Representative, Informative, and De-Amplifying: Requirements for Robust Bayesian Active Learning Under Model Misspecification

Roubing Tang[1] **Sabina J. Sloman**[1] **Samuel Kaski**[1,2,3]
[1] Department of Computer Science, University of Manchester, Manchester, UK
[2] ELLIS Institute Finland, Helsinki, Finland
[3] Department of Computer Science, Aalto University, Helsinki, Finland
{roubing.tang, sabina.sloman, samuel.kaski}@manchester.ac.uk

Abstract

In many science and industry settings, a central challenge is designing experiments under time and budget constraints. *Bayesian Optimal Experimental Design (BOED)* is a paradigm to pick maximally informative designs that has been widely applied to such problems. During training, BOED selects inputs according to a pre-determined acquisition criterion to target *informativeness*. During testing, the model learned during training encounters a naturally occurring distribution of test samples. This leads to an instance of covariate shift, where the train and test samples are drawn from different distributions (the training samples are not *representative* of the test distribution). Prior work has shown that in the presence of model misspecification, covariate shift amplifies generalization error. Our first contribution is to provide a mathematical analysis of generalization error in the presence of model misspecification, revealing that, beyond covariate shift, generalization error is also driven by a previously unidentified phenomenon we term *error (de-)amplification*. We then develop a new acquisition function that mitigates the effects of model misspecification by including terms for representativeness, informativeness, and de-amplification (R-IDeA). Our experimental results demonstrate that the proposed method performs better than methods that target only informativeness, only representativeness, or both.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Bayesian modeling is a principled approach to making inferences when data is scarce or costly. Most Bayesian machine learning methods are developed under the assumption that the true data-generating process (DGP) is included in the chosen model family (Bernardo and Smith, 2009). However, in complex real-world environments, this assumption rarely holds: The true DGP often lies outside of the assumed model family (Uppal and Wang, 2003). The inevitability of the phenomenon of *model misspecification* (Walker, 2013) is captured by the saying that "all models are wrong" (Box., 1976; Box, 1980). Common causes of model misspecification include omitted variables (Wooldridge, 2010), mistaken beliefs about the structure of the error term (e.g., a failure to account for heteroskedasticity or autocorrelation; Greene 2003; Grünwald and Van Ommen 2017), or the choice of a misinformed or underexpressive model class (Wooldridge, 2010; Dubova et al., 2025). The consequences of model misspecification range from biased inferences (Greene, 2003; Müller, 2013; Caprio et al., 2023; Bonhomme and Weidner, 2022), unreliable approximations (e.g., in simulation-based inference methods; Frazier et al. 2020; Lintusaari et al. 2017; Huang et al. 2023), to suboptimal decision-making (Sutton et al., 1998; Rainforth et al., 2024).

There is a substantial literature on the effects of model misspecification on Bayesian inference when data is independently and identically distributed (i.i.d.), or "passively" collected from the distribution to which the learner wants their inferences to generalize (Kleijn and van der Vaart, 2006, 2012; Knoblauch et al., 2022; Walker, 2013; Nott et al., 2023; Kelly et al., 2025). However, in part because of the widespread availability of large datasets, *active* learning methods have become much more prevalent (Settles, 2009). These methods select the training data to tailor it to a specified learning objective (Silberman, 1996; Farquhar et al.,

2021). Active learning methods rely on the specified model twice: once to make inferences to fit training data, and again to select the data (Konyushkova et al., 2017). Model misspecification thus has a double impact on these methods, introducing potential bias in both the acquisition function and the resulting inferences. In particular, in the context of active learning, model misspecification can produce poor quality datasets (Sugiyama, 2005; Bach, 2006; Ali et al., 2014; Vincent and Rainforth, 2017; Farquhar et al., 2021). Understanding the consequences of model misspecification is of paramount importance to developing robust active learning methods.

In a Bayesian setting, Bayesian Optimal Experimental Design (BOED) is a natural and frequently used active learning method (Rainforth et al., 2024; Huan et al., 2024). BOED selects the optimal design by maximizing an acquisition function known as *the expected information gain* (Rainforth et al., 2024; Chaloner and Verdinelli, 1995), enabling budget and time efficiency in many applications, such as drug discovery (Park et al., 2013), clinical trial design (Chaloner and Verdinelli, 1995), chemistry (Walker and Ravisankar, 2019; Hickman et al., 2022), biology (Kreutz and Timmer, 2009; Thompson et al., 2023), and psychology (Cavagnaro et al., 2010; Myung et al., 2013). While the limitations of BOED in the presence of model misspecification have been acknowledged in the literature, characterizing and proposing methods to overcome this limitation is an area of on-going research (Overstall and McGree, 2022; Sloman et al., 2022; Catanach and Das, 2023; Schmitt et al., 2023; Ivanova et al., 2024; Barlas et al., 2025; Forster et al., 2025; Overstall et al., 2025).

We provide a novel theoretical analysis of generalization error in the presence of model misspecification. Our analysis reveals that training datasets that lead to robustness to model misspecification have two properties: They are *representative* of the target DGP, and they are *de-amplifying*. The expected information gain does not include a term for either of these, and standard BOED can lead to training datasets that have neither of these characteristics. In this sense, standard BOED is not robust to model misspecification.

Unrepresentative Training Data. BOED selects samples to achieve a particular objective, and these samples likely do not reflect the distribution to which the learner would like to generalize. In other words, BOED induces a form of distribution shift, whereby the distribution used for (active) learning is different than the distribution used for evaluation. Recent work on the interaction between model misspecification and distribution shift has introduced the concept of *misspecification amplification* (Amortila et al., 2024), whereby the generalization error attributable to misspecification is "amplified" by the density ratio between the test and training input distributions. A similar phenomenon has been observed in the context of BOED: In the presence of model misspecification, the generalization error in some settings has been shown to depend on the degree of model misspecification and the extent of distribution shift (Sloman et al., 2022).

De-amplifying Training Data. As our novel decomposition of generalization error shows, generalization performance depends on not only the representativeness of the training data, but also on the way it interacts with model (mis)specification (which will be defined in Section 3): Generalization performance is enhanced when training data is in regions where the direction in which the model will tend to adjust on the basis of these data opposes the direction in which the model is misspecified. We refer to this property as error "de-amplification" to stress that the effect is to counteract, rather than amplify, the effect of the misspecification.

Contributions. In this work, we explore the problem of BOED under model misspecification and make the following contributions:

- **Theoretical Decomposition of Generalization Error.** Prior work has primarily explored the effects of misspecification and distribution shift, overlooking the role of de-amplifying designs. We formally decompose generalization error into three components: (1) misspecification bias, (2) estimation bias, and (3) a novel term we introduce, *error (de-)amplification*. We also derive an upper bound on generalization error that characterizes its dependence on the representativeness of the training data, the degree to which these data are de-amplifying, and model misspecification.

- **Novel Acquisition Function Incorporating Representativeness and De-amplification.** We propose a novel acquisition function designed to mitigate the effects of model misspecification by identifying designs that not only are informative, but are additionally representative and de-amplifying. Our empirical results show that the new acquisition outperforms traditional BOED in the presence of misspecification.

2 PRELIMINARIES

2.1 Problem Setting

A modeler aims to predict an observed variable $y \in \mathbb{R}$ which depends on a fully observable input (design) $\xi \in \Xi \subseteq \mathbb{R}^d$. The relationship between the observed variable y and the input ξ is governed by a conditional

distribution $y|\xi \sim P^\star$, referred to as the *true data-generating process (DGP)*, which depends on the output of a true (possibly unknown) regression function $f^\star(\xi)$ and observation noise. Let $\{(\xi_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples drawn from the true DGP $P^\star$. To approximate the true DGP, the modeler proposes a hypothetical model $f(\xi, \theta) : \Xi \mapsto \mathbb{R}$, where $\theta \in \Theta$ represents the parameters within the fixed parameter space Θ. The model class is denoted $\mathcal{F}(\xi, \Theta) = \{f(\xi, \theta) : \theta \in \Theta\}$. *Model misspecification* arises when the assumed model class $\mathcal{F}(\xi, \Theta)$ fails to capture the true DGP (Walker, 2013; Kleijn and van der Vaart, 2012). Let $\hat{f}^{(n)}(\xi)$ be a learned predictor, depending on training designs $\{\xi_1, \ldots, \xi_n\}$. Let $\bar{f} \in \mathcal{F}$ be the predictor that best approximates the true data-generating function $f^\star$, i.e., $\bar{f} = \arg\min_{f \in \mathcal{F}} R_{\text{test}}(f)$.

Definition 1 (Model misspecification)**.** *Model misspecification occurs when the assumed model class $\mathcal{F}(\xi, \Theta) = \{f(\xi, \theta) : \theta \in \Theta\}$ cannot mimic the true output $f^\star(\xi)$ for any parameter $\theta \in \Theta$. That is, the model is misspecified if*

$$f^\star(\xi) \notin \mathcal{F}(\xi, \Theta). \tag{1}$$

In Bayesian inference, the modeler additionally specifies a prior distribution over the model parameters. According to this prior, the probability that the data the learner will encounter is generated by a value θ is $p(\theta)$. Model fitting is carried out by updating the prior distribution using Bayes' rule. The result is a posterior distribution which assigns to a θ a probability $p(\theta \mid y, \xi) \propto p(\theta)p(y \mid \theta, \xi)$. This process depends on both the assumed prior and the specified likelihood model. When the model is misspecified, i.e., the likelihood does not reflect the true DGP, the updated posterior becomes unreliable or biased (Frazier et al., 2023; Oberauer et al., 2025).

2.2 Bayesian Optimal Experimental Design

Bayesian Optimal Experimental Design (BOED) is a model-based framework to select the optimal design ξ by maximizing the expected information gained about the parameter θ, enabling budget and time efficiency (Rainforth et al., 2024; Chaloner and Verdinelli, 1995). The expected information gain (EIG) is (Dong et al., 2024; Lindley, 1956):

$$\begin{aligned} \text{EIG}(\xi) &= \mathbb{E}_{p(y|\xi)}[\text{IG}_\theta(\xi, y))] \\ &= \mathbb{E}_{p(\theta, y|\xi)}[\log p(y \mid \theta, \xi) - \log p(y \mid \xi)] \end{aligned} \tag{2}$$

The optimal design $\xi^\star$ is the design in the set of candidate designs Ξ that maximizes the EIG:

$$\xi^\star = \text{argmax}_{\xi \in \Xi} \text{EIG}(\xi). \tag{3}$$

Traditional BOED methods (Foster et al., 2019; Sebastiani and Wynn, 1997), also called Bayesian Adaptive Design (BAD), iterate between making design decisions by evaluating Equation (3), and updating the underlying model through Bayesian inference to condition on data obtained so far. Traditional BOED is computationally expensive, due to the substantial costs required to both estimate and optimize $\text{EIG}(\xi)$ and update the model at each step.

2.3 Distribution Shift

Distribution shift is a well-known challenge in machine learning. It refers to the setting where the data distribution differs between the training and test phases. In BOED, training designs are selected via an acquisition criterion, while the model's performance at test-time is evaluated on a given test distribution of interest. This mismatch induces a specific form of distribution shift known as *covariate shift*, where the distribution of inputs shifts (i.e., $p_{\text{train}}(\xi) \neq p_{\text{test}}(\xi)$) while the conditional output distribution remains unchanged (i.e., $p_{\text{train}}(y \mid \xi) = p_{\text{test}}(y \mid \xi)$). Prior work has studied the covariate shift induced by BOED (Sugiyama, 2005; Ali et al., 2014; Sloman et al., 2022). To address the potential resultant biases, density ratio estimation — whereby the training data are reweighted according to the estimated ratio between test and training input distributions — has proven effective (Ge et al., 2023).

3 THEORETICAL RESULTS

3.1 Decomposition of Generalization Error

Recent work has demonstrated that generalization error depends on an interaction between the degree of covariate shift (the degree to which the training data are unrepresentative of the test distribution) and of model misspecification (Amortila et al., 2024; Ge et al., 2023; Wen et al., 2014). In this section, we show that generalization error additionally depends on the degree of presence of a phenomenon we term *error (de-)amplification*. We show that generalization error can be decomposed into three terms, reflecting separate contributions of the degree of misspecification bias, of estimation bias, and of error (de-)amplification.

Prior work (Hastie, 2009; Sugiyama, 2005) has provided decompositions of generalization error in the context of linear regression—where the error (de-)amplification term vanishes under the orthogonality assumption between model bias and estimation error. We extend such analyses to general (nonlinear) models under misspecification, where this orthogonality no longer holds.

Definition 2 (Generalization error (R_{test}))**.** *Let d_{test} be the test data distribution. The generalization error*

is defined as

$$R_{\text{test}}(\hat{f}^{(n)}) := \mathbb{E}_{\xi\sim d_{test}}\left[(\hat{f}^{(n)}(\xi) - f^\star(\xi))^2\right]. \quad (4)$$

Proposition 1 (Generalization Error Decomposition)**.** *Equation* (4) *can be decomposed into the following*

$$\begin{aligned} R_{\text{test}}(\hat{f}^{(n)}) = &\underbrace{\mathbb{E}_{\xi\sim d_{test}}\left[(\bar{f}(\xi) - f^\star(\xi))^2\right]}_{\textit{Misspecification Bias } (\mathbb{B})} \\ &+ \underbrace{\mathbb{E}_{\xi\sim d_{test}}\left[(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))^2\right]}_{\textit{Estimation Bias } (\mathbb{C})} \\ &+ \underbrace{2\,\mathbb{E}_{\xi\sim d_{test}}\left[(\bar{f}(\xi) - f^\star(\xi))(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))\right]}_{\textit{Error (de-)amplification } (\mathbb{A})}. \end{aligned} \quad (5)$$

In the *well-specified* case where the true function lies within the model class, $\bar{f}(\xi) = f^\star(\xi)$. In this case, both the bias and interaction terms vanish, and the generalization error reduces to:

$$R_{\text{test}}(\hat{f}^{(n)}) = \mathbb{E}_{\xi\sim d_{\text{test}}}\left[(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))^2\right], \quad (6)$$

where the only quantity that depends on the training sample is $\hat{f}^{(n)}(\xi)$, since the test data distribution d_{test} and the best predictor $\bar{f}$ are fixed.

In the *misspecified* case where the true function lies outside the model class, $\bar{f}(\xi) \neq f^\star(\xi)$. In this case, all three terms in Equation (5) contribute to generalization error. The terms have the following interpretations:

- **Misspecification Bias ($\mathbb{B}$)** captures the discrepancy between the best predictor and the true data-generating function, and reflects the degree of model misspecification. This term is fixed and unaffected by the training data.
- **Estimation Bias ($\mathbb{C}$)** captures the discrepancy between the best predictor $\bar{f} \in \mathcal{F}$ and the predictor arrived at on the basis of finite training data. In the BOED setting, this training data depends on the modeler's sequential evaluations of the EIG.
- **Error (De-)amplification ($\mathbb{A}$)** measures the correlation between the extent and direction of the model's bias and the extent and direction of the estimation error over the *test distribution*. This term can either amplify or mitigate the overall generalization error:
  - A positive correlation indicates that the directions of misspecification and estimation bias tend to coincide, which *amplifies* (increases) the generalization error.
  - A negative correlation indicates that the directions of misspecification and estimation bias tend to counteract each other, which *de-amplifies* (decreases) the generalization error.

3.2 An Upper Bound on Generalization Error with Error (De-)amplification

While Proposition 1 provides valuable insights into the various contributors to generalization error, computing these terms requires evaluating the outputs of $f^\star$ and $\bar{f}$ in expectation over the test samples. Of course, this is infeasible in practice.

To understand and control generalization error during training, the learner requires a formulation that explicitly relates it to quantities available during training. Theorem 1 provides such a formulation. Theorem 1 shows that the learner can control generalization error by selecting designs that (i) are *representative* of the test distribution (reduce the degree of covariate shift), and (ii) are *de-amplifying* (have the potential to counteract the misspecification bias). We apply these insights in our development of a novel acquisition function in Section 4.

Theorem 1 builds on a result from Amortila et al. (2024). In particular, we tighten the upper bound introduced by Amortila et al. (2024) to depend on the degree of (de-)amplification of the training samples.

We use d_{train} (resp., d_{test}) to refer to the training (resp., test) data distribution. Throughout, we assume that $d_{\text{train}}(\xi) > 0$ for all $\xi \in \Xi$, i.e., that each candidate design has some positive probability of being encountered during training.

We make use of the following definitions introduced by Amortila et al. (2024):

Definition 3 (The degree of covariate shift)**.** *The density ratio between the test and training input distributions is (Amortila et al., 2024; Sugiyama et al., 2007):*

$$\mathbb{C}_\infty := \sup_{\xi\in\Xi}\left|\frac{d_{test}(\xi)}{d_{train}(\xi)}\right|. \quad (7)$$

The degree of covariate shift $\mathbb{C}_\infty$ measures the worst-case distance between the selected training samples and the test samples. A more representative design (i.e., one that reduces covariate shift) helps control estimation bias.

Definition 4 (The degree of misspecification)**.** *The discrepancy between the predictive distribution induced by $\bar{f}$ and that of the true data-generating function $f^\star$ is (Amortila et al., 2024):*

$$\mathbb{B}_\infty := \|\bar{f} - f^\star\|_\infty = \sup_{\xi\in\Xi} |\ \bar{f}(\xi) - f^\star(\xi)\ | \quad (8)$$

where $\mathbb{B}_\infty \geq 0$.

The degree of misspecification $\mathbb{B}_\infty$ measures the worst-case discrepancy between the data-generating function and the best predictor in the model class. In other words, it is an upper bound on the misspecification bias $\mathbb{B}$. Notice that if $\mathbb{B}_\infty = 0$, the model is well-specified (i.e., $f^\star \in \mathcal{F}$). On the other hand, if $\mathbb{B}_\infty > 0$, the model is misspecified (i.e., $f^\star \notin \mathcal{F}$) and $\mathbb{B}_\infty$ quantifies the degree of misspecification.

We also require the following assumption:

Assumption 1 (Boundedness of model class and outcomes (Amortila et al., 2024)). *For all* $\xi \in \Xi$,

$$\sup_{f \in \mathcal{F}} \|f\|_\infty \leq y_\infty,\ \|f^\star\|_\infty \leq y_\infty,\ \textit{and}\ |y| \leq y_\infty$$

for some $0 < y_\infty < \infty$ *and where* $\|f\|_\infty = \sup_{\xi \in \Xi} \mid f(\xi) \mid$.

The finite setting ensures that the bound in Theorem 1 is non-vacuous.

Our Result. Theorem 1 extends the result of Amortila et al. (2024) by explicitly characterizing the behavior of generalization error in a way that accounts for error (de-)amplification. While Theorem 1 characterizes generalization performance given a data set (i.e., in the data-fitting phase), it can also inform data selection by revealing properties of those data that facilitate generalization. The insights from Theorem 1 motivate us to incorporate error (de-)amplification and representativeness into the decision-making (design selection) phase, thereby reducing the generalization error in the data-fitting phase, particularly in settings with a limited number of training samples.

Theorem 1 (Generalization Error Bound under Covariate Shift with Amplification). *Let* $\mathcal{F}$ *be a finite model class, and let* $f^\star$ *denote the true regression function. Let* $\hat{f}^{(n)}$ *be the empirical risk minimizer over training data drawn from the distribution* d_{train}. *Then, with probability at least* $1-\delta$, *the generalization error under covariate shift satisfies:*

$$R_{test}(\hat{f}^{(n)}) \leq \mathbb{C}_\infty \cdot \begin{cases} \mathbb{B}_\infty^2 + \dfrac{224 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - 2\widehat{\mathbb{A}}(\hat{f}^{(n)}), \\ \qquad\qquad \textit{if}\ \widehat{\mathbb{A}}(\hat{f}^{(n)}) < 0 \\ \mathbb{B}_\infty^2 + \dfrac{128 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - \sqrt{3}\widehat{\mathbb{A}}(\hat{f}^{(n)}), \\ \qquad\qquad \textit{if}\ 0 \leq \widehat{\mathbb{A}}(\hat{f}^{(n)}) \end{cases} \tag{9}$$

where $\widehat{\mathbb{A}}(\hat{f}^{(n)}) \coloneqq \mathbb{E}_{\xi \sim d_{train}}\left[(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi))\right]$. *The proof can be found in Appendix B.*

Remark 1 (Connection to Proposition 1). *This bound is consistent with the full generalization error decomposition structure, including the cross-term* $\widehat{\mathbb{A}}(f)$, *which captures the interaction between properties of the model (misspecification bias) and of the sampling strategy (estimation bias). Proposition 1 does not explicitly reveal how the representativeness and (de-)amplifying properties of the training data interact with model bias. In contrast, Theorem 1 makes this interaction explicit, providing a more interpretable perspective for understanding and controlling the generalization error during the training phase.*

Remark 2. *Proposition 1 defines the generalization error as a function of the learned function* $\hat{f}^{(n)}$ *and the true data-generating function* $f^\star$, *in expectation over the test distribution. However, since the true data-generating function* $f^\star$ *is unknown and outside the model class, it cannot easily offer guidance for decision-making. In contrast, although the* $\widehat{\mathbb{A}}$ *term that appears in Theorem 1 depends on the unknown* $\bar{f}$, $\bar{f}$ *is within the model class* $\mathcal{F}$, *which suggests that it can be better approximated in practice. We leverage this in our construction of a novel acquisition function in Section 4.2.*

Remark 3. *Theorem 1 requires the assumption that* $\hat{f}^{(n)}$ *is the member of* $\mathcal{F}$ *that minimizes risk in the training data. Although our experiments adopt the Bayesian learning framework, in which the learner predicts on the basis of a distribution over members of* $\mathcal{F}$, *Theorem 1 still provides valuable insights about the role of representative and (de-)amplifying training samples.*

Theorem 1 reveals that the following factors contribute to generalization error:

- **Representativeness of the Training Data ($\mathbb{C}_\infty$).** The multiplicative factor of $\mathbb{C}_\infty$ implies that, by reducing $\mathbb{C}_\infty$, choosing more representative training data can reduce generalization error.
- **Misspecification Bias ($\mathbb{B}_\infty$).** Misspecification bias cannot be reduced by the training data. However, because of the multiplicative effect of $\mathbb{C}_\infty$, the effect of model misspecification on generalization error can be amplified by unrepresentative training samples (Amortila et al., 2024).
- **Error (De-)amplification ($\widehat{\mathbb{A}}$).** This term captures a key component of generalization error: the interaction between the learner's misspecification and estimation errors. The term $\widehat{\mathbb{A}}(\hat{f}^{(n)})$ can be interpreted as an average, across the training samples, of the (signed) estimation errors weighted by the (signed) misspecification errors. Where $\widehat{\mathbb{A}}(\hat{f}^{(n)})$ is large (error de-amplification), the learner's misspecification and estimation biases tend to agree, and so sampling at the given design (reducing the estimation error) has a *de-amplifying* effect.

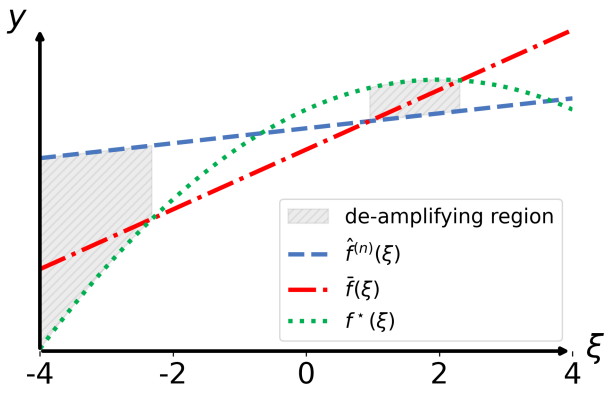


Figure 1: Illustration of error amplification and de-amplification. The green curve denotes the true data-generating function $f^\star(\xi)$, the red line represents the best-in-class approximation $\bar{f}(\xi)$, and the blue line shows the learned predictor $\hat{f}^{(n)}(\xi)$. The grey dashed shading highlights *de-amplifying* regions, where $\bar{f}(\xi)$ lies between $f^\star(\xi)$ and $\hat{f}^{(n)}(\xi)$, so that estimation error and misspecification error partially offset each other. In contrast, the remaining regions indicate *amplifying* regions, where estimation error reinforces misspecification, leading to larger overall prediction error.

For any $\xi \in \Xi$, the misspecification error $\bar{f}(\xi) - f^\star(\xi)$ is fixed and cannot be removed by additional data. In selecting the de-amplifying design, the objective is to induce a negative correlation between the estimation error $\hat{f}(\xi) - \bar{f}(\xi)$ and the misspecification error in regions where the misspecification error is large. In other words, when the misspecification error is large, the estimation error should not exacerbate it; ideally, the estimation error offsets (de-amplifies) the misspecification. Figure 1 illustrates the (de-)amplifying regions in a simple example.

4 TWO NOVEL ACQUISITION FUNCTIONS

Leveraging the insights from Theorem 1, we design two novel acquisition functions. R-I identifies designs that are both *representative* and *informative* about the parameter of interest. R-IDeA identifies designs that are *representative*, *informative*, and *de-amplifying*.

4.1 R-I: A Representative and Informative Acquisition Function

To account for covariate shift, we modify the standard EIG acquisition function by introducing a maximum mean discrepancy (MMD)-based correction term. The idea is to encourage the selection of design points that not only have high information gain but also help reduce the difference between the distributions of training and test points. Specifically, we use the following form:

$$\text{R-I}(\xi_t) = \text{EIG}(\xi_t) \cdot \underbrace{\left(1 - \lambda \frac{\text{MMD}(h_{t-1} \cup \{\xi_t\}, d_{\text{test}})}{\text{MMD}(h_{t-1}, d_{\text{test}})}\right)}_{\text{Robust Ratio}}) \tag{10}$$

where h_{t-1} is the history of selected designs before time step t. The motivation and expression for MMD can be found in Appendix D.

This robust acquisition function penalizes designs that are only representative or only informative; in other words, the designs selected by R-I are both representative and informative. The hyperparameter λ controls the tradeoff between informativeness and representativeness. When λ tends to zero, the selected designs are informative, and R-I selects similar designs to traditional BOED.

4.2 R-IDeA: A Representative, Informative, and De-amplifying Acquisition Function

Theorem 1 shows that larger $\widehat{\mathbb{A}}(\hat{f}^{(n)})$ implies that designs on which $\hat{f}^{(n)}$ was trained *de-amplify* generalization error. We refer to a design ξ as de-amplifying whenever $(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0$ for a given threshold τ_0. Ideally, we would design an acquisition function that selects only designs in the de-amplifying region $\Xi_{\mathbb{A}+}(\tau_0) := \left\{\xi \in \Xi : (\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0\right\}$. However, as discussed in Remark 2, this cannot be determined exactly since $f^\star$ and $\bar{f}$ are unknown. We instead construct a subset of the de-amplifying region, the *approximate de-amplifying region*:

Theorem 2 (Approximate de-amplifying region)**.** *Let $\bar{f}$ be the predictor that best approximates the true data-generating function $f^\star$. Then,*

$$\widehat{\Xi}_{\mathbb{A}+}(\tau_1) \subseteq \Xi_{\mathbb{A}+}(\tau_0)$$

where the approximate de-amplifying region $\widehat{\Xi}_{\mathbb{A}+}(\tau_1) := \left\{\xi \in \Xi : |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq \tau_1\right\}$, $\tau_1 = \tau_0/\mathbb{B}_\infty + c\mathbb{B}_\infty$, $c \geq 2$, *and* $\tau_0 \geq 0$. *A detailed derivation can be found in Appendix C.*

τ_0 and c are constants that determine the minimum threshold required for a design to be considered part of the approximate de-amplifying region.

Unlike $\Xi_{\mathbb{A}+}$, $\widehat{\Xi}_{\mathbb{A}+}$ does not depend on $f^\star$. However, it does depend on $\bar{f}$. As discussed in Remark 2, $\bar{f}$ is within the model class $\mathcal{F}$. Below, we leverage this in

construction of the *proxy approximate de-amplification region*, which depends on a trainable proxy g:

Lemma 1 (Proxy approximate de-amplification region $\widehat{\Xi}_{\mathbb{A}+}(\tau)$)**.** *Given a* proxy function $g : \Xi \mapsto \mathbb{R}$ *such that* $\sup_{\xi\in\Xi} |g(\xi) - \bar{f}(\xi)| \leq \tau_2$ *for some* $\tau_2 \geq 0$*,*

$$\widehat{\Xi}^{g}_{\mathbb{A}+}(\tau) \coloneqq \big\{\xi \in \Xi : |\hat{f}^{(n)}(\xi) - g(\xi)| \geq \tau\big\} \subseteq \widehat{\Xi}_{\mathbb{A}+}(\tau_1).$$

where $\tau = \tau_1 + \tau_2$*. The proof can be found in Appendix C.*

Heuristically, a good proxy function g (i) is aligned with $\bar{f}$, and (ii) maintains disagreement with $\hat{f}^{(n)}$ thereby ensuring the proxy approximate de-amplification region is sufficiently large (notice that in the extreme case where $g = \hat{f}^{(n)}$, $\widehat{\Xi}^{g}_{\mathbb{A}+}$ is empty). At time step t, g is trained to (i) fit the observations collected up until the previous time step $\{y_i : i \in \{1 \dots t-1\}\}$, and (ii) maintain disagreement with $\hat{f}^{(n)}$. This leads to the following objective:

$$g = \arg\min_{g} \mathcal{L}(g) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(g(\xi_i) - y_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\max\big(0,\ \tau - |\hat{f}^{(n)}(\xi_i) - g(\xi_i)|\big). \tag{11}$$

Leveraging the *proxy approximate de-amplifying region* and learned g, we introduce an acquisition function for the decision-making phase that balances de-amplification with informativeness and representativeness. The novel acquisition function is given by

$$\text{R-IDeA}(\xi_t) = \text{R-I}(\xi_t)\,\text{DeA}(\xi_t),$$
$$\text{where} \quad \text{DeA}(\xi_t) = \text{Sigmoid}\left(\frac{|\hat{f}^{(n)}(\xi_t) - g(\xi_t)| - \tau}{\kappa}\right), \tag{12}$$

where τ is as defined in Theorem 2 and further details of DeA derivation are provided in Appendix C. R-IDeA requires two hyperparameters: τ corresponds to the de-amplifying threshold, i.e., to the minimum separation level required for a design to be considered in the proxy de-amplifying region, and κ corresponds to the optional smoothing parameter[1]. The hyperparameter τ controls the tradeoff between de-amplification and the requirements of informativeness and representativeness: When τ is large, the proxy approximate de-amplifying region is conservative in the sense that it includes only designs with very high $\widehat{\mathbb{A}}$; DeA's downweighting of most designs reflects their effective exclusion from this region.

More generally, Equation (12) can be framed as a framework, or family, of acquisition functions. While we specified R-I using the EIG and robust ratio shown in Equation (10) to measure the degrees of informativeness and representativeness, respectively, of a design, one could easily substitute these terms with problem- or application-specific measures.

[1]We fixed $\kappa = 1$ in our experiments.

5 EXPERIMENTS

This section contains comparative experiments and analysis to explore which algorithm performs best in the presence of model misspecification in three experimental paradigms: a polynomial regression experiment, a source location paradigm, and a pharmacokinetic setting. We also empirically validate the theoretical results of Section 3. The code to reproduce our experiments is available at https://github.com/TrbingWY/robust-boed.git

We compare the following methods: A **Random** strategy selects designs from the test distribution at random. **Bayesian adaptive design (BAD)** (Foster et al., 2019) selects designs according to the traditional BOED strategy, i.e., according to the EIG (Equation (2)). **Representative and informative BAD (R-I)** selects designs according to our novel representative acquisition function (Section 4.1). **Representative, informative and de-amplifying BAD (R-IDeA)** selects designs according to our novel de-amplifying acquisition function (Section 4.2). To explore how the hyperparameters affect our proposed acquisition function, we conduct experiments with different values of λ in R-I and τ in R-IDeA. These results are given in Appendix E.

The generalization performance of each method is evaluated using the Mean Squared Error (MSE), while the degree of covariate shift is measured by the Maximum Mean Discrepancy (MMD). Further details are provided in Appendix D.

5.1 Polynomial Regression Experiments

In the **polynomial regression setting**, the DGP is a degree-two polynomial regression model, $y = 1 + 2x - 0.5x^2 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.1)$. In the *misspecified case*, we use a linear model to fit the data this model generates. In the *well-specified case*, we use a quadratic model to fit the data this model generates. More implementational details are given in Appendix D.1, and more experimental results (one of which is in the well-specified case) can be found in Appendix E.1. In particular, we also explore the performance of our proposed acquisition functions under different misspecification degrees in Appendices E.1.3 and E.1.4.

Figure 2 shows that under model misspecification, R-I

outperforms both BAD and Random, indicating that incorporating informativeness and representativeness leads to more effective design selection. R-IDeA further improves generalization performance and achieves the best results overall, demonstrating that jointly accounting for informativeness, representativeness, and de-amplification is most effective. Finally, we compare the performance of R-IDeA and R-IDeA-oracle, which uses the true $\bar{f}$ instead of the proxy g. R-IDeA performs comparably to R-IDeA-oracle, implying that R-IDeA is effective even while relying on the proxy g to approximate $\bar{f}$.

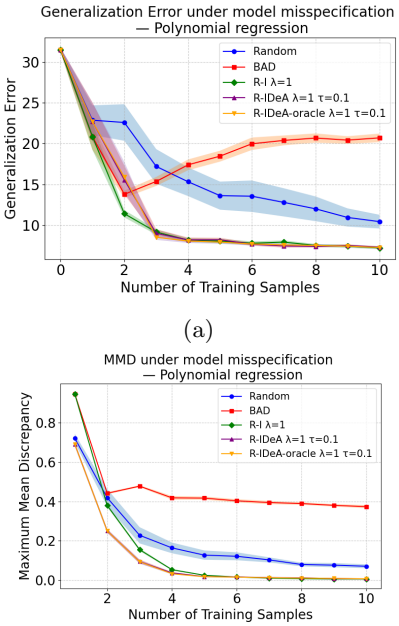


Figure 2: **Polynomial regression experiments (misspecified case).** Comparison of different design strategies (Random, BAD, proposed R-I, proposed R-IDeA, and R-IDeA-oracle, which uses the true $\bar{f}$ instead of the proxy g) under misspecified models in the polynomial regression experiments. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

5.2 Source Localization Experiments

The **acoustic energy attenuation model** simulates the total intensity at location ξ of a signal emitted from multiple sources at locations $\theta = \{\theta_k\}_{k=1}^K$. The objective of the design problem is to strategically select points at which to observe the total signal to infer the locations of the source effectively. More implementational details can be found in Appendix D.2, and more experimental results are in Appendix E.2.

Interestingly, under model misspecification, we observe that the generalization error of BAD and R-I slightly increases (Figure 3a) while the degree of covariate shift these methods induce decreases (Figure 3b). We speculate that this discrepancy is due to the amplifying properties of the designs selected by these methods. In the presence of model misspecification, R-I both outperforms BAD and induces less covariate shift, highlighting the effectiveness of representative designs.

R-IDeA performs better than other methods and induces the highest covariate shift. This implies that covariate shift is not the only factor influencing generalization error, and that R-IDeA may perform better than other methods due to its selection of de-amplifying training data.

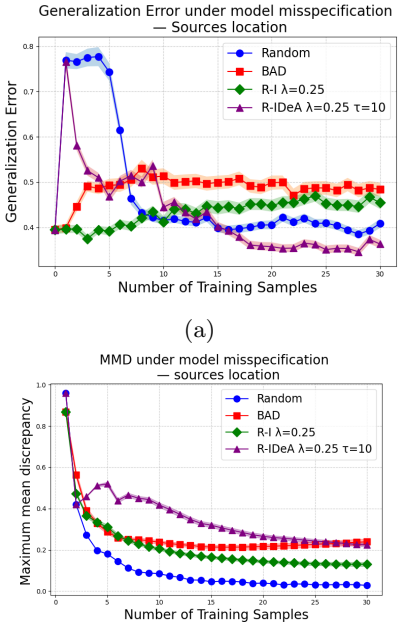


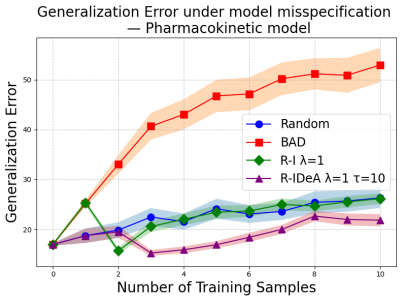
Figure 3: **Source localization experiments (misspecified case).** Comparison of baseline methods (Random, BAD) and our proposed R-I and R-IDeA in the source localization experiments. *Top*: Generalization error across methods. *Bottom*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

5.3 Pharmacokinetic Experiments

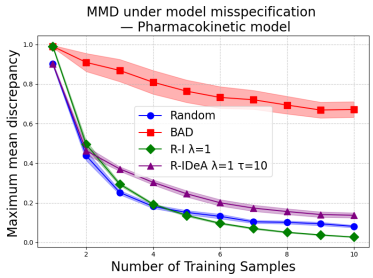
According to the **pharmacokinetic model**, the distribution of an administered drug in the body is determined by three key parameters: the absorption rate k_α, the elimination rate k_e, and the volume V. These define the parameter vector of interest, $\theta = (k_\alpha, k_e, V)$. The design task is to adaptively select blood sampling times $0 \leq \xi_t \leq 24$ hours for each patient, measured

from the moment of drug administration (with patient 2 receiving the drug only after collecting a sample from patient 1, and so on). More implementational details can be found in Appendix D.3, and more experimental results can be found in Appendix E.3.

Figure 4a shows that, under model misspecification, R-IDeA exhibits the best performance, suggesting that the selection of de-amplifying and representative designs can help reduce generalization error, consistent with the theoretical result established in Theorem 1. Figure 4b illustrates that in addition to exhibiting the best generalization performance, R-IDeA induces the largest degree of covariate shift, further demonstrating that designs that are de-amplifying in addition to representative contribute to robustness under model misspecification.



(a)



(b)

Figure 4: **Pharmacokinetic model experiments (misspecified case).** Comparison of baseline methods (Random, BAD) and our proposed R-I and R-IDeA in the Pharmacokinetic model experiments *Top*: Generalization error across methods. *Bottom*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

6 CONCLUSION

When models are correctly specified, powerful experimental designs are informative about a parameter of interest. When models are misspecified, effective experimental designs are informative and robust to the misspecification. Our detailed analysis unpacks what is required for robustness. Our analysis reveals that robustness is a function of designs' representativeness of the test distribution and de-amplification of misspecification errors. We leverage these insights to propose a novel method for BOED in the presence of potential model misspecification. Our empirical results demonstrate the effectiveness of the proposed method.

Limitations and Future Work Our proposed method is informed by insights from Theorem 1, which provides an upper bound on generalization performance. The degree to which Theorem 1 reflects actual generalization performance depends on the tightness of this bound. Assessing the tightness of this bound is therefore an important direction for future work. Moreover, additional empirical results in Appendix E suggest that the properties of de-amplification, representativeness, and informativeness are not independent. Consequently, tuning a hyperparameter to control one property will implicitly affect the others. This highlights the importance of selecting hyperparameters automatically and appropriately, rather than simply increasing or decreasing their values in a heuristic manner. Developing principled methods for automatic hyperparameter selection is an important direction for future work.

Acknowledgements

The authors thank Zhang Wan and Xiaomei Mi for their helpful discussion. This work was supported by EU grant (101120237 and ERC ODD-ML 101201120) and the Research Council of Finland Flagship programme: Finnish Center for Artificial Intelligence FCAI and decisions 359207, 359567, 358958. SJS and SK were supported by the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1].

References

Alnur Ali, Rich Caruana, and Ashish Kapoor. Active learning with model selection. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 130–160. PMLR, 2024.

Francis Bach. Active learning for misspecified generalized linear models. Technical Report N15/06/MM, Ecole des mines de Paris, 2006.

Yasir Zubayr Barlas, Sabina J Sloman, and Samuel Kaski. Robust experimental design via generalised bayesian inference. *arXiv preprint arXiv:2511.07671*, 2025.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Ayush Bharti, Masha Naslidnyk, Oscar Key, Samuel Kaski, and François-Xavier Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, pages 2289–2312. PMLR, 2023.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

Stéphane Bonhomme and Martin Weidner. Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3):907–954, 2022.

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2286841`.

George EP Box. Sampling and bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 143 (4):383–404, 1980.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal bayesian deep learning. *arXiv e-prints*, pages arXiv–2302, 2023.

Tommie A Catanach and Niladri Das. Metrics for bayesian optimal experiment design under model misspecification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7707–7714. IEEE, 2023.

Daniel R Cavagnaro, Jay I Myung, Mark A Pitt, and Janne V Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, 22(4):887–905, 2010.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

Jiayuan Dong, Christian Jacobsen, Mehdi Khalloufi, Maryam Akram, Wanjiao Liu, Karthik Duraisamy, and Xun Huan. Variational bayesian optimal experimental design with normalizing flows. *arXiv preprint arXiv:2404.13056*, 2024.

Marina Dubova, Suyog Chandramouli, Gerd Gigerenzer, Peter Grünwald, William Holmes, Tania Lombrozo, Marco Marelli, Sebastian Musslick, Bruno Nicenboim, Lauren N. Ross, Richard Shiffrin, Martha White, Eric-Jan Wagenmakers, Paul-Christian Bürkner, and Sabina J. Sloman. Is ockham's razor losing its edge? new perspectives on the principle of model parsimony. *PNAS*, 2025.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

Alexander J Forster, Desi R Ivanova, and Tom Rainforth. Improving robustness to model misspecification in bayesian experimental design. In *Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference*, 2025.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32, 2019.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *International conference on machine learning*, pages 3384–3395. PMLR, 2021.

David T Frazier, Christian P Robert, and Judith Rousseau. Model misspecification in approximate bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):421–444, 2020.

David T Frazier, Robert Kohn, Christopher Drovandi, and David Gunawan. Reliable bayesian inference in misspecified models. *arXiv preprint arXiv:2302.06031*, 2023.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you

need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Peter Grünwald and Thijs Van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. 2017.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

Riley J Hickman, Matteo Aldeghi, Florian Häse, and Alán Aspuru-Guzik. Bayesian optimization with known experimental and design constraints for chemistry applications. *Digital Discovery*, 1(5):732–744, 2022.

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024.

Daolang Huang, Ayush Bharti, Amauri Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. *Advances in Neural Information Processing Systems*, 36:7289–7310, 2023.

Desi R Ivanova, Marcel Hedman, Cong Guan, and Tom Rainforth. Step-dad: Semi-amortized policy-based bayesian experimental design. 2024.

Ryan P Kelly, David J Warne, David T Frazier, David J Nott, Michael U Gutmann, and Christopher Drovandi. Simulation-based bayesian inference under model misspecification. *arXiv preprint arXiv:2503.12315*, 2025.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, 34(2), 2006.

B.J.K. Kleijn and A.W. van der Vaart. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6, 2012.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23, 2022.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. *Advances in neural information processing systems*, 30, 2017.

Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *The FEBS journal*, 276(4): 923–942, 2009.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1): e66–e82, 2017.

Ulrich K Müller. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

Jay I Myung, Daniel R Cavagnaro, and Mark A Pitt. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67, 2013.

David J Nott, Christopher Drovandi, and David T Frazier. Bayesian inference for misspecified generative models. *Annual Review of Statistics and Its Application*, 11, 2023.

Klaus Oberauer, Frederik Aust, and Philipp Musfeld. Variance, bias, and computational cost of estimating the bayes factor using bridge sampling and the savage-dickey density ratio. 2025.

Antony Overstall and James McGree. Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Analysis*, 17(4):1021–1041, 2022.

Antony M Overstall, Jacinta Holloway-Brown, and James M McGree. Gibbs optimal design of experiments. *arXiv preprint arXiv:2310.17440*, 2025.

Mijung Park, Marcel Nassar, and Haris Vikalo. Bayesian active learning for drug combinations. *IEEE transactions on biomedical engineering*, 60(11): 3248–3255, 2013.

A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Detecting model misspecification in amortized bayesian inference with neural networks. In *DAGM German Conference on Pattern Recognition*, pages 541–557. Springer, 2023.

Paola Sebastiani and Henry P Wynn. Bayesian experimental design and shannon information. In *Proceedings of the Section on Bayesian Statistical Science*, volume 44, pages 176–181. The Association, 1997.

Burr Settles. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*, 2009.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Mel Silberman. *Active learning: 101 strategies to teach any subject.* ERIC, 1996.

Sabina J Sloman, Daniel M Oppenheimer, Stephen B Broomell, and Cosma Rohilla Shalizi. Characterizing the robustness of bayesian adaptive experimental designs to active learning bias. *arXiv preprint arXiv:2205.13698*, 2022.

Masashi Sugiyama. Active learning for misspecified models. *Advances in neural information processing systems*, 18, 2005.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Jaron C Thompson, Victor M Zavala, and Ophelia S Venturelli. Integrating a tailored recurrent neural network with bayesian experimental design to optimize microbial community functions. *PLOS Computational Biology*, 19(9):e1011436, 2023.

Raman Uppal and Tan Wang. Model misspecification and underdiversification. *The Journal of Finance*, 58(6):2465–2486, 2003.

Thomas Viehmann. Partial wasserstein and maximum mean discrepancy distances for bridging the gap between outlier detection and drift detection. *arXiv preprint arXiv:2106.12893*, 2021.

Benjamin T Vincent and Tom Rainforth. The darc toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design. *PsyArXiv. October*, 20, 2017.

Eric A Walker and Kishore Ravisankar. Bayesian design of experiments: Implementation, validation and application to chemical kinetics. *arXiv preprint arXiv:1909.03861*, 2019.

Stephen G Walker. Bayesian inference with misspecified models. *Journal of statistical planning and inference*, 143(10):1621–1633, 2013.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR, 2014.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PROOF of Proposition 1

$$
\begin{aligned}
R_{\text{test}}(f) &:= \mathbb{E}_{\xi\sim d_{\text{test}}}\left[(f(\xi)-f^\star(\xi))^2\right] \\
&= \mathbb{E}_{\xi\sim d_{\text{test}}}\left(\left(\left[f(\xi)-\bar{f}(\xi)\right]+\left[\bar{f}(\xi)-f^\star(\xi)\right]\right)^2\right) \\
&= \underbrace{\mathbb{E}_{\xi\sim d_{\text{test}}}\left[(\bar{f}(\xi)-f^\star(\xi))^2\right]}_{\text{Misspecification Bias}} + \underbrace{\mathbb{E}_{\xi\sim d_{\text{test}}}\left[(f(\xi)-\bar{f}(\xi))^2\right]}_{\text{Estimation Bias}} \\
&\quad + \underbrace{2\,\mathbb{E}_{\xi\sim d_{\text{test}}}\left[(\bar{f}(\xi)-f^\star(\xi))(f(\xi)-\bar{f}(\xi))\right]}_{\text{Error (De-)amplification}}.
\end{aligned}
\tag{13}
$$

B PROOF of Theorem 1

Theorem: with probability at least $1-\delta$, the generalization error under covariate shift satisfies:

$$
R_{\text{test}}(\hat{f}^{(n)}) \le \mathbb{C}_\infty \cdot \begin{cases} \mathbb{B}^2 + \dfrac{224 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - 2\widehat{\mathbb{A}}(\hat{f}^{(n)}), & \text{if } \widehat{\mathbb{A}} < 0 \\ \mathbb{B}^2 + \dfrac{128 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - \sqrt{3}\widehat{\mathbb{A}}(\hat{f}^{(n)}), & \text{if } \widehat{\mathbb{A}} \ge 0 \end{cases}
\tag{14}
$$

Proof of Theorem We provide the full derivation of the generalization error bound that preserves the error (de-)amplification term. Our analysis, which handles misspecification, is adapted from the proof of Proposition 2.1 in (Amortila et al., 2024). In the proof, we first analyse the empirical error considering the training data distribution d_{train}, and then use the bounded density ratio to analyse the generalization error in the testing data distribution d_{test}.

The goal is to bound the generalization risk

$$
R(f) := \mathbb{E}_\xi[(f(\xi)-f^\star(\xi))^2],
$$

for any $f \in \mathcal{F}$. And the empirical risk is defined as:

$$
\widehat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\xi_i\right)-y_i\right)^2,
$$

Observe that conditional on any ξ we have:

$$
\begin{aligned}
&\widehat{R}(f)-\widehat{R}(\bar{f}) \\
&= \mathbb{E}[(f\ (\xi)-y)^2-(\bar{f}(\xi)-y)^2 \mid \xi] \\
&= \mathbb{E}\left[(f(\xi)-y)^2-\left(\bar{f}(\xi)-f^\star(\xi)+f^\star(\xi)-y\right)^2 \mid \xi\right] \\
&= \mathbb{E}\left[(f(\xi)-y)^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2-2\left(\bar{f}(\xi)-f^\star(\xi)\right)\left(f^\star(\xi)-y\right)-\left(f^\star(\xi)-y\right)^2 \mid \xi\right] \\
&= \mathbb{E}\left[(f(\xi)-y)^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2-\left(f^\star(\xi)-y\right)^2 \mid \xi\right] \\
&= f(\xi)^2-f^\star(\xi)^2-2\mathbb{E}_{\text{train}}\left[y \mid \xi\right]\left(f(\xi)-f^\star(\xi)\right)-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2 \\
&= \left(f(\xi)-f^\star(\xi)\right)^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2 \quad (\text{since} \quad \mathbb{E}_{\text{train}}\left[y \mid \xi\right]=f^\star(\xi))
\end{aligned}
\tag{15}
$$

thus

$$\mathbb{E}(\widehat{R}(f)-\widehat{R}(\bar{f}))=\mathbb{E}((f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2)=R(f)-R(\bar{f})$$

And

$$\begin{aligned}
&\mathrm{Var}\left[(f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right]\\
&=\mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)^2\right]-\left(\mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)\right]\right)^2 \quad (\text{since} \quad \mathrm{Var}(X)=\mathbb{E}(X^2)-(\mathbb{E}X)^2)\\
&\le\mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)^2\right]\\
&=\mathbb{E}\left[\left(f^2(\xi)-2f(\xi)y-\bar{f}^2(\xi)+2y\bar{f}(\xi)\right)^2\right]\\
&=\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))^2(f(\xi)+\bar{f}(\xi)-2y)^2\right]\\
&\le 16y_\infty^2\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))^2\right] \quad (\text{since } |f(\xi)|,|\bar{f}(\xi)|,|y|\le y_\infty)\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi)-(\bar{f}(\xi)-f^\star(\xi)))^2\right]\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2+\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]-32y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))(\bar{f}(\xi)-f^\star(\xi))\right]\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2+2\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]-32y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))(\bar{f}(\xi)-f^\star(\xi))\right]\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2+2\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]\\
&\quad-32y_\infty^2\mathbb{E}\left[(f(\xi)-\bar{f}(\xi)+\bar{f}(\xi)-f^\star(\xi))(\bar{f}(\xi)-f^\star(\xi))\right]\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]+32y_\infty^2\mathbb{E}\left(\bar{f}(\xi)-f^\star(\xi)^2\right)\\
&\quad-32y_\infty^2\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))(\bar{f}(\xi)-f^\star(\xi))\right]-32y_\infty^2\mathbb{E}[(\bar{f}(\xi)-f^\star(\xi))^2]\\
&\le 16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]-32y_\infty^2\underbrace{\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))(\bar{f}(\xi)-f^\star(\xi))\right]}_{\widehat{\mathbb{A}}(f)}\\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]-32y_\infty^2\widehat{\mathbb{A}}(f)\\
&=16y_\infty^2(R(f)-R(\bar{f}))-32y_\infty^2\widehat{\mathbb{A}}(f)
\end{aligned}\tag{16}$$

Application of Bernsteins Inequality and Union Bound Based on Bernsteins inequality (Shalev-Shwartz and Ben-David, 2014) and union bound (see Lemma 2.2 in (Shalev-Shwartz and Ben-David, 2014)),

Bernsteins inequality becomes:

$$\mathbf{P}\left(\mathbf{E}Z^{(f)}-\frac{1}{n}\sum_{i=1}^n Z_i^{(f)}<\frac{c}{3n}\log(|\mathcal{F}|/\delta)+\sqrt{\frac{2(\mathrm{Var}\,Z^{(f)})\log(|\mathcal{F}|/\delta)}{n}}\right)\ge 1-\delta$$

Setting $Z_i^{(f)}=(f(\xi)-y_i)^2-(\bar{f}(\xi)-y_i)^2$, thus $\frac{1}{n}\sum_{i=1}^n Z_i^{(f)}=\widehat{R}(f)-\widehat{R}(\bar{f})$; $\mathbb{E}Z^{(f)}=R(f)-R(\bar{f})$, $|Z^{(f)}-EZ^{(f)}|\le 2\sup|Z^{(f)}|\le 2\sup(f(\xi)-y_i)^2-(\bar{f}(\xi)-y_i)^2=2\sup(f(\xi)-y_i+\bar{f}(\xi)-y_i)(f(\xi)-\bar{f}(\xi))=2\cdot(4y_\infty)(2y_\infty)=c$, thus $c=16y_\infty^2$

Now, by using Bernstein's inequality and a union bound over $f\in\mathcal{F}$ that with probability at least $1-\delta$

$$\forall f\in\mathcal{F}: R(f)-R(\bar{f})-(\widehat{R}(f)-\widehat{R}(\bar{f}))\le\sqrt{\frac{\left(16(R(f)-R(\bar{f}))-32\widehat{\mathbb{A}}(f)\right)2y_\infty^2\log(|\mathcal{F}|/\delta)}{n}}+\frac{16y_\infty^2\log(|\mathcal{F}|/\delta)}{3n}.\tag{17}$$

Supppose the

$$\widehat{R}\left(\hat{f}^{(n)}\right)-\widehat{R}(\bar{f})\le 0\tag{18}$$

we have

$$R\left(\hat{f}^{(n)}\right) - R(\bar{f}) \leq \sqrt{\frac{\left(16\left(R\left(\hat{f}^{(n)}\right) - R(\bar{f})\right) - 32\widehat{\mathbb{A}}(\hat{f}^{(n)})\right) 2y_\infty^2 \log(|\mathcal{F}|/\delta)}{n}} + \frac{16y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} \tag{19}$$

To simplify, make

$$\begin{aligned} A &:= R(\hat{f}^{(n)}) - R(\bar{f}) \\ C &:= y_\infty^2 \frac{\log(|\mathcal{F}|/\delta)}{n} > 0 \end{aligned} \tag{20}$$

Why $(16A - 32\widehat{\mathbb{A}}) \geq 0$

$$\begin{aligned} A := R(\hat{f}^{(n)}) - R(\bar{f}) &= \mathbb{E}_\xi[(\hat{f}^{(n)}(\xi) - f^\star(\xi))^2] - \mathbb{E}_\xi[(\bar{f}(\xi) - f^\star(\xi))^2] \\ &= \mathbb{E}_\xi[(\hat{f}^{(n)}(\xi) - f^\star(\xi) + \bar{f}(\xi) - f^\star(\xi))(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))] \end{aligned} \tag{21}$$

Then

$$\begin{aligned} 1/2A - \widehat{\mathbb{A}} &= 1/2\mathbb{E}_\xi[(\hat{f}^{(n)}(\xi) - f^\star(\xi) + \bar{f}(\xi) - f^\star(\xi))(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))] - \mathbb{E}_\xi\left[(\hat{f}^{(n)} - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi))\right] \\ &= \mathbb{E}_\xi[(1/2\hat{f}^{(n)}(\xi) - 1/2f^\star(\xi) + 1/2\bar{f}(\xi) - 1/2f^\star(\xi) - \bar{f}(\xi) + f^\star(\xi))(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))] \\ &= \mathbb{E}_\xi[(1/2\hat{f}^{(n)}(\xi) - 1/2\bar{f}(\xi))(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))] \geq 0 \end{aligned} \tag{22}$$

B.1 Suppose that $\widehat{\mathbb{A}} < 0$

So the above inequality function equals:

$$\begin{aligned} A &\leq \sqrt{(16A - 32\widehat{\mathbb{A}})2C} + \frac{16C}{3} \\ &\leq \sqrt{32AC} + \sqrt{-64\widehat{\mathbb{A}}C} + \frac{16C}{3} \quad (\text{since}\sqrt{a+b} < \sqrt{a} + \sqrt{b}) \\ &= \sqrt{A \cdot 32C} + \sqrt{-2\widehat{\mathbb{A}} \cdot 32C} + \frac{16C}{3} \\ &\leq \frac{1}{2}A + 16C - \widehat{\mathbb{A}} + 16C + \frac{16C}{3} \quad (\sqrt{ab} \leq a/2 + b/2) \\ &= \frac{1}{2}A - \widehat{\mathbb{A}} + \frac{112C}{3} \end{aligned} \tag{23}$$

so that

$$A \leq -2\widehat{\mathbb{A}} + \frac{224C}{3} \tag{24}$$

So

$$R\left(\hat{f}^{(n)}\right) - R(\bar{f}) \leq -2\widehat{\mathbb{A}}(\hat{f}^{(n)}) + \frac{224y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} \tag{25}$$

Re-arranging and using that $R_{\text{train}}(\bar{f}) \leq \mathbb{B}^2$, thus we have

$$R_{\text{train}}\left(\hat{f}^{(n)}\right) \leq \mathbb{B}^2 - 2\widehat{\mathbb{A}}(\hat{f}^{(n)}) + \frac{224y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} \tag{26}$$

Finally, let $\mathbb{C}_\infty := \sup_{\xi \in \Xi} \left|\frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)}\right|$, the upper bound of generalization error can be expressed via the density ratio:

$$R_{\text{test}}\left(\hat{f}^{(n)}\right) = \mathbb{E}_{\text{train}}\left[\frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)}\left(\hat{f}^{(n)}(\xi) - f^*(\xi)\right)^2\right] \leq \mathbb{C}_\infty \cdot \left(\mathbb{B}^2 - 2\widehat{\mathbb{A}}(\hat{f}^{(n)}) + \frac{224y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n}\right) \tag{27}$$

where $\widehat{\mathbb{A}}(\hat{f}^{(n)}) = \mathbb{E}_{\text{train}}\left[(\hat{f}^{(n)} - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi))\right]$

B.2 Suppose that $\widehat{\mathbb{A}} \geq 0$

Although the second solution, based on solving a quadratic equation, yields a tighter upper bound, it involves more complex derivations. In contrast, the first solution offers a simpler and more interpretable decomposition, while preserving the same theoretical insights into the behavior of the error (de-)amplification term $\widehat{\mathbb{A}}(\hat{f}^{(n)})$. Therefore, we adopt the first formulation in the main text and provide the alternative quadratic-based bound in the appendix for completeness.

B.2.1 Solution 1: First Order Based Upper Bound Derivation

Suppose that $\widehat{\mathbb{A}} \geq 0$ and $A > 0$. We use the first-order upper bound for concave functions (Boyd and Vandenberghe, 2004):

$$\sqrt{a+b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}} \quad \text{for } a > 0,\ a+b \geq 0.$$

Since $(16A - 32\widehat{\mathbb{A}}) \geq 0$, thus $0 \leq \widehat{\mathbb{A}} \leq 1/2A$

$$\begin{aligned}
A &\leq \sqrt{(16A - 32\widehat{\mathbb{A}})2C} + \frac{16C}{3} \\
&\leq \sqrt{32AC} - \frac{64\widehat{\mathbb{A}}C}{2\sqrt{32AC}} + \frac{16C}{3} \\
&= \sqrt{A \cdot 32C} - \frac{4\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{A}} + \frac{16C}{3} \\
&\leq \frac{1}{2}A + 16C - \frac{4\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{A}} + \frac{16C}{3} \quad (\text{since} \sqrt{ab} \leq a/2 + b/2) \\
&= \frac{1}{2}A - \frac{4\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{A}} + \frac{64C}{3} \\
&\leq \frac{1}{2}A - \frac{4\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{A_{max}}} + \frac{64C}{3}
\end{aligned} \tag{28}$$

so that

$$\begin{aligned}
A &\leq -\frac{8\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{A_{max}}} + \frac{128C}{3} \\
&= -\frac{8\sqrt{2}\widehat{\mathbb{A}}\sqrt{C}}{\sqrt{\frac{128C}{3}}} + \frac{128C}{3} \quad (\text{make} \widehat{\mathbb{A}} = 0, \text{thus} A_{max} < \frac{128C}{3}) \\
&= \frac{128C}{3} - \sqrt{3}\widehat{\mathbb{A}}
\end{aligned} \tag{29}$$

thus:

$$R\left(\hat{f}^{(n)}\right) - R(\bar{f}) \leq \frac{128 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - \sqrt{3}\widehat{\mathbb{A}} \tag{30}$$

Re-arranging and using that $R_{\text{train}}(\bar{f}) \leq \mathbb{B}^2$, thus we have

$$R_{train}\left(\hat{f}^{(n)}\right) \leq \mathbb{B}^2 + \frac{128 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - \sqrt{3}\widehat{\mathbb{A}} \tag{31}$$

Finally, let $\mathbb{C}_\infty := \sup_{\xi \in \Xi} \left| \frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)} \right|$, the upper bound of generalization error can be expressed via the density ratio:

$$R_{test}\left(\hat{f}^{(n)}\right) \leq \mathbb{C}_\infty \cdot \left(\mathbb{B}^2 + \frac{128 y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} - \sqrt{3}\widehat{\mathbb{A}}) \right) \tag{32}$$

where $\widehat{\mathbb{A}}(\hat{f}^{(n)}) = \mathbb{E}_{\text{train}}\left[(\hat{f}^{(n)} - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \right]$

B.2.2 Solution 2: Quadratic-based Upper Bound Derivation

Or solving the quadratic inequality $A \leq \sqrt{(16A - 32\widehat{\mathbb{A}})2C} + \frac{16C}{3}$, thus

$$\begin{aligned}
&(A - \frac{16C}{3})^2 \leq 32AC - 64\widehat{\mathbb{A}}C \\
&\Rightarrow A^2 - \frac{32AC}{3} + (\frac{16C}{3})^2 \leq 32AC - 64\widehat{\mathbb{A}}C \\
&\Rightarrow A^2 - \frac{128AC}{3} + (\frac{16C}{3})^2 + 64\widehat{\mathbb{A}}C \leq 0
\end{aligned} \tag{33}$$

thus,

$$A \leq \frac{64C}{3} + \sqrt{\frac{1280C^2}{3} - 64\widehat{\mathbb{A}}C} \tag{34}$$

here exists a solution if and only if$\frac{1280C^2}{3} - 64\widehat{\mathbb{A}}C \geq 0$, which equals $\widehat{\mathbb{A}} \leq \frac{20C}{3}$

thus:

$$R\left(\hat{f}^{(n)}\right) - R(\bar{f}) \leq \frac{64y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} + \sqrt{\frac{1280y_\infty^4[\log(|\mathcal{F}|/\delta)]^2}{3n^2} - \frac{64\widehat{\mathbb{A}}y_\infty^2 \log(|\mathcal{F}|/\delta)}{n}} \tag{35}$$

Re-arranging and using that $R_{\text{train}}(\bar{f}) \leq \mathbb{B}^2$, thus we have

$$R_{train}\left(\hat{f}^{(n)}\right) \leq \mathbb{B}^2 + \frac{64y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} + \sqrt{\frac{1280y_\infty^4[\log(|\mathcal{F}|/\delta)]^2}{3n^2} - \frac{64\widehat{\mathbb{A}}y_\infty^2 \log(|\mathcal{F}|/\delta)}{n}} \tag{36}$$

Finally, let $\mathbb{C}_\infty := \sup_{\xi \in \Xi} \left| \frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)} \right|$, the upper bound of generalization error can be expressed via the density ratio:

$$R_{test}\left(\hat{f}^{(n)}\right) \leq \mathbb{C}_\infty \cdot \left(\mathbb{B}^2 + \frac{64y_\infty^2 \log(|\mathcal{F}|/\delta)}{3n} + \sqrt{\frac{1280y_\infty^4[\log(|\mathcal{F}|/\delta)]^2}{3n^2} - \frac{64\widehat{\mathbb{A}}y_\infty^2 \log(|\mathcal{F}|/\delta)}{n}}) \right) \tag{37}$$

where $\widehat{\mathbb{A}}(\hat{f}^{(n)}) = \mathbb{E}_{\text{train}}\left[(\hat{f}^{(n)} - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi))\right]$

B.3 Loosen the $\widehat{\mathbb{A}}$

We reproduce this derivation for completeness, following the approach in (Amortila et al., 2024), although it is not directly used in our main results.

In the inequality 16, we keep the interaction term $\widehat{\mathbb{A}}$. and if we loose the $\widehat{\mathbb{A}}$ by using the AM-GM inequality $\left((a+b)^2 \leq 2a^2 + 2b^2\right)$, the above equation becomes :

$$
\begin{aligned}
&\mathrm{Var}\left[(f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right] \\
&=\mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)^2\right]-\left(\mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)\right]\right)^2 \\
&\leq \mathbb{E}\left[\left((f(\xi)-y)^2-(\bar{f}(\xi)-y)^2\right)^2\right] \\
&=\mathbb{E}\left[\left(f^2(\xi)-2f(\xi)y-\bar{f}^2(\xi)+2y\bar{f}(\xi)\right)^2\right] \\
&=\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))^2(f(\xi)+\bar{f}(\xi)-2y)^2\right] \\
&\leq 16y_\infty^2\mathbb{E}\left[(f(\xi)-\bar{f}(\xi))^2\right] \quad (\text{since } |f(\xi)|,|\bar{f}(\xi)|,|y|\leq y_\infty) \\
&=16y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi)+f^\star(\xi)-\bar{f}(\xi))^2\right] \\
&\leq 32y_\infty^2\mathbb{E}\left[((f(\xi)-f^\star(\xi))^2+(\bar{f}(\xi)-f^\star(\xi))^2)\right] \quad (\text{since } (a+b)^2\leq 2a^2+2b^2) \\
&=32y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2+2\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right] \\
&\leq 32y_\infty^2\mathbb{E}\left[(f(\xi)-f^\star(\xi))^2-\left(\bar{f}(\xi)-f^\star(\xi)\right)^2\right]+64y_\infty^2\mathbb{B}^2 \\
&=32y_\infty^2(R(f)-R(\bar{f}))+64y_\infty^2\mathbb{B}^2
\end{aligned}
\tag{38}
$$

Using the same logic in the above derivation, based on the Equation (38), we have:

$$
\begin{aligned}
A &\leq \sqrt{(32A+64B)2C}+\frac{16C}{3} \\
&\leq \sqrt{64AC}+\sqrt{128BC}+\frac{16C}{3} \quad (\text{since}\sqrt{a+b}<\sqrt{a}+\sqrt{b}) \\
&=\sqrt{A\cdot 64C}+\sqrt{2B\cdot 64C}+\frac{16C}{3} \\
&\leq \frac{1}{2}A+32C+B+32C+\frac{16C}{3} \quad (\text{since}\sqrt{ab}\leq a/2+b/2) \\
&=\frac{1}{2}A+B+\frac{208C}{3}
\end{aligned}
\tag{39}
$$

so that

$$
A\leq 2B+\frac{416C}{3} \tag{40}
$$

So

$$
R\left(\hat{f}^{(n)}\right)-R(\bar{f})\leq 2\mathbb{B}^2+\frac{416y_\infty^2\log(|\mathcal{F}|/\delta)}{3n} \tag{41}
$$

Re-arranging and using that $R_{\text{train}}(\bar{f})\leq\mathbb{B}^2$, thus we have

$$
R_{\text{train}}\left(\hat{f}^{(n)}\right)\leq 3\mathbb{B}^2+\frac{416y_\infty^2\log(|\mathcal{F}|/\delta)}{3n} \tag{42}
$$

Let $\mathbb{C}_\infty:=\sup_{\xi\in\Xi}\left|\frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)}\right|$, the upper bound of generalization error can be expressed via the density ratio:

$$
R_{\text{test}}\left(\hat{f}^{(n)}\right)=\mathbb{E}_{\text{train}}\left[\frac{d_{\text{test}}(\xi)}{d_{\text{train}}(\xi)}\left(\hat{f}^{(n)}(\xi)-f^*(\xi)\right)^2\right]\leq\mathbb{C}_\infty\cdot\left(3\mathbb{B}^2+\frac{416y_\infty^2\log(|\mathcal{F}|/\delta)}{3n}\right) \tag{43}
$$

C CONSTRUCTING THE APPROXIMATE/PROXY DE-AMPLIFYING REGION

C.1 PROOF of Theorem 2

Theorem 2 [Approximate de-amplifying region] Let $\bar{f}$ be the predictor that best approximates the true data-generating function $f^\star$ in expectation with respect to the test distribution. Then,

$$
\widehat{\Xi}_{\mathbb{A}+}(\tau_1)\subseteq\Xi_{\mathbb{A}+}(\tau_0)
$$

where the approximate de-amplifying region $\widehat{\Xi}_{\mathbb{A}+}(\tau_1) \coloneqq \left\{\xi : |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq \tau_1\right\}$, $\tau_1 = \tau_0/\mathbb{B}_\infty + c\mathbb{B}_\infty$, $c \geq 2$, and $\tau_0 \geq 0$.

Proof Sketch: To compute a subset of $\left\{\xi : (\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0\right\}$ and remove the dependence on $f^\star$, we instead construct a subset of the de-amplifying region, the *approximate de-amplifying region*:

$$\left\{\xi : |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq c\mathbb{B}_\infty + \tau_0/\mathbb{B}_\infty,\ c \geq 2, \tau_0 \geq 0\right\} \subseteq \left\{\xi : (\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0\right\}. \tag{44}$$

Below, we detail the derivation of each of the following subset relations:

$$\begin{aligned}
\widehat{\Xi}_{\mathbb{A}+}(\tau_1) &= \left\{\xi : |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq c\mathbb{B}_\infty + \tau_0/\mathbb{B}_\infty,\ c \geq 2, \tau_0 \geq 0\right\} \\
&\subseteq^{(i)} \left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty + \mathbb{B}_\infty, \tau_0 \geq 0\right\} \\
&\subseteq^{(ii)} \left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| - |\bar{f}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty, \tau_0 \geq 0\right\} \\
&\subseteq^{(iii)} \left\{\xi : (\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0\right\} \\
&= \Xi_{\mathbb{A}+}(\tau_0)
\end{aligned}$$

Proof of subset relation (iii) We would design an acquisition function that selects only designs in the de-amplifying region $\Xi_{\mathbb{A}+}(\tau_0)$, where $\Xi_{\mathbb{A}+}(\tau_0) \coloneqq \left\{\xi : (\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \geq \tau_0\right\}$. And the left side of the inequality can be expressed as below:

$$\begin{aligned}
&(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \\
&= (\hat{f}^{(n)}(\xi) - f^\star(\xi) + f^\star(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) \\
&= (\hat{f}^{(n)}(\xi) - f^\star(\xi))(\bar{f}(\xi) - f^\star(\xi)) - (\bar{f}(\xi) - f^\star(\xi))^2
\end{aligned} \tag{45}$$

Thus, we refer to a design ξ as de-amplifying whenever $(\hat{f}^{(n)}(\xi) - \bar{f}(\xi))(\bar{f}(\xi) - f^\star(\xi)) = (\hat{f}^{(n)}(\xi) - f^\star(\xi))(\bar{f}(\xi) - f^\star(\xi)) - (\bar{f}(\xi) - f^\star(\xi))^2 \geq \tau_0$ for a given threshold τ_0.

Make an assumption that $\tau_0 \geq 0$, and based on $|\bar{f}(\xi) - f^\star(\xi)| < \mathbb{B}_\infty$, thus we have

$$|\hat{f}^{(n)}(\xi) - f^\star(\xi)| - |\bar{f}(\xi) - f^\star(\xi)| \geq \tau_0/|\bar{f}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty,$$

the subset (iii) follows.

Proof of subset relation (ii) From $\left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| - |\bar{f}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty\right\}$ we obtain

$$\left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty + |\bar{f}(\xi) - f^\star(\xi)|\right\}.$$

Since $|\bar{f}(\xi) - f^\star(\xi)| \leq \mathbb{B}_\infty$, it follows that

$$\left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty + \mathbb{B}_\infty\right\} \subseteq \left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| - |\bar{f}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty\right\}.$$

Proof of subset relation (i) Making a assumption that $\{\xi : |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq c\mathbb{B}_\infty + \tau_0/\mathbb{B}_\infty\}$ holds, then

$$|\hat{f}^{(n)}(\xi) - f^\star(\xi)| = |\hat{f}^{(n)}(\xi) - \bar{f}(\xi) + \bar{f}(\xi) - f^\star(\xi)| \geq |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| - |f^\star(\xi) - \bar{f}(\xi)| \geq (c-1)\mathbb{B}_\infty + \tau_0/\mathbb{B}_\infty.$$

Combining this with $\left\{\xi : |\hat{f}^{(n)}(\xi) - f^\star(\xi)| \geq \tau_0/\mathbb{B}_\infty + \mathbb{B}_\infty\right\}$, we see that for $c \geq 2$, subset (i) follows.

C.2 PROOF of Lemma 1

Lemma 1 [Proxy region approximates $\widehat{\Xi}_{\mathbb{A}+}(\tau)$] Given a proxy function $g : \Xi \mapsto \mathbb{R}$ such that $\sup_{\xi\in\Xi} |g(\xi)-\bar{f}(\xi)| \leq \tau_2$ for some $\tau_2 \geq 0$,

$$\widehat{\Xi}^{g}_{\mathbb{A}+}(\tau) \coloneqq \left\{\xi \in \Xi : |\hat{f}^{(n)}(\xi) - g(\xi)| \geq \tau\right\} \subseteq \widehat{\Xi}_{\mathbb{A}+}(\tau_1).$$

where $\tau = \tau_1 + \tau_2$.

Proof. By the triangle inequality,

$$|\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq |\hat{f}^{(n)}(\xi) - g(\xi)| - |g(\xi) - \bar{f}(\xi)| \geq |\hat{f}^{(n)}(\xi) - g(\xi)| - \tau_2$$

Thus when

$$|\hat{f}^{(n)}(\xi) - g(\xi)| \geq \tau_1 + \tau_2,$$

we have

$$|\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq |\hat{f}^{(n)}(\xi) - g(\xi)| - \tau_2 \geq \tau_1 \Rightarrow |\hat{f}^{(n)}(\xi) - \bar{f}(\xi)| \geq \tau_1$$

So

$$\left\{\xi : |\hat{f}^{(n)}(\xi) - g(\xi)| \geq \tau_1 + \tau_2\right\} \subseteq \widehat{\Xi}_{\mathbb{A}+}(\tau_1).$$

D ADDITIONAL DETAILS OF THE NUMERICAL EXPERIMENTS

Our experiments were implemented using PyTorch (Paszke, 2019) and Pyro (Bingham et al., 2019). All experiments were conducted on a shared computing cluster using NVIDIA A100 GPUs. Each job was allocated one A100 GPU, 8 CPU cores, and 32GB of RAM.

Measuring the degree of covariate shift To measure the distance between the distributions or datasets, we use the *Maximum mean discrepancy* (MMD). A number of advantages of this distance are put forward in the literature: 1) it is more robust to outliers than other discrepancy measurements (like KL divergence) (Gretton et al., 2012); 2) it can be approximated on the basis of differently-sized samples from the distributions being compared (Gretton et al., 2012); 3) the measurement is robust to repeated samples (unlike the Wasserstein distance) (Viehmann, 2021); 4) the measurement can be computed efficiently using samples (Bharti et al., 2023; Huang et al., 2023).

We compute the squared Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) between two empirical distributions $\mathcal{P} = \{p_i\}_{i=1}^m$ and $\mathcal{Q} = \{q_j\}_{j=1}^n$ using the Gaussian (RBF) kernel:

$$\mathrm{MMD}^2(\mathcal{P}, \mathcal{Q}) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{i'=1}^{m} k(p_i, p_{i'}) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n} k(q_j, q_{j'}) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(p_i, q_j),$$

where $k(p, q) = \exp\left(-\frac{\|p-q\|^2}{2\sigma^2}\right)$ is the RBF kernel with fixed bandwidth $\sigma = 1$.

Measuring the generalization performance To evaluate the generalization performance, we compute the *Mean Squared Error (MSE)* between the model predictions $\hat{f}^{(n)}$ and the corresponding true observations from the data-generating process (DGP) y, given the test samples.

$$\mathrm{MSE} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N}\sum_{i=1}^{N}(\hat{f}_{d,i}^{(n)} - y_{d,i})^2$$

where D is the number of test samples, and N is the sampling number.

D.1 Polynomial Regression Experiments

Both well-specified and misspecified models were run across 20 runs. For each run, the design is adaptively selected, and the total number of designs is $T = 10$.

Well-specified case

- **DGP**: The DGP is a degree-two polynomial regression model, $y = 1 + 2x - 0.5x^2 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.1)$
- **well-specified model**: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ So the feature functions are $\phi(x) = [1, x, x^2]^\top$
- **Test distribution** (arbitrary ξ^{test}): $\xi^{test} \sim \mathcal{U}(-4, 4)$

Misspecified case

- **DGP**: The DGP is a degree-two polynomial regression model, $y = 1 + 2x - 0.5x^2 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.1)$
- **misspecified model**: $f(x) = \beta_0 + \beta_1 x$ So the feature functions are $\phi(x) = [1, x]^\top$
- **Test distribution** (arbitrary ξ^{test}): $\xi^{test} \sim \mathcal{U}(-4, 4)$

D.2 Source Localization Experiments

We take $T = 30$ iteration steps for selecting training samples ξ^{train} and 200 test samples for the observed $f^\star(\xi^{test})$ and predicted output $f(\xi^{test})$. We set $K = 2$ sources and limit the range of design to $\xi^{train} \in [-4, 4]$. We use 200 samples drawn from the uniform distribution over $\xi^{test} \in [-4, 4]$ as the test distribution to estimate the MMD.

In this experiment, we have K sources with unknown location parameter is $\theta = \{\theta_k\}_{k=1}^K$. We assume the number of sources K is known.

Acoustic energy attenuation model The total intensity, a superposition of the individual ones, at location ξ from K sources is considered in the following equation (Foster et al., 2021; Ivanova et al., 2024):

$$\mu(\theta, \xi) = b + \sum_{k=1}^{K} \frac{\alpha_k}{m + \|\theta_k - \xi\|^2} \tag{46}$$

The prior distribution of each location parameter θ_k is a normal distribution, and the observation noise is Gaussian noise. Therefore, the prior and likelihood are in the following:

$$\theta_k \overset{\text{i.i.d.}}{\sim} N\left(0_d, I_d\right), \log y \mid \theta, \xi \sim N(\log \mu(\theta, \xi), \sigma)$$

Well-specified case The hyperparameters of the true DGP used in our well-specification experiments can be found in the table below:

Parameter	Value
Number of sources, K	2
Base signal, b	10^{-1}
Max signal, m	10^{-4}
α_1, α_2	1
Signal noise, σ	0.1

The model hyperparameters used in our experiments are the same in the above table.

Misspecfied case The hyperparameters of the true DGP used in our mis-specification experiments can be found in the table below:

Parameter	Value
Number of sources, K	2
Base signal, b	$4 * 10^{-1}$
Max signal, m	$4 * 10^{-4}$
α_1, α_2	0.4
Signal noise, σ	0.1

The model hyperparameters used in our experiments are the same as the table in the well-specified case.

D.3 Pharmacokinetic Model

We take $T = 10$ iteration steps for selecting training samples ξ^{train} and 200 test samples for the observed $f^\star(\xi^{test})$ and predicted output $f(\xi^{test})$. We limit the range of design to $\xi^{train} \in [0, 24]$. We use 200 samples drawn from the uniform distribution over $\xi^{test} \in [0, 24]$ as the test distribution to estimate the MMD. In the experiment, we set the true theta as $\theta_{\text{real}} = [1.5, 0.15, 15.0]$

Well-specified case. The drug concentration z, measured ξ hours after administration, and the corresponding noisy observation y, are modeled as

$$z(\xi;\theta) = \frac{D_V}{V} \cdot \frac{k_\alpha}{k_\alpha - k_e} \left[e^{-k_e\xi} - e^{-k_\alpha\xi}\right], \qquad y(\xi;\theta) = z(\xi;\theta)(1+\epsilon) + \eta, \tag{47}$$

where $\theta = (k_\alpha, k_e, V)$, $D_V = 400$ is a constant, $\epsilon \sim \mathcal{N}(0, 0.01)$ is multiplicative noise (to capture heteroscedasticity), and $\eta \sim \mathcal{N}(0, 0.1)$ is additive observation noise.

The prior distribution for the parameters θ is specified as

$$\log\theta \sim \mathcal{N}\left(\begin{bmatrix}\log 1\\ \log 0.1\\ \log 20\end{bmatrix}, \begin{bmatrix}0.05 & 0 & 0\\ 0 & 0.05 & 0\\ 0 & 0 & 0.05\end{bmatrix}\right). \tag{48}$$

Since both noise sources are Gaussian, the observation likelihood and DGP are the same and are also Gaussian:

$$y(\xi;\theta) \sim \mathcal{N}\big(z(\xi;\theta),\ 0.01\, z(\xi;\theta)^2 + 0.1\big), \tag{49}$$

Misspecified case. To introduce model misspecification, DGP comes from Equation (49). And a dualabsorption model with two parallel absorption rates is generated to make a prediction. Let $k_{a1} = k_\alpha$ and $k_{a2} = \rho\, k_{a1}$ with $\rho \in (0, 1)$, and let $f \in (0, 1)$ denote the fraction of the fast pathway. The mean concentration is

$$z_{\text{pre}}(\xi;\theta,\rho,f) = \tfrac{D_V}{V}\left[f \cdot \tfrac{k_{a1}}{k_{a1}-k_e}\big(e^{-k_e\xi} + e^{-k_{a1}\xi}\big) + (1-f)\cdot \tfrac{k_{a2}}{k_{a2}-k_e}\big(e^{-k_e\xi} + e^{-k_{a2}\xi}\big)\right]. \tag{50}$$

We set $\rho = 0.25$ and $f = 0.6$. The observation in the assumed model is,

$$y_{\text{pre}}(\xi;\theta) \sim \mathcal{N}\big(z_{\text{pre}}(\xi;\theta,\rho,f),\ 0.02\, z_{\text{pre}}(\xi;\theta,\rho,f)^2 + 0.2\big).$$

E ADDITIONAL RESULTS OF THE NUMERICAL EXPERIMENTS

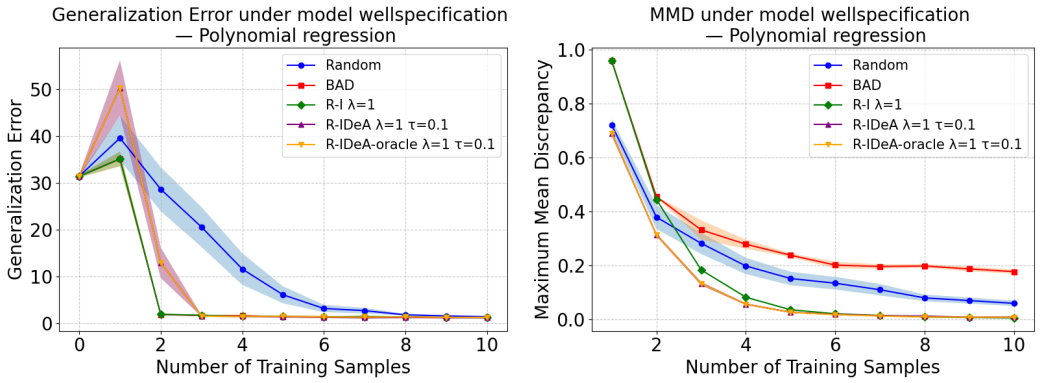
E.1 Polynomial Regression Experiments

E.1.1 Model Well-specification

Figure 5 shows that, in the well-specified case, all methods yield similar generalization error (Figure 5a), regardless of the degree of covariate shift (Figure 5b). This indicates that covariate shift does not significantly impact generalization performance when the model is well-specified. Moreover, the proposed R-I acquisition function, which combines informativeness and representativeness, converges more quickly than random selection. R-IDeA has comparable results to R-IDeA-oracle, which uses the true value of $\bar{f}$ instead of the proxy g, demonstrating the effectiveness of our method for selecting the proxy g. For clarity, we present results for $\lambda = 1$ and $\tau = 0.1$, as other settings exhibit similar behavior.

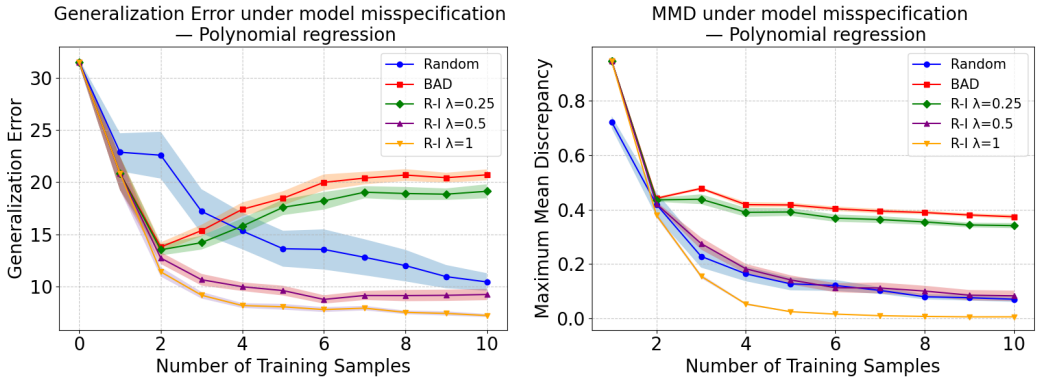
E.1.2 Model Misspecification

R-I - varying λ Figures 6a and 6b show the performance of the proposed R-I acquisition function under various values of λ. For larger values of λ, we expect the representativeness term to dominate the acquisition function, resulting in a design distribution that resembles the test distribution. Figure 6b shows that when designs are more representative, generalization error is reduced (Figure 6a), consistent with the theoretical prediction introduced in Theorem 1. These results demonstrate again that representative designs effectively reduce estimation bias and improve generalization performance.



(a) (b)

Figure 5: **Polynomial regression experiments (well-specified case).** Comparison of different design strategies (Random, BAD, proposed R-I, proposed R-IDeA, and R-IDeA-oracle under well-specified models in polynomial regression. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.



(a) (b)

Figure 6: **Polynomial regression experiments (effect of λ).** Comparison of baseline methods (Random, BAD) and our proposed R-I with varying λ in the polynomial regression experiments. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

R-IDeA - varying τ Figure 7 shows the performance of our R-IDeA under various learned adversarial proxies with different values of τ. Over all values of τ, R-IDeA has a higher generalization performance than R-I, suggesting that de-amplifying designs indeed increase generalization performance.

From Equations (11) and (12), we see that larger values of τ result in design distributions with stricter de-amplifying constraints. In other words, when τ is large, we expect the de-amplifying term to dominate the acquisition function. For smaller values of τ, the de-amplifying constraint is loosened, resulting in more candidate designs being effectively considered as part of the de-amplifying region. Therefore, τ should be chosen to balance de-amplification and other properties (informativeness and representativeness).

Figures 7a, 7c and 7e show that when $\tau = 0.5$, R-IDeA performs better than under other values of τ we tested. Also, R-IDeA with $\tau = 0.5$ induces less covariate shift (lower MMD) than R-IDeA with other values of τ (e.g., $\tau = 10$), showing that increasing the degree of de-amplification could reduce the degree of representativeness. These results show that the design properties of de-amplification and representativeness are interrelated; therefore, balancing de-amplification with other properties is important. We leave the development of systematic methods for selecting the hyperparameters in R-I and R-IDeA for future work.

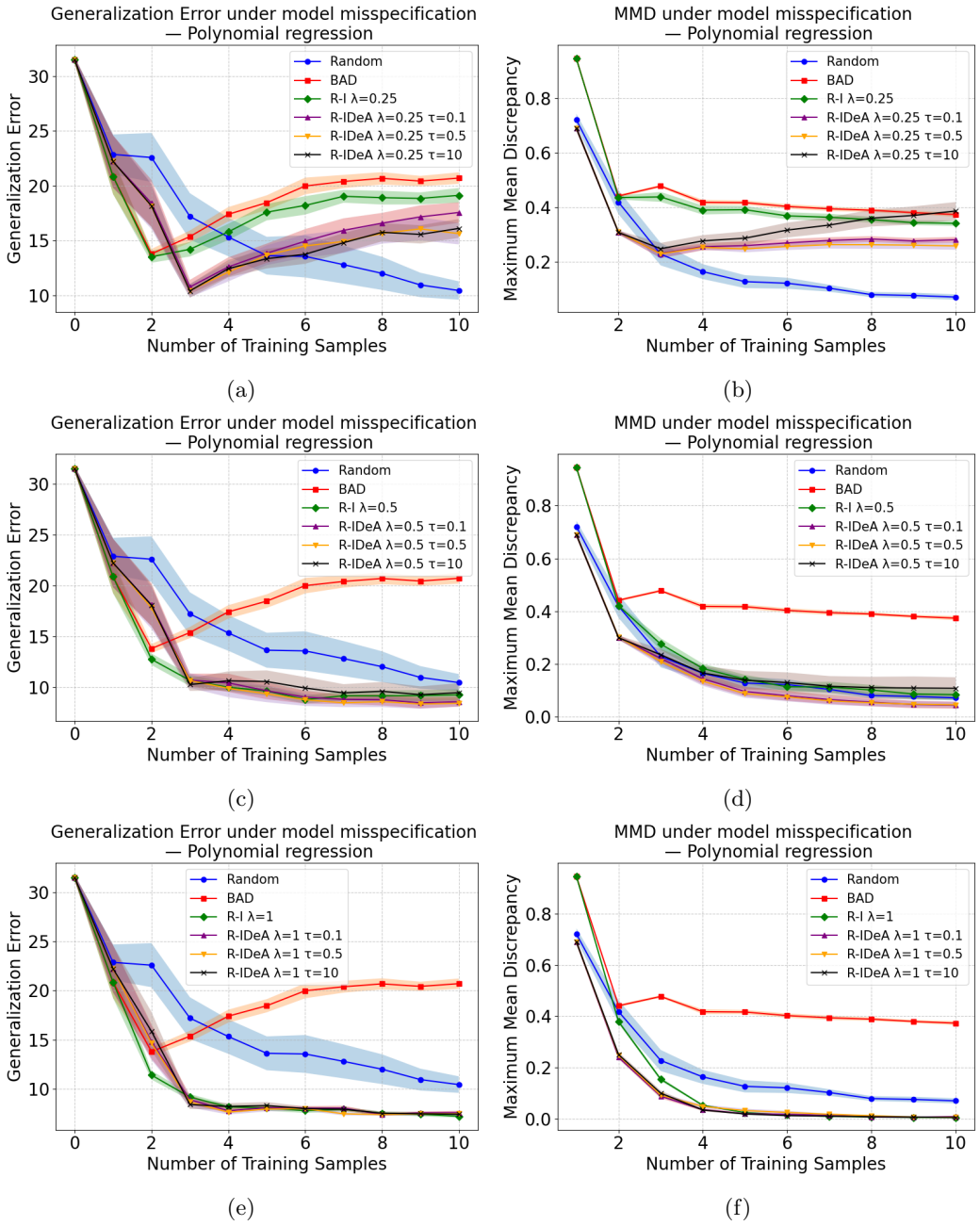
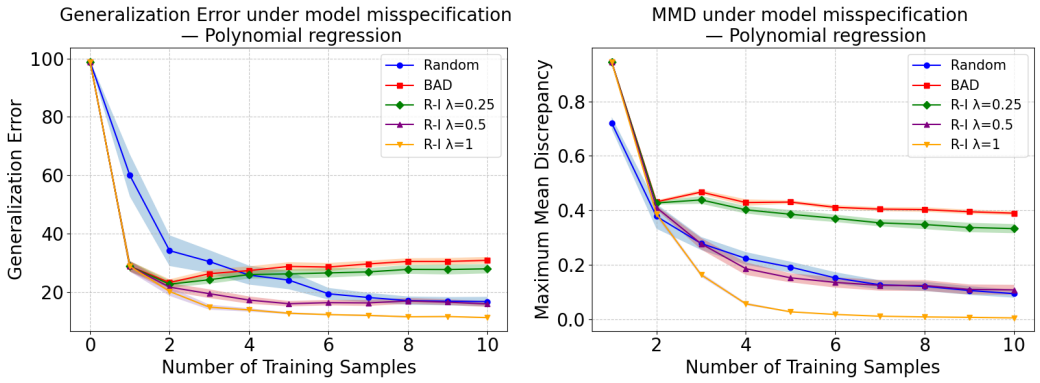


Figure 7: **Polynomial regression experiments (effect of τ).** Comparison of baseline methods (Random, BAD), our proposed R-I and our proposed R-IDeA with varying τ in the polynomial regression experiments. Rows correspond to variation in λ. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

E.1.3 High Degrees of Misspecification

In the discussion before, we showed the results of experiments where the DGP was linear (Appendix E.1.1) and quadratic (Appendix E.1.2), corresponding to no and mild misspecification, respectively.

To investigate the impact of the proposed acquisition function under different levels of misspecification, we ran the same experiments with a higher degree of misspecification by adding a cubic term to the DGP (the DGP is $y = 1 + 2x - 0.5x^2 + 0.2x^3 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.1)$). The results, shown in Figures 8 and 9, show a similar trend: R-I and R-IDeA continue to provide stable and competitive performance even as misspecification increases. We



(a) (b)

Figure 8: **Polynomial regression experiments (severely misspecified case).** Comparison of baseline methods (Random, BAD) and our proposed R-I with varying λ. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

note that when $\lambda = 0.25$ (i.e., informativeness is weighted much more highly than representativeness), both R-I and R-IDeA perform worse than Random (Figures 9a and 9b), demonstrating the importance of representativeness again.

E.1.4 Comparison of Different Degrees of Misspecification

We also compared the relative performance of R-IDeA under different degrees of misspecification. Figure 10 shows different degrees of misspecification (x-axis) and the ratio of generalization error resulting from using a given method and from a random design strategy (y-axis). This ratio is equal to 1 in the well-specified case, showing that each method yields similar generalization error under model well-specification. Under model misspecification, Figure 10 shows that the performance of the proposed R-IDeA increases slightly as misspecification grows but that the ratio remains below 1. These results indicate that the relative advantage of R-IDeA becomes smaller under severe misspecification, but remains positive. We speculate this slight reduction may be because of an inappropriate choice of τ (we keep the same τ regardless of the degree of misspecification). For BAD, as the misspecification degree grows, the error ratio still remains above 1, suggesting that BAD is not robust to the degree of misspecification. The large variation in the relative performance of BAD and Random indicates the instability of the BAD method.

E.2 Source Localization Experiments

E.2.1 Model Well-specification

Like in the toy example, Figure 11 illustrates that covariate shift does not impact generalization performance when the model is well-specified. In the well-specified model, Random and BAD result in similar generalization errors. This may be due to poor estimation of the expected information gain (EIG) in high-dimensional spaces. R-IDeA has the best and fastest performance in the well-specified case compared to other methods.

E.2.2 Model Misspecification

R-I - varying λ Figure 12a shows that our novel R-I acquisition function achieves a smaller generalization error than BAD. Figure 12b further demonstrates that the designs selected by this robust acquisition function are more representative than those selected by BAD. These results show that representative designs effectively reduce estimation bias and improve generalization performance. However, varying the value of λ leads to similar generalization performance and covariate shift, suggesting that performance is not sensitive to λ and the ability of R-I to improve representativeness in high-dimensional settings is limited.

R-IDeA - varying τ Figure 13 illustrates that across different values of τ, R-IDeA consistently outperforms R-I in terms of generalization performance. These results support the conclusion we drew from the polynomial

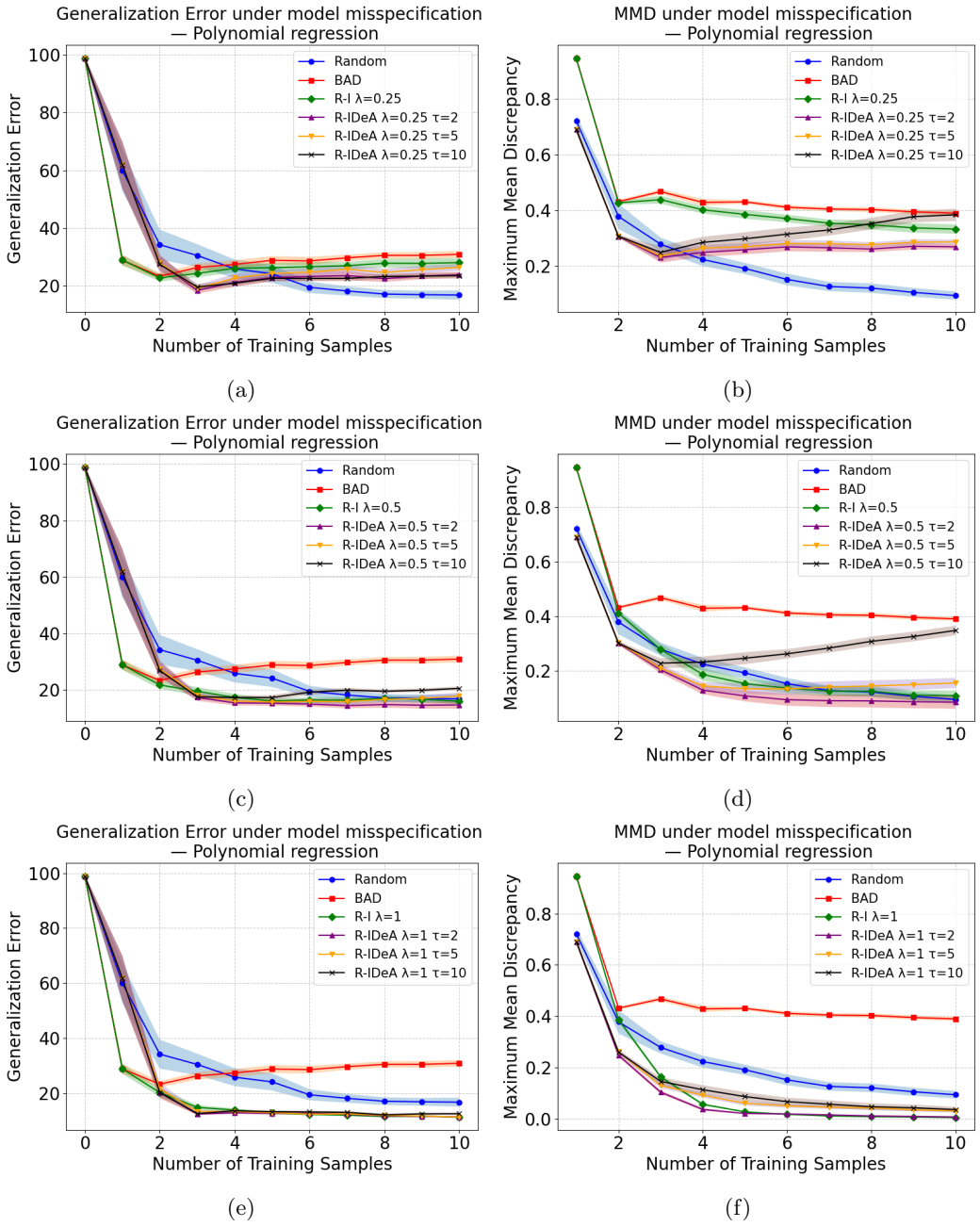


Figure 9: **Polynomial regression experiments (effects of λ and τ in the severely misspecified case).** Comparison of baseline methods (Random, BAD), our proposed R-I and our proposed R-IDeA with varying τ. Rows correspond to variation in λ. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

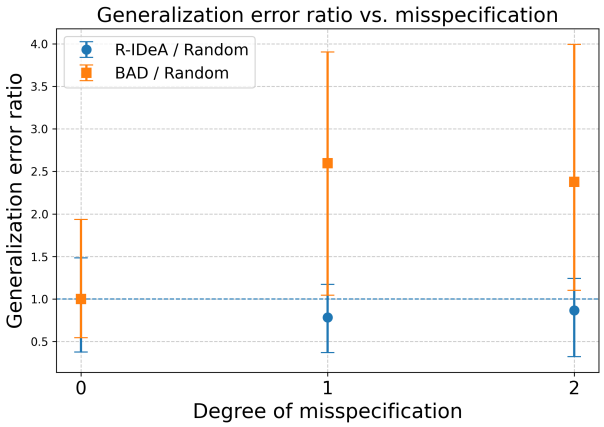


Figure 10: **Polynomial regression experiments (effect of degree of misspecification).** Comparison of the BAD method and our proposed R-IDeA across different degrees of misspecification. The x-axis represents the degree of misspecification. At degree 0, the model is well-specified. At degree 1, the DGP is $y = 1 + 2x - 0.5x^2 + \epsilon$ while the assumed model is linear. At degree 2, the DGP is $y = 1 + 2x - 0.5x^2 + 0.2x^3 + \epsilon$ while the assumed model is linear. The y-axis shows the ratio between the generalization error resulting from each method and a random design selection strategy with the 10 selected designs.

regression experiments that incorporating error de-amplification improves generalization. For $\lambda = 0.25$ and $\lambda = 0.5$, R-IDeA can lead to designs that, depending on the value of τ, exhibit more or less covariate shift than BAD (Figure 13b and Figure 13d). However, our experimental results suggest that even when less representative, these designs lead to higher generalization performance (Figure 13a and Figure 13c). When $\lambda = 1$, R-IDeA performs worse than Random but still better than BAD (Figure 13e), showing that both representativeness and de-amplification contribute to robustness.

E.3 Pharmacokinetic Model

E.3.1 Model Well-specification

Like in the polynomial regression experiments (Figure 5) and source localization experiments (Figure 11), Figure 14 illustrates that covariate shift does not impact generalization performance when the model is well-specified.

BAD, R-I and R-IDeA decrease error more quickly than random in the well-specified case, suggesting that the expected information gain (EIG) leads to informativeness designs.

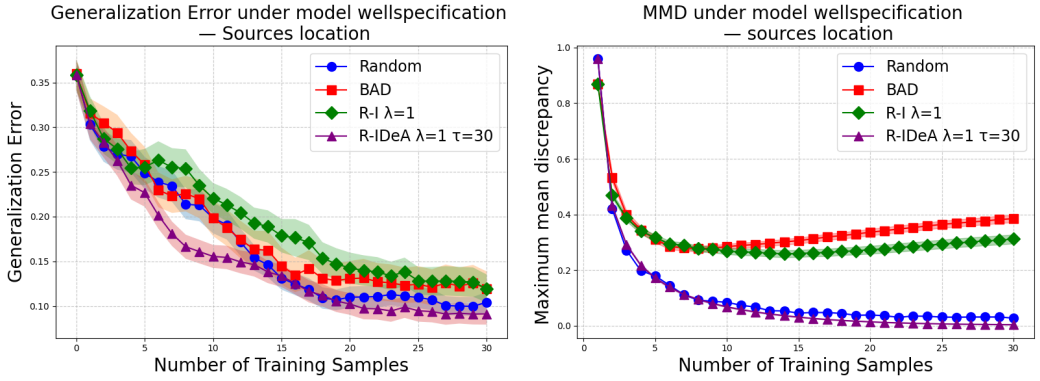
E.3.2 Model Misspecification

R-I - varying λ Figure 15 presents the performance of our R-I acquisition function under different values of λ. As shown in Figure 15b, more representative designs lead to lower generalization error (Figure 14a), consistent with the theoretical prediction in Theorem 1. The effect of varying λ follows the same trend as in the Polynomial Regression experiment in Figure 6, further demonstrating the robustness of the R-I.

R-IDeA - varying τ To explore how the hyperparameter τ affects our proposed acquisition function, Figure 16 shows the performance of our novel R-IDeA acquisition function with different values of τ.

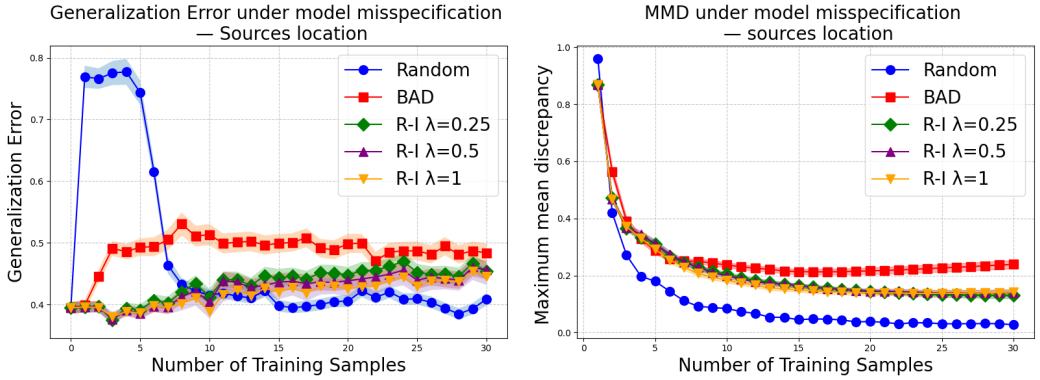
When $\lambda = 1$, Figure 16e illustrates that across different values of τ, R-IDeA consistently outperforms R-I and Random in terms of generalization performance, while some value of τ leads to a larger degree of covariate shift (Figure 16f). These findings support our earlier conclusion in Theorem 1 that incorporating error de-amplification improves generalization.

Interestingly, for $\lambda = 0.25$ (Figure 16a), R-IDeA performs worse than both Random and R-I. We speculate that this is due to an inappropriate choice of τ. When $\lambda = 0.5$ (Figure 16c), R-IDeA with $\tau = 0.5$ outperforms R-I,



(a) (b)

Figure 11: **Sources location experiments (well-specified case).** Comparison of different design strategies (Random, BAD, proposed R-I, proposed R-IDeA under well-specified models in sources location. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.



(a) (b)

Figure 12: **Sources location experiments (effect of λ).** Comparison of baseline methods (Random, BAD) and our proposed R-I with varying λ in the sources location experiments *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

whereas other values of τ lead to worse performance, highlighting the importance of properly choosing τ. For $\lambda = 1$ (Figure 16e), R-IDeA outperforms both Random and R-I. Moreover, R-IDeA with $\tau = 10$ achieves better results than with other values of τ, illustrating that selecting an appropriate hyperparameter cannot be achieved by simply increasing or decreasing its value in a heuristic manner. From Equation (11) and Equation (12), larger values of τ cause the de-amplifying term to dominate the acquisition function, enforcing a strict de-amplifying constraint, whereas smaller values loosen this constraint and yield more candidate designs. Therefore, τ should be carefully chosen to balance de-amplification with other properties such as informativeness and representativeness. How to choose the hyperparameters in our R-I and R-IDeA is an avenue for future work.

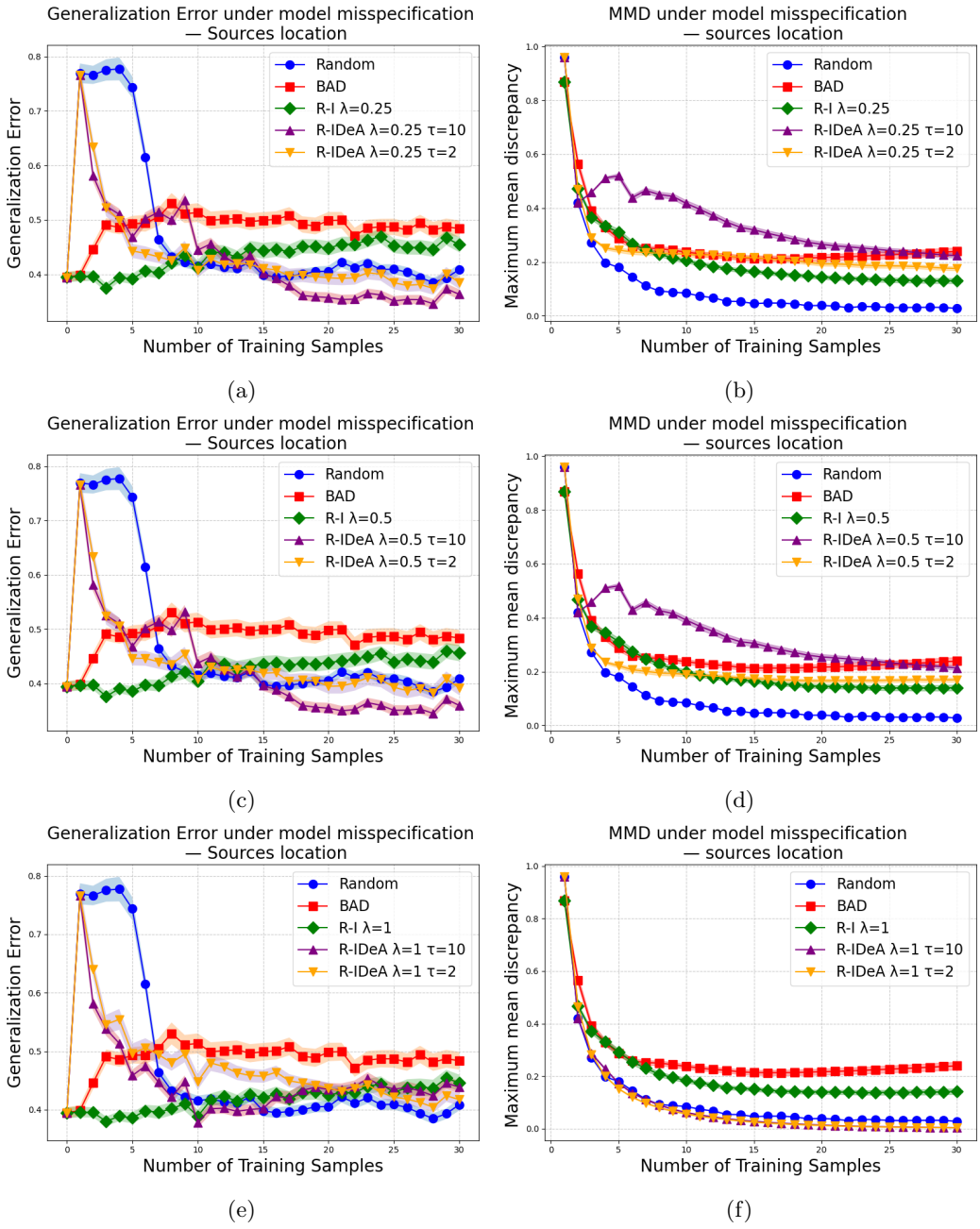
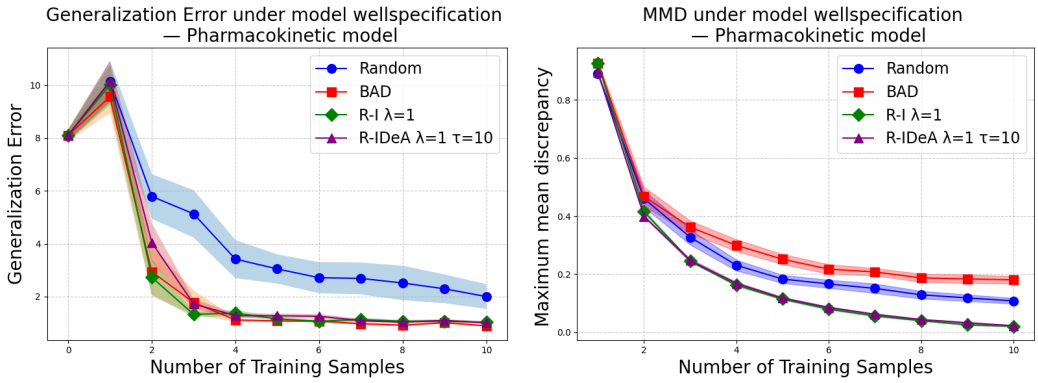
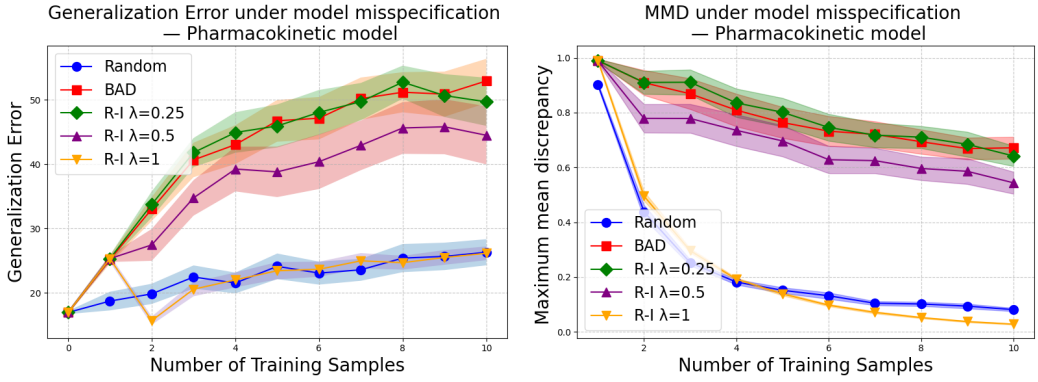


Figure 13: **Sources location experiments (effect of τ)** Comparison of baseline methods (Random, BAD), our proposed R-I and our proposed R-IDeA with varying τ in the source location experiments. Rows correspond to variation in λ. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.



(a) (b)

Figure 14: **Pharmacokinetic model experiments (well-specified case).** Comparison of different design strategies (Random, BAD, proposed R-I, proposed R-IDeA and R-IDeA-oracle which uses the $\bar{f}$ instead of the proxy g) under well-specified models in the Pharmacokinetic model experiment. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.



(a) (b)

Figure 15: **Pharmacokinetic model experiments (effect of λ).** Comparison of baseline methods (Random, BAD) and our proposed R-I with varying λ in the Pharmacokinetic model experiments *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.

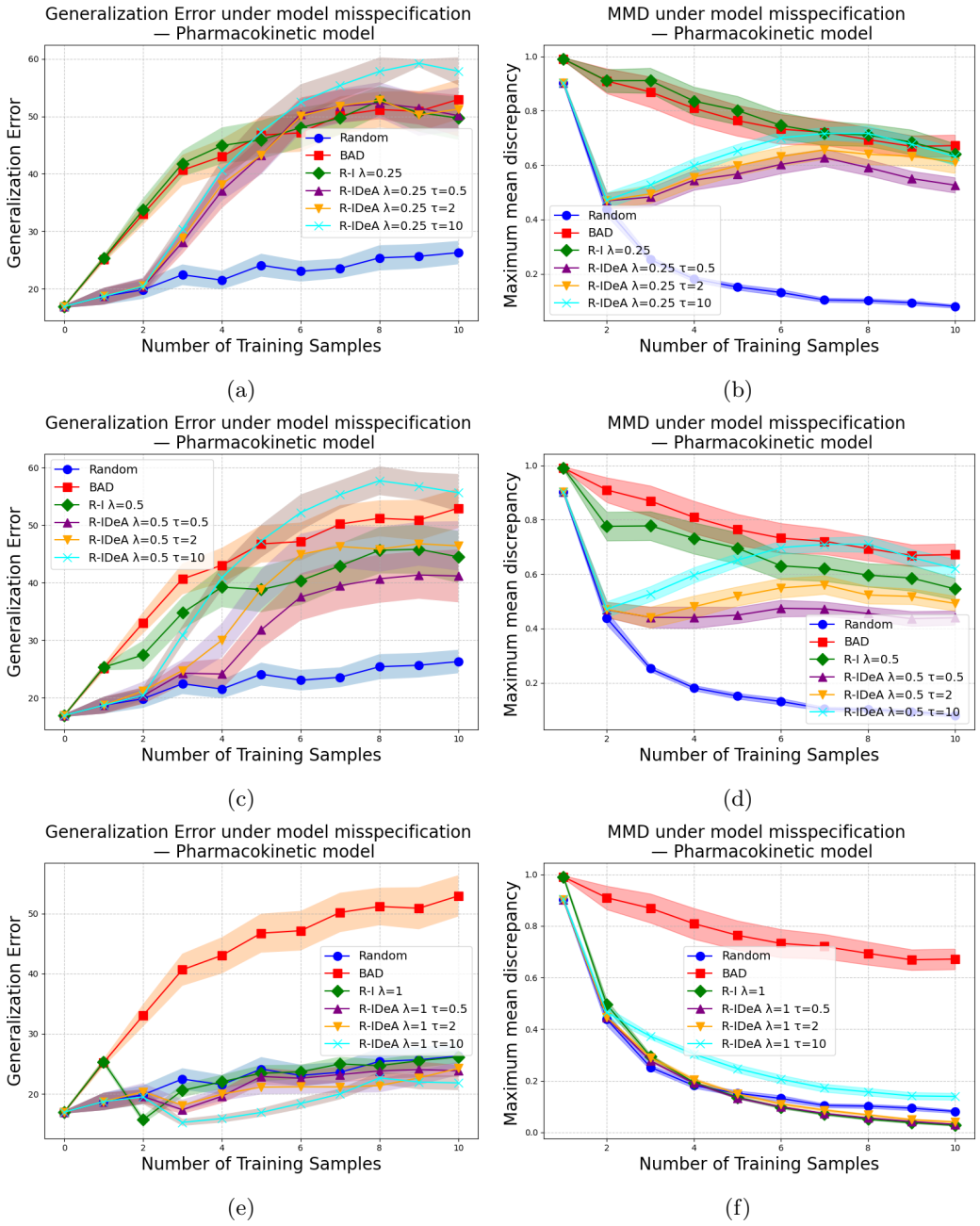


Figure 16: **Pharmacokinetic model experiments (effect of τ).** Comparison of baseline methods (Random, BAD), our proposed R-I and our proposed R-IDeA with varying τ in the Pharmacokinetic model experiments. Rows correspond to variation in λ. *Left*: Generalization error across methods. *Right*: MMD distance across methods; higher values indicate a greater degree of covariate shift.