MIR-QBIC: A HIGH-Q ALL-DIELECTRIC METASURFACE BENCHMARK FOR MID-INFRARED BIOSENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning is advancing nanophotonic device design, yet learning the sharp, high-quality-factor (Q) resonances relevant to molecular sensing remains constrained by data: these features are computationally demanding to resolve and annotate, and public datasets capturing them remain scarce. We introduce MIR-QBIC, an open dataset of 13,741 rigorous coupled-wave analysis (RCWA) simulations of Ge/CaF_2 quasi-bound-state-in-the-continuum (quasi-BIC) metasurfaces operating over 5–8 μm. Each structure is described by four interpretable geometric parameters and paired with broadband and adaptively resolved transmission spectra, together with resonance wavelength, radiative quality factor, dip depth, and Fano descriptors. The responses span a broad linewidth range with a sparse high-Q tail, and a substantial fraction of resonances lie in the amide-I region motivating mid-infrared biosensing. Numerical verification, frozen quality-factor-stratified splits, and a physics-aware evaluation protocol support systematic study of scalar prediction, adaptive-window reconstruction, and high-Q generalization. Six classical and neural baselines demonstrate the benchmark's utility and difficulty: similar quality-factor accuracy can accompany substantially different adaptive-window errors, while high-Q training coverage improves generalization without resolving the sharpest responses. MIR-QBIC provides a shared, physically grounded resource for advancing narrow-feature, high-dynamic-range spectral learning.

1 INTRODUCTION

High-quality-factor (Q) resonances in all-dielectric metasurfaces provide a platform for mid-infrared molecular spectroscopy, where resonance position, linewidth, and lineshape govern the optical response near molecular absorption bands (Tittl et al., 2018; Leitis et al., 2019). Learning these responses requires data that resolve their narrow spectral structure. A coarse wavelength grid can miss a sharp resonance, whereas uniformly dense simulation becomes costly across geometries with widely varying linewidths. Consistent resonance extraction and numerical verification add further computational demands, making suitable training data a central barrier to surrogate development in this regime.

Public nanophotonic benchmarks have advanced spectral prediction and design optimization (Nadell et al., 2019; Ren et al., 2022; Kim et al., 2023), while resonance-focused resources connect optical spectra to modal properties (Lilja et al., 2026). A remaining need is an open geometry-to-spectrum collection that combines broadband context, adaptively resolved quasi-BIC profiles, and explicit physical annotations across a broad radiative-Q range. Such a resource would allow researchers to investigate narrow-feature learning without independently rebuilding a demanding simulation and annotation pipeline.

We introduce MIR-QBIC, an open dataset of 13,741 RCWA simulations of Ge/CaF_2 quasi-BIC metasurfaces over 5–8 μm, motivated by mid-infrared biosensing. Each four-parameter geometry is paired with broadband and resonance-adaptive transmission spectra, resonance wavelength, radiative quality factor, dip depth, and Fano descriptors. Discretization studies, independent-solver comparisons, and dense spectral checks assess numerical fidelity. Figure 1 summarizes the resource.

MIR-QBIC enables systematic study of how surrogate accuracy changes with resonance linewidth. A geometry-only prediction task separates scalar accuracy from adaptive-window reconstruction, while frozen splits support both in-distribution benchmarking and high-Q extrapolation. Six classical

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.

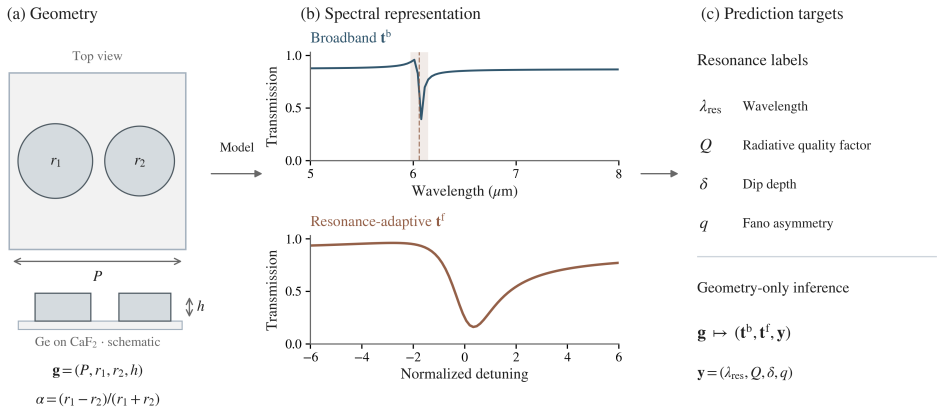

Figure 1: MIR-QBIC workflow: geometry-to-spectrum simulation, adaptive sampling, physical annotation, and separate evaluation of scalar properties, spectral shape, and resonance localization.

and neural baselines demonstrate the value of these distinctions: neural models achieve similar quality-factor accuracy, yet the coordinate model reduces adaptive-window RMSE by approximately 18% relative to the strongest fixed-output network. Adding high-Q training examples substantially improves performance on the same held-out high-Q samples, although the sharpest resonances remain challenging. The dataset thus provides both a reproducible basis for method comparison and a concrete setting for advancing spectral learning across sparsely represented linewidth regimes.

2 Related Work

Public nanophotonic benchmarks. The all-dielectric metasurface benchmark examined by Deng et al. (2021); Ren et al. (2022) maps a 14-dimensional geometry of four SiC elliptical resonators to densely sampled absorptivity spectra, supporting comparisons of forward and inverse models. Kim et al. (2023) broaden this resource landscape with parametric nanophotonic optimization benchmarks across multiple device families. The closest public resonance-focused resource is QNM-Net (Lilja et al., 2026), which includes 10,000 near-infrared photonic-crystal designs with uniformly sampled spectra dominated by a Fano resonance. Its quasinormal-mode formulation connects spectral prediction to modal frequencies and decay rates. These resources establish substantial precedents for dielectric spectral learning; MIR-QBIC focuses on explicitly annotated, adaptively resolved mid-infrared quasi-BIC responses across a broad radiative-Q range. Table 1 distinguishes the closest public resources from related quasi-BIC design studies.

Learning narrow resonances and quasi-BIC sensors. Prior work addresses both the representation and design of sharp resonances. Ma et al. (2022) separate smooth spectral backgrounds from narrow features and use adaptive acquisition to resolve high-Q BIC responses. Wang et al. (2024) learn resonance wavelength and quality factor for forward and inverse design. For sensing, Metaformer (Gao et al., 2024) models symmetry-broken silicon metasurfaces supporting quasi-BIC resonances for fingerprint and refractive-index detection. Most directly, Wei & Geng (2026) develop transformer forward and inverse models for mid-infrared molecular fingerprinting, reporting designed resonances with $Q \approx 200$. These studies establish the relevance of learning-based quasi-BIC design and adaptive spectral resolution to the application motivating MIR-QBIC.

Dataset gap and contribution. The remaining data need is a reusable collection that pairs metasurface geometry with both broadband context and resolved narrow-resonance profiles, together with explicit physical annotations across a wide linewidth range. Such data are needed to study how learning accuracy changes as resonances sharpen, without requiring each study to repeat costly electromagnetic simulation and resonance extraction. MIR-QBIC supplies this resource for Ge/CaF_2 quasi-BIC metasurfaces in the mid-infrared, with paired spectral representations and physical annotations spanning broad responses and a sparse high-Q tail.

Table 1: Positioning MIR-QBIC against public nanophotonic resources and closely related resonance-learning studies.

Resource or study	Physical system and learning target	Data contribution and study emphasis
Public learning resources		
ADM (Deng et al., 2021; Ren et al., 2022)	Four SiC elliptical resonators; 14 geometry parameters	Dense, fixed-grid absorptivity spectra for forward and inverse design
Kim et al. (Kim et al., 2023)	Multiple nanophotonic device families	Optical-response datasets and parametric optimization benchmarks
QNM-Net (Lilja et al., 2026)	Near-infrared photonic-crystal slab; five geometry parameters	10,000 designs with fixed-grid Fano spectra; quasinormal-mode learning
Closely related resonance-learning studies		
RIDL (Ma et al., 2022)	Photonic BICs with narrow resonances	Background–resonance decomposition and adaptive acquisition
Metaformer (Gao et al., 2024)	Symmetry-broken Si metasurfaces on fused silica	Quasi-BIC sensor design and experiments
Wei and Geng (Wei & Geng, 2026)	Rotated elliptical Ge dimers on CaF_2	Transformer forward and inverse design with reflection spectrum
MIR-QBIC	Ge/CaF_2 dimers, 5–8 μm; four geometry parameters	13,741 designs; broadband and adaptive spectra; explicit $\lambda_{\mathrm{res}}, Q, \delta, q$ labels; numerical verification and frozen splits

Table 2: Geometric sampling ranges; $r_{1,2} = \bar{r}(1 \pm \alpha)$.

Parameter	Symbol	Range	Unit
Lattice period	P	2800–3500	nm
Mean pillar radius	$\bar{r}$	500–750	nm
Pillar height	h	1000–2600	nm
Asymmetry	α	0.005–0.30	—

3 The MIR-QBIC Dataset

3.1 Physical System and Generation

Each structure comprises two circular Ge pillars on a CaF_2 substrate in a square lattice under normally incident, x-polarized illumination. The materials are modeled as lossless over 5–8 μm, with refractive indices approximately 4.0 and 1.4, respectively; the reported Q therefore characterizes radiative decay. The geometry is $\mathbf{g} = (P, r_1, r_2, h)$, with lattice period P, radii r_1, r_2, and height h. We parameterize the radii as $r_{1,2} = \bar{r}(1 \pm \alpha)$, where $\alpha = (r_1 - r_2)/(r_1 + r_2)$ (Table 2). The quasi-BIC scaling $Q \propto \alpha^{-2}$ (Koshelev et al., 2018) makes asymmetry a physical control over spectral sharpness. Overlapping geometries and those violating the substrate sub-diffraction condition are excluded before simulation.

Transmission is computed using GPU-accelerated RCWA with `torcwa` (Kim & Lee, 2023), Fourier order 11, periodic boundary conditions, and a nominal real-space spacing of 30 nm, subject to a minimum grid size of 128 points per axis. A coarse sweep locates the transmission dip; adaptive refinement and Fano fitting characterize the resonance. The released population retains resonances within $5000 \leq \lambda_{\mathrm{res}} \leq 8000$ nm and transmission satisfying $\max T \leq 1.01$, yielding the 13,741-sample benchmark.

3.2 Spectral Representation and Annotations

Each sample contains a 50-point broadband transmission spectrum on a shared 5–8 μm grid. For $Q \geq 15$, a 40-point adaptive spectrum samples a nominal window of ± 6 linewidths around λ_{res}, clipped

Table 3: Numerical verification over the evaluated subsets. Entries are median relative discrepancies, except resonance-window RMSE, reported as the median per-spectrum RMSE in transmission units.

Comparison	Quantity	Discrepancy
Fourier order 11 vs. 17	Resonance wavelength	0.017%
	Quality factor	5.36%
	Window peak-to-trough depth	0.81%
	Resonance-window RMSE	0.021
Grid n_x vs. $2n_x$ (order 11)	Resonance wavelength	0.089%
	Quality factor	2.79%
Independent solver (`grcwa`)	Resonance wavelength	0.082%
	Quality factor	5.93%

to the operating band where necessary. The corresponding physical wavelengths and normalized detuning coordinates are stored with each sample.

The baseline scalar targets are resonance wavelength λ_{res}, radiative quality factor Q, and minimum sampled transmission $T_{\min}$. The latter is the minimum of the adaptive spectrum when available and of the broadband spectrum otherwise. It is distinct from the peak-to-trough depth used for profile evaluation.

The scalar metadata additionally provide the Fano asymmetry parameter q and original extraction-window depth δ. These annotations are distinct from the baseline target $T_{\min}$ and from depth recomputed on the released adaptive spectrum.

The 40 adaptive transmission samples are stored together with their physical wavelength coordinates. For each record, reference and predicted adaptive values are compared pointwise on the same stored wavelength grid, $\Lambda_i = \{\lambda_{ij}\}_{j=1}^{40}$; no interpolation or cross-sample resampling is required for the main benchmark. The normalized-detuning coordinates are retained as auxiliary metadata for coordinate-based modeling and shape analyses. Boundary clipping changes the extent of a record's adaptive window but not the one-to-one physical correspondence between its target and prediction entries. Appendix E uses a restricted canonical-detuning cohort only for a separate decomposition of normalized shape accuracy and placement-aware wavelength-space reconstruction. Reference resonance parameters are used to construct targets, but prediction uses geometry alone. A released Q-reliability flag identifies high ($Q < 200$), medium ($200 \leq Q < 1000$), and low ($Q \geq 1000$) reliability, reflecting the increasing numerical sensitivity of sharp resonances.

3.3 Numerical Verification

We assess numerical fidelity through discretization refinement, independent-solver comparison, and direct validation of the released spectral sampling.

Solver verification. We select 48 fine-eligible geometries for the Fourier-order comparison, using four samples per Q decile and eight additional high-Q samples. Orders $\{7, 9, 11, 13, 15, 17\}$ are compared at fixed real-space grid and extraction settings, with order 17 as reference. Grid refinement uses 10 selected geometries, and the independent implementation is evaluated on 12. Comparison with `grcwa` (Jin, 2020) provides an independent implementation check, with an energy-conservation residual $\leq 6.6 \times 10^{-11}$ over the checked off-resonance points ($T > 0.5$). Table 3 summarizes the discrepancies in resonance parameters and transmission. Wavelength estimates are comparatively stable, whereas quality factors exhibit greater sensitivity for the sharpest resonances, motivating the released reliability flags. Order 11 balances the observed numerical discrepancies against the computational cost of dataset generation; physical-consistency diagnostics and theoretical computational scaling are detailed in Appendix B.

Spectral-sampling validation. To assess how faithfully the compact representation retains the simulated response, we select 80 test geometries, with eight samples from each Q decile among those with adaptive spectra. For each geometry, a 241-point RCWA sweep provides a dense reference over

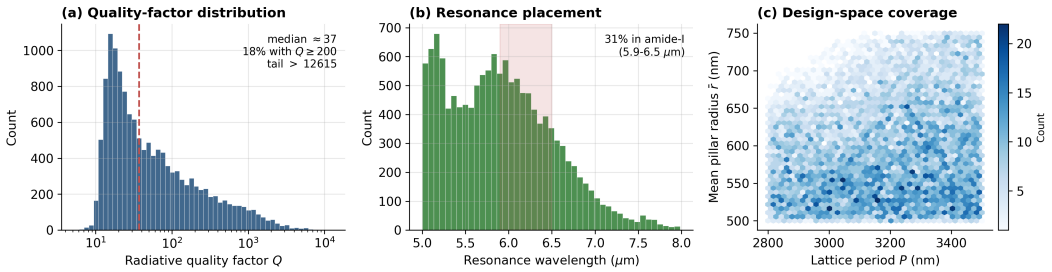

Figure 2: Distributions of quality factor, geometry, and resonance wavelength. The sparse high-Q tail defines a demanding regime within the broader design space.

the same resonance window at the production Fourier order. Piecewise-linear interpolation of the released 40-point spectra onto these grids yields a pooled transmission RMSE of 0.00913; the median and 90th-percentile per-spectrum RMSE are 0.00987 and 0.01153, respectively. This comparison directly supports the reconstruction fidelity of the adaptive representation across the sampled low-, intermediate-, and high-Q regimes. Appendix B.3 reports the evaluated Q range, stratified results, and interpolation procedure.

3.4 Dataset Composition and Access

Radiative quality factors range from approximately 5.3 to 1.26×10^4, with median $Q \approx 37$. Approximately 18% of samples have $Q \geq 200$, and 30% of resonances fall within the targeted amide-I interval 5.9–6.5 μm (Figure 2). The population therefore supports both broad-response learning and evaluation in a comparatively scarce, sharp-resonance regime. Samples are distributed as compressed NumPy records with scalar metadata, solver provenance, and frozen split indices. Appendix A describes the schema.

Availability. An anonymized review snapshot provides the dataset, six baseline implementations, shared loader and evaluator, numerical-verification and supplementary-analysis code, and the documented split files at Dataset Repo. Data are distributed under CC-BY-4.0 and code under the MIT license. Permanent repositories on GitHub and Hugging Face, including the original generation pipeline, are planned upon acceptance. Versioned releases will preserve the benchmark snapshot (Appendix A).

4 Benchmark Protocol

4.1 Prediction Task

The forward task maps geometry to two spectra and three scalar targets:

$$f_\theta : \mathbf{g} \mapsto \left(\mathbf{t}^{\mathrm{b}}, \mathbf{t}^{\mathrm{f}}, \mathbf{y}\right), \qquad \mathbf{y} = (\log_{10} Q, \lambda_{\mathrm{res}}, T_{\mathrm{min}}), \tag{1}$$

where $\mathbf{t}^{\mathrm{b}} \in [0,1]^{50}$ and $\mathbf{t}^{\mathrm{f}} \in [0,1]^{40}$ use the grids defined in Section 3.2. Multi-task regression supervises all available targets; fine-spectrum loss is masked for samples without an adaptive window. Predicting $\log_{10} Q$ gives comparable weight to multiplicative errors across the dynamic range.

4.2 Frozen Splits

Two fixed partitions distinguish prediction within the sampled distribution from generalization to sharper resonances. The in-distribution benchmark uses a nominal $70/15/15$ training–validation–test split, stratified by Q decile. The frozen partitions contain 9,611 training, 2,060 validation, and 2,070 test samples.

The high-Q benchmark tests how access to sharp resonances during training affects generalization. A threshold at the 80th percentile ($\tau \approx 178$) separates 10,992 low-Q samples from 2,749 high-Q samples spanning $Q \in [178, 1.26 \times 10^4]$. The low-$Q$ population provides 7,692 training, 1,650

validation, and 1,650 test samples. The high-Q population is partitioned with Q stratification into 1,374 training-accessible samples and 1,375 held-out test samples.

The released coverage partitions are reproduced from the original deterministic splitting procedure and stored separately from the in-distribution split.

We evaluate two training regimes on this identical high-Q test set: extrapolation, which is trained exclusively on low-Q samples, and a seen regime augmented with the training-accessible high-Q samples. Architectures and hyperparameters selected on the in-distribution split remain fixed. This design quantifies the benefit of adding high-Q training data without increasing model capacity or introducing high-Q validation feedback. All partitions use the same fixed seed, with test sets reserved for final evaluation. Additional thresholds at the 70th and 90th percentiles assess sensitivity to the generalization boundary (Appendix F.2).

4.3 Evaluation

The evaluation separates scalar prediction, spectral reconstruction, and resonance localization, providing complementary measures of surrogate fidelity.

Scalar and spectral accuracy. For reference values y_k and predictions $\hat{y}_k$,

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{k=1}^{M}(y_k - \hat{y}_k)^2}, \tag{2}$$

where M is the number of evaluated values. Scalar targets are assessed using RMSE and R^2, with MAE additionally reported for $\log_{10} Q$ and resonance wavelength. Quality-factor errors are computed in $\log_{10} Q$ space, with RMSE and MAE reported in dex.

Broadband RMSE pools the 50 wavelength samples across all test geometries. Adaptive-window RMSE pools pointwise errors over the 40 stored physical wavelength locations for all fine-eligible samples. For sample i, t^{f}_{ij} and $\hat{t}^{\mathrm{f}}_{ij}$ correspond to the same stored wavelength λ_{ij} by construction, so the score is directly aligned on the released adaptive physical grid and requires no resampling. Both spectral metrics are expressed in transmission units. The main adaptive score does not use predicted resonance position or linewidth; Appendix E separately evaluates placement-aware wavelength reconstruction after remapping with predicted resonance parameters.

Profile diagnostics and stratification. Additional diagnostics measure absolute MAE in peak-to-trough depth and the historical reference-grid half-depth span. The latter is a normalized-output-grid diagnostic rather than a physical linewidth error. Stratified analyses examine quality-factor dependence in the generalization and loss studies. Appendix C specifies the extraction and aggregation procedures and the supplementary localization metrics.

5 Baselines and Results

Matched protocol. Six models share training-partition standardization of geometry and scalar targets, a 20-trial random hyperparameter search with validation-based selection, and the same held-out test set. Spectra remain in transmission units. Gradient-trained models, including Linear, are evaluated over five final training seeds; SVR is deterministic and reported from one final fit. Every model predicts the scalar and spectral targets, enabling comparison of geometry encodings and spectral decoders under a common task.

Model families. Linear uses an affine backbone with linear scalar heads and sigmoid-bounded spectral heads. SVR fits one radial-basis support-vector regressor per output. MLP uses a fully connected trunk and separate heads. CNN expands geometry into a sequence decoded through one-dimensional convolutions. Transformer applies self-attention to geometric parameter tokens. Coordinate combines a geometry encoder with separate broadband and adaptive coordinate decoders using random Fourier features (Tancik et al., 2020). Each decoder shares weights across its own output coordinates; the two decoders have separate parameters. Appendix D summarizes the training protocol, and presents seed variability.

Table 4: In-distribution results. Gradient-trained entries are means over five seeds; SVR uses one deterministic fit. Spectral errors are pooled transmission RMSE. Conditional wavelength RMSE uses scalar predictions with absolute error at most 100 nm. Detection is the fraction of fine-eligible samples whose predicted broadband minimum lies within 50 nm of the reference resonance. Best values are bold.

Method	$\log_{10} Q$ RMSE (dex)	Conditional λ_{res} RMSE (nm)	Broadband RMSE	Adaptive-window RMSE	Detection ± 50 nm
Linear	0.391	56.4	0.220	0.228	0.017
SVR	0.175	**32.3**	0.131	0.160	0.101
MLP	0.132	41.4	0.102	0.117	0.168
CNN	**0.124**	38.5	0.109	0.125	0.164
Transformer	0.133	37.6	0.138	0.146	0.113
Coordinate	0.127	40.8	**0.057**	**0.096**	**0.369**

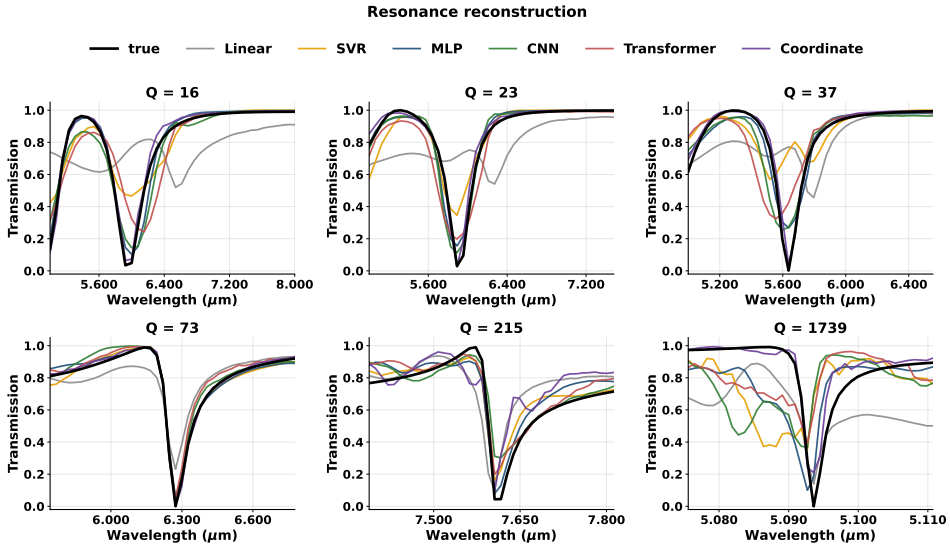

Figure 3: Reference and predicted resonance profiles across increasing quality factor. The examples illustrate the growing difficulty of reconstructing sharp spectral features.

Scalar accuracy does not determine reconstruction quality. The four neural models achieve similar $\log_{10} Q$ RMSE, while adaptive-window RMSE ranges from 0.096 to 0.146 (Table 4). Coordinate improves upon the strongest fixed-output network, MLP, by approximately 18% on adaptive-window RMSE and 44% on broadband RMSE; its detection rate also increases from 0.168 to 0.369. CNN gives the lowest quality-factor error, whereas Coordinate gives the lowest broadband and adaptive-window errors and the highest broadband-minimum detection rate. The rankings therefore depend on the predicted quantity and evaluation criterion.

Sharp resonances remain difficult. Reconstruction error increases with Q, and all six baselines struggle in the highest decile. The examples in Figure 3 illustrate the loss of shape fidelity as linewidth decreases. This remaining difficulty makes the high-Q tail useful for distinguishing progress beyond aggregate performance and motivates the controlled coverage study below.

6 Analysis

6.1 High-Q Generalization and Training Coverage

On the low-Q test set, the four neural models achieve $\log_{10} Q$ RMSE of 0.064–0.074 dex ($R^2 \approx 0.95$). Training on low-Q data alone increases error on the held-out high-Q set to 0.56–0.62 dex, with R^2

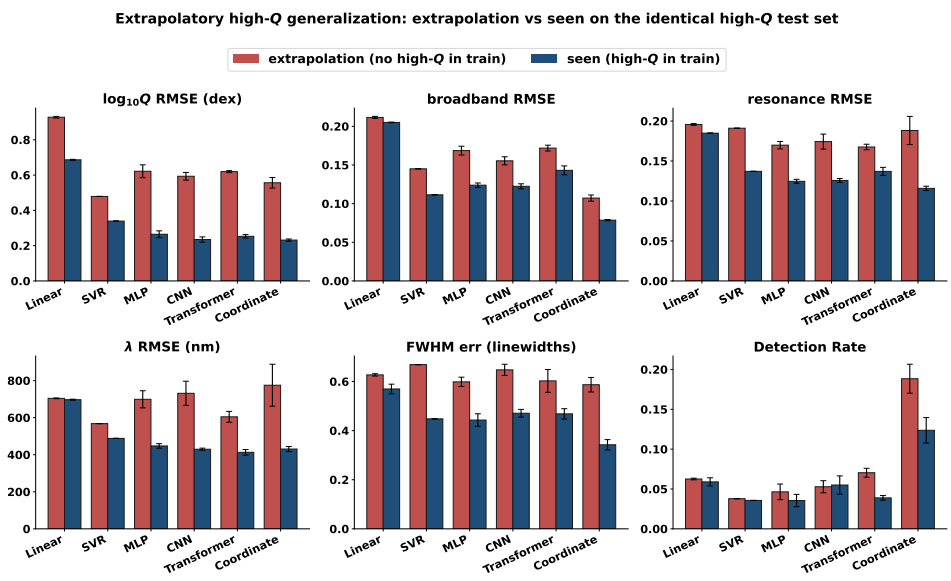

Figure 4: Effect of high-Q training augmentation. Adding high-Q examples reduces most reported errors on the same held-out high-Q test set, while model architectures and hyperparameters remain unchanged.

between -2.2 and -1.5 (Appendix F.1, Figure 8). Adding the high-Q seen partition reduces error to 0.23–0.27 dex and restores R^2 to 0.43–0.56 on the same high-Q test samples (Figure 4). This improvement occurs while model selection remains restricted to low-Q validation. The result shows that augmenting the training distribution with high-Q examples markedly reduces error on the held-out high-Q set, although augmentation also increases training-set size (Appendix F.2). Substantial error remains after augmentation.

The training regime also changes model rankings. Under extrapolation, Coordinate retains the lowest broadband RMSE (0.107, versus 0.145 for the next-best model), but its adaptive-window reconstruction advantage returns only with high-Q training coverage: high-Q adaptive-window RMSE falls from 0.188 to 0.116. SVR has the lowest extrapolatory $\log_{10} Q$ error (0.48 dex), yet benefits least from augmentation and is overtaken by the neural models. Architecture comparisons thus depend on both the predicted quantity and the training distribution.

6.2 Shape Accuracy and Physical Reconstruction

A supplementary single-seed analysis on 829 test samples selected for a canonical normalized-detuning comparison distinguishes shape prediction from placement-aware reconstruction using predicted resonance position and linewidth. This cohort restriction is specific to the supplementary decomposition and is not required by the main adaptive-window RMSE. The coordinate model achieves the lowest shape RMSE among geometry-only predictors (0.0783), whereas SVR achieves the lowest physical-wavelength RMSE (0.2158, compared with 0.2258 for the coordinate model). Improved shape prediction therefore does not necessarily translate into improved wavelength-space reconstruction. Appendix E reports the complete comparison and specifies the subset, interpolation, and endpoint-extension procedure.

6.3 Representation and Training Choices

We examine whether reconstruction benefits from constraints on the spectral representation and from physically motivated target parameterization. Additional configuration and sensitivity studies are reported in Appendix F.

Consistency between spectral representations. The paired broadband and adaptive targets provide a natural constraint: their predictions should remain consistent where they describe the same response.

Relative to multi-task MSE, a continuity term combined with resonance weighting reduces broadband RMSE by approximately 10% and adaptive-window RMSE by 4%, without modifying the architecture (Figure 10). Improvements extend across all ten Q deciles for broadband reconstruction and eight of nine evaluated deciles for the adaptive spectrum. The benefit therefore extends beyond the aggregate score, although the sharpest resonances remain challenging.

Continuity uses each record's stored physical wavelengths; the historical wing weights use the shared reference encoding of adaptive output positions.

The choice of physical constraint matters. A symmetric Lorentzian-tail prior degrades reconstruction, consistent with its mismatch to asymmetric Fano profiles, whereas generic smoothness provides no appreciable benefit. These contrasting outcomes indicate that physical motivation alone is insufficient: the constraint must reflect the structure of the modeled response. Here, enforcing consistency between spectral representations and emphasizing the resonance region improves fidelity without prescribing a symmetric lineshape.

Target and input parameterization. Regressing raw Q produces approximately eight times greater error in dex than regressing $\log_{10} Q$, consistent with the disproportionate influence of large quality factors under a linear-scale squared-error objective. Adding the derived asymmetry α to the four geometric inputs provides no measurable improvement for the tested models. Coordinate reconstruction also varies little across the tested Fourier scales, feature counts, and decoder activations (Appendix F.3). Together, these ablations support logarithmic targets and representation-consistent training while establishing robustness to the tested decoder configurations.

MIR-QBIC describes simulated radiative responses of one material system and a four-parameter geometry family. The lossless model excludes absorption, finite-array effects, roughness, and fabrication variation; its high radiative Q values do not establish experimentally attainable performance. Numerical uncertainty also increases in the sharpest resonances, as reflected by the reliability flags. Generalization to other materials and geometries, and experimental sensing performance, remain untested. Appendix H provides additional physical context. The main adaptive-spectrum benchmark compares prediction and reference pointwise at identical stored physical wavelengths within each record. Its pooled RMSE therefore contains no coordinate mismatch. Because each adaptive window is resonance centered, however, the aggregate score averages errors over different local wavelength intervals across samples rather than over one global 5–8 μm axis. Appendix E separately decomposes normalized shape accuracy and placement-aware wavelength-space reconstruction. Open reuse can reduce duplicated simulation effort; downstream sensing devices require experimental validation, and diagnostic applications require additional clinical validation.

7 Conclusion

MIR-QBIC provides a numerically verified benchmark for learning narrow quasi-BIC responses in all-dielectric mid-infrared metasurfaces. The results show that similar quality-factor accuracy can coexist with substantially different spectral reconstruction: Coordinate achieves the lowest broadband and adaptive-window RMSE, whereas CNN achieves the lowest quality-factor error. Increasing high-Q training coverage substantially improves prediction on the same held-out high-Q population, although the sharpest responses remain challenging and numerically sensitive. Together with frozen splits and resonance-aware evaluation, MIR-QBIC provides a reproducible setting for developing and evaluating spectral surrogates across a broad range of resonance linewidths.

AI Use Statement

A large language model was used to refine grammar and phrasing. The model did not generate data or scientific interpretations. The authors reviewed all resulting revisions and accept full responsibility for the final manuscript.

REFERENCES

Yang Deng, Juncheng Dong, Simiao Ren, Omar Khatib, Mohammadreza Soltani, Vahid Tarokh, Willie Padilla, and Jordan Malof. Benchmarking data-driven surrogate simulators for artificial electromagnetic materials. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/8c19f571e251e61cb8dd3612f26d5ecf-Abstract-round2.html.

Yuan Gao, Wei Chen, Fajun Li, Mingyong Zhuang, Yiming Yan, Jun Wang, Xiang Wang, Zhaogang Dong, Wei Ma, and Jinfeng Zhu. Meta-attention deep learning for smart development of metasurface sensors. *Advanced Science*, 11(42):2405750, 2024. doi: 10.1002/advs.202405750.

Weiliang Jin. grcwa: Rigorous coupled-wave analysis with automatic differentiation. https://github.com/weiliangjinca/grcwa, 2020. Software repository.

Changhyun Kim and Byoungho Lee. TORCWA: Gpu-accelerated fourier modal method and gradient-based optimization for metasurface design. *Computer Physics Communications*, 282:108552, 2023. doi: 10.1016/j.cpc.2022.108552.

Jungtaek Kim, Mingxuan Li, Oliver Hinder, and Paul W. Leu. Datasets and benchmarks for nanophotonic structure and parametric design simulations. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 4685–4715, 2023.

Kirill Koshelev, Sergey Lepeshov, Mingkai Liu, Andrey Bogdanov, and Yuri Kivshar. Asymmetric metasurfaces with high-Q resonances governed by bound states in the continuum. *Physical Review Letters*, 121(19):193903, 2018. doi: 10.1103/PhysRevLett.121.193903.

Aleksandrs Leitis, Andreas Tittl, Mingkai Liu, Bang Hyun Lee, Man Bock Gu, Yuri S. Kivshar, and Hatice Altug. Angle-multiplexed all-dielectric metasurfaces for broadband molecular fingerprint retrieval. *Science Advances*, 5(5):eaaw2871, 2019. doi: 10.1126/sciadv.aaw2871.

Viktor A. Lilja, Albin J. Svärdsby, Timo Gahlmann, and Philippe Tassin. A general framework for knowledge integration in machine learning for electromagnetic scattering using quasinormal modes. *Laser & Photonics Reviews*, 20(11):e02769, 2026. doi: 10.1002/lpor.202502769. arXiv:2509.06130.

Xuezhi Ma, Yuan Ma, Preston Cunha, Qiushi Liu, Kaushik Kudtarkar, Da Xu, Jiafei Wang, Yixin Chen, Zi Jing Wong, Ming Liu, M. Cynthia Hipwell, and Shoufeng Lan. Strategical deep learning for photonic bound states in the continuum. *Laser & Photonics Reviews*, 16(10):2100658, 2022. doi: 10.1002/lpor.202100658.

Christian C. Nadell, Bohao Huang, Jordan M. Malof, and Willie J. Padilla. Deep learning for accelerated all-dielectric metasurface design. *Optics Express*, 27(20):27523–27535, 2019. doi: 10.1364/OE.27.027523.

Simiao Ren, Ashwin Mahendra, Omar Khatib, Yang Deng, Willie J. Padilla, and Jordan M. Malof. Inverse deep learning methods and benchmarks for artificial electromagnetic material design. *Nanoscale*, 14(10):3958–3969, 2022. doi: 10.1039/D1NR08346E.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Andreas Tittl, Aleksandrs Leitis, Mingkai Liu, Filiz Yesilkoy, Duk-Yong Choi, Dragomir N. Neshev, Yuri S. Kivshar, and Hatice Altug. Imaging-based molecular barcoding with pixelated dielectric metasurfaces. *Science*, 360(6393):1105–1109, 2018. doi: 10.1126/science.aas9768.

Lanfei Wang, Wenqi Wang, Qiao Dong, Lianhui Wang, and Li Gao. Deep learning enabled inverse design of bound states in the continuum with ultrahigh Q factor. *Journal of the Optical Society of America B*, 41(2):A146–A151, 2024. doi: 10.1364/JOSAB.499287.

Yinian Wei and Zhaoxin Geng. Transformer-enabled intelligent design of high-Q quasi-BIC metasurfaces for molecular vibrational fingerprinting. *Photonics Research*, 14(4):1441–1454, 2026. doi: 10.1364/PRJ.578302.

A Data Format and Release Documentation

The MIR-QBIC release comprises compressed NumPy records (`.npz`), scalar metadata, and frozen split indices. Each record stores the four geometric parameters in nanometres, resonance annotations, transmission spectra and their coordinates, and numerical provenance. The per-record JSON configuration stores spectral-sampling constants, Fourier order, and nominal grid spacing. Material indices, grid-size limits, and solver conventions are specified in the accompanying documentation and verification code. The metadata include the Q-reliability classification, Fano asymmetry parameter `fano_q`, and original extraction-window depth `dip_depth`. Asymmetry is derived from the two stored radii. The metadata field `dip` stores $T_{\min}$: the minimum sampled adaptive transmission when `fine_used` is true, otherwise the minimum sampled broadband transmission. The field name is retained for compatibility with the baseline code. Original extraction-window depth is distinct from $T_{\min}$ and from depth computed on the released 40-point spectrum. Two original depth annotations outside $[0, 1.01]$ are preserved and identified by `dip_depth_out_of_range`; neither is used as the baseline minimum-transmission target. When no adaptive spectrum is available, `fine_T`, `fine_detuning`, and `fine_wl_nm` are stored as empty arrays.

Broadband transmission contains 50 floating-point values on the shared wavelength grid. The supplied spectral arrays are stored in `float64`; the baseline loader converts training tensors to `float32`. Records with `fine_used` set to true additionally contain 40 adaptive transmission values, normalized detuning coordinates, and their corresponding physical wavelengths. The detuning convention is

$$u = \frac{(\lambda - \lambda_{\mathrm{res}})Q}{\lambda_{\mathrm{res}}}, \qquad \lambda = \lambda_{\mathrm{res}} \left(1 + \frac{u}{Q}\right). \tag{3}$$

The adaptive window nominally spans ± 6 linewidths and is clipped to the operating band where necessary. Physical wavelength coordinates support reconstruction and evaluation; they are not supplied as inference inputs.

The six main baselines predict three scalar targets, $(\log_{10} Q, \lambda_{\mathrm{res}}, T_{\min})$, in addition to the two spectra; q is available for further modeling studies.

The in-distribution partitions are stored in `splits.npz`. The coverage partitions are exported from the original deterministic procedure to `splits_highq_coverage.npz`, with separate low-Q training, validation, and test indices and high-Q seen and test indices. Additional files provide the 70th- and 90th-percentile partitions. The separate $Q = 500$ split is not used for the reported coverage study.

The reference loader and evaluator implement the common record schema and benchmark definitions. The review snapshot includes a README describing the record schema, baseline execution, licensing, available split files, and supplementary artifacts. The maintenance plan uses tagged snapshots and repository issue tracking to document corrections and extensions while preserving the original 13,741-sample benchmark and its frozen split indices.

B Numerical Verification and Spectral Sampling

B.1 Physical Consistency

Radiative quality factor increases as geometric asymmetry decreases, with a pooled log–log slope of -1.37 (Figure 5). This association is consistent with symmetry-controlled radiative coupling. Because the pooled population also varies in period, mean radius, and height, the fitted slope is not a controlled estimate of the fixed-family scaling $Q \propto \alpha^{-2}$ (Koshelev et al., 2018).

Median dip transmission is approximately 0.006. For single-dip samples, the median separation between the fitted resonance wavelength and the reconstructed spectral minimum is approximately 17 nm. A Fano resonance's fitted center and transmission minimum need not coincide; this separation characterizes the profiles rather than providing a standalone estimate of wavelength-label error.

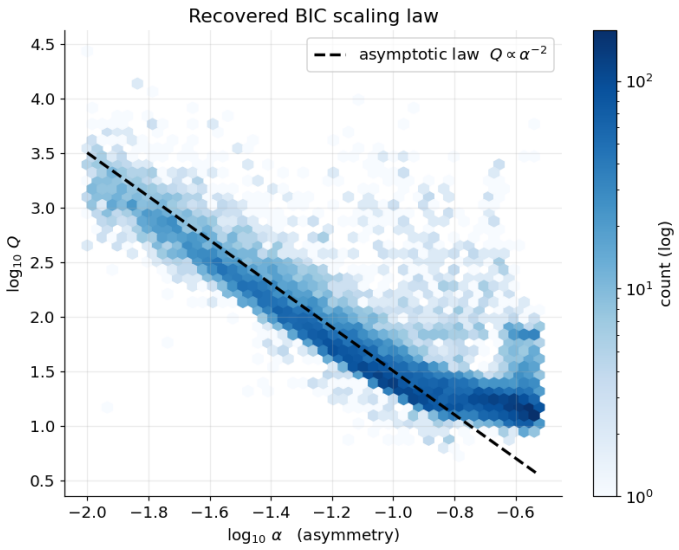

Figure 5: Radiative quality factor versus geometric asymmetry. The pooled fitted slope is -1.37, compared with the ideal fixed-family inverse-square slope of -2.

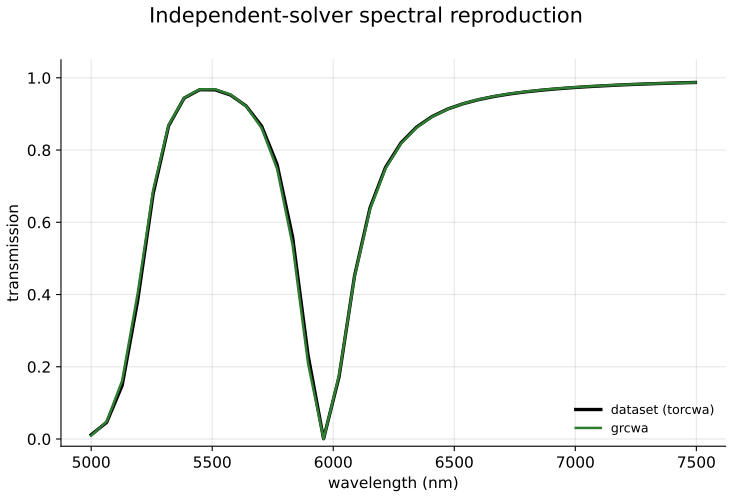

Figure 6: Transmission from `torcwa` and `grcwa` for a representative geometry.

B.2 Independent Implementation and Computational Cost

Figure 6 compares transmission from `torcwa` and `grcwa` for a representative geometry. It complements the quantitative Fourier-order, grid-refinement, and independent-implementation comparisons in Table 3. The aggregate comparisons assess numerical sensitivity, while the overlay illustrates agreement in the reconstructed spectral response.

The released comparison script evaluates both polarization settings and retains the response whose sampled minimum is closest to the stored resonance wavelength. The comparison therefore includes this selection procedure; the selected polarization is recorded for each sample.

Computational scaling. For square Fourier truncation at order m, the number of retained harmonics is $H = (2m+1)^2$. Increasing the order from 11 to 17 raises H from 529 to 1,225. For dense matrices whose dimensions scale with H, quadratic storage and cubic arithmetic imply approximately $5.4\times$ and $12.4\times$ greater requirements, respectively. These theoretical ratios explain the computational

Table 5: Released 40-point spectra compared with 241-point RCWA references after linear interpolation. RMSE is pooled over wavelengths and samples within each group and is reported in transmission units.

Reference Q	Samples	RMSE
$Q < 200$	63	0.00839
$200 \le Q < 1000$	14	0.01124
$Q \ge 1000$	3	0.01243
All evaluated samples	80	0.00913

incentive for using order 11 alongside the reported refinement checks; they are not measured end-to-end runtime ratios.

B.3 Validation of Spectral Sampling

We assess the released adaptive representation using 80 test geometries, selected as eight samples from each Q decile of the fine-eligible test population. For each geometry, transmission is computed at 241 uniformly spaced wavelengths between the endpoints of its released adaptive window, using Fourier order 11. The dense and released records agree in geometry and window endpoints.

Piecewise-linear interpolation maps the stored 40-point spectrum onto the dense grid without extrapolation, smoothing, or refitting. For the evaluated subset $\mathcal{V}$, the pooled discrepancy is

$$\mathrm{RMSE}_{40\to241} = \sqrt{\frac{1}{241|\mathcal{V}|}\sum_{i\in\mathcal{V}}\sum_{j=1}^{241}\left[\widetilde{T}_i^{40}(\lambda_{ij}) - T_i^{241}(\lambda_{ij})\right]^2}, \tag{4}$$

where $\widetilde{T}_i^{40}$ denotes the interpolated released spectrum and T_i^{241} the dense RCWA response.

Table 5 reports pooled errors within the reference Q bands. The overall RMSE is 0.00913; the median and 90th-percentile per-spectrum RMSE are 0.00987 and 0.01153, respectively. Average discrepancies increase toward higher Q, and the largest absolute pointwise discrepancy is approximately 0.123. Equal sampling across deciles yields unequal counts within the three reliability bands.

The subset spans $Q = 15.1$–1357.2. The comparison supports reconstruction fidelity within the selected resonance windows at the production solver settings. Its scope does not include full-band resonance completeness, preservation of fitted Q, or sampling accuracy beyond the evaluated quality-factor range.

C Evaluation Definitions

C.1 Scalar and Spectral Accuracy

For the main baseline targets $y \in \{\log_{10} Q, \lambda_{\mathrm{res}}, T_{\min}\}$, RMSE follows Section 4.3. The coefficient of determination is reported for all three scalar targets; MAE is additionally reported for $\log_{10} Q$ and resonance wavelength:

$$\mathrm{MAE}(y) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \qquad R^2(y) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{5}$$

where $\bar{y}$ is the reference test-set mean. Quality-factor RMSE and MAE are expressed in dex, wavelength errors in nanometres, and minimum-transmission errors in dimensionless transmission units. Negative R^2 indicates greater squared error than prediction by the reference test-set mean.

Let $\mathcal{S}$ contain test samples with `fine_used` set to true, and let $\Lambda_i = \{\lambda_{ij}\}_{j=1}^{40}$ denote the stored physical wavelength grid of adaptive sample i. Spectral accuracy is measured by

$$\mathrm{RMSE_b} = \sqrt{\frac{1}{50N}\sum_{i=1}^{N}\sum_{j=1}^{50}(t_{ij}^{\mathrm{b}} - \hat{t}_{ij}^{\mathrm{b}})^2}, \tag{6}$$

$$\mathrm{RMSE_f} = \sqrt{\frac{1}{40|\mathcal{S}|}\sum_{i\in\mathcal{S}}\sum_{j=1}^{40}(t_{ij}^{\mathrm{f}} - \hat{t}_{ij}^{\mathrm{f}})^2}. \tag{7}$$

Both spectral metrics pool squared errors before taking the square root and are reported in transmission units. Missing adaptive spectra are excluded. For the fine-spectrum metric, t_{ij}^{f} and $\hat{t}_{ij}^{\mathrm{f}}$ correspond to the same stored physical wavelength λ_{ij} by construction; no interpolation or resampling is required, and all fine-eligible samples are evaluated. Appendix E uses a canonical normalized-detuning cohort only for the auxiliary decomposition of shape accuracy and placement-aware wavelength reconstruction.

C.2 Conditional Wavelength Error and Spectral Detection

Conditional scalar wavelength error. Define

$$\mathcal{C}_{100} = \left\{ i : |\hat{\lambda}_{\mathrm{res},i} - \lambda_{\mathrm{res},i}| \le 100\ \mathrm{nm} \right\}. \tag{8}$$

The conditional error is

$$\mathrm{RMSE}_{\lambda,\mathrm{cond}} = \sqrt{\frac{1}{|\mathcal{C}_{100}|}\sum_{i\in\mathcal{C}_{100}}(\hat{\lambda}_{\mathrm{res},i} - \lambda_{\mathrm{res},i})^2}. \tag{9}$$

Both conditioning and scoring use the predicted scalar wavelength. This metric characterizes errors among predictions within the specified tolerance and does not measure the frequency of larger misses.

Spectral detection. The predicted broadband minimum is defined by

$$j_i^\star = \operatorname*{arg\,min}_{1\le j\le 50} \hat{t}_{ij}^{\mathrm{b}}, \qquad \hat{\lambda}_i^{\min} = \lambda_{j_i^\star}^{\mathrm{b}}. \tag{10}$$

For tolerance ϵ, the detection rate is

$$\mathrm{Detection}_\epsilon = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \mathbf{1}\left[|\hat{\lambda}_i^{\min} - \lambda_{\mathrm{res},i}| \le \epsilon \right], \qquad \epsilon \in \{50, 100\}\ \mathrm{nm}. \tag{11}$$

The denominator is the fine-eligible test population, matching the released evaluator. The broadband minimum is the raw argmin of the 50-point prediction, without interpolation. Detection therefore depends on broadband-grid resolution and is distinct from conditional scalar wavelength accuracy.

C.3 Profile Depth, Linewidth, and Stratification

Reference and predicted peak-to-trough depths are

$$d_i = \max_j t_{ij}^{\mathrm{f}} - \min_j t_{ij}^{\mathrm{f}}, \qquad \hat{d}_i = \max_j \hat{t}_{ij}^{\mathrm{f}} - \min_j \hat{t}_{ij}^{\mathrm{f}}. \tag{12}$$

Their absolute mean error is

$$\mathrm{MAE_{depth}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} |d_i - \hat{d}_i|. \tag{13}$$

This peak-to-trough diagnostic is evaluated separately from prediction of the scalar minimum transmission $T_{\min}$.

For a sampled profile $t(u)$, the half-depth level is

$$t_{1/2} = \min_j t(u_j) + \frac{\max_j t(u_j) - \min_j t(u_j)}{2}. \tag{14}$$

Starting at the sampled minimum, the implementation identifies the contiguous sequence of grid points whose transmission does not exceed the half-depth level. The reported width is the coordinate span between its first and last points, without interpolation. Two observed threshold crossings are not required: a boundary-truncated profile can return a finite span, and a single qualifying point returns zero. Profiles with peak-to-trough depth below 10^{-6} return a non-finite estimate. The reference and predicted spans are denoted by w_i and $\hat{w}_i$. For samples $\mathcal{W} \subseteq \mathcal{S}$ with finite reference and predicted estimates,

$$\mathrm{MAE}_{\mathrm{width}} = \frac{1}{|\mathcal{W}|} \sum_{i \in \mathcal{W}} |w_i - \hat{w}_i|. \tag{15}$$

Depth errors are in transmission units. The historical evaluator computes half-depth spans on its fixed reference detuning axis; the resulting error is a normalized-output-grid span diagnostic rather than a physical linewidth error. Pairs with non-finite estimates are excluded.

For any reported subgroup, aggregation is restricted to that subgroup while retaining the metric's eligibility rule: available adaptive spectra for profile errors and detection, the scalar-error tolerance for conditional wavelength error, and finite span pairs for width error.

D Training Protocol and Extended Results

D.1 Preprocessing and Optimization

The main baselines map $\mathbf{g} = (P, r_1, r_2, h)$ to $(\log_{10} Q, \lambda_{\mathrm{res}}, T_{\min})$, a 50-point broadband spectrum, and a 40-point adaptive spectrum.

Gradient-trained models use linear scalar heads and sigmoid-bounded spectral outputs. Their objective is

$$\mathcal{L} = \mathcal{L}_{\log Q} + \mathcal{L}_{\lambda} + 0.5\mathcal{L}_{T_{\min}} + \mathcal{L}_{\mathrm{b}} + \mathcal{L}_{\mathrm{f}}, \tag{16}$$

where each component is a mean-squared error and $\mathcal{L}_{\mathrm{f}}$ is masked to samples with an available adaptive spectrum.

Each baseline receives a 20-trial random hyperparameter search, selected by the lowest weighted validation loss. Gradient-trained trials run for at most 150 epochs; the selected configuration is retrained for at most 300 epochs per seed. Optimization uses AdamW, cosine learning-rate annealing over the epoch budget, and validation-based early stopping with patience 25.

In the implementation used for the reported results, MLP, CNN, and Transformer use dropout probability 0.1. Their search records contain a sampled `dropout` field, but this field was not forwarded to the backbone constructors and did not change their effective dropout. Coordinate applies its sampled setting; Linear has no dropout layer. The release preserves this behavior to retain the evaluated configurations.

The split seed and fixed training seed during search are 20260916. Final gradient-trained runs use seeds $\{0, 1, 2, 3, 4\}$, with test results summarized by their mean and standard deviation. All models use the same frozen partitions.

SVR fits independent RBF-kernel regressors for each output and uses the same weighted validation criterion. Its 20 search trials use a 4,000-sample training subsample. The selected configuration is refit on the full training partition, with fine-spectrum regressors restricted to eligible samples. The deterministic SVR baseline is reported from one final fit.

D.2 Architectures and Selected Configurations

Linear uses an affine backbone. MLP uses fully connected layers with GELU activations and dropout. CNN uses one-dimensional convolutional blocks with kernel size 5. Transformer represents the four geometric parameters as tokens and applies pre-normalization encoder layers with GELU feed-forward blocks.

Coordinate combines an MLP geometry encoder and scalar heads with separate broadband and fine coordinate decoders. Fixed, seeded random Fourier features encode the queried coordinates. Each decoder shares weights across its own spectral coordinates; the broadband and fine decoders have separate parameters.

Table 6: Hyperparameter search spaces. Each baseline receives 20 random trials. Common settings apply to gradient-trained models.

Scope	Hyperparameter	Candidate values
Common	Learning rate	$\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$
	Weight decay	$\{0, 10^{-4}, 10^{-3}\}$
	Batch size	$\{128, 256, 512\}$
Linear	Width	$\{32, 64, 128\}$
MLP	Width	$\{64, 128, 256, 512\}$
	Depth	$\{2, 3, 4, 5\}$
CNN	Channels	$\{32, 64, 128\}$
	Depth	$\{2, 3, 4\}$
Transformer	Model width	$\{32, 64, 128\}$
	Attention heads	$\{2, 4, 8\}$; must divide model width
	Depth	$\{2, 3, 4\}$
Coordinate	Encoder width	$\{64, 128, 256\}$
	Encoder depth	$\{2, 3, 4\}$
	Decoder width	$\{64, 128, 256\}$
	Decoder depth	$\{2, 3, 4\}$
	Fourier-feature count	$\{16, 32, 64\}$
	Fourier scale σ	$\{1, 2, 4, 8\}$
	Dropout	$\{0, 0.1, 0.2\}$
SVR	C	$\{0.3, 1, 3, 10, 30\}$
	γ	$\{\texttt{scale}, 0.03, 0.1, 0.3\}$
	ϵ	$\{0.01, 0.03, 0.05, 0.1\}$

Table 7: Selected architecture and kernel configurations. Validation loss is the weighted multi-task selection objective.

Model	Selected configuration	Val. loss
Linear	Width 64	1.8220
MLP	Width 256; depth 4	0.3921
CNN	128 channels; depth 3; kernel size 5	0.4019
Transformer	Width 128; 4 heads; depth 4	0.4546
Coordinate	Encoder width 128, depth 3; decoder width 256, depth 4; 64 Fourier features, $\sigma = 4$	0.4168
SVR	$C = 30$; $\gamma = \texttt{scale}$; $\epsilon = 0.01$	0.9476

Table 6 specifies the search spaces. Tables 7 and 8 report the configurations selected from the saved search results.

D.3 Compute and Seed Variability

Gradient-trained models use a single NVIDIA RTX 4090 GPU; SVR is fitted on CPU. Recorded script completion times are 221.0 minutes for the main forward-model run, 65.5 minutes for the coordinate-model script, and 5.4 minutes for the SVR script. These invocation-level times may include cached work, evaluation, and output generation. They are not per-model training times or a measured cost of reproducing all experiments from scratch.

Figure 7 reports the extended metric comparison and variability across final training seeds. The supplementary normalized-shape and placement-aware physical-wavelength study follows the separate single-seed protocol in Appendix E.

Table 8: Selected optimization settings and effective dropout probabilities used in the reported experiments.

Model	Learning rate	Weight decay	Dropout	Batch size
Linear	10^{-3}	10^{-4}	—	256
MLP	3×10^{-3}	10^{-3}	0.1	256
CNN	10^{-3}	10^{-4}	0.1	128
Transformer	3×10^{-4}	0	0.1	256
Coordinate	3×10^{-3}	0	0.0	256

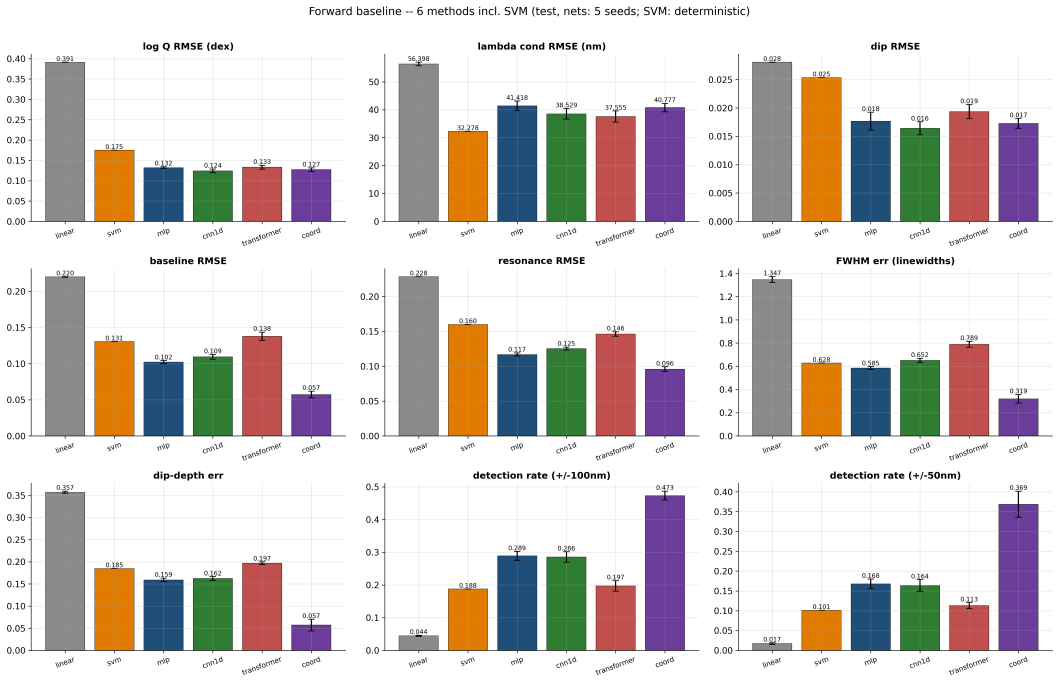

Figure 7: Extended baseline comparison. Error bars show standard deviations across five seeds for gradient-trained models; SVR is deterministic.

E Supplementary Physical-Wavelength Reconstruction

E.1 Cohort and Evaluation Protocol

This single-seed supplementary analysis uses seed 0 and the frozen in-distribution split. Of 1,805 fine-eligible test samples, 829 satisfy the canonical detuning-template requirements used for this auxiliary shape/placement decomposition. This restriction is specific to the supplementary analysis and is not required by the main adaptive-window RMSE, which evaluates all 1,805 fine-eligible samples on their aligned stored physical wavelength grids. The same 829-sample cohort is used across methods and is reported separately from the five-seed main benchmark.

Normalized shape RMSE compares predicted and reference profiles on the canonical detuning grid. Physical-wavelength evaluation additionally incorporates the predicted resonance position and linewidth, thereby measuring placement-aware reconstruction rather than the directly aligned 40-point error used by the main adaptive-window metric. For predicted $(\hat{\lambda}_i, \hat{Q}_i)$, the coordinate mapping is

$$\hat{u}_i(\lambda) = \frac{(\lambda - \hat{\lambda}_i)\hat{Q}_i}{\hat{\lambda}_i}. \tag{17}$$

Tabulated profiles are evaluated by linear interpolation at $\hat{u}_i(\lambda)$, with constant endpoint extension outside their predicted support.

For the common cohort $\mathcal{E}$, physical reconstruction error is

$$\mathrm{RMSE}_{\mathrm{phys}} = \sqrt{\frac{1}{40|\mathcal{E}|}\sum_{i\in\mathcal{E}}\sum_{j=1}^{40}\left[T_i(\lambda_{ij}) - \hat{T}_i(\lambda_{ij})\right]^2}, \tag{18}$$

where λ_{ij} are the stored RCWA wavelengths. The metric is expressed in transmission units and combines profile, placement, linewidth-scaling, and endpoint-extension errors. Endpoint extension affects approximately 36.8% of evaluated points for Coordinate and 29.1% for SVR.

This analysis uses the existing reference spectra. It is distinct from the dense RCWA sampling study in Appendix B.3.

E.2 Reference Fits and Exploratory Fano Models

Reference Fano fit. The analytic transmission profile is

$$T(\omega) = \frac{I_0}{1+q^2}\frac{(q+\epsilon)^2}{1+\epsilon^2} + I_{\mathrm{bg}}, \qquad \epsilon = \frac{2(\omega-\omega_0)}{\gamma}, \qquad Q = \frac{\omega_0}{\gamma}, \tag{19}$$

with $\omega = 2\pi c/\lambda$. A bounded, single-start fit estimates the five parameters directly from each reference test spectrum and returns numerical solutions for all 829 samples. The resulting residual measures this fitting procedure's agreement with the reference data; it is neither a geometry-only prediction score nor a theoretical error bound.

FanoNet and residual correction. FanoNet predicts $(\lambda_{\mathrm{res}}, \log_{10} Q, q, I_0, I_{\mathrm{bg}})$ from geometry using four hidden layers of 256 units, GELU activations, and dropout 0.1. Parameter targets are standardized. Training uses full-batch Adam for 500 epochs, initial learning rate 10^{-3}, weight decay 10^{-5}, and cosine learning-rate decay.

Fano + residual adds an MLP with the same hidden architecture, trained for 400 epochs on profile residuals. Residual targets are constructed using the fitted profile coordinates, with points outside measured support masked.

For normalized shape evaluation, analytic profiles are queried at

$$\hat{\lambda}_i(u) = \hat{\lambda}_i\left(1 + \frac{u}{\hat{Q}_i}\right). \tag{20}$$

For physical reconstruction, the analytic component is evaluated directly at the stored wavelengths; the residual is interpolated using the predicted position and linewidth, with constant endpoint extension. Analytic and residual outputs are not clipped to $[0, 1]$.

Of 3,875 successful training fits on compatible grids, one cannot define positive wavelengths over the complete detuning interval and is excluded from canonical-profile training. The corresponding constraint $Q \geq 6.0006$ is imposed on predicted parameters and is active for one evaluated test sample.

These Fano models use fixed configurations rather than the main benchmark's 20-trial hyperparameter search. They are exploratory comparisons and do not constitute a matched-search ranking against the six main baselines.

E.3 Results and Interpretation

Table 9 separates normalized shape accuracy from reconstruction in physical wavelength space. Coordinate achieves the lowest shape RMSE, improving on MLP by 12.8%, while SVR achieves the lowest physical-wavelength RMSE. The residual model reduces FanoNet's physical error by approximately 1.8% without improving its shape error.

The comparison demonstrates that lower normalized shape error need not yield lower physical reconstruction error under predicted placement and scaling. The rankings are conditional on the evaluated cohort, training seed, and interpolation procedure; they do not establish a general limitation of parameter-based spectral models.

Table 9: Single-seed supplementary comparison on 829 test samples, using seed 0. Entries are pooled RMSE in transmission units. Bold marks the lowest error among geometry-only predictors. The reference Fano fit uses test spectra directly.

Method	Normalized shape	Physical wavelength
Linear	0.1536	0.2650
SVR	0.1014	**0.2158**
MLP	0.0898	0.2210
CNN	0.0963	0.2229
Transformer	0.1044	0.2249
Coordinate	**0.0783**	0.2258
FanoNet (exploratory)	0.1242	0.2409
Fano + residual (exploratory)	0.1246	0.2365
Reference Fano fit	Fit residual: 0.0730	

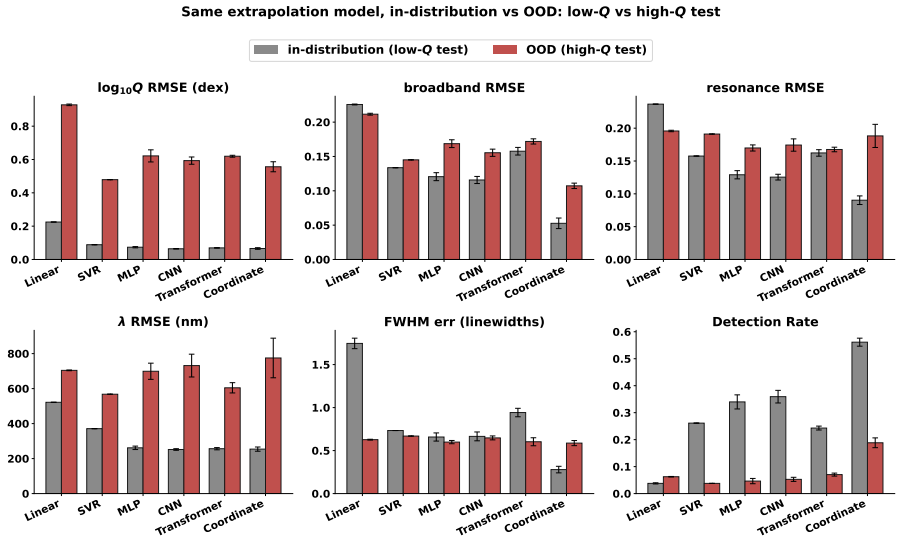

Figure 8: Low-Q testing and high-Q extrapolation for models trained exclusively on low-Q samples. Quality-factor prediction deteriorates on the held-out high-Q population.

F Generalization and Training Ablations

F.1 Low-Q Accuracy and High-Q Extrapolation

Figure 8 compares low-Q and held-out high-Q test performance for models trained exclusively on the low-Q partition. The deterioration under extrapolation motivates the augmentation comparison in Section 6.1, which holds the high-Q test population fixed while changing training coverage.

F.2 Sensitivity to the High-Q Threshold

Extrapolation is additionally evaluated with boundaries at the 70th and 90th percentiles of Q. These runs assess sensitivity to the definition of the held-out high-Q population. The paired extrapolation-versus-seen coverage comparison is conducted at the 80th-percentile boundary.

Changing the boundary alters both the size of the low-Q training population and its maximum quality factor. The threshold sweep therefore assesses robustness to the definition of extrapolation. Similarly, augmentation changes both training coverage and sample count; its benefit is measured at fixed architecture and hyperparameters, rather than at fixed training-set size.

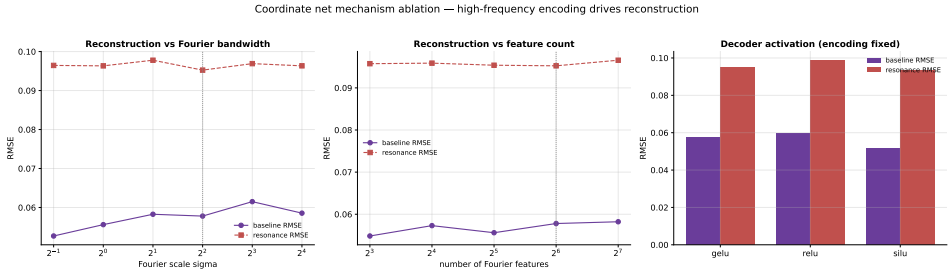

Figure 9: Sensitivity of coordinate reconstruction to Fourier-feature scale, feature count, and decoder activation.

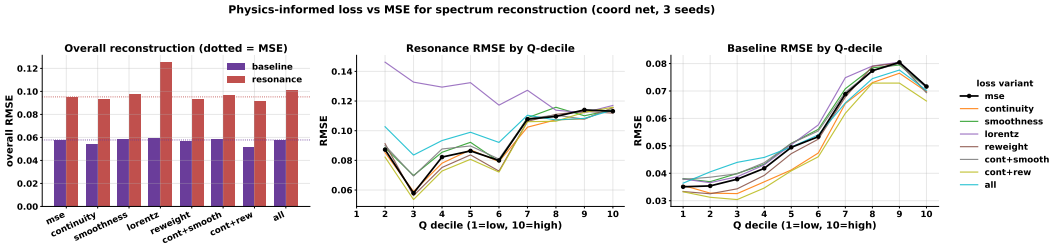

Figure 10: Training-loss comparison for the coordinate model. Continuity with resonance weighting improves reconstruction across most evaluated Q deciles. A symmetric Lorentzian-tail prior degrades performance, while generic smoothness offers no appreciable benefit.

F.3 Coordinate-Decoder Configuration

Reconstruction varies little across the tested Fourier-feature scales, feature counts, and decoder activations (Figure 9). The coordinate model's performance is therefore robust within the examined configurations. This sensitivity study does not independently isolate the contribution of Fourier features or establish that they are unnecessary.

F.4 Loss, Target, and Input Ablations

The loss study in Figure 10 compares additions to the coordinate model's multi-task MSE without modifying its architecture. Continuity between broadband and fine predictions, combined with resonance weighting, improves reconstruction. A symmetric Lorentzian-tail prior instead degrades performance, while generic smoothness provides little benefit. These comparisons support consistency between the two spectral representations without imposing a symmetric resonance profile.

The historical continuity term uses per-record physical wavelengths, whereas wing and lineshape weights are defined on the shared reference encoding of output positions. This implementation is retained in the released ablation code.

The target and input ablations use three seeds $(0, 1, 2)$. Their backbone is selected from the four fixed-output models using the aggregate test-score comparison, so these are exploratory follow-up analyses rather than independent confirmatory tests.

Target parameterization has a larger effect: direct regression of Q produces approximately eight times greater error in dex than regression of $\log_{10} Q$. This behavior is consistent with the disproportionate influence of large targets under a linear-scale squared-error objective. Adding the derived asymmetry α to the four geometric inputs provides no measurable improvement for the tested models.

G Limitations and Broader Impact

H Physical Scope and Interpretation

MIR-QBIC describes lossless, periodic structures and reports radiative quality factors. In a fabricated device, absorption, finite-array effects, and fabrication variation introduce additional loss channels. When independent decay rates add, their contributions can be expressed as

$$\frac{1}{Q_{\mathrm{tot}}} = \frac{1}{Q_{\mathrm{rad}}} + \frac{1}{Q_{\mathrm{abs}}} + \frac{1}{Q_{\mathrm{other}}}. \tag{21}$$

Consequently, increasing the ideal radiative quality factor does not imply an equivalent increase in the total quality factor of an experimental device.

The Ge/CaF_2 sensing platform demonstrated by Leitis et al. (2019) motivates the application context. MIR-QBIC itself contains simulated optical responses rather than fabricated devices or experimental sensing measurements. Its benchmark therefore assesses surrogate fidelity within the specified material system and geometric family; experimental sensing performance and transfer to other platforms remain separate questions.