DECISION-AWARE RESIDUAL CALIBRATION: BRIDGING THE PREDICTION–DECISION GAP IN ELECTROMAGNETIC DESIGN WITH SPARSE FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse response feedback provides direct evidence for correcting surrogate predictions during electromagnetic design. Converting such feedback into reliable design decisions poses two coupled challenges. First, errors observed at only a few response locations must be propagated to reconstruct the complete response. Second, errors that remain after correction must be accounted for when evaluating nonlinear objectives, as they can alter candidate rankings. We propose Decision-Aware Residual Calibration (DARC), a gradient-free framework that uses sparse feedback to correct surrogate responses and account for remaining errors in nonlinear design scoring, without updating the surrogate. DARC first uses historical cross-location error dependencies to propagate observed discrepancies across the full response. DARC further identifies residual errors that may still affect design scoring after response correction. It applies the feedback-propagation operator used for the current prediction to each historical error pattern. The unrecovered component represents a possible remaining error and is added to the corrected response to form a scenario. Together, these scenarios yield a design score that accounts for how post-correction errors affect the nonlinear design objective. Across all six task–surrogate combinations, DARC achieves the lowest average response-reconstruction error and selection regret among the baselines. In particular, DARC achieves stronger performance with fewer feedback observations. Controlled ablations further show that accounting for post-correction errors improves selection under nonlinear objectives beyond scoring a single corrected response.

1 INTRODUCTION

Modern wireless communication relies on electromagnetic structures (EMS), including Wi-Fi antennas (Zhang et al., 2022), RF filters (Luo et al., 2024), and metasurfaces (Yang et al., 2025), to transmit electromagnetic waves. Their design is historically difficult because performance evaluation requires numerical simulations, which are computationally intensive. To address this, surrogate models are used to rapidly predict electromagnetic responses (Peurifoy et al., 2018) and efficiently screen promising designs from large candidate pools (Lindemans et al., 2025). However, their training still depends on data from costly simulations. Under limited data budgets, surrogates typically exhibit reduced accuracy (Ma et al., 2023), and their outputs generally require additional calibration.

Recent studies have increasingly focused on using partial observations to recover complete electromagnetic responses at reduced simulation cost (Peumans et al., 2024; Åkerstedt et al., 2025). These observations provide a practical source of feedback for correcting surrogate predictions under limited evaluation budgets. Existing methods include (1) data reconstruction methods (Gustavsen & Semlyen, 1999; De Ridder et al., 2019), which fit response models to reconstruct broadband responses from partial observations; and (2) model tuning methods, which update surrogate model parameters (Zhu et al., 2023) or latent variables (Kim et al., 2024) using sparse observations. However, two challenges remain. First, corrections inferred from observed frequencies must generalize to unobserved frequencies, where improved accuracy is not guaranteed. Second, design decisions rely on nonlinear performance metrics such as bandwidth and extrema, yet improved overall response accuracy does not necessarily lead to more accurate estimates of these metrics.

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.

To address these challenges, we propose *Decision-Aware Residual Calibration* (DARC), a two-stage framework that refines response predictions and nonlinear objective estimates while keeping the surrogate model fixed. First, leveraging cross-frequency correlations in historical prediction errors, we use sparse observations to correct the response at unobserved frequencies. This extends local feedback to the complete response without updating surrogate parameters. Second, with the corrected response fixed, we refine design scores by accounting for the errors that may remain after correction. We apply the same correction rule to historical errors to construct empirical scenarios around the corrected response. Each scenario is evaluated under the nonlinear objective, and its contribution is weighted by the similarity between the corresponding historical errors at observed frequencies and the current feedback. This incorporates the effects of remaining errors into design evaluation without changing the corrected response.

Our key contributions are as follows:

1. **A Unified Framework for Response and Decision Calibration:** We propose Decision-Aware Residual Calibration (DARC), a framework that uses sparse feedback to refine both response predictions and nonlinear design scores. A shared model of historical prediction errors connects these two tasks, producing a corrected response and a separate score for candidate selection without updating surrogate parameters.
2. **Accounting for Remaining Errors in Design Scoring:** We construct empirical scenarios from the historical error patterns left by the same correction rule used for the current response. By evaluating nonlinear objectives within these scenarios and weighting the results according to their relevance to current feedback, we make design scores account for errors that may persist after correction.
3. **Systematic Evaluation with Broader Implications for AI4Science:** We evaluate DARC across 90 settings covering three electromagnetic tasks, three predictor architectures, and varying training and feedback budgets. DARC achieves lower mean reconstruction error and selection regret than the evaluated baselines within each task–architecture combination. Controlled ablations show that residual-aware scoring improves selection while keeping the corrected response fixed. These findings motivate a decision-aware approach to scientific surrogate calibration, accounting for how remaining prediction errors affect nonlinear design objectives.

2 Problem Formulation

We consider electromagnetic design selection using surrogate predictions and sparse response observations acquired at test time.

Surrogate Predictions. A structure is described by parameters $x \in \mathcal{X}$, where $\mathcal{X}$ is the design space. Its response at F frequencies is a real vector $\boldsymbol{y} \in \mathbb{R}^d$, with complex responses represented through their real and imaginary components. A pretrained surrogate $f_\theta : \mathcal{X} \to \mathbb{R}^d$ predicts $\boldsymbol{p} = f_\theta(x)$.

Available Information. For a single test sample, responses are partially observed at a set $\mathcal{O}$ of k frequencies, where $1 \le k < F$. Let $H \in \mathbb{R}^{m \times d}$ select the m observed response coordinates. The observed response and prediction error are

$$\boldsymbol{y}_{\mathcal{O}} = H\boldsymbol{y}, \qquad \boldsymbol{r}_{\mathcal{O}} = \boldsymbol{y}_{\mathcal{O}} - H\boldsymbol{p} \in \mathbb{R}^m. \tag{1}$$

We also have a reference dataset $\mathcal{V} = \{(x_j, \boldsymbol{y}_j)\}_{j=1}^{M}$ with complete responses, which may reuse the surrogate's validation set. These samples are not used to fit the surrogate parameters. The test sample's response at unobserved frequencies is unavailable to calibration.

Prediction and Scoring. The task is to estimate the full response as $\boldsymbol{q} \in \mathbb{R}^d$ and the design score as $\widehat{s} \in \mathbb{R}$. The score estimates the true objective value $s = g(\boldsymbol{y})$, where $g : \mathbb{R}^d \to \mathbb{R}$ is a specified performance objective, such as bandwidth or a band extremum, taken to be maximized without loss of generality. We do not require $\widehat{s} = g(\boldsymbol{q})$.

Design Decision. Given a candidate index set $\mathcal{C}$, let $\widehat{s}_i$ and s_i denote the estimated and true objective values of candidate i, respectively. For a selection budget $L \in \{1, \dots, |\mathcal{C}|\}$, the selected set and the

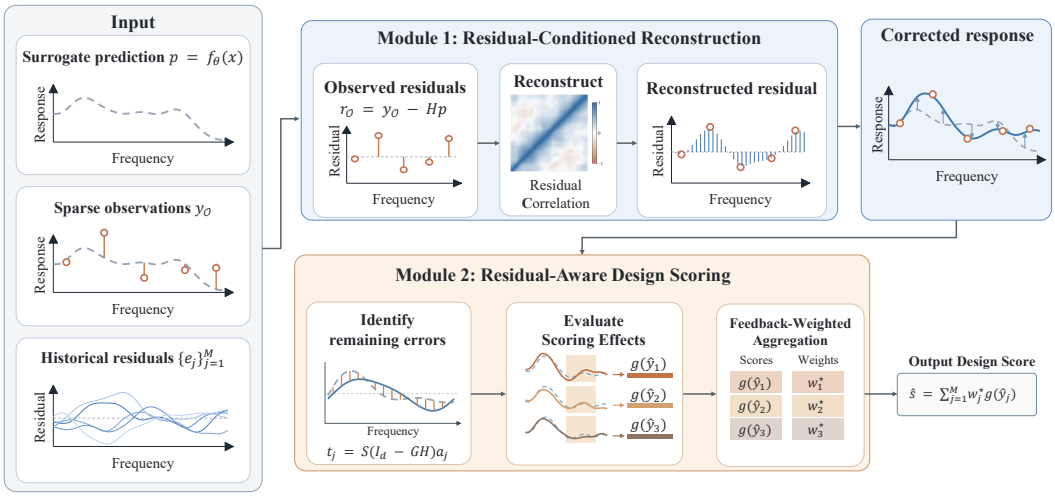


Figure 1: Overview of DARC. Sparse observations guide response correction through historical residual correlations. The corrected response is combined with possible remaining errors, producing alternative curves whose nonlinear objective values are aggregated using feedback-dependent weights. This yields a design score that accounts for remaining errors without changing the corrected response.

optimal set under the true scores are

$$\widehat{\mathcal{A}} \in \arg\max_{\substack{\mathcal{A}\subseteq\mathcal{C}\\|\mathcal{A}|=L}} \sum_{i\in\mathcal{A}} \widehat{s}_i, \qquad \mathcal{A}^\star \in \arg\max_{\substack{\mathcal{A}\subseteq\mathcal{C}\\|\mathcal{A}|=L}} \sum_{i\in\mathcal{A}} s_i. \tag{2}$$

Our goal is to minimize selection regret, defined as the gap in average true objective value:

$$\mathrm{Regret}(\widehat{\mathcal{A}}) = \frac{1}{L}\sum_{i\in\mathcal{A}^\star} s_i - \frac{1}{L}\sum_{i\in\widehat{\mathcal{A}}} s_i. \tag{3}$$

The true scores are used only for evaluation and are unavailable when selecting candidates.

3 Method

We present Decision-Aware Residual Calibration (DARC) in two stages: response correction using sparse feedback, followed by design scoring that accounts for the errors remaining after correction. Figure 1 summarizes the workflow.

3.1 Residual-Conditioned Reconstruction

We turn the surrogate's errors on reference samples into a reusable model for correcting new predictions. Their cross-frequency dependencies provide information about errors beyond the observed locations, while sparse feedback identifies the correction for the current sample.

Historical Error Model. Let $S \in \mathbb{R}^{d\times d}$ be a positive diagonal matrix of response scales computed during surrogate training and retained for calibration. For each reference sample, we compute the standardized error $\boldsymbol{e}_j = S^{-1}(\boldsymbol{y}_j - f_\theta(x_j))$. With reference weights $\pi_j > 0$ satisfying $\sum_{j=1}^M \pi_j = 1$, its empirical mean and covariance are

$$\boldsymbol{\mu} = \sum_{j=1}^M \pi_j \boldsymbol{e}_j, \qquad \boldsymbol{a}_j = \boldsymbol{e}_j - \boldsymbol{\mu}, \qquad C = \sum_{j=1}^M \pi_j \boldsymbol{a}_j \boldsymbol{a}_j^\top. \tag{4}$$

Each structure has equal weight, divided equally among its records. $\pi_j = 1/M$. The mean $\boldsymbol{\mu}$ captures systematic prediction bias, while C captures how errors vary together across frequencies.

Algorithm 1 Decision-Aware Residual Calibration (DARC)

Input: Fixed surrogate f_θ, reference set $\mathcal{V}$, reference weights $\{\pi_j\}_{j=1}^M$, scales S; test sample x, observation selector H, observations $\boldsymbol{y}_\mathcal{O}$; objective g, hyperparameters $\gamma, \tau > 0$ and $\rho \in (0, 1]$.
Output: Corrected response $\boldsymbol{q}$ and design score $\widehat{s}$.

```
/* Offline Preparation */
```
Compute reference errors $\boldsymbol{e}_j = S^{-1}(\boldsymbol{y}_j - f_\theta(x_j))$ for $j = 1, \dots, M$.
Compute $\boldsymbol{\mu}$, $\{\boldsymbol{a}_j\}_{j=1}^M$, and C using Eq. (4).
```
/* Response Correction */
```
Predict $\boldsymbol{p} = f_\theta(x)$ and compute $\boldsymbol{r}_\mathcal{O} = \boldsymbol{y}_\mathcal{O} - H\boldsymbol{p}$.
Set $S_\mathcal{O} = HSH^\top$ and $\boldsymbol{b}_\mathcal{O} = S_\mathcal{O}^{-1}\boldsymbol{r}_\mathcal{O} - H\boldsymbol{\mu}$.
Compute A, G, and the corrected response $\boldsymbol{q}$ using Eq. (5).
```
/* Residual-Aware Design Scoring */
```
Construct remaining errors $\boldsymbol{t}_j = S(I_d - GH)\boldsymbol{a}_j$ for $j = 1, \dots, M$.
Compute feedback-based similarities $\{v_j\}_{j=1}^M$ using Eq. (7).
Obtain $\boldsymbol{w}^\star$ with $N_{\text{eff}}(\boldsymbol{w}^\star) \geq \rho M$ using the projection in Eq. (8).
Evaluate scenarios and aggregate: $\widehat{s} = \sum_{j=1}^M w_j^\star g(\boldsymbol{q} + \boldsymbol{t}_j)$.
return $(\boldsymbol{q}, \widehat{s})$.

Feedback Propagation. We instantiate the correction with a regularized covariance-based estimator following optimal interpolation (Gandin, 1963). For the current sample, let $S_\mathcal{O} = HSH^\top$. The observed error from Eq. (1), after standardization and mean removal, is $\boldsymbol{b}_\mathcal{O} = S_\mathcal{O}^{-1}\boldsymbol{r}_\mathcal{O} - H\boldsymbol{\mu}$. With $\gamma > 0$ and average residual variance $c = \operatorname{tr}(C)/d > 0$, define

$$A = HCH^\top + c\gamma I_m, \qquad G = CH^\top A^{-1}, \qquad \boldsymbol{q} = \boldsymbol{p} + S(\boldsymbol{\mu} + G\boldsymbol{b}_\mathcal{O}), \tag{5}$$

where I_m is the m-dimensional identity matrix and $G \in \mathbb{R}^{d\times m}$ propagates observed errors to the full response. Regularization stabilizes this propagation. The historical error correlations determine how each observed discrepancy affects unobserved frequencies. We retain G to construct the remaining error scenarios used in the next stage.

3.2 Residual-Aware Design Scoring

We next use the historical errors to inform design scoring after response correction. Applying the same operator G to each historical pattern isolates the part left uncorrected; the current feedback then determines its relevance to the test sample. This connects the correction already performed to the error scenarios used for nonlinear objective evaluation.

Remaining Error Scenarios. Consider a historical centered error $\boldsymbol{a}_j$ under the current observation layout. Its observed part is $H\boldsymbol{a}_j$, from which the correction rule recovers $GH\boldsymbol{a}_j$. The remaining part yields a physical response error $\boldsymbol{t}_j$ and a scenario $\widetilde{\boldsymbol{y}}_j$:

$$\boldsymbol{t}_j = S(I_d - GH)\boldsymbol{a}_j, \qquad \widetilde{\boldsymbol{y}}_j = \boldsymbol{q} + \boldsymbol{t}_j, \tag{6}$$

where I_d is the d-dimensional identity matrix. Thus, the scenarios reflect errors left by the same operator used to produce $\boldsymbol{q}$, rather than uncorrected historical errors. Since $\sum_j \pi_j \boldsymbol{a}_j = 0$, their reference-weighted mean is $\sum_j \pi_j \widetilde{\boldsymbol{y}}_j = \boldsymbol{q}$. These are empirical error scenarios obtained by transforming historical residuals.

Feedback-Based Weights. To identify relevant historical patterns, we compare their observed errors $H\boldsymbol{a}_j$ with the current evidence $\boldsymbol{b}_\mathcal{O}$. Define a regularized Mahalanobis distance and similarity weights by

$$D_j = \frac{1}{m}(H\boldsymbol{a}_j - \boldsymbol{b}_\mathcal{O})^\top A^{-1}(H\boldsymbol{a}_j - \boldsymbol{b}_\mathcal{O}), \qquad v_j = \frac{\pi_j \exp[-D_j/(2\tau)]}{\sum_{\ell=1}^M \pi_\ell \exp[-D_\ell/(2\tau)]}, \tag{7}$$

where $\tau > 0$ controls how strongly the weights favor patterns matching the feedback. To retain a sufficiently broad set of scenarios, we constrain the effective sample size $N_{\text{eff}}(\boldsymbol{w}) = 1/\sum_j w_j^2$. For

a retention fraction $\rho \in (0, 1]$, the final weights are

$$\boldsymbol{w}^\star = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^M} \sum_{j=1}^{M} w_j \log \frac{w_j}{v_j} \quad \text{subject to} \quad w_j \geq 0, \quad \sum_{j=1}^{M} w_j = 1, \quad \sum_{j=1}^{M} w_j^2 \leq \frac{1}{\rho M}. \tag{8}$$

This convex projection retains the feedback-based preferences while maintaining an effective sample size of at least ρM. Appendix A.4 details its numerical solution.

Nonlinear Objective Evaluation. We evaluate each scenario before combining its contribution:

$$\widehat{s} = \sum_{j=1}^{M} w_j^\star g(\boldsymbol{q} + \boldsymbol{t}_j). \tag{9}$$

For nonlinear g, this generally differs from evaluating the weighted mean response, $g(\boldsymbol{q}+\sum_j w_j^\star \boldsymbol{t}_j)$. For example, remaining errors can move a threshold crossing or change which frequency determines a band extremum. Evaluating complete scenarios retains these effects in the score. The response output remains $\boldsymbol{q}$: changing the scenario weights changes the design score without changing the corrected curve.

Algorithm 1 summarizes the offline preparation and test-time calibration for a single sample.

4 Experiments

We evaluate how sparse feedback improves response prediction and structure selection across three electromagnetic design tasks. Our experiments examine performance across feedback budgets and isolate the contribution of scenario scoring beyond correcting a single response curve.

4.1 Experimental Setup

Tasks and Surrogates. We select three representative engineering design tasks to evaluate DARC across diverse inputs and outputs: high-gain antennas (HGA), random-structure RF filters (RFF), and third-order hairpin filters (HP). They span binary layouts and continuous design parameters, real- and complex-valued responses, and design objectives with different forms of nonlinearity (Table 1). To assess how predictor architecture affects performance, we evaluate an MLP and a residual network on each task. Task definitions, architectures, and training details appear in Appendix A.

Table 1: Design tasks. F counts frequency positions; a complex response has two real components at each position.

Task	Structural input	Response	Frequency grid
HGA	15×20 binary grid	Real gain	2.4–5.9 GHz; $F = 71$
RFF	15×16 binary grid	S_{21} in dB	10–30 GHz; $F = 1001$
HP	9 geometric parameters	Complex S_{21}	3–7 GHz; $F = 41$

Data and Feedback. We train surrogates on 128, 256, or 512 structures. A separate set of 256 structures supports checkpoint selection and reference statistics, another 128 supports calibration parameter selection, and 4,096 are held out for evaluation. These roles are disjoint at the structure level. Five joint data-split and training-seed repetitions yield 90 surrogates. At test time, we reveal $k \in \{3, 5, 10, 15, 20\}$ frequencies using three random layouts with nested budgets. All methods receive the same observations, without added noise; complex feedback supplies both real and imaginary components. The five-repeat averages give 90 task–architecture–training-size–budget settings.

Evaluation. Our primary metric is selection regret from Eq. (3), selecting 13 of 256 structures in each of 100 shared candidate pools. We also measure unobserved response MSE (UMSE) in physical coordinates and objective MAE. Pool and layout results are averaged within each model before computing means and sample standard deviations over five repetitions. The main table additionally averages training sizes and feedback budgets within each repetition. For structures with multiple response records, scalar objective values are averaged before selection. HP scores are negated to follow the maximization convention of Section 2.

Table 2: Selection regret (↓) across tasks and surrogate architectures. Entries are means with standard deviations over five repetitions, after averaging training sizes and budgets within each repetition. Best and second-best means are bold and underlined. Response GPs do not use the surrogate, so their values repeat across architectures.

	HGA		RFF		HP	
Method	MLP	ResNet	MLP	ResNet	MLP	ResMLP
Source	9.53 ± 1.46	10.00 ± 2.50	12.60 ± 0.33	12.62 ± 1.02	3.77 ± 0.20	3.60 ± 0.28
Model tuning						
Fine-tuning	4.64 ± 0.41	4.91 ± 0.98	3.42 ± 0.20	3.33 ± 0.30	$\underline{2.87} \pm 0.11$	$\underline{2.78} \pm 0.16$
Latent adaptation	7.16 ± 1.40	8.97 ± 2.60	10.54 ± 0.22	11.96 ± 0.93	3.34 ± 0.13	3.36 ± 0.30
Gappy AE	5.12 ± 0.48	5.33 ± 0.84	2.18 ± 0.13	2.06 ± 0.09	3.20 ± 0.14	3.17 ± 0.19
Data reconstruction						
Residual spline	$\underline{2.83} \pm 0.11$	$\underline{2.93} \pm 0.17$	$\underline{1.40} \pm 0.03$	1.45 ± 0.03	3.79 ± 0.13	3.72 ± 0.19
Response GP (emp.)	3.01 ± 0.05	3.01 ± 0.05	1.42 ± 0.03	$\underline{1.42} \pm 0.03$	3.41 ± 0.06	3.41 ± 0.06
Response GP (Matérn)	4.31 ± 0.46	4.31 ± 0.46	1.47 ± 0.03	1.47 ± 0.03	4.80 ± 0.04	4.80 ± 0.04
DARC (ours)	$\mathbf{2.46} \pm 0.10$	$\mathbf{2.53} \pm 0.16$	$\mathbf{1.19} \pm 0.01$	$\mathbf{1.17} \pm 0.02$	$\mathbf{2.04} \pm 0.07$	$\mathbf{2.00} \pm 0.10$

4.2 Compared Methods

We select baselines to cover two ways of exploiting sparse feedback: adapting model parameters or latent representations, and reconstructing responses directly from observations. Together with the uncorrected surrogate (**Source**), these comparisons assess the value of feedback and whether effective calibration requires model adaptation. Scoring variants of DARC are evaluated separately in Section 4.4 to isolate the contribution of decision scoring beyond response correction.

Model Tuning. We include **Fine-tuning** (Zhu et al., 2023), which updates all surrogate parameters for each test sample; **Latent adaptation**, a latent-space counterpart that optimizes the sample's representation with the prediction head fixed; and **Gappy AE** (Kim et al., 2024), which learns a response autoencoder offline and fits its latent variables to the observations through a fixed decoder initialized from the source prediction. Each method adapts independently to each sample.

Data Reconstruction. We include **Residual spline** to test whether interpolating observed residuals is sufficient for effective correction. It fits a natural cubic spline to these residuals and adds it to the source response. We also compare two response-space GPs to examine reconstruction using data-driven correlations versus a prescribed smoothness kernel: **Response GP (emp.)** uses an empirical response covariance following optimal interpolation (Gandin, 1963), whereas **Response GP (Matérn)** uses a Matérn-$3/2$ frequency kernel (Rasmussen & Williams, 2006). Both reconstruct responses from reference data and sparse observations without using a source prediction.

Hyperparameters are selected on the separate set; full details are provided in Appendix A.

4.3 Main Results

Selection Quality. As shown in Table 2, DARC achieves the lowest mean selection regret across all electromagnetic tasks and surrogate architectures, with low variability across repeated runs. Notably, even when hyperparameters are selected separately for each setting, the strongest baseline varies by task. DARC, by contrast, consistently outperforms the baselines. Compared with the strongest baseline for each architecture, it reduces regret by an average of approximately 13%, 16%, and 28% on HGA, RFF, and HP, respectively. These results demonstrate DARC's effectiveness across diverse tasks and surrogate architectures, supporting its broad applicability to electromagnetic design selection with sparse feedback.

Feedback Efficiency under Varying Observation Budgets. To assess whether DARC's advantage requires abundant feedback, Figure 2 compares all methods across feedback budgets from $k = 3$ to 20. At the smallest budget, DARC outperforms every baseline on all three tasks, reducing selection regret by 14.5%–18.6% relative to the strongest baseline for each task. Data reconstruction methods

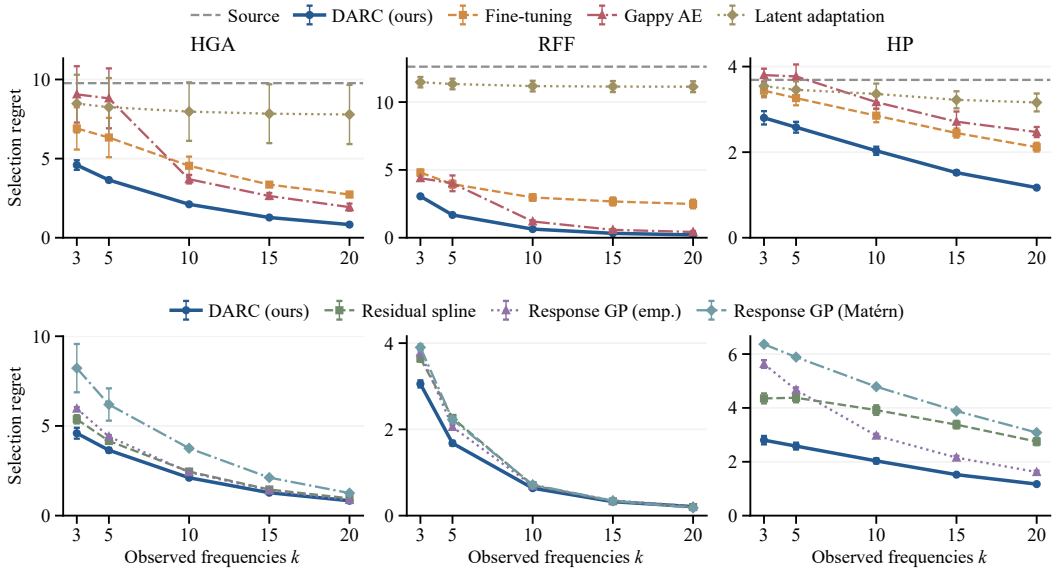


Figure 2: Selection regret versus feedback budget for DARC against model-tuning (top) and data-reconstruction (bottom) baselines. Averages span architectures and training sizes; error bars show standard deviations over five repetitions. Gray lines denote Source; axes vary by panel.

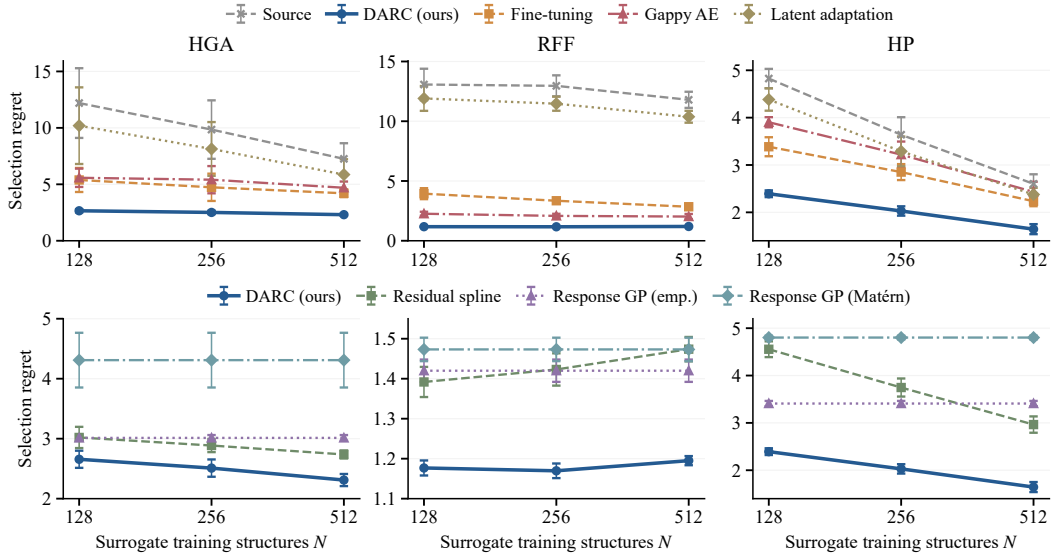


Figure 3: Selection regret versus surrogate training size. Plotting conventions follow Figure 2, with averages over feedback budgets and architectures. Response GPs are training-size independent.

become comparable only at the largest budget on HGA and RFF. These results show that DARC uses limited feedback more efficiently, enabling reliable structure selection from fewer local observations.

Effect of Surrogate Training Size. Figure 3 shows that DARC maintains its advantage across all three surrogate training sizes. Notably, on HGA and RFF, DARC with only 128 training structures achieves lower regret than every baseline with 512 training structures, averaged over feedback budgets and architectures. Thus, increasing surrogate training data alone does not close the gap to DARC within the evaluated range. These results highlight the effectiveness of DARC even when surrogate training data are limited.

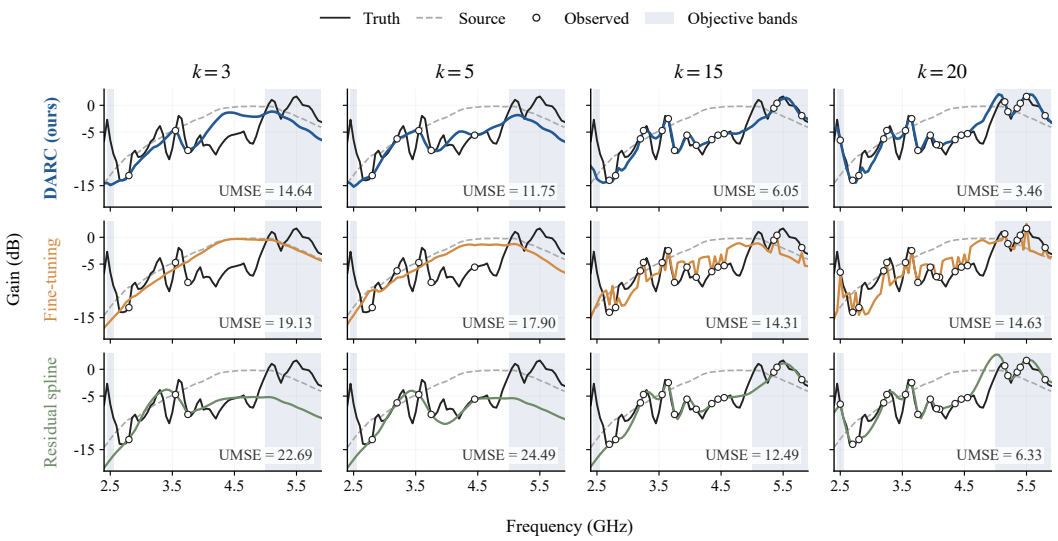


Figure 4: Response correction for HGA structure. Columns show $k = 3, 5, 15, 20$ observed frequencies; rows compare DARC, fine-tuning, and residual spline under the same nested observations. Shaded regions mark the objective frequency bands. Each panel reports mean squared error at unobserved frequencies (UMSE). Case-selection details appear in Appendix B.1.

Response Accuracy. DARC also outperforms fine-tuning, latent adaptation, and Gappy AE in unobserved response MSE across all 90 settings (Appendix B). Thus, the improvement is not confined to design rankings: DARC produces more accurate responses beyond the measured frequencies. Together with the selection results, this shows that sparse feedback can improve both response prediction and downstream design decisions without updating the surrogate.

Case Study on Response Correction. Sparse feedback is useful only insofar as it improves predictions beyond the observed frequencies. Figure 4 illustrates this distinction on an HGA structure. With few observations, fine-tuning mainly shifts the source response; with more observations, its adjustments concentrate around the observed frequencies, leaving substantial deviations in between. Residual spline extends corrections through interpolation, but without historical error information, it requires more observations to recover the response shape. In contrast, DARC uses historical cross-frequency error correlations to guide corrections at unobserved frequencies, achieving lower unobserved error across all four budgets shown. This highlights the value of historical error structure in turning local observations into broader response improvements.

4.4 Scoring Ablations

Table 3: Scoring ablations: selection regret (↓). All variants share the same base response $\boldsymbol{q}$ and scenario offsets. The weighted-mean control and DARC also share the weights, selected for DARC and held fixed for this comparison. Means and standard deviations are over five repetitions after averaging architectures, training sizes, and budgets.

Scoring rule	Estimate	HGA	RFF	HP
Direct	$g(\boldsymbol{q})$	2.533 ± 0.141	1.281 ± 0.023	2.424 ± 0.077
Fixed-weight scenarios	$\sum_j \pi_j g(\boldsymbol{q} + \boldsymbol{t}_j)$	$\mathbf{2.490} \pm 0.140$	1.276 ± 0.015	2.099 ± 0.073
Weighted mean	$g(\boldsymbol{q} + \sum_j w_j^\star \boldsymbol{t}_j)$	2.549 ± 0.096	1.198 ± 0.012	2.385 ± 0.077
DARC	$\sum_j w_j^\star g(\boldsymbol{q} + \boldsymbol{t}_j)$	2.492 ± 0.096	$\mathbf{1.180} \pm 0.008$	$\mathbf{2.022} \pm 0.084$

Scoring beyond Reconstruction. Table 3 shows that DARC improves over direct scoring on all three tasks, despite using exactly the same corrected response $\boldsymbol{q}$. Fixed-weight scenarios also improve over direct scoring, indicating that the remaining error patterns carry useful information for

design evaluation. Adapting the weights further lowers regret on RFF and HP, while yielding nearly the same performance as fixed weights on HGA. These comparisons separate the two contributions: scenario evaluation improves on using $\boldsymbol{q}$ alone, and feedback-based weighting further refines which historical patterns contribute to the score. DARC therefore improves selection without requiring an additional change to the response prediction. A case study in Appendix B.3 illustrates how scenario scoring corrects a candidate-ordering error made by directly evaluating the reconstructed responses.

Evaluating Scenarios or Their Mean. DARC also outperforms the weighted-mean control in Table 3, with a 15.21% regret reduction on HP and smaller gains on HGA and RFF. Both methods use the same responses, offsets, and weights selected for DARC; the control is not separately retuned. Interestingly, the weighted-mean response has lower UMSE on all three tasks, yet produces worse aggregate selection regret. We hypothesize that averaging the scenarios suppresses differences that matter to nonlinear objectives, particularly the worst-band deviations used for HP. By evaluating each scenario, DARC retains these differences in the design score. The comparison thus links its selection gains to how remaining errors are evaluated, rather than to a more accurate mean response.

5 Related Work

Surrogate-Based Electromagnetic Design. Surrogates accelerate forward prediction, inverse design, and optimization for microwave filters, antennas, and photonic or electromagnetic structures (Yu et al., 2022; Peurifoy et al., 2018; Qian et al., 2025; Lindemans et al., 2025). Recent work also studies electromagnetic design when only a limited number of high-fidelity evaluations is available (Zheng et al., 2025). These methods motivate our research. DARC is designed to use a small amount of response feedback to improve the final selection without retraining the surrogate.

Response Refinement from Partial Observations. Frequency-domain responses can be approximated from selected samples through curve fitting, adaptive sampling, and macromodeling (Gustavsen & Semlyen, 1999; Deschrijver et al., 2011; De Ridder et al., 2019). More broadly, missing fields can also be recovered (Everson & Sirovich, 1995; Kim et al., 2024), and sparse observations can adapt a learned operator at test time (Zhu et al., 2023). These approaches primarily produce a completed or adapted response. They motivate DARC's reconstruction stage, but do not by themselves specify how errors that remain after correction should affect a nonlinear design objective.

Residual Conditioning and Data Assimilation. Model-discrepancy formulations explicitly represent the gap between a simulator and the physical system (Kennedy & O'Hagan, 2001). Classical data assimilation likewise combines partial observations with prior error statistics through covariance-based updates (Evensen, 2003). DARC's response stage adapts this conditioning principle to the cross-frequency errors of a fixed surrogate. The key distinction is that DARC reuses the same correction operator on historical errors to expose the patterns it would leave unresolved, then carries these post-correction patterns into design evaluation.

Decision-Aware Prediction. Decision-focused learning shows that minimizing predictive error need not optimize downstream decisions (Elmachtoub & Grigas, 2022), while the optimizer's curse characterizes the selection bias created by choosing alternatives from uncertain estimates (Smith & Winkler, 2006). DARC addresses this prediction–decision mismatch without retraining. It conditions remaining-error scenarios on the current feedback, evaluates the nonlinear objective within each scenario, and aggregates objective values to produce the final score.

6 Conclusion

We propose Decision-Aware Residual Calibration (DARC), a gradient-free framework that uses sparse response feedback to improve electromagnetic response prediction and design selection without updating the surrogate. By coupling historical-error-based response correction with nonlinear objective evaluation that accounts for remaining errors, DARC accounts for both the observed discrepancies and the errors that may remain after correction. Experiments across three electromagnetic tasks show that DARC improves response accuracy and achieves the lowest average selection regret among the compared methods, enabling effective design selection even when feedback is scarce.

AI USE STATEMENT

The authors led the development of the core research ideas, hypotheses, methodology, and primary code implementation. Generative AI tools assisted with manuscript polishing, code diagnostics, and statistical analysis of experimental data. During exploratory development, these tools also provided candidate methodological and experimental suggestions. The authors have reviewed and verified the AI-assisted content and take full responsibility for the final manuscript, including its methods, results, claims, and accompanying artifacts.

REFERENCES

Lucas Åkerstedt, Darwin Blanco, and B. Lars G. Jonsson. On adaptive frequency sampling for data-driven model order reduction applied to antenna responses. *IEEE Transactions on Antennas and Propagation*, 73(8):5741–5750, 2025. doi: 10.1109/TAP.2025.3555890.

Simon De Ridder, Dirk Deschrijver, Domenico Spina, Tom Dhaene, and Dries Vande Ginste. Adaptive frequency sampling using linear bayesian vector fitting. *Electronics Letters*, 55(2):74–76, 2019.

Dirk Deschrijver, Karel Crombecq, Huu Minh Nguyen, and Tom Dhaene. Adaptive sampling algorithm for macromodeling of parameterized s -parameter responses. *IEEE Transactions on Microwave Theory and Techniques*, 59(1):39–45, 2011. doi: 10.1109/TMTT.2010.2090407.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003. doi: 10.1007/s10236-003-0036-9.

Richard Everson and Lawrence Sirovich. Karhunen–loève procedure for gappy data. *Journal of the Optical Society of America A*, 12(8):1657–1664, 1995. doi: 10.1364/JOSAA.12.001657.

Lev Semenovich Gandin. Objective analysis of meteorological fields. *Israel Program for Scientific Translations*, 242, 1963.

B. Gustavsen and A. Semlyen. Rational approximation of frequency domain responses by vector fitting. *IEEE Transactions on Power Delivery*, 14(3):1052–1061, 1999. doi: 10.1109/61.772353.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

Youngkyu Kim, Youngsoo Choi, and Byounghyun Yoo. Gappy AE: A nonlinear approach for gappy data reconstruction using auto-encoder. *Computer Methods in Applied Mechanics and Engineering*, 426:116978, 2024. doi: 10.1016/j.cma.2024.116978.

Yens Lindemans, Thijs Ullrick, Ivo Couckuyt, Tim Pattyn, Dirk Deschrijver, Dries Vande Ginste, and Tom Dhaene. Spectral bayesian optimization using a physics-informed rational szegö kernel for microwave design. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 15(9):1836–1846, 2025. doi: 10.1109/TCPMT.2025.3592441.

Sha Luo, Jiteng Ma, Chuanting Zhang, Shuping Dang, and Raed Shubair. Federated learning for microwave filter behavior prediction. *IEEE Microwave and Wireless Technology Letters*, 34(3): 255–258, 2024. doi: 10.1109/LMWT.2024.3353066.

Jiteng Ma, Shuping Dang, Peizheng Li, Gavin Watkins, Kevin Morris, and Mark Beach. Transfer learning for the behavior prediction of microwave structures. *IEEE Microwave and Wireless Technology Letters*, 33(2):126–129, 2023. doi: 10.1109/LMWC.2022.3214467.

Dries Peumans, Sander De Keersmaeker, Cedric Busschots, Yves Rolain, and Francesco Ferranti. Adaptive frequency sampling based on local rational modeling for microwave electromagnetic simulations. *IEEE Transactions on Microwave Theory and Techniques*, 72(8):4568–4578, 2024. doi: 10.1109/TMTT.2024.3355270.

John Peurifoy, Yichen Shen, Li Jing, Yi Yang, Fidel Cano-Renteria, Brendan G DeLacy, John D Joannopoulos, Max Tegmark, and Marin Soljačić. Nanophotonic particle simulation and inverse design using artificial neural networks. *Science advances*, 4(6):eaar4206, 2018.

Chao Qian, Ido Kaminer, and Hongsheng Chen. A guidance to intelligent metamaterials and metamaterials intelligence. *Nature Communications*, 16(1):1154, 2025.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. URL https://gaussianprocess.org/gpml/.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Han Qing Yang, Jun Yan Dai, Hui Dong Li, Lijie Wu, Zi Hang Shen, Qun Yan Zhou, Meng Zhen Zhang, Si Ran Wang, Zheng Xing Wang, Jun Wei Wu, et al. Adaptively programmable metasurface for intelligent wireless communications in complex environments. *Nature communications*, 16(1):6070, 2025.

Yang Yu, Zhen Zhang, Qingsha S Cheng, Bo Liu, Yi Wang, Cheng Guo, and Terry Tao Ye. State-of-the-art: Ai-assisted surrogate modeling and optimization for microwave filters. *IEEE transactions on microwave theory and techniques*, 70(11):4635–4651, 2022.

Weiquan Zhang, Yue Li, Kunpeng Wei, and Zhijun Zhang. A two-port microstrip antenna with high isolation for wi-fi 6 and wi-fi 6e applications. *IEEE Transactions on Antennas and Propagation*, 70(7):5227–5234, 2022.

Shijian Zheng, Fangxiao Jin, Shuhai Zhang, Quan Xue, and Mingkui Tan. Deep electromagnetic structure design under limited evaluation budgets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78170–78189. PMLR, 2025. URL https://proceedings.mlr.press/v267/zheng25g.html.

Min Zhu, Handi Zhang, Anran Jiao, George Em Karniadakis, and Lu Lu. Reliable extrapolation of deep neural operators informed by physics or sparse observations. *Computer Methods in Applied Mechanics and Engineering*, 412:116064, 2023.

A Experimental Details

This appendix specifies the task objectives, data handling, and calibration searches used in the experiments. All methods use the same test observations and candidate pools.

A.1 Data and Design Objectives

Data. HGA contains 37,655 structures. RFF contains 68,241 structures and 68,640 response records after removing 428 exact input–output duplicates. Records sharing a geometry remain in the same data role, and each structure receives equal total weight. HP contains 10,000 structures from the public third-order hairpin dataset accompanying *Federated Learning for Microwave Filter Behavior Prediction*. We use the `hairpin_BPF_3th_gap_combined.csv` file from the authors' repository, version `726e39b`. Its S21M and S21P columns represent real and imaginary components, respectively. The nine geometric parameters form the input; sample identifiers are excluded. All tasks retain their original frequency grids.

HGA. The antenna task originates from Zheng et al. (2025). We use a collection of simulated responses for this task with the data partition and score specified here. For a gain response $y(f)$, the score is the sum of two band minima,

$$g_{\mathrm{HGA}}(y) = \min_{f\in[2.45,2.55]} y(f) + \min_{f\in[5.0,5.9]} y(f). \tag{10}$$

All frequencies are in GHz, and band endpoints are included. The second interval is the intersection of the original specification $[5, 6]$ with the available grid.

RFF. Each filter is represented by a 15×16 binary geometry array, with transmission S_{21} recorded in dB at 1,001 frequencies from 10 to 30 GHz. The CST model uses a Rogers RO4350B substrate ($\epsilon_r = 3.66$, $\tan\delta = 0.0037$) and annealed-copper conductors ($\sigma = 5.8\times10^7$ S/m). Two single-mode waveguide ports terminate the feed lines. The simulation project uses the CST 2024.4 frequency-domain solver with second-order tetrahedral elements, adaptive mesh refinement, a solver accuracy setting of 10^{-4}, and an S-parameter mesh-convergence threshold of 0.01 over two checks. The score combines mean transmission in two bands with a high-frequency suppression term:

$$g_{\mathrm{RFF}}(y) = \mathrm{mean}_{f\in[10,15]}\, y(f) + \mathrm{mean}_{f\in[20,25]}\, y(f) + B(y). \tag{11}$$

Here y is measured in dB. In the band $[25, 30]$, $B(y)$ is the width between the first and last frequencies of the longest consecutive run satisfying $y(f) \le -20$. Isolated qualifying points have zero width. If fewer than two points in the band satisfy the threshold, $B(y) = -\min_{f\in[25,30]} |y(f)+20|$. We apply this fixed numerical composite without interpolating threshold crossings. Figure 5 illustrates the filter's design region and fixed feed structure.

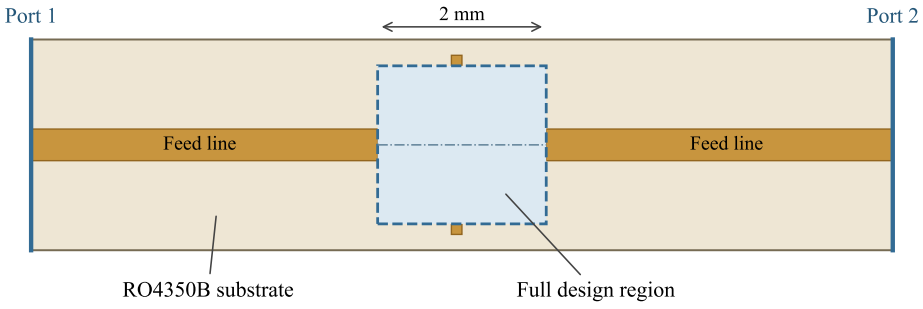


Figure 5: RFF design domain (top view). Gold denotes fixed copper structures on the RO4350B substrate; the blue dashed rectangle marks the full design region, symmetric about the feed centerline. The blue fill indicates the design domain, not the copper pattern of a sampled filter.

HP. For complex transmission $z(f)$, let $a(f) = 20\log_{10}|z(f)|$. We maximize the negative of the worst specification deviation,

$$g_{\mathrm{HP}}(z) = -\max\left\{-\min_{f\in[4.8,5.2]} a(f) - 3,\ \max_{f\in[3,4]} a(f) + 20,\ \max_{f\in[6,7]} a(f) + 20\right\}. \tag{12}$$

This corresponds to a 3 dB passband insertion-loss requirement and 20 dB attenuation in both stopbands. Negative deviations are retained to distinguish designs by their margins. Reconstruction operates on interleaved real and imaginary components; conversion to dB occurs only during objective evaluation.

A.2 Implementation and Parameter Selection

Surrogate Architectures. The MLP has three hidden layers of width 256 with GELU activations. The HGA and RFF two-dimensional ResNets reshape standardized inputs in row-major order into single-channel grids of size 15×20 and 15×16, respectively. A 3×3 stem convolution and GELU lift the input to C channels, with $C = 68$ for HGA and $C = 87$ for RFF. Each of two residual blocks computes $(h + \mathrm{Conv}_2(\mathrm{GELU}(\mathrm{Conv}_1(h))))/\sqrt{2}$, using C-channel 3×3 convolutions with stride 1 and zero padding of width 1. Adaptive average pooling to 3×4 precedes a $12C$-to-C linear layer, GELU, and an output layer with 71 coordinates for HGA or 1,001 for RFF.

For HP, a linear layer and GELU map the nine inputs to width 185. Each of two residual blocks computes $(h + W_2\,\mathrm{GELU}(W_1 h + b_1) + b_2)/\sqrt{2}$ with width-185 linear maps. The skips are identities, without projection or post-addition activation, and a linear head produces 82 coordinates. Neither residual architecture uses normalization or dropout, and all convolutional and linear layers include biases. The HGA, RFF, and HP residual networks contain 227,871, 452,705, and 154,742 trainable parameters, respectively.

Surrogate Training. We train with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 32, a maximum of 250 epochs, and an early-stopping patience of 30 epochs. The output-layer weights and biases are initialized to zero. Training minimizes standardized response MSE; checkpoint selection uses physical response MSE on the validation set. Only training data determine the normalization statistics. Real and imaginary components at a given frequency share a response scale. Architectures remain fixed across training-set sizes and repetitions.

Shared Protocol. Within each repetition, the architectures share data roles and the training sizes use nested subsets. The validation/reference set and calibration selection set contain 256 and 128 structures, respectively, in addition to the stated training budget. Thus, training size is not the total number of offline labeled structures. We generate feedback by randomly ordering the native frequencies of each structure and taking the first k positions; records of the same structure share these positions. Each candidate pool contains 256 distinct structures, while different pools may overlap. Ties are resolved by fixed structure-identifier order. For complex responses, UMSE averages the squared complex error over unobserved frequencies, without an extra factor of $1/2$.

Model Tuning. Fine-tuning uses Adam on standardized observed response MSE and resets the model for each test sample. Latent adaptation freezes the source prediction head and optimizes the latent vector using physical observed response MSE. Both search learning rates $\{10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ and update counts $\{1, 5, 20\}$, together with a no-update option. Gappy AE has one width-128 SiLU hidden layer in each of its encoder and decoder. It is trained on complete training responses with Adam, learning rate 10^{-3}, batch size 128, at most 1,000 epochs, and validation patience 100. Its latent dimension is selected from $\{2, 3, 4, 6\}$ and its maximum Gauss–Newton iteration count from $\{1, 4, 10\}$, with a no-update option. Convergence can stop the latent solve earlier. The source prediction initializes its latent representation. All three methods select adaptation parameters by objective MSE on the separate calibration selection set.

Reconstruction Baselines. The spline fits observed residuals with natural boundary conditions and keeps the residual constant at the nearest observed endpoint outside the observed range. Real and imaginary components are handled separately. Response GPs use the validation/reference responses, not the surrogate training set or source predictions. Their regularization search is $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, with the prior mean as an additional option. The Matérn-$3/2$ GP also searches length scales $\{0.03, 0.1, 0.3, 1\}$ on normalized frequency coordinates. GP parameters minimize unobserved response MSE on the calibration selection set. Identical GP predictions are reused across source architectures and training sizes within a task and repetition.

DARC. Reconstruction uses the same six regularization values and a no-correction option, selected by unobserved response MSE. Scoring then searches $\rho \in \{0.1, 0.25, 0.5, 0.75\}$ and $\tau \in \{3, 1, 0.3, 0.1\}$, together with the fixed reference weights π_j, using objective MSE. Reference mass

is equal across structures and divided among each structure's records. Selecting no correction returns the original prediction and its score, without scenario scoring. All searches average the three calibration feedback layouts and freeze their choices before evaluation. Search spaces and computational costs differ across methods; the shared controls are data roles, feedback, and evaluation pools.

Metric Aggregation. For a structure with multiple records, we average its true and estimated scalar objectives separately before computing objective MAE or MSE. Response error is first averaged over records within each structure. Both response and objective errors are then averaged across structures in each 256-candidate pool, followed by the 100 pools and three layouts. Thus, these metrics use the same pool weighting as regret, rather than a separate uniform average over the evaluation library.

A.3 Labeled Data Budgets and Result Availability

Table 4 separates surrogate training from the additional complete responses used for validation, reference statistics, and parameter selection. Counts refer to unique structures, not response records. The three offline roles are disjoint; their labels are reused across methods rather than acquired separately for each method. At test time, k denotes observed frequency positions per response, not complete labeled structures.

Table 4: Offline labeled structures used by each method. $N \in \{128, 256, 512\}$. Validation and reference uses of the same 256 structures are counted only once. Source uses no test feedback; the other methods use the shared k observations.

Method	Training	Validation/reference	Selection	Total
Source; residual spline	N	256	0	$N + 256$
Model tuning	N	256	128	$N + 384$
Response GPs	0	256	128	384
DARC	N	256	128	$N + 384$

For DARC and model tuning, the total offline counts are therefore 512, 640, and 896 for the three training sizes. Response GPs are fitted directly to reference responses and do not use the surrogate training set. The 4,096 evaluation structures form a separate library: only the selected frequencies are exposed to calibration, while complete responses provide evaluation scores. This protocol compares methods under shared data roles and feedback, rather than equalizing their total label consumption.

Numerical Results. The accompanying numerical supplement reports all 90 task–architecture–training-size–feedback-budget settings for selection regret, unobserved response MSE, and objective MAE. The machine-readable results include 4,500 method–setting–repetition rows and 900 method–setting summaries, together with the scoring-ablation results and random seeds. Standard deviations use the five repetitions, with denominator four. File hashes are provided in a manifest. These experimental results are distinct from the underlying geometry–response datasets. The full datasets and implementation are planned for release upon acceptance; only a small subset of dataset samples is intended for the submission supplement.

A.4 Solving the Effective Sample Size Projection

We solve Eq. (8) in double precision using normalized log similarities, avoiding explicit exponentiation of very small similarities. If $\sum_j v_j^2 \le 1/(\rho M)$, the solution is $\boldsymbol{w}^\star = \boldsymbol{v}$, where $\boldsymbol{v} = (v_1, \dots, v_M)^\top$. For $\rho = 1$ or $M = 1$, the feasible solution is uniform. Otherwise the quadratic constraint is active. The stationarity conditions give

$$\log w_j + t w_j = \log v_j + a, \qquad \sum_j w_j = 1, \qquad \sum_j (w_j - 1/M)^2 = \frac{1-\rho}{\rho M}, \tag{13}$$

where $t > 0$ is twice the multiplier of the quadratic constraint and a enforces normalization. For fixed t, the positive solution is $w_j = W(t v_j e^a)/t$, where W is the principal Lambert function. We evaluate this relation in log coordinates by solving $x + e^x = \log v_j + a + \log t$ and setting $\log w_j = x - \log t$.

The implementation uses nested safeguarded Newton iterations with bisection. The inner normalization solve brackets a in $[0, t]$ and uses at most 48 iterations, targeting a normalization residual of 8×10^{-14}. The outer solve brackets t between zero and

$$(1 + 10^{-12}) \frac{\| \log \boldsymbol{v} - \overline{\log v}\, \mathbf{1} \|_2}{\sqrt{(1-\rho)/(\rho M)}}, \qquad \overline{\log v} = M^{-1} \sum_j \log v_j.$$

It uses at most 80 iterations. With $\delta = (1-\rho)/(\rho M)$ and $h = \sum_j (w_j - 1/M)^2$, convergence requires $|h - \delta| \le 2 \times 10^{-13}/M$, together with scaled stationarity and complementarity residuals at most 2×10^{-11}. The scale is $1 + \max_j |\log v_j|$. Final checks verify normalization, nonnegativity, the effective sample size bound, stationarity, and complementarity; a failed check raises an error rather than silently returning an unconstrained fallback. The fixed-reference-weight option uses π_j directly and does not invoke this projection.

B Additional Results

Tables 5 and 6 report response and objective errors using the aggregation of the main results. Figure 6 includes all external baselines across feedback budgets, and Figure 8 shows the five paired repetitions behind the scoring comparison. All error bars and reported standard deviations refer to model repetitions, not to overlapping pools.

Table 5: Unobserved response MSE (↓), with the same aggregation as Table 2. HP values are scaled by 10^2 for display; errors are measured in the original response coordinates.

	HGA		RFF		HP ($\times 10^{-2}$)	
Method	MLP	ResNet	MLP	ResNet	MLP	ResMLP
Source	17.04 ±0.10	16.19 ±0.41	35.27 ±0.38	32.06 ±0.37	3.71 ±0.16	3.57 ±0.20
Model tuning						
Fine-tuning	11.81 ±0.69	11.91 ±0.54	12.25 ±1.47	12.51 ±1.51	2.82 ±0.09	2.76 ±0.19
Latent adaptation	15.02 ±0.20	15.55 ±0.45	25.92 ±0.33	29.64 ±0.29	3.18 ±0.14	3.25 ±0.27
Gappy AE	12.41 ±0.38	12.22 ±0.62	11.06 ±4.43	10.03 ±1.89	3.59 ±0.32	3.33 ±0.34
Data reconstruction						
Residual spline	10.44 ±0.02	10.15 ±0.12	5.26 ±0.06	5.29 ±0.05	10.86 ±0.38	10.62 ±0.33
Response GP (emp.)	7.76 ±0.02	7.76 ±0.02	5.26 ±0.07	5.26 ±0.07	3.27 ±0.03	3.27 ±0.03
Response GP (Matérn)	9.62 ±0.04	9.62 ±0.04	6.66 ±0.10	6.66 ±0.10	10.66 ±0.02	10.66 ±0.02
DARC (ours)	**6.53** ±0.02	**6.34** ±0.09	**4.27** ±0.07	**4.26** ±0.03	**2.02** ±0.05	**1.96** ±0.03

Table 6: Objective MAE (↓), with the same aggregation as Table 2.

	HGA		RFF		HP	
Method	MLP	ResNet	MLP	ResNet	MLP	ResMLP
Source	6.42 ±0.26	5.95 ±0.10	4.82 ±0.13	4.83 ±0.11	2.89 ±0.08	2.76 ±0.09
Model tuning						
Fine-tuning	4.56 ±0.15	4.60 ±0.09	2.17 ±0.18	2.24 ±0.11	2.11 ±0.03	2.07 ±0.05
Latent adaptation	5.53 ±0.23	5.64 ±0.09	3.96 ±0.10	4.55 ±0.09	2.71 ±0.07	2.67 ±0.09
Gappy AE	4.60 ±0.10	4.46 ±0.04	1.57 ±0.08	1.54 ±0.06	2.47 ±0.06	2.46 ±0.07
Data reconstruction						
Residual spline	3.60 ±0.09	3.48 ±0.03	1.24 ±0.02	1.23 ±0.01	2.30 ±0.05	2.26 ±0.04
Response GP (emp.)	3.96 ±0.05	3.96 ±0.05	**1.21** ±0.01	1.21 ±0.01	2.40 ±0.05	2.40 ±0.05
Response GP (Matérn)	4.66 ±0.03	4.66 ±0.03	1.54 ±0.06	1.54 ±0.06	2.78 ±0.11	2.78 ±0.11
DARC (ours)	**3.14** ±0.04	**3.04** ±0.04	1.22 ±0.03	**1.20** ±0.03	**1.23** ±0.03	**1.22** ±0.01

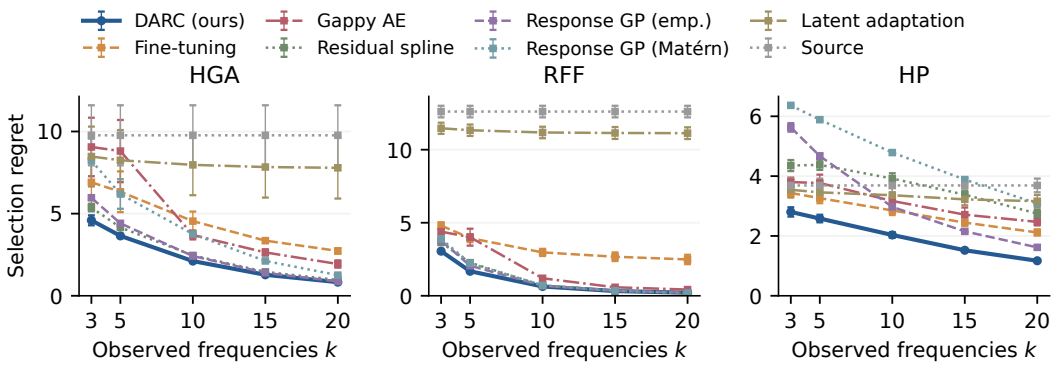


Figure 6: All baseline selection-regret curves. Architectures and training sizes are averaged within each repetition; error bars indicate one standard deviation across five repetitions.

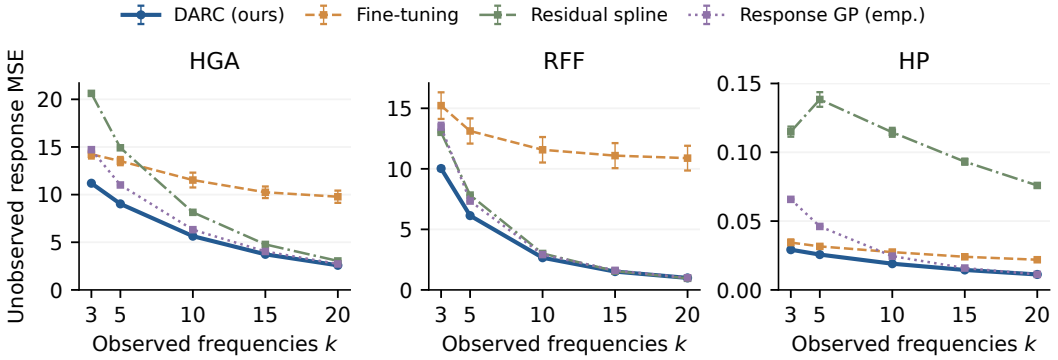


Figure 7: Unobserved response error across feedback budgets for the methods displayed in the main budget figure. HP error is measured in complex response coordinates, not in dB amplitude.

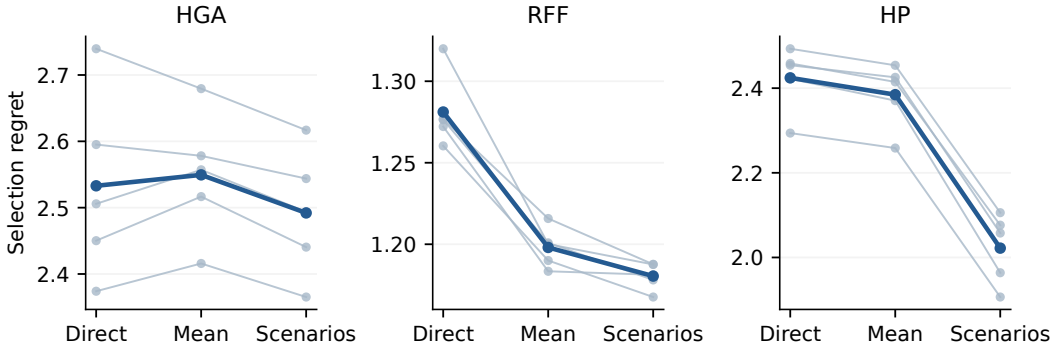


Figure 8: Paired scoring comparison across five repetitions (thin lines), with their mean highlighted. Direct, Mean, and Scenarios denote $g(\boldsymbol{q})$, $g(\boldsymbol{q} + \sum_j w_j^\star \boldsymbol{t}_j)$, and $\sum_j w_j^\star g(\boldsymbol{q} + \boldsymbol{t}_j)$, respectively. Architectures, training sizes, and budgets are averaged within each repetition. Each task uses its own vertical scale.

B.1 HGA Reconstruction Case

Figure 4 uses HGA structure 24068, the ResNet trained on 512 structures in repetition 0, and feedback layout 0. Fine-tuning uses objective-MSE parameter selection. All curves use the native frequency grid without smoothing. The case was selected post hoc from a gallery of reconstruction

examples: its DARC UMSE was at least 5% lower than the best included comparator at three or more of the five budgets, and no more than 10% higher at the other budgets. Gallery comparators included Source, residual spline, empirical response GP, and both objective-MSE- and UMSE-selected versions of fine-tuning, latent adaptation, and Gappy AE. Within qualifying cases, examples near the median relative advantage were retained. The displayed fine-tuning and spline methods are the strongest HGA baselines in their respective families in the aggregate main table, not necessarily for this individual structure. The displayed budgets are $k = 3, 5, 15, 20$; UMSE is evaluated on each budget's unobserved frequencies, whose membership changes with k. This example illustrates reconstruction rather than establishing an improvement in selection regret.

B.2 Extended Response-Correction Comparisons

Figures 9–11 extend the response-level comparison to six methods at $k = 3, 10, 20$. Each case uses identical nested observations across methods, and all curves retain their native frequency grids without smoothing. Source predictions are included in every panel. Fine-tuning, latent adaptation, and Gappy AE use the cached objective-MSE-selected variants; the response GP uses the empirical kernel. The available case caches do not include the Matérn response GP.

HGA uses the same structure as Figure 4. RFF structure 55077 and HP structure 7369 come from the same gallery's stable-advantage category, with the eligibility and near-median selection rule described in Appendix B.1. The RFF case uses an MLP trained on 256 structures in repetition 4; the HP case uses a ResMLP trained on 256 structures in repetition 3. Both use feedback layout 0. These outcome-conditioned examples illustrate correction behavior, rather than estimating its frequency across the test population. UMSE is evaluated only at unobserved coordinates. For HP, curves show dB magnitude, whereas UMSE is computed in the original real and imaginary response coordinates.

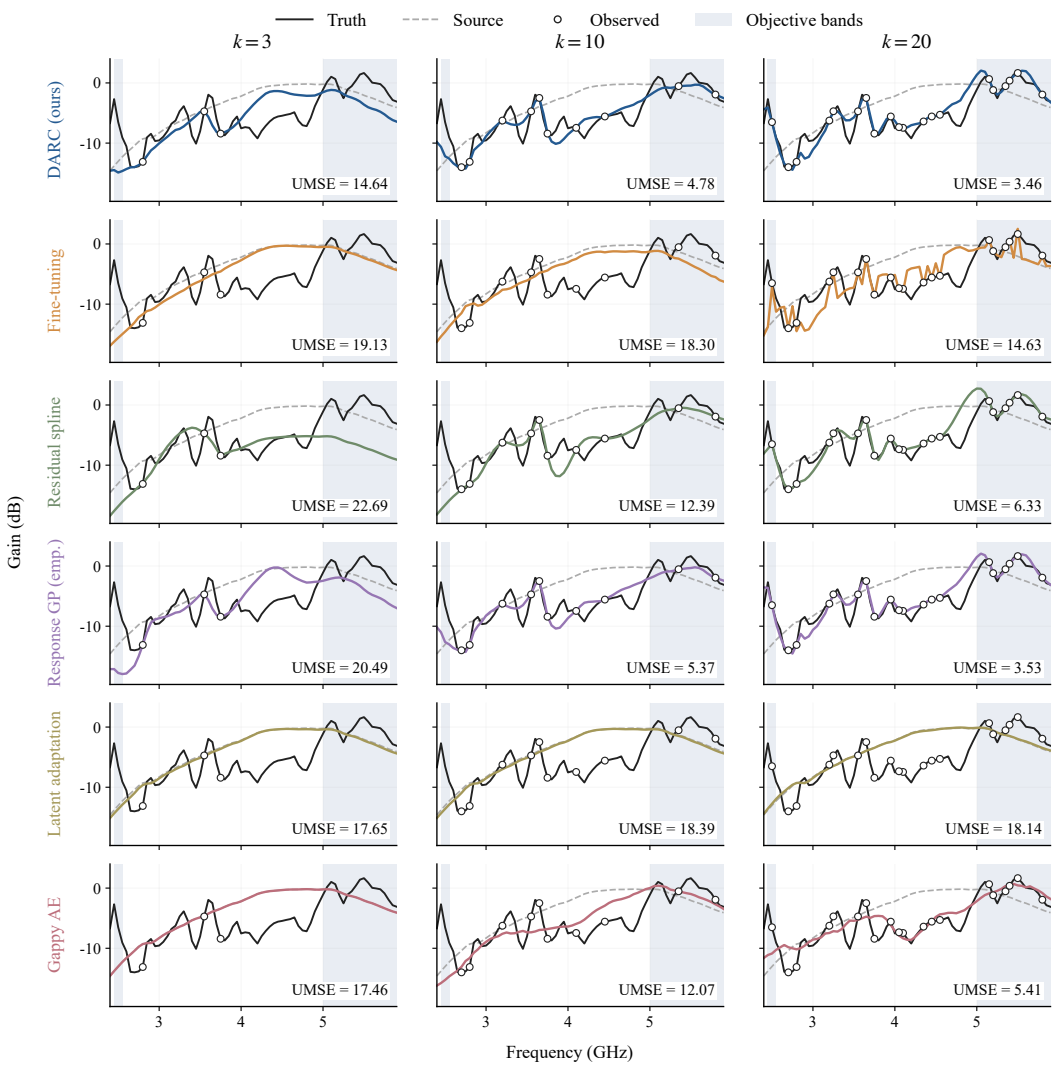


Figure 9: Extended response-correction comparison for HGA structure 24068 (ResNet, 512 training structures, repetition 0). Rows show six methods; columns show three feedback budgets. Shading marks objective bands, and each panel reports UMSE.

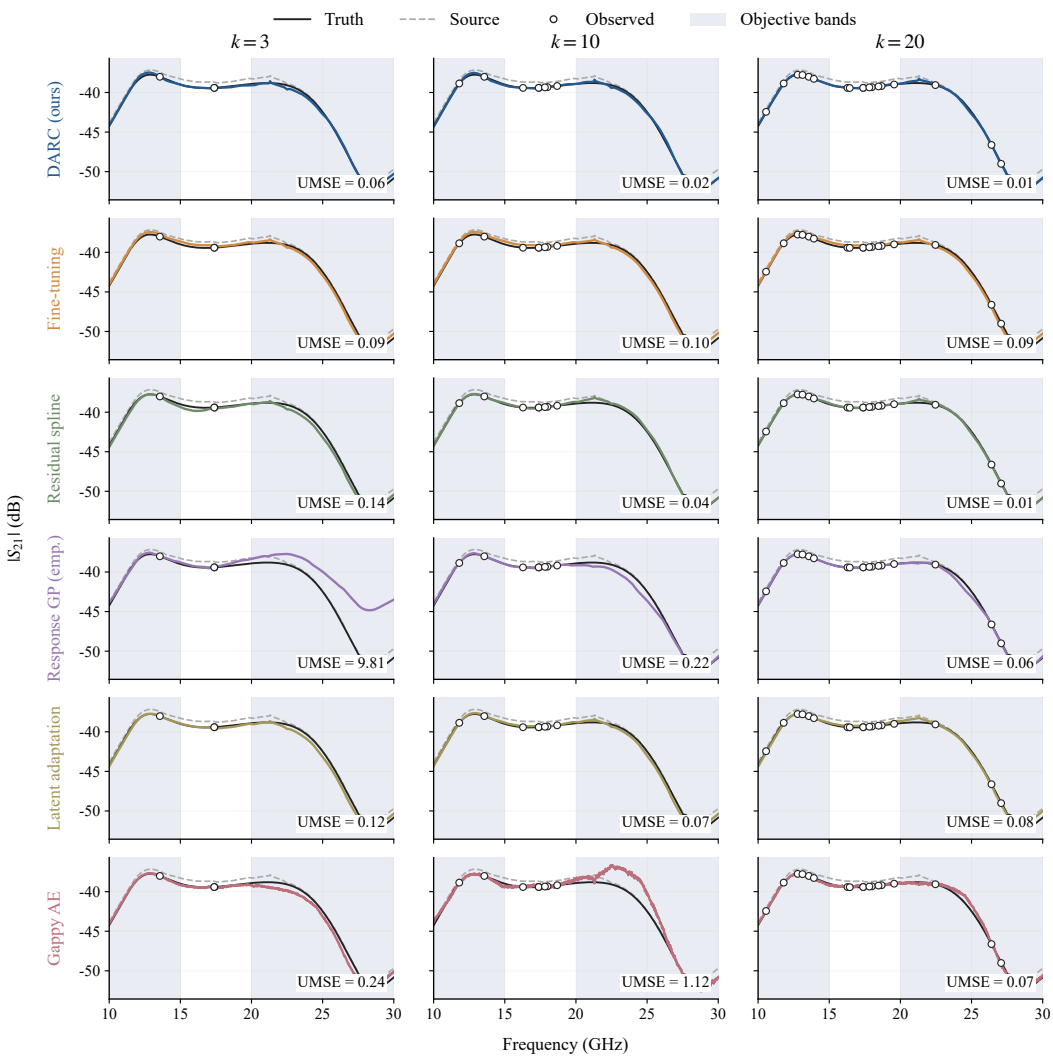


Figure 10: Extended response-correction comparison for RFF structure 55077 (MLP, 256 training structures, repetition 4). All methods receive the same observations within each column. Shading marks objective bands; UMSE measures error at unobserved frequencies.

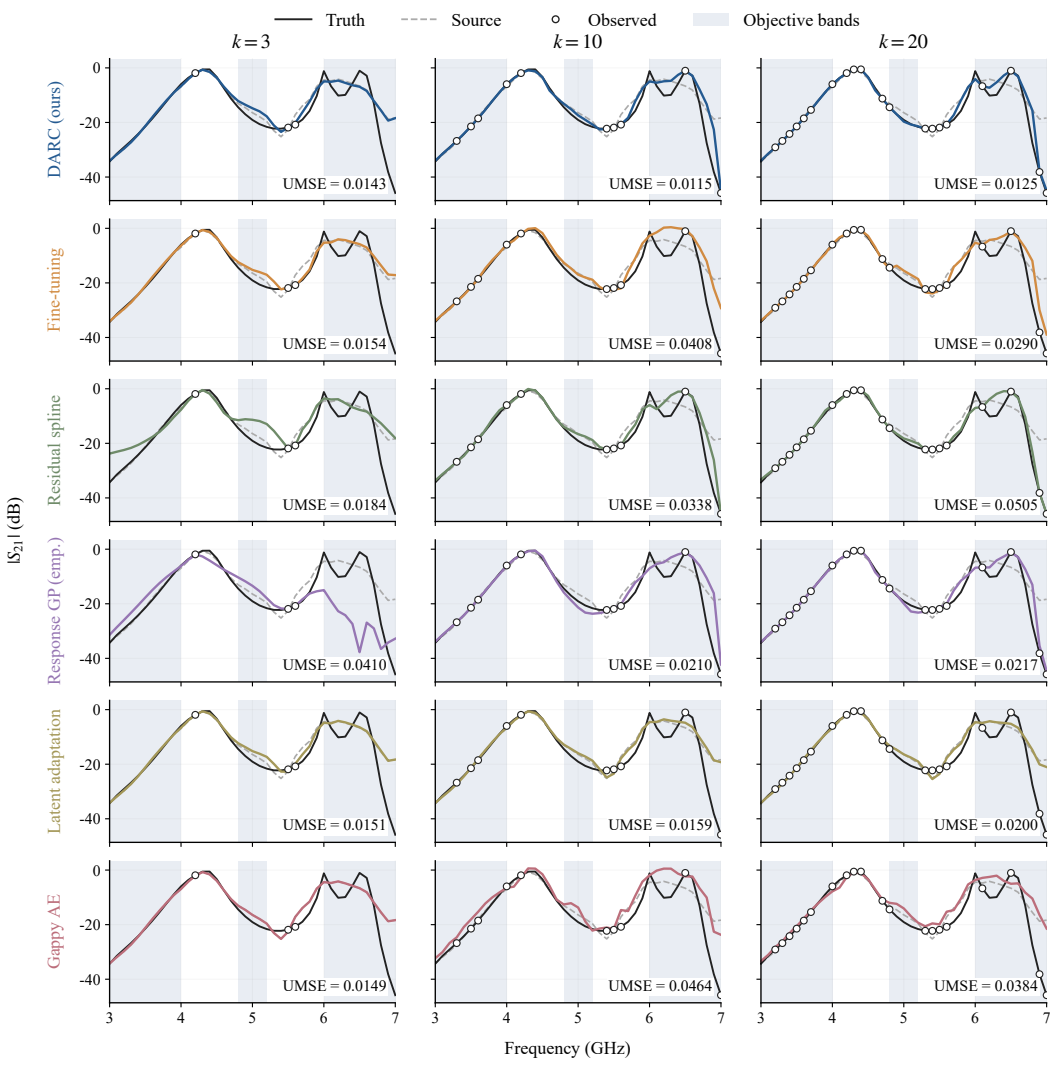
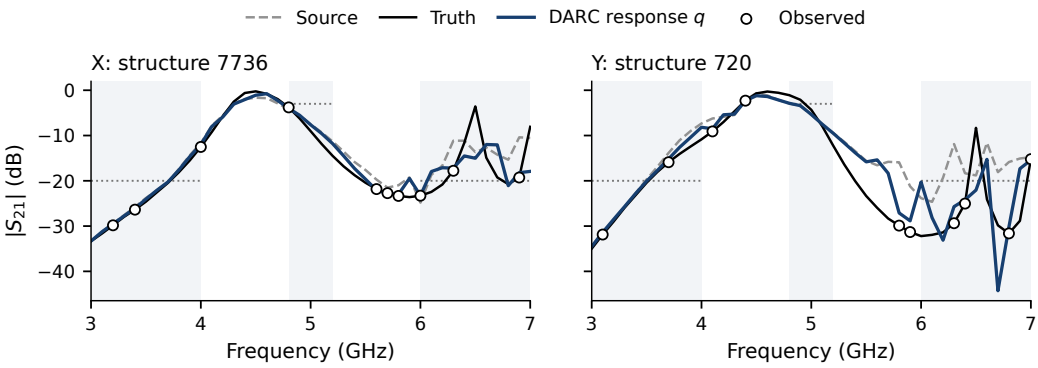


Figure 11: Extended response-correction comparison for HP structure 7369 (ResMLP, 256 training structures, repetition 3). Shading marks objective bands. Curves show transmission magnitude in dB; UMSE is computed over unobserved real and imaginary coordinates, not dB magnitudes.

B.3 Response and Selection Case Study

We use the first original HP candidate pool for the MLP trained on 256 structures in repetition 0, with feedback layout 0 and $k = 10$. This fixes a pool of 256 structures without drawing a new evaluation set. Figure 13 shows every candidate under direct scoring, weighted-mean-response scoring, and scenario scoring. Their Top-13 regrets are 1.674, 2.304, and 0.876, respectively. These are single-pool results, separate from the aggregate evidence in Table 3.



Score / rank	True	$g(q)$	Scenario score
X	-16.43 / 54	-8.78 / 7	-13.28 / 14
Y	-11.70 / 7	-11.82 / 20	-12.04 / 6

Figure 12: An illustrative HP selection case with ten observed frequencies. Shading marks the objective bands; dotted segments mark the -3 and -20 dB limits. The table beneath the curves gives scores (higher is better) and ranks within the same 256-candidate pool. Direct scoring selects X and misses Y; scenario scoring reverses their order and selects Y instead. The pair is chosen post hoc to illustrate a corrected selection.

From Curves to Choices. In Figure 12, direct scoring favors X because the corrected response misses its sharp upper-stopband peak. DARC instead ranks the genuinely better Y above X, moving Y into the Top-13 set and X out of it. The response curves remain unchanged; only their design scores differ. Notably, Y's score becomes slightly less accurate in absolute terms, but its position relative to X is corrected. This example illustrates how scenario scoring can improve the final design choice by correcting an overly optimistic assessment, even when individual score errors do not uniformly decrease.

For the pair in Figure 12, we identify structures incorrectly selected by direct scoring but excluded by scenario scoring, and structures missed by direct scoring but correctly selected by scenario scoring. We first seek pairs with both true ranks between 7 and 26. None satisfy this boundary restriction, so we consider all six eligible pairs and select the pair whose true-score gap is closest to the median. This yields X (structure 7736, true rank 54) and Y (structure 720, true rank 7). The pair is an outcome-conditioned explanatory example, not a random sample or an estimate of overall success frequency. Scores use the maximization convention throughout; negating the original HP minimization objective preserves ranks and regret.

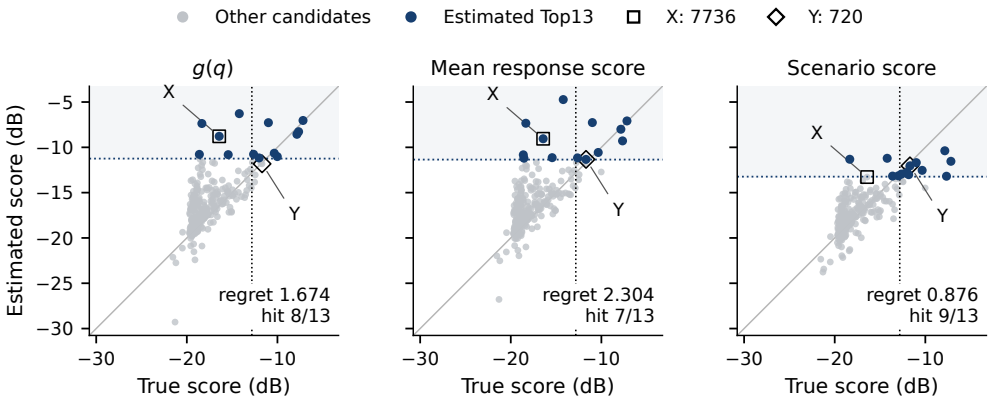


Figure 13: Objective estimates and selection in the fixed HP pool. Dark points are each rule's selected 13 candidates. Vertical and horizontal dotted lines mark the true and estimated selection cutoffs; the diagonal marks exact score agreement. Higher scores are better. The hit count measures overlap with the true Top-13 set. Open markers identify X and Y from Figure 12.