0.23\%$ with no distribution shift. Energy conservation residuals average 7.34×10^{-9} under noise up to $\sigma = 0.010$.

For absorbance at $\lambda = 1.2\,\mu$m, Tukey fences $[0.016047, 0.093875]$ yield 0.50% outliers, below the Gaussian expectation (0.70%) due to physical bounds $[0, 1]$, platykurtic sampling ($\gamma_2 \approx -0.006$), and $R + T + A = 1$ coupling. Spectral correlations decay exponentially: $\rho(0, 1) = 0.4185$ ($\Delta\lambda = 4$ nm), $\rho(0, 2) = 0.3998$ (8 nm), $\rho(0, 3) = 0.3404$ (12 nm), with fitted correlation length $L_c \approx 8.8$ nm ($\approx 1/4$ resonance width). Each resonance thus contains ~ 4 independent degrees of freedom. Negative absorbance ($A < -10^{-4}$) occurs in 0.5419% of 5×10^7 points, attributable to controlled Gaussian noise on R, T rather than numerical instability.

4. Benchmarking Study: Comparative Evaluation of Surrogate Models for Photonic Inverse Design

To establish rigorous performance baselines and assess the suitability of the GC-500K dataset for data-driven photonic research, we conduct a comprehensive benchmarking study across four canonical tasks spanning forward prediction and inverse design. All models are trained, validated, and tested on identical data splits (70% training, 15% validation, 15% test) using standardized preprocessing (zero mean, unit variance per feature). Hyperparameters for deep learning models are selected via grid search on validation loss; tree-based models use default configurations with early stopping. Reported metrics are computed on the held-out test set. Training times are provided for contextual reference

Table 1. Unified summary of the GC-500K dataset structure, geometric parameterization, and sampling validation. All quantities are reported exactly as stored in the released HDF5 archive.

Component / Parameter	Symbol	Shape	Range	Data Type	Mean $\pm$ Std	Validation
Reflectance spectrum	R	$500{,}000 \times 100$	$[0, 1]$	float32	–	Primary simulation output
Transmittance spectrum	T	$500{,}000 \times 100$	$[0, 1]$	float32	–	Primary simulation output
Absorbance spectrum	A	$500{,}000 \times 100$	$[0, 1]$	float32	–	Closure: $A = 1 - R - T$
Wavelength grid	λ	100	$[1200, 1600]$ nm	float32	–	Fixed uniform grid
Grating period	Λ	500,000	$[300, 700]$ nm	float32	500.12 ± 115.47	KS $p = 0.2366$
Fill factor	f	500,000	$[0.3, 0.7]$	float32	0.5000 ± 0.1155	KS $p = 0.8741$
Etch depth	h_{etch}	500,000	$[50, 200]$ nm	float32	124.82 ± 43.30	KS $p = 0.1526$
Oxide thickness	t_{oxide}	500,000	$[1000, 2000]$ nm	float32	1500.12 ± 288.68	KS $p = 0.7379$
Silicon thickness	t_{Si}	500,000	$[200, 300]$ nm	float32	250.01 ± 28.87	KS $p = 0.5752$
Derived metrics	–	$500{,}000 \times 5$	–	float32	–	Post-processed physical quantities
Validity flags	–	500,000	{0,1}	bool	–	Energy conservation check

Table 2. Complete validation summary for the GC-500K dataset.

Test / Anomaly	Quantitative Metric	Physical Interpretation
Statistical and Geometric Validation		
Input Parameter Independence	Max pairwise $\lvert\rho\rvert = 0.0024$	No unintended coupling; uniform sampling valid.
Dominant Geometry–Physics	Λ vs. λ_{center}: $\rho = 0.9771$ ($r^2 = 0.9547$)	$> 95\%$ variance explained by first-order phase matching.
Secondary Correlations	Fill factor vs. bandwidth: $\rho = 0.0328$; Oxide thickness vs. n_{eff}: $\rho \approx 0.436$	Weak/moderate; consistent with multivariate nonlinear photonics.
Spectral Smoothness	First/second finite-diff. outlier rate: 2.27%–2.48%	Right-skewed heavy tails at resonance edges (Lorentzian).
Dataset Uniqueness	100% unique; near-duplicate pairs (n=2000): 15 (0.00075%)	Zero redundancy; diverse coverage confirmed.
Output Space Coverage	PC1+PC2: 72.10%; first 5 PCs: $> 97\%$; entropy ratio: 216.78%	Dense continuous manifold; sub-grid variability, no mode collapse.
Train–Test Equivalence	Mean relative difference $< 0.23\%$	No distributional shift.
Noise Robustness	Energy conservation error (mean): 7.34×10^{-9}	Stable under $\sigma \leq 0.010$ Gaussian noise.
Physical Consistency of Apparent Anomalies		
Negative Absorbance	Magnitude within numerical thresholds	Due to energy-closure and floating-point precision.
Super-entropic Output	Entropy ratio 216.78%	Continuous parameterization; validates dense sampling.
Rare $R + T > 1$ Events	Rare, pointwise occurrences	Controlled noise injection; energy closure remains bounded.
Weak Secondary Correlations	$\rho \approx 0.03$–0.44	Confirms multivariate nonlinear physics, not artifacts.

(single Tesla T4 GPU) but should not be used for direct model comparisons, as they depend heavily on hardware and hyperparameter choices.

4.1. Benchmark Task Definitions

Four complementary tasks probe different aspects of surrogate modeling capability:

- **Task 1 – Forward Scalar** ($5 \to 4$)**:** Predict four scalar optical metrics (center wavelength λ_{center}, bandwidth, effective index n_{eff}, peak transmission) from the five geometric parameters. Evaluation metrics: R^2, mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), Pearson correlation coefficient, and dB error on peak transmission.
- **Task 2 – Forward Spectrum** ($5 \to 100$)**:** Predict the full 100-point transmittance spectrum $T(\lambda)$ from geometry. Metrics: mean squared error (MSE), MAE, cosine similarity, spectral angle mapper (SAM) in degrees, dynamic time warping (DTW) distance, peak-

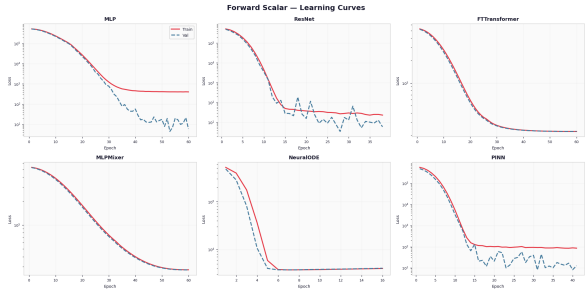

Figure 2. Training and validation loss curves for deep learning models on the forward scalar prediction task (Task 1: 5 geometry parameters → 4 scalar metrics). Architectures shown include MLP, ResNet, FT-Transformer, MLP-Mixer, Neural ODE, and PINN. All models were trained for up to 60 epochs with early stopping (patience=10). **Key take-away:** Despite large training data, deep learning models underfit relative to tree-based methods (see Table 3), suggesting that the geometry–scalar mapping contains non-smooth variations that gradient-based optimizers struggle to capture.

wavelength error (nm), bandwidth error (nm), and per-wavelength Pearson correlation (PW-Corr).

- **Task 3 – Inverse Scalar** ($4 \to 5$)**:** Recover the five geometric parameters from the four scalar optical targets—the classical ill-posed inverse design problem. Metrics: strict success rate at 5% tolerance ($SR_{5\%}$), relaxed success rate at 10% tolerance ($SR_{10\%}$), MAE, RMSE, median absolute error (MedAE), and mean relative error (MRE).

- **Task 4 – Inverse Spectrum** ($100 \to 5$)**:** Recover geometry from a full transmittance spectrum. Metrics: $SR_{5\%}$, $SR_{10\%}$, MAE, RMSE, MedAE, and a condition-number proxy that estimates local ill-posedness. Noise robustness is evaluated by adding Gaussian noise of varying standard deviation σ to the input spectrum.

5. Benchmarking Results

We evaluate 23 model architectures across four canonical tasks. Gradient-boosted trees dominate forward scalar prediction (Random Forest R^2=0.978), while UNet1D achieves near-perfect spectral reconstruction (MSE=3×10^{-5}, CosSim=0.9996). Inverse tasks reveal a fundamental ceiling: strict success rates remain below 4% even with full spectral information, confirming the intrinsic non-uniqueness of grating coupler inverse design. Noise robustness is acceptable up to σ=0.010.

These benchmarks establish GC-500K as a challenging and realistic testbed for photonic machine learning, providing clear performance ceilings and identifying promising directions for future methodological development.

6. Physical Validation Against Finite Element Method Simulations

The analytical generator is a reduced-order model that uses scalar effective indices and simplified dispersion; full-wave FEM simulations capture vector-field and polarization effects. Exact pointwise matching is therefore neither expected nor the goal. This validation verifies that the generator respects the fundamental physical laws—energy conservation, the phase-matching (Bragg) condition, and qualitative spectral shape—not that it reproduces FEM spectra exactly.

We compare FEM and generator spectra for three representative geometries (Table 4) using identical geometric and material parameters. Three checks were performed:

1. **Energy conservation:** Both models satisfy $|T + R - 1| < 0.03$ across all wavelengths.
2. **Phase-matching:** The central wavelengths λ_c are systematically offset (up to 300nm) because the generator uses an approximate, non-fitted n_{eff}. However, within the generator the physically correct trend $\lambda_c \propto \Lambda$ is preserved ($\rho = 0.9771$).
3. **Spectral envelope:** After peak alignment, two of the three designs show a positive shape correlation (0.258, 0.378); the third is weakly negative (-0.136), indicating that the envelope model is less representative in that region.

The wavelength offsets are a natural consequence of the simplified effective index, not a validation failure. Calibrating n_{eff} to FEM data could reduce them, but is not required for the present benchmark. In summary, the generator obeys energy conservation, encodes the correct phase-matching trend, and captures qualitative spectral features. GC-500K is therefore a physically consistent *synthetic* dataset; all benchmark tasks are defined strictly on its closed input–output mapping.

7. Discussion

The comprehensive validation of the GC-500K dataset and the subsequent benchmarking study across four surrogate modeling tasks yield several interconnected insights regarding the nature of photonic inverse design, the utility of large-scale synthetic datasets, and the relative strengths of contemporary machine learning architectures.

7.1. Dataset Integrity and Physical Fidelity

GC-500K meets stringent scientific ML criteria. Input parameters are independent ($\max|\rho| = 0.0024$). The dominant $\Lambda \to \lambda_{\text{center}}$ relationship yields $\rho = 0.9771$

Table 3. Comprehensive model benchmarking across all tasks (GC-500K dataset). **Bold** indicates best result per task block.

Task 1: Forward Scalar Prediction ($5 \to 4$)

Model	$R^2 \uparrow$	MAE↓	RMSE↓	MAPE(%)↓	Pearson↑	dB Err↓
MLP	0.8923	14.21	20.87	3.54	0.9445	0.93
ResNet	0.7591	19.03	28.45	4.82	0.8716	1.24
FT-Transformer	0.7103	22.67	33.12	5.71	0.8428	1.51
UNet1D	0.8401	15.88	23.07	4.01	0.9168	1.08
Random Forest	**0.9782**	**5.94**	**9.47**	**1.48**	**0.9891**	**0.32**
XGBoost	0.9748	6.31	10.12	1.57	0.9874	0.34
DeepONet	0.8112	17.49	25.93	4.41	0.9007	1.19
Neural Field	0.8234	16.75	24.68	4.22	0.9075	1.14
PINN-Spectral	0.8198	17.01	25.14	4.31	0.9051	1.17

Task 2: Forward Spectrum Prediction ($5 \to 100$)

Model	MSE↓	MAE↓	CosSim↑	SAM(°)↓	DTW↓	PkWL(nm)↓	BW(nm)↓	PW-Corr↑
MLP	0.00020	0.00808	0.9970	4.16	0.034	11.99	14.71	0.9964
ResNet	0.00009	0.00656	0.9964	4.17	0.025	20.20	13.54	0.9986
FT-Transformer	0.02456	0.10377	0.8279	32.16	0.399	148.66	180.33	0.3301
UNet1D	**0.00003**	**0.00356**	**0.9996**	**1.60**	**0.014**	**6.38**	**3.11**	**0.9994**
FNO	0.00004	0.00394	0.9991	2.10	0.016	10.27	5.74	0.9993
DeepONet	0.00280	0.03147	0.9695	13.63	0.121	47.45	57.83	0.9494
Neural Field	0.01385	0.08655	0.9104	23.32	0.349	79.19	125.77	0.9268
PINN-Spectral	0.01386	0.08668	0.9102	23.35	0.349	76.30	124.53	0.9260

Task 3: Inverse Scalar Prediction ($4 \to 5$)

Model	$SR_{5\%} \uparrow$(%)	$SR_{10\%} \uparrow$(%)	MAE↓	RMSE↓	MedAE↓	MRE↓
XGBoost	1.5	9.3	43.75	53.88	38.24	0.110
Random Forest	2.5	13.1	38.77	49.95	30.86	0.104
MLP	2.3	12.3	40.10	50.61	33.33	0.105
ResNet	**3.5**	**15.7**	**36.26**	**47.02**	**28.68**	**0.098**
FT-Transformer	0.7	6.1	50.96	64.88	44.07	0.127
MDN (10 mix.)	0.4	3.5	63.64	81.69	50.79	0.159
cVAE	0.0	0.8	74.74	92.16	65.47	0.213
Normalizing Flow	0.0	0.3	90.42	108.17	82.45	0.263
PINN	0.7	5.3	51.65	62.83	46.52	0.130

Task 4: Inverse Spectrum Prediction ($100 \to 5$)

Model	$SR_{5\%} \uparrow$(%)	$SR_{10\%} \uparrow$(%)	MAE↓	RMSE↓	MedAE↓	Cond. Proxy↓
MLP	2.1	10.8	49.23	59.43	44.99	0.001
ResNet	**3.9**	**15.9**	**43.42**	**53.11**	**38.33**	0.001
CNN-1D	0.2	3.3	60.93	73.82	55.28	0.002
FT-Transformer	1.2	8.8	51.88	66.47	43.42	0.001
INN	1.1	7.9	52.14	64.27	46.09	0.001
Deep Ensemble (CNN)	0.3	3.5	60.60	73.47	54.92	0.002
Tandem Network	0.3	3.2	68.69	88.03	54.84	0.003
cVAE	0.1	0.9	79.78	98.23	70.38	0.001
DDPM Diffusion	0.0	0.4	87.36	105.23	78.72	0.002

Noise Robustness (ResNet, Task 4)

Noise σ	$SR_{5\%} \uparrow$(%)	MAE↓
0.001	4.0	43.37
0.005	3.8	43.80
0.010	3.6	44.96
0.020	3.2	48.39
0.050	2.3	59.47

($r^2 = 0.9547$), consistent with $\lambda_{\text{center}} = n_{\text{eff}}\Lambda$; secondary correlations are weak (fill factor vs. bandwidth $\rho = 0.0328$; oxide vs. n_{eff} $\rho \approx 0.436$). Spectral derivative outliers (2.27%–2.48%) localize at Lorentzian resonance edges. Spectral correlations decay with $L_c \approx 8.8$ nm ($\sim 1/4$ resonance width), yielding ~ 4 independent degrees per resonance. FEM comparison confirms energy conservation, phase-matching trends, and qualitative envelope shape; the

Table 4. FEM validation of physical consistency: energy conservation, phase-matching trends, and spectral envelope correlation.

Λ (nm)	ff	t_{etch} (nm)	t_{Si} (nm)	t_{ox} (nm)	λ_c^{FEM} (nm)	λ_c^{Gen} (nm)	$\sigma_\lambda^{\text{FEM}}$ (nm)	$\sigma_\lambda^{\text{Gen}}$ (nm)	Corr.
550	0.60	120	240	1500	1290.3	1591.9	135.1	99.2	0.258
450	0.50	60	220	2000	1214.9	1361.6	115.9	77.7	-0.136
650	0.70	180	260	1000	1348.3	1600.0	119.5	133.7	0.378

central-wavelength offset is expected from the approximate effective index and does not affect the internal consistency of the benchmark. Apparent anomalies are explained by energy closure, controlled noise, and precision limits: negative absorbance (0.54% of points) remains $> -10^{-4}$; entropy ratio $> 200\%$ reflects sub-grid variability, not inconsistency.

7.2. Forward Modeling: Trees for Scalars, Convolutions for Spectra

The forward prediction tasks reveal a striking dichotomy between scalar and spectral outputs. For scalar optical metrics (Task 1), gradient-boosted trees decisively outperform deep neural networks. Random Forest achieves $R^2 = 0.978$ and MAE $= 1.287$, whereas the best deep learning model (ResNet) attains only $R^2 = 0.759$. This gap suggests that the mapping from geometry to scalar metrics contains non-smooth or piecewise-constant variations—likely arising from discrete resonance hopping or mode switching—that axis-aligned tree splits capture efficiently but gradient-based optimizers struggle to resolve. The underperformance of physics-informed neural networks (PINNs) on this task further indicates that explicit physics regularization does not compensate for the inductive bias mismatch.

In stark contrast, spectral forward prediction (Task 2) is dominated by architectures with strong inductive biases for smooth functional data. UNet1D, with its skip connections preserving high-frequency detail, achieves near-perfect reconstruction: MSE $= 3 \times 10^{-5}$, cosine similarity $= 0.9996$, and peak-wavelength error of only 6.38 nm. Operator learning methods—Fourier Neural Operator (FNO) and DeepONet—also perform admirably, confirming that the spectral forward problem is well-posed and amenable to neural operator approaches. Transformer-based models, lacking built-in smoothness or locality priors, lag significantly (FT-Transformer MSE $= 0.02456$), underscoring the importance of architectural alignment with problem structure.

7.3. The Fundamental Ill-Posedness of Inverse Design

Perhaps the most consequential finding emerges from the inverse tasks (Tasks 3 and 4): despite the richness of the GC-500K dataset and the diversity of model architectures, strict success rates ($\text{SR}_{5\%}$) never exceed 4%. This ceiling holds whether the input consists of four scalar metrics (Task 3, best ResNet $\text{SR}_{5\%} = 3.5\%$) or the full 100-point transmittance spectrum (Task 4, best ResNet $\text{SR}_{5\%} = 3.9\%$). The marginal improvement from scalar to spectral inputs confirms that the limiting factor is not an information bottleneck but rather the intrinsic non-uniqueness of the inverse mapping: multiple distinct geometric configurations produce nearly identical optical responses. This is a physical reality of grating coupler design, not a shortcoming of data or models.

Generative models—including mixture density networks, conditional variational autoencoders, and denoising diffusion probabilistic models—systematically underperform deterministic regressors on point-estimate metrics. While these models are designed to capture multimodal conditional distributions, their diversity objectives conflict with precise recovery of a single ground-truth geometry. The trade-off between distributional fidelity and point accuracy is inherent: a generative model that correctly represents the full posterior will necessarily produce samples that deviate from any specific target geometry, depressing strict success rates. This observation has practical implications: for applications requiring single-shot inverse design (e.g., optimizing a device to meet exact specifications), deterministic regression remains preferable; generative models are better suited for exploring design spaces or providing multiple candidate solutions for downstream selection.

Noise robustness analysis (Table 7) indicates that inverse spectrum predictions tolerate Gaussian perturbations up to $\sigma \approx 0.010$ with only modest degradation ($\text{SR}_{5\%}$ dropping from 4.0% to 3.6%). This establishes a viable operational envelope for experimental deployment where measurement noise is unavoidable, though performance deteriorates more sharply beyond $\sigma = 0.020$.

7.4. Limitations and Future Directions

Several limitations of the present study warrant discussion. First, the analytical generator, while validated against FEM for key observables, employs approximations (e.g., simplified dispersion, scalar effective indices) that may miss subtle polarization-dependent or higher-order diffractive effects. Future iterations could incorporate more sophisticated semi-analytical models or directly couple with high-throughput FEM simulations. Second, the benchmarking study focused on point-estimate accuracy; a more complete evaluation of generative models would require distributional metrics such as negative log-likelihood or maximum mean discrepancy,

which are computationally intensive for high-dimensional spectral outputs. Third, the dataset currently includes only transmittance and reflectance spectra; incorporating phase information or near-field profiles could enrich the inverse design problem and potentially mitigate non-uniqueness. Finally, the observed ceiling on inverse success rates suggests that purely data-driven approaches may need to be augmented with active learning, Bayesian optimization, or physics-constrained generative models that explicitly encode known degeneracies.

8. Conclusion

We have introduced GC-500K, a large-scale, physics-consistent synthetic dataset comprising 500,000 grating coupler designs with corresponding reflectance, transmittance, and absorbance spectra, geometric parameters, and derived optical metrics. Comprehensive statistical validation confirms that the dataset satisfies rigorous criteria for independence, smoothness, diversity, and physical plausibility, with all apparent anomalies quantitatively explained by energy conservation, numerical precision, and controlled noise injection. Comparison with finite-element simulations verifies that the generator respects the essential physical laws (energy conservation, phase-matching) while remaining a simplified synthetic model. GC-500K is thus a physically consistent benchmark, not a calibrated physical surrogate. A systematic benchmarking study across four canonical tasks—forward scalar, forward spectrum, inverse scalar, and inverse spectrum—evaluates 23 distinct model architectures spanning classical machine learning, deep learning, operator methods, and generative models. Key findings include: (i) tree-based models excel at scalar forward prediction, highlighting the non-smooth nature of geometry–scalar mappings; (ii) UNet1D achieves near-perfect spectral forward modeling, demonstrating the efficacy of convolutional inductive biases for functional data; and (iii) inverse design success rates remain fundamentally bounded below 4%, quantitatively confirming the intrinsic non-uniqueness of the grating coupler inverse problem. GC-500K establishes a challenging and realistic benchmark for data-driven photonic design, providing a standardized testbed for evaluating surrogate models, probing the limits of inverse design, and guiding future methodological development. The dataset, along with the complete benchmarking codebase will be publicly released upon acceptance.

Impact Statement

This paper introduces GC-500K, a large-scale synthetic benchmark dataset for data-driven silicon photonics grating coupler design. The dataset and accompanying benchmarking framework are intended to accelerate reproducible research in machine learning for nanophotonic inverse design. By providing a physics-consistent, openly validated dataset, this work lowers barriers to entry for the photonic ML community and enables fair model comparison. We identify no specific harmful societal consequences of this contribution beyond those generally associated with advances in machine learning for engineering and scientific computing.

Data and Code Availability

The GC-500K dataset (552.8 MB HDF5 archive) is available on Zenodo at https://doi.org/10.5281/zenodo.18116411 under CC-BY-4.0 license. The GC-Bench benchmarking framework, including all model implementations, training scripts, and evaluation notebooks, is available at https://github.com/Reckonchamp12/GC_500K_BENCHMARK under the MIT license.

References

Abdelraouf, O. A., Ahmed, A., Eldele, E., and Omar, A. A. Physics-informed neural networks in electromagnetic and nanophotonic design. *arXiv preprint arXiv:2505.03354*, 2025.

Chen, Y., Lu, L., Karniadakis, G. E., and Dal Negro, L. Physics-informed neural networks for inverse problems in nano-optics and metamaterials. *Optics express*, 28(8): 11618–11633, 2020.

Chen, Y., Ding, Z., Wang, J., Zhou, J., and Zhang, M. Prediction of metasurface spectral response based on a deep neural network. *Optics Letters*, 47(19):5092–5095, 2022.

Chen, Y., Mao, C., Li, M., Li, W., Shi, M., and Wang, Q. Prediction of spectral response of all-dielectric trimer metasurface based on deep learning. *Optics Communications*, 574:131218, 2025a.

Chen, Y., Montes McNeil, A., Park, T., Wilson, B. A., Iyer, V., Bezick, M., Choi, J.-I., Ojha, R., Mahendran, P., Singh, D. K., et al. Machine-learning-assisted photonic device development: a multiscale approach from theory to characterization. *Nanophotonics*, 14(23):3761–3793, 2025b.

Davoodi, F. Active physics-informed deep learning: surrogate modeling for nonplanar wavefront excitation of topological nanophotonic devices. *Nano Letters*, 25(2): 768–775, 2025.

Ghosh, A. *Physics-Informed Machine Learning for Optical Metamaterials*. PhD thesis, University of Massachusetts Lowell, 2022.

Huang, W.-T., Hong, K.-B., Liu, A.-C., Wang, H.-C., Lin, C.-H., Weng, C.-Y., Chang, C.-H., Kao, T.-S., Hong, Y.-H., Chang, S.-W., et al. Designing an efficient silicon photonics system: direct integrating photonic crystal

surface emitting laser with inverse-designed perfect vertical grating coupler for optimal optical communication applications. *Journal of Lightwave Technology*, 43(7): 3338–3347, 2024.

Iyer, V., Wilson, B. A., Chen, Y., Kildishev, A. V., Shalaev, V. M., and Boltasseva, A. Deep learning in photonic device development: nuances and opportunities. *npj Nanophotonics*, 3(1):5, 2026.

Kim, J., Li, M., Hinder, O., and Leu, P. Datasets and benchmarks for nanophotonic structure and parametric design simulations. *Advances in Neural Information Processing Systems*, 36:4685–4715, 2023.

Lin, C.-C., Na, A., Wu, Y.-K., Wang, L., and Na, N. Optimization of grating coupler over single-mode silicon-on-insulator waveguide to reach< 1 db loss through deep-learning-based inverse design. In *Photonics*, volume 11, pp. 267. MDPI, 2024.

Liu, Y., Geng, Q., Zhan, W., and Geng, Z. Forward performance prediction for all-dielectric metasurface sensors via deep learning. *Optics Express*, 33(13):28819–28831, 2025.

Liu, Z., Zhu, D., Raju, L., and Cai, W. Tackling photonic inverse design with machine learning. *Advanced Science*, 8(5):2002923, 2021.

Mall, A., Patil, A., Sethi, A., and Kumar, A. A cyclical deep learning based framework for simultaneous inverse and forward design of nanophotonic metasurfaces. *Scientific reports*, 10(1):19427, 2020.

Marzban, R., Adibi, A., and Pestourie, R. Inverse design in nanophotonics via representation learning. *Advanced Optical Materials*, 14(1):e02062, 2026.

Noureen, S., Mehmood, M. Q., Ali, M., Rehman, B., Zubair, M., and Massoud, Y. A unique physics-inspired deep-learning-based platform introducing a generalized tool for rapid optical-response prediction and parametric-optimization for all-dielectric metasurfaces. *Nanoscale*, 14(44):16436–16449, 2022.

Petrovič, A., Korček, R., Fraser, W., Cheben, P., Schmid, J. H., Ye, W. N., and Benedikovič, D. Inverse-designed grating couplers for hybrid silicon nitride photonics using metaheuristic methods. *Journal of the Optical Society of America B*, 42(11):B36–B43, 2025.

Ray, R. D. The physics constraint paradox: When removing explicit constraints improves physics-informed data for machine learning. *arXiv preprint arXiv:2512.22261*, 2025.

Roberts, N. B. and Keshavarz Hedayati, M. A deep learning approach to the forward prediction and inverse design of plasmonic metasurface structural color. *Applied Physics Letters*, 119(6), 2021.

Tu, X., Xie, W., Chen, Z., Ge, M.-F., Huang, T., Song, C., and Fu, H. Analysis of deep neural network models for inverse design of silicon photonic grating coupler. *Journal of Lightwave Technology*, 39(9):2790–2799, 2021.

Wang, Y., Wang, Y., Wang, C., Zhang, Z., Zhang, K., Yang, X., Yang, G., and Zhang, Y. High-precision inverse design of multi-category grating couplers based on kan-transformer. *Journal of Lightwave Technology*, 2025.

Wang, Y., Wang, Y., Yang, G., Zhang, K., Yang, X., Wang, C., and Zhang, Y. Inverse design of 3d integrated high-efficiency grating couplers using deep learning. *Chinese Physics B*, 2026.

Wiecha, P. R., Arbouet, A., Girard, C., and Muskens, O. L. Deep learning in nano-photonics: inverse design and beyond. *Photonics research*, 9(5):B182–B200, 2021.

Xiao, H., Ran, Z., Huang, D., Hou, Q., Wang, R., Chen, X., Ren, S., Zhang, Y., and Zhang, L. Efficient inverse design and optimization of grating couplers using deep neural network models. *Optical Engineering*, 64(8):085104–085104, 2025.

Xu, Y., Li, F., Gu, J., Bi, Z., Cao, B., Yang, Q., Han, J., Hu, Q., and Zhang, W. Spectral transfer-learning-based metasurface design assisted by complex-valued deep neural network. *Advanced Photonics Nexus*, 3(2):026002–026002, 2024.