DISTRIBUTIONALLY ROBUST MACHINE INTELLIGENCE FOR MEDICINE AND SCIENTIFIC DISCOVERY

Michael S Yao

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2025

Supervisor of Dissertation	Co-Supervisor of Dissertation
James C Gee	Osbert Bastani
Professor of Radiologic Science in Radiology	Associate Professor of Computer and Information Science

Graduate Group Chairperson

Yale E Cohen, Professor of Bioengineering and Otorhinolaryngology

Dissertation Committee

Kevin B Johnson, David L. Cohen University Professor of Pediatrics and Biomedical Informatics
M Dylan Tisdall, Research Assistant Professor of Radiology
Mark Yatskar, Assistant Professor of Computer and Information Science
Walter R Witschey, Associate Professor of Radiology

DISTRIBUTIONALLY ROBUST MACHINE INTELLIGENCE

FOR MEDICINE AND SCIENTIFIC DISCOVERY

COPYRIGHT

2025

Michael Steven Yu-Shuan Yao

To my wife, Boo.

ACKNOWLEDGEMENT

This thesis could not be complete without first thanking my co-advisors—Jim Gee and Osbert Bastani—for their incredible mentorship and guidance throughout my graduate studies. My co-advisors have always encouraged and empowered me to pursue my research interests to the fullest extent possible, and I thank them for pushing and encouraging me in my graduate training. I have also been fortunate to benefit from many other mentors and collaborators in my research endeavors at Penn. First, I thank Kevin Johnson, Dylan Tisdall, Mark Yatskar, and Walter Witschey for their invaluable mentorship as members of my dissertation committee. I also thank my collaborators Hersh Sagreiya, Hamsa Bastani, Jacob Gardner, Charles Kahn, Yimeng Zeng, Yifan Wu, Mona Gandhi, and Yufei Wang for their generous contributions and willingness to work with me on an array of (exciting!) research projects together.

My personal path to seriously considering a career in the sciences started well before graduate school. I was incredibly privileged to benefit from amazing mentors even in high school, and especially would like to thank my teachers Ethan Schnell and Minu Basu for inspiring me to dream and pursue the path that I am fortunate to walk today. Separately, I am also grateful for my previous research mentors—Shingo Suzuki, Dieter Gruenert, and Mikhail Shapiro among many others—that collectively built the foundation of my budding career as a researcher. I also thank the many external mentors I have benefited from during my graduate training: Laura Sacolick, Michael Hansen, Alma Andersson, Aïcha Bentaieb, and Claudia Iriondo.

I am also grateful for the people who ensured graduate school was never lonely. To Kristen Park, Kevin Shen, Mimi Kim, Juliana Qin, Noah Cho, and Katie Choi: I have you to thank for helping me learn how important friends can be. I also thank my wonderful family for their never-ending support: Jane, Yé ye, Năi nai, Peter, Joyce, Lynn, Kevin, Elaine, Zach, Kiwi, Lychee, and Cherry.

Finally (and most importantly), I thank my incredible and loving wife Allison. From soothing my frustrations when code bases were not compiling; to enabling me to be the best person I can be, I am grateful and in awe every day that I am able to be supported by a life partner like you.

ABSTRACT

DISTRIBUTIONALLY ROBUST MACHINE INTELLIGENCE FOR MEDICINE AND SCIENTIFIC DISCOVERY

Michael S Yao

James C Gee

Osbert Bastani

Machine learning systems are becoming increasingly adopted in high-stakes applications from clinical medicine to scientific discovery. In these settings, the predictions made by learned algorithms can have profound consequences. While modern machine learning models have achieved impressive empirical performance, they often behave unpredictably outside their training distribution, raising concerns about their reliability, fairness, and safety. These limitations are especially pronounced in domains where failure can be costly or irreversible, such as healthcare and scientific discovery. As a result, there is a growing need for AI systems that are not only performant, but also **safe** and **generalizable** when faced with new, diverse, and unforeseen inputs in the wild.

This dissertation investigates how we can design such ML systems to make reliable predictions across the range of inputs they might encounter in the real world. We explore this question through two complementary hypotheses. First, by incorporating **structured priors** generated from natural language and domain knowledge of biomedical systems directly into model architectures, we can build systems that are more generalizable. We show how such ML systems that are interpretable-by-design are better aligned with human reasoning to solve challenging domain-specific tasks. Second, we show how leveraging **adversarial supervision** from auxiliary neural networks can help us estimate when and where black-box model predictions can be trusted. We demonstrate how this framework can be readily adapted to solve a wide range of optimization problems in medicine and science. In summary, this dissertation provides a principled framework for making machine learning systems more aligned, robust, and actionable in safety-critical biomedical applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

Certainly! Below is a draft of a **Preface** for your dissertation written in a formal yet personal tone.

—

Most of you will have seen such a prefix in your conversations with ChatGPT; I hope it illustrates not only how ubiquitous AI has become in our everyday lives, but also how it shapes our trust in and relationships with one another. More importantly, tools such as (but certainly not limited to) AI have the potential to *erode trust* between individuals if not used **safely** and **responsibly**. To this end, I was inspired to pursue my dissertation research in this field because I believed (and still do) that trustworthy AI is one of the most important problems to work on—now more than ever.

This dissertation is oriented around how we can ensure the *safe* and *responsible* usage of AI tools. I specifically look at this question through the lens of *interpretability* and *generalizability*—how can we ensure that the ML systems we use in real-world pipelines yield reliable predictions for the many different possible inputs to the system? In this work, we explore two central hypotheses to answer this question: (1) if we use **prior knowledge** to build better ML models that align with how we as humans think, then those ML models might better generalize like humans do; and (2) we can better learn from **prior data** to determine when and where to trust black-box ML predictions.

I initially started my research career at the lab bench working on projects like targeted gene-editing therapies and designing targeted cancer treatments (shamelessly citing Suzuki et al. (2022) and Abedi et al. (2022)). While I should no longer be trusted with a pipette, these early experiences were formative in my decision to research trustworthy AI as it specifically pertains to the biomedical problems, such as scientific discovery and clinical medicine. During my graduate training, I have been fortunate to think critically about how AI can affect domains such as medical imaging, emergency medicine, scientific innovation, and clinical trial design: asking and answering scientific questions shaped by my own lived and ongoing experiences. I look forward to sharing the fruits of these works with you in the remainder of this dissertation.

CHAPTER 1

INTRODUCTION

Modern machine learning (ML) systems have demonstrated extraordinary capabilities across a wide range of domains—from materials synthesis (Szymanski et al., 2023; MacLeod et al., 2020) and drug discovery (Li et al., 2015; Brown et al., 2019) to robotics (Ahn et al., 2020; Ma et al., 2023; Radosavovic et al., 2022; Huang et al., 2022) and healthcare (Yang et al., 2024c; Adams et al., 2022; Pyrros et al., 2023; Singhal et al., 2023). These advances have largely been driven by algorithms that learn strictly from data: given a set of inputs x and outputs y, we are often satisfied with learning an ML model that is able to identify patterns and correlations in the data without explicit programming. In this traditional data-centric paradigm, we may not need to know (or even care) how or why a model prediction was generated. Instead, large and diverse datasets combined with powerful learning architectures are frequently enough to yield robust performance in many consumer applications (AI, 2024b; Adams et al., 2022; Szymanski et al., 2023).

While this has been a relatively successful paradigm, it is important to recognize that this is *not* how humans learn (Spelke and Kinzler, 2007; Zaadnoordijk et al., 2022; Collins et al., 2024). Human learning is richly contextual and structured by innate priors, abstract world models, and a continuous interplay between perception, action, and reflection (Ho et al., 2022; Hafner et al., 2025; Gottlieb et al., 2013; Agrawal et al., 2016). We do not simply memorize correlations or optimize for predictive accuracy; we interpret, reason, and build models of how the world works. These internal models allow us to generalize from limited experience, adapt quickly to new environments, and navigate uncertainty with resilience and creativity.

The divergence between how machines and humans learn is not merely an academic curiosity; recent evidence has shown that it can have real and pressing consequences. Systems trained purely from data can be susceptible to spurious correlations (DeGrave et al., 2021; Antony et al., 2023; Glocker et al., 2023), adversarial input examples (Goodfellow et al., 2015; Mehrotra et al., 2024; Chang et al., 2025; Han et al., 2024), and failure in out-of-distribution settings (Futoma et al., 2020;

Yang et al., 2022, 2024b) that can degrade the **trustworthiness** of such applications. Worse, they often lack the capacity for causal reasoning and ethical judgment (Campbell et al., 2024; Omiye et al., 2023; Tennant et al., 2025; Joshi et al., 2024). In critical applications, such as education, healthcare, and science, these limitations can lead to outcomes that are not only suboptimal, but also harmful (Bastani et al., 2025; Caruana et al., 2015; Zink et al., 2024).

How do we define trustworthiness? We know it when we see it, although achieving an objective, concrete definition continues to elude us (Floridi and Cowls, 2022; Ibáñez and Olmeda, 2022; Morley et al., 2021; Jobin et al., 2019). However, the consensus among recent work (Li et al., 2025; Mucsányi et al., 2023; Wang et al., 2023; Eshete, 2021) includes the following core tenets:

Interpretability. The notion of interpretability is a notoriously challenging property of machine learning systems to define (Lipton, 2018; Agarwal et al., 2024a; Madsen et al., 2024). For the purposes of this thesis, we define interpretable models as those with semantically meaningful internal representations of the model's inputs. These internal *feature representations* can be abstracted as *concepts* that are able to be understood by the model's human users. A growing body of work has explored how to build interpretable models from scratch by considering a specialized hypothesis class of learnable functions, such as linear models, concept bottleneck models, (Wu et al., 2025; Koh et al., 2020; Yang et al., 2023b; Srivastava et al., 2024; Sun et al., 2025) and generalized additive models (Hastie and Tibshirani, 1986; Yang et al., 2023a; McLean et al., 2014; Caruana et al., 2015). A common critique with such approaches is that by restricting the hypothesis class to functions that are interpretable to humans, we may adversely impact model complexity and therefore performance. I argue in this dissertation that this claim need not be true; put simply, it is possible to design ML pipelines to be both performant and interpretable: for example, we demonstrate this in the setting of opportunistic diabetes screening in **Chapter 4** (**Fig. 1.1**).

A separate body of work has looked at post-hoc analysis of black-box models (Ribeiro et al., 2016; Adebayo et al., 2018; Turbé et al., 2023; Yuksekgonul et al., 2023; Lundberg and Lee, 2017). For example, Ribeiro et al. (2016) introduced Local Interpretable Model-agnostic Explanations (**LIME**) to fit an interpretable model to a black-box model in a local neighborhood in the input space,

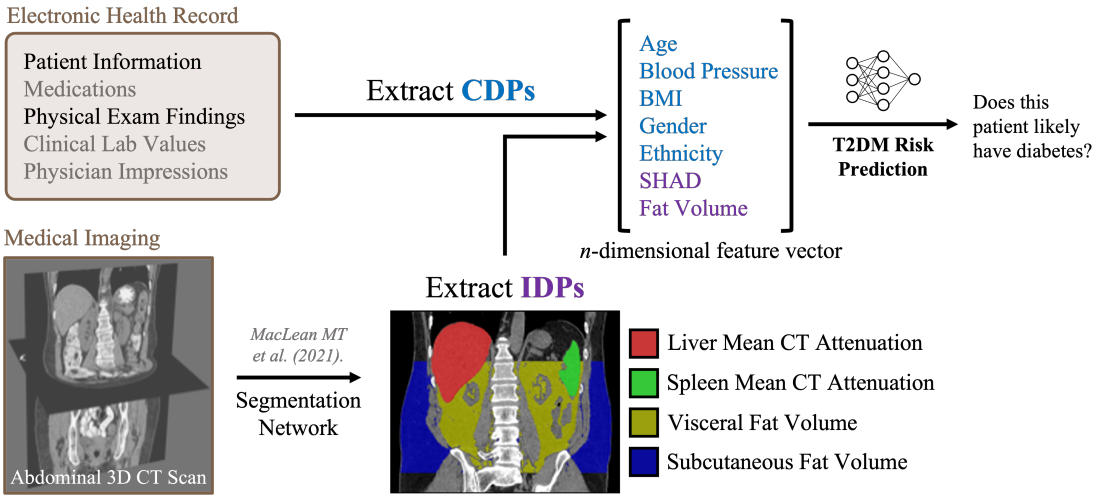

Figure 1.1: **Synthetic HbA1c lab values derived from multimodal patient data enables interpretable and generalizable opportunistic diabetes screening.** Raw patient data can be high-dimensional, multimodal, and therefore difficult to interpret. We leverage clinical knowledge to determine which clinically derived phenotypic features (**CDPs**) and image-derived phenotypic features (**IDPs**) are relevant for opportunistic diabetes screening. Using the IDP extraction pipeline from MacLean et al. (2021), we estimate quantitative IDPs from abdominal computed tomography (CT) scans associated with an increased risk for diabetes. We use these interpretable IDPs and CDPs from health record data to train generalizable diabetes risk prediction models (**Chapter 4**).

approximating (and therefore explaining) the local behavior of the black-box model. SHapley Additive exPlanations (**SHAP**) was introduced by Lundberg and Lee (2017), and estimates the average marginal contribution of each feature to the observed model output. However, recent work have found that such methods can be brittle in practice (Crabbé and van der Schaar, 2023; Laugel et al., 2019; Ragodos et al., 2024) leading to arbitrarily derived explanations, and have no guarantee that highlighted features causally influence the model's decision making (Chou et al., 2022; Adebayo et al., 2022).

Recent advancements in *large language models* (LLMs) (OpenAI et al., 2024; Anthropic, 2024) have also sparked recent work interrogating the interpretability of textual autoregressive models. For example, Wei et al. (2022) introduced **chain-of-thought** (CoT) prompting to elicit human-readable reasoning traces, which may help illustrate *how* an LLM arrives at a final answer (Dutta et al., 2024; Wei Jie et al., 2024; Zhao et al., 2024). Chain-of-thought and the associated reasoning steps can therefore offer a glimpse into the model's inner decision process beyond the classical view of 'sim-

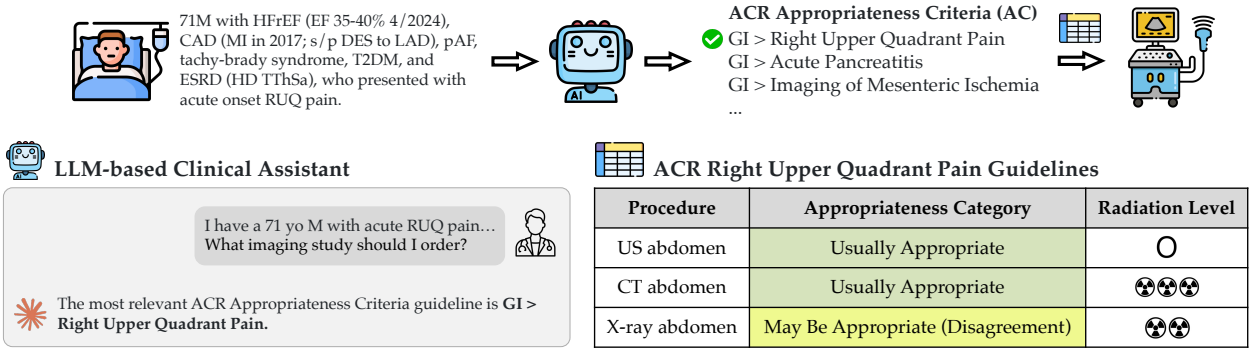

Figure 1.2: **Adapting generalist LLMs as clinical assistants for medical image ordering.** In **Chapter 3**, we show that traditional large language model (LLM) systems struggle with recommending evidence-based imaging studies to order for patients. To overcome this limitation, we explicitly enforce the LLM to predict the most relevant medical guideline from a corpus for the patient. We can then directly look up the most appropriate imaging study in the guidelines document to recommend a final imaging study. This simple, interpretable zero-shot strategy allows us to construct an LLM pipeline that outperforms even fine-tuned biomedical models, enabling consumer-grade LLMs to generalize to a real-world clinical task.

ple' next-token prediction (Lyu et al., 2023). Separately, a growing body of literature challenges whether CoT truly reflects internal reasoning (Chen et al., 2025b). Turpin et al. (2023) found that LLMs could be manipulated towards a wrong answer in question answering (QA) evaluations *and also* plausible CoT traces that hide known underlying biases. Agarwal et al. (2024a) discuss empirical tradeoffs observed between the faithfulness and plausibility of LLM-generated thought processes. In **Chapter 3**, we build on recent efforts (Hicks et al., 2022) to evaluate the real-world utility of chain-of-thought prompting in eliciting interpretability in autoregressive models (**Fig. 1.2**).

Generalizability. Generalizability is arguably *the* core desiderata at the heart of modern machine learning, as we often seek to train an ML model on one dataset in order to generalize to new, previously unseen datasets at inference time. Classical strategies to prevent overfitting of ML models include imposing inductive biases (Gatmiry et al., 2023; Helmbold and Long, 2015; Morwani and Ramaswamy, 2022), bootstrap aggregation (Ngo et al., 2022; Debeire et al., 2024), and data augmentation (Goceri, 2023; Shorten and Khoshgoftaar, 2019; Mumuni and Mumuni, 2022). These methods often enforce certain priors over the model weight space or leverage empirical training heuristics to enable the development of more robust ML models. Recent work

on neural architecture search (NAS) has also helped discover novel model architectures that are able to generalize better across different input distributions (Oymak et al., 2021; White et al., 2021; Zoph and Le, 2017; Liu et al., 2025b). Interpretable-by-design ML models discussed above can be thought of as a 'human-in-the-loop' method of NAS to constrain model features as compositional, modular, and/or semantically meaningful representations to improve their generalizability. Importantly, while some have argued that such strategies are no longer of critical concern in the era of large foundation models trained on Internet-scale data (Singhal et al., 2023; OpenAI et al., 2024; Zhang et al., 2024), many domain-specific tasks bottlenecked by expert knowledge are still limited in their ability to generalize to related task environments (Yao et al., 2025a; Trabucco et al., 2022).

However, a crucial limitation of these methods is that they assume total control over the design, training, and deployment of the machine learning model. This is a strict assumption that does not hold in many real-world applications: for example, leading domain-specific expert models are often proprietary and made available only by limited application programming interface (API) endpoints, and it may not be feasible to retrain computationally expensive models to optimally perform on different input distributions. In these settings, we may only think of ML models as **black-box functions**, where the only permissible interaction with f is querying it with inputs x to observe outputs $f(x)$. In this most general setting, the above methods to improve generalizability are not applicable and it is not possible to ensure any generalizable correctness guarantees.

Instead, recent work has looked at the generalization behavior of black-box models subject to certain assumptions about the environment. For example, domain adaptation methods (Ganin et al., 2016; Sun and Saenko, 2016; Bousmalis et al., 2016; Tzeng et al., 2017; Bousmalis et al., 2017) assume that samples from the test distribution are accessible to align feature representations of datums from the source and test distributions. Trabucco et al. (2021); Yu et al. (2021); and related work assume a smoothness prior over the black-box function to implicitly enforce a constraint on the Lipschitz norm of the function over the local domain of interest. In this dissertation, we consider a separate environmental setup—namely, the experimental environment typical of *offline optimization* problems described below—and investigate how to improve the generalizability of

black-box models as applied to this setting.

Offline Optimization. We have primarily limited our discussion of the interpretability and generalizability of ML models in isolation; how can we generally build machine learning models that are interpretable and generalizable? However, the requirements for these attributes of ML systems differ based on the underlying application. For example, models deployed in healthcare settings almost always benefit from trading accuracy to increase their interpretability (Caruana et al., 2015; Chae et al., 2024), and may therefore be hyper-specialized to a target patient population. In contrast, consumer recommendation systems may care less about the interpretability of model predictions, but must critically generalize to a wide variety of user preferences.

The latter half of this dissertation proposes strategies to improve the generalizability of ML systems in the context of **offline optimization**. Formally, we consider problems of the form

$$x^\star = \arg\max_{x \in \mathcal{X}} f(x)$$

where the goal is to find a design x that maximizes a black-box function $f : \mathcal{X} \to \mathbb{R}$, with the additional restriction that $f(x)$ is not evaluable during optimization. To overcome this limitation, a common approach is to instead learn a (parametrized) *surrogate* function approximation $f_\theta : \mathcal{X} \to \mathbb{R}$ trained to approximate f by fitting on a static offline dataset $\mathcal{D} = \{(x_i, f(x_i))\}_{i=1}^n$ of previously observed designs. We may then solve the related optimization problem

$$x^* = \arg\max_{x \in \mathcal{X}} f_\theta(x) \tag{1.1}$$

with the hope that $x^* \approx x^\star$. Importantly, while a sufficiently well-trained surrogate may approximate the true objective well, there is no guarantee of the correctness of f_θ on designs $x \in \mathcal{X} \setminus \mathcal{D}$ necessarily encountered during optimization. Put simply, naïvely constructed surrogate functions used in offline optimization frequently lead to suboptimal proposed designs (Trabucco et al., 2021, 2022; Yu et al., 2021; Kumar and Levine, 2019; Fu and Levine, 2021) due to a failure to ***generalize*** to newly proposed designs (**Fig. 1.3**).

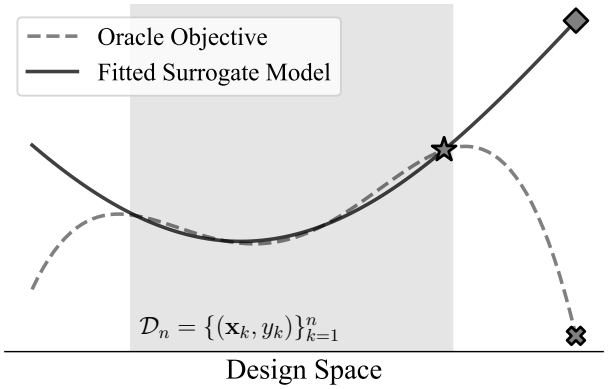

Figure 1.3: **Poor model generalizability limits the utility of traditional optimization methods in the offline setting.** Consider a black-box machine learning model $f_\theta : \mathcal{X} \to \mathbb{R}$ (i.e., a 'fitted surrogate model') trained on a fixed dataset $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ (shaded region) to approximate a true function in nature $f : \mathcal{X} \to \mathbb{R}$ (i.e., an 'oracle objective'). Evaluation of f_θ on inputs that are grossly out-of-distribution compared to $\mathcal{D}_n$ (e.g., cross) can result in inaccurate model predictions (e.g., diamond) compared to in-distribution inputs to f_θ (e.g., star). In **Chapter 5**, we address black-box model generalizability in the context of offline optimization using adversarial feedback.

In this problem setting, we make a number of key observations. First, we note that solving the optimization problem in (1.1) does not require that f_θ well-approximates the true black-box function f everywhere in the domain; rather, we only care that f_θ preserves the *ranking* of inputs x with respect to f (Tan et al., 2025). Second, the space of possible 'target' distributions over $\mathcal{X}$ is large and diverse, and importantly not known prior to optimization. This limits the utility of domain adaptation techniques and similar methodology previously described above. Finally, we assume that the labeled dataset $\mathcal{D}$ used to train f_θ is accessible. This assumption enables us to analyze the statistical properties of inputs with respect to $\mathcal{D}$, and incorporate this information into the optimization process itself. We discuss this further in the following **Chapters 5-6**.

1.1. Dissertation Statement

This dissertation is motivated by the belief that if we want to build and use machines that are trustworthy, we must critically re-examine the foundations of how we build and use them. I propose a series of novel algorithms to help make ML systems more generalizable, interpretable, and robust. These contributions are based in the belief that we must move beyond the notion that training a single ML model from data alone is sufficient, and move instead toward approaches that incor-

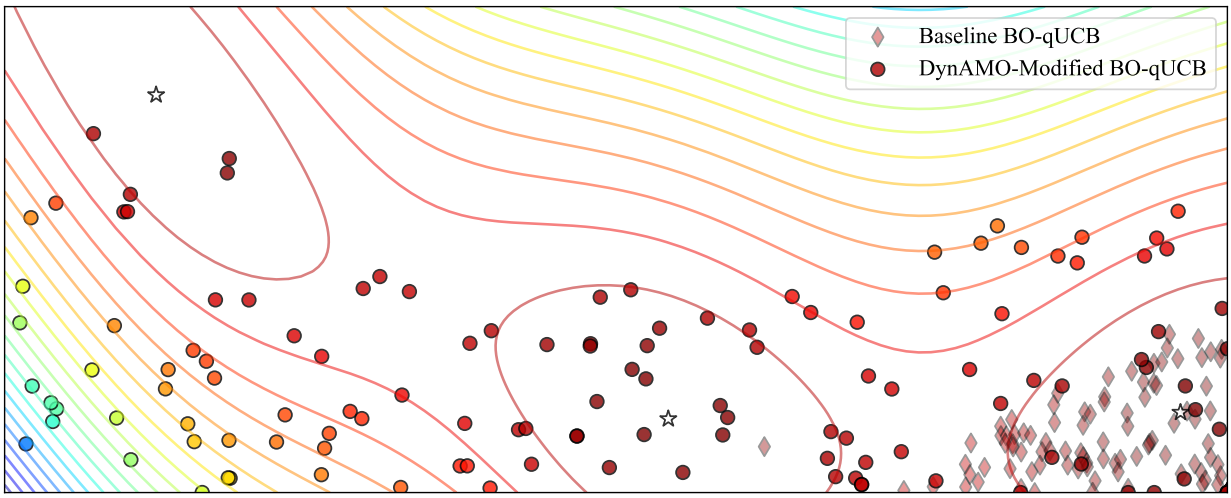

Figure 1.4: **Improving the diversity of designs proposed in offline optimization.** Traditional model-based optimization (MBO) (Trabucco et al., 2021) techniques can generate high-scoring designs, although often at the expense of the *diversity* of proposed designs. Ideally, the final set of candidates should be of high quality while capturing multiple 'modes of goodness.' For example, although there are 3 unique global maxima (stars) in the 2D Branin (Branin, 1972) optimization problem, traditional Bayesian optimization (BO-qUCB) proposes designs clustered around only a single optima (diamonds). In contrast, we show in **Chapter 6** how to modify the MBO objective to discover diverse *and* high-quality designs (circles).

porate additional learning signals—including structured priors and adversarial supervision—that mirror those available to us in human cognition. Such signals can offer a scaffold for more building more trustworthy ML systems, improving their readiness for real-world critical applications.

In this work, I explore new directions for aligning machine learning with human-like learning. I investigate both theoretical and practical strategies for incorporating structured knowledge into learning algorithms, analyze the limitations of purely data-driven models, and propose methods for bridging the gap between statistical learning and cognitive modeling. The overarching goal is to move toward a framework of machine learning that is not only more powerful and reliable but also more human-aligned in its assumptions, behaviors, and outcomes.

1.2. Dissertation Contributions

The contributions of this thesis are organized as follows:

In **Chapter 2**, I introduce the core background knowledge and preliminary concepts that lay the foundation for the remainder of the dissertation. This chapter motivates the key problem formulations that the work introduced discussed in this dissertation seeks to solve, and also the technical methodology we use in the remainder of the text.

In **Chapters 3-4**, I discuss methods to build ***interpretable-by-design* ML systems** to improve their ability to solve challenging tasks in clinical medicine. **Chapter 3** outlines a zero-shot strategy (Yao et al., 2025a) to prompt generalist large language models for guidance on ordering medical imaging studies aligned with evidence-based guidelines (**Fig. 1.2**). We achieve this by enforcing an intermediate representation space of patient input data that is explicitly constructed from medical guidelines. Separately, **Chapter 4** demonstrates how we can predict 'synthetic lab values' from multimodal clinical data (**Fig. 1.1**) (Yao et al., 2023). We use our method for interpretable and generalizable opportunistic screening of Type 2 Diabetes using real-world patient data.

In **Chapters 5-6**, I consider the problem of building generalizable ML systems in the setting of offline model-based optimization, where we may not have the control to choose the underlying model architecture. In this setting, **Chapter 5** first introduces a principled method to leverage ***adversarial feedback*** from source critic models to regularize how black-box models are used during offline optimization (**Fig. 1.3**). This method enables us to better solve offline optimization tasks across a wide variety of different scientific domains. In **Chapter 6**, we extend our method to consider the secondary problem of *diversity* in offline optimization: in the setting where multiple final designs can be proposed, it is often desirable to propose candidates that collectively cover a greater proportion of the overall design space (**Fig. 1.4**). We show how adversarial feedback can be naturally incorporated into a modified problem formulation that considers both the quality and diversity of final design proposals in a range of scientific discovery tasks.

In **Chapter 7**, I conclude this dissertation by summarizing the major findings and discussing future research directions to continue building more generalizable, safe, and robust machine learning systems for challenging domain-specific tasks.

1.3. Relevant Publications

This thesis discusses the following first or co-first author publications.

1. (Yao et al., 2025a) **Michael S Yao**, Allison Chae, Piya Saraiya, Charles E Kahn, Jr, Walter R Witschey, James C Gee, Hersh Sagreiya†, Osbert Bastani†. Evaluating acute image ordering for real-world patient cases via language model alignment with radiological guidelines. (**Communications Medicine** 2025)

2. (Chae et al., 2024) Allison Chae*, **Michael S Yao***, Hersh Sagreiya, Ari D Goldberg, Neil Chatterjee, Matthew T MacLean, Jeffrey Duda, Ameena Elahi, Arijitt Borthakur, Marylyn D Ritchie, Daniel Rader, Charles E Kahn, Jr, Walter Witschey†, James C Gee†. Strategies for implementing machine learning algorithms in the clinical practice of radiology. (**Radiology** 2024)

3. (Yao et al., 2023) **Michael S Yao***, Allison Chae*, Matthew T MacLean, Anurag Verma, Jeffrey Duda, James C Gee, Drew A Torigian, Daniel Rader, Charles E Kahn, Jr., Walter R Witschey†, Hersh Sagreiya†. SynthA1c: Towards clinically interpretable patient representations for diabetes risk stratification. (**MICCAI PRIME Workshop** 2023)

4. (Yao et al., 2024) **Michael S Yao**, Yimeng Zeng, Hamsa Bastani, Jacob R Gardner, James C Gee, Osbert Bastani. Generative adversarial model-based optimization via source critic regularization. (**NeurIPS** 2024)

5. (Yao et al., 2025b) **Michael S Yao**, James Gee, Osbert Bastani. Diversity by design: Leveraging distribution matching for offline model-based optimization. (**ICML** 2025)

Here, * denotes the co-first authorship and † denotes co-senior authorship. These publications are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Therefore, parts of this thesis are quoted directly from the above publications with the explicit approval of all co-authors and my thesis committee. None of these aforementioned works have been or will be extensively discussed in any of my collaborators' theses. Detailed statements describing my

individual contributions to the above projects can be found at the beginning of each chapter.

In the following works, I made a secondary contribution as a co-author; the below publications are therefore discussed only briefly in this dissertation in **Appendix B**.

1. (Wu et al., 2025) Yifan Wu, Wang Liu, Yue Yang, **Michael S Yao**, Wenli Yang, Xuehui Shi, Lihong Yang, Dongjun Li, Yueming Liu, Shiyi Yin, Chunyan Lei, Meixia Zhang, James C Gee, Xuan Yang, Wenbin Wei, Shi Gu. A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. (**Nature Communications** 2025)

2. (Yang et al., 2024c) Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, **Michael S Yao**, James C Gee, Chris Callison-Burch, Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. (**NeurIPS** 2024)

Finally, the following first-author publications describe significant efforts during my dissertation research that are outside the scope of the main body of this thesis:

1. (Yao et al., 2025c) **Michael S Yao**, Lawrence Huang*, Emily Leventhal*, Clara Sun, Steve J Stephen, Lathan Liou. Leveraging datathons to teach AI in undergraduate medical education: Case study. (**JMIR Medical Education** 2025)

2. (Yao and Hansen, 2022) **Michael S Yao**, Michael S Hansen. A path towards clinical adaptation of accelerated MRI. (**Machine Learning for Health Symposium** 2022)

CHAPTER 2

BACKGROUND AND PRELIMINARIES

In this section, we first formalize a notion of out-of-distribution (OOD) evaluation of machine learning systems, and establish that OOD evaluation is not merely a theoretical concern but a provably *unavoidable component* of many real-world, high-dimensional environments. As we demonstrate, when data is sparse relative to the dimensionality of the configuration space—as is common in many scientific and biomedical scenarios—standard generalization assumptions break down, and trained models are almost surely used to extrapolate to regions not well represented in the training set. This phenomenon is especially pronounced in cases where the data-generating distribution exhibits complex structure or where coverage guarantees (e.g., from Gaussian sampling or low-entropy priors) do not sufficiently capture the range of possible inputs. Recognizing the inevitability of such OOD generalization, we provide relevant background discussion to inform our contributions in mitigating the impact of this problem.

First, let us consider a motivating example. Suppose that the function $f(x) = \max(0, ||x|| - 1) + \varepsilon$ describes the relationship between inputs $x \in \mathbb{R}^d$ and outputs $f(x) \in \mathbb{R}$, where ε is the aleatoric uncertainty independent of x. We would like to train a predictive model $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ from a hypothesis class $\mathcal{H}$ that we believe contains the true underlying function $f(x)$. The standard practice in machine learning is to sample n i.i.d. observations $\mathcal{D} := \{(x_i, f(x_i)\}_{i=1}^n$ from nature and then learn the parameters of $\hat{f}$ to maximize the likelihood of observing $\mathcal{D}$.

However, such approaches are rarely so straightforward in practice. For illustrative purposes, suppose that the exact form of $f(x)$ is not known *a priori*—a common limitation in real-world problems. We might consider a hypothesis class $\mathcal{H} = \left\{x \mapsto w^\top x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ of linear models as a naïve ansatz. Furthermore, we may only be able to sample our training dataset from the unit hypersphere volume $S^d_{\text{train}} := \{x \in \mathbb{R}^d \mid ||x||_2 \leq 1\}$, even though we would like our trained model to predict on arbitrary inputs from $\mathbb{R}^d$. It is easy to see that empirical risk minimization yields the solution $\hat{f}(x) = 0$ from $\mathcal{H}$, which fits the true function $f(x)$ perfectly on S^d_{train}. However, suppose

that we now sample a test datum $X \sim \mathcal{N}(0, I_d)$. Observe first that $||X||_2^2$ follows a χ^2-distribution, whose well-known cumulative distribution function gives $\Pr(||X||_2 \geq 1) = \Pr(||X||_2^2 \geq 1) = \Gamma(d/2, 1/2)/\Gamma(d/2)$. In the limit that $d \gg 1$, we have $\Gamma(d/2, 1/2) \approx \Gamma(d/2, 0) = \Gamma(d/2)$ and so the norm of $||X||_2$ almost surely is greater than 1 with large d, meaning $X \notin S^d$ with high probability. Secondly, note that if $||X||_2 \geq 1$, then the prediction error $|f(X) - \hat{f}(X)| = |\max(0, ||X||_2 - 1) - 0| = ||X||_2 - 1$, which we show scales with $\mathcal{O}(\sqrt{d})$ in **Theorem 3** below. In other words, the predictive error scales with $\sqrt{d}$ with high probability in high-dimensional input spaces!

While this toy thought experiment is somewhat contrived, it crucially illustrates a number of key features of real machine learning systems:

1. **Epistemic Uncertainty.** Broadly speaking, epistemic uncertainty arises from a lack of knowledge or information about a task-specific domain. In the above example, this form of uncertainty is manifested by our 'incorrect' choice of the hypothesis class $\mathcal{H}$, which did not contain the true underlying function $f(x)$. While this may be evident in hindsight, choosing a hypothesis class that is large enough to contain $f(x)$ but small enough to avoid overfitting a finite training dataset is often difficult in practice.[1]

2. **Covariate Shift.** Covariate shift refers to the difference between the distribution of covariates $p(x)$ in the training dataset of $\hat{f}(x)$ and in the dataset of covariates we are interested in using the learned model on at test time. In the above example, we were only able to construct our training dataset by sampling from S^d_{train}, but wanted to use our predictive model on inputs sampled from $\mathcal{N}(0, I_d)$ with non-zero support over all of $\mathbb{R}^d$. In practical applications, we might only have access to training data from one patient population, but want to still generalize learned insights to a new set of patients, for example.[2]

3. **Error Scaling in d.** Together, epistemic uncertainty and covariate shift lead to poor prediction error scaling with the dimensionality of the input space. We observed that the predictive er-

[1] This thesis does not significantly discuss *aleatoric uncertainty*, which instead stems from (usually random) measurement noise of the input covariates and/or output observations.

[2] The work presented in this thesis primarily focuses on addressing covariate shift, and only briefly mentions techniques to address other types of distribution shift, such as label shift and concept shift (Yao et al., 2022).

ror of machine learning models is significantly worse in the high-dimensional setting of our toy example above. For many real-world problems with high-dimensional intrinsic dimensions and even higher input dimensions, error scaling can thus lead to catastrophic failures in machine learning systems.

The overarching goal of this thesis is to propose methods to address **(1) Epistemic Uncertainty** and **(2) Covariate Shift** to reduce the **(3) Error Scaling in** d in real-world machine learning systems. More specifically, we introduce generalizable algorithms to address these problems as they pertain to scientific discovery and clinical medicine, where these issues can lead to deleterious consequences if not properly addressed.

Is the problem of covariate shift unavoidable in high-dimensional problems? One might argue that obtaining more training data with better coverage of the input space and designing more intelligent model architectures mitigate this issue. However, consider the following theorem:

Theorem 1 (Necessary Extrapolation in Higher Dimensions)**.** *Suppose that we have a dataset of size n of d-dimensional points $\{X_i\}_{i=1}^n$ sampled i.i.d. according to $X_i \sim \mathcal{N}(0, I_d)$, and let $X \sim \mathcal{N}(0, I_d)$ be an independent test point. Then for any finite $\varepsilon > 0$,*

$$\lim_{d \to +\infty} Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) = 0 \tag{2.1}$$

Proof. The standard union bound gives

$$\Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) \leq \sum_{i=1}^{n} \Pr(||X - X_i||_2 \leq \varepsilon) = n \cdot \Pr(||X - X'|| \leq \varepsilon)$$

for any $X' \in \{X_i\}_{i=1}^n$. It is easy to see that $X - X' \sim \mathcal{N}(0, 2I_d)$. Park (1961) has shown that the random variable $||X - X'||_2$ is distributed according to the χ-distribution with the probability density function

$$f(z) = \frac{(z/\sqrt{2})^{d-1} e^{-(z/\sqrt{2})^2/2}}{2^{d/2-1}\Gamma(d/2)} = \frac{z^{d-1}e^{-z^2/4}}{2^{d-1}\Gamma(d/2)}$$

Therefore,

$$n \cdot \Pr(||X - X'||_2 \leq \varepsilon) = \frac{n}{2^{d-1}\Gamma(d/2)} \int_0^\varepsilon dz\, z^{d-1} e^{-z^2/4}$$

This integral is well-known and can be written in terms of the incomplete Γ function:

$$\begin{aligned} n \cdot \Pr(||X - X'||_2 \leq \varepsilon) &= \frac{n\left[\Gamma(d/2) - \Gamma(d/2, \varepsilon^2/4)\right]}{\Gamma(d/2)} = n\left(1 - \frac{\Gamma(d/2, \varepsilon^2/4)}{\Gamma(d/2)}\right) \\ &= \frac{n\varepsilon^d}{d \cdot 2^{d-1}\Gamma(d/2)} + \mathcal{O}(\varepsilon^{d+2}) \end{aligned}$$

in the limit of small $\varepsilon/d \ll 1$. The denominator asymptotically scales like $\mathcal{O}(d!)$ while the numerator is only exponential in d, meaning that the right hand side quickly converges to 0 for any finite choice of n, ε using Stirling's formula. We therefore have

$$\begin{aligned} \lim_{d\to+\infty} \Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) \leq \lim_{d\to+\infty} n \cdot \Pr(||X - X'||_2 \leq \varepsilon) &= \lim_{d\to+\infty} \left(\frac{n\varepsilon^d}{d \cdot 2^{d-1}\Gamma(d/2)}\right) \\ &= 0 \end{aligned} \tag{2.2}$$

The claim follows. □

Remark 1 (Generalization of Theorem 1 to Out-of-Distribution Scenarios)**.** *Consider the same setup as in* ***Theorem 1*** *except that the point X is now drawn from an arbitrary distribution with Lebesgue density $f_X(x)$ satisfying $\sup_{x\in\mathcal{X}\subseteq\mathbb{R}^d} f_X(x) \leq M$ finite.*[3] *Then*

$$\lim_{d\to+\infty} Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) = 0$$

for any choice of finite ε.

[3]The assumption that f_X has a bounded Lebesgue density is relatively weak—it is well-known that most probability distributions in the real world satisfy this property.

Proof. Using the same standard union bound as in our approach above,

$$\begin{aligned}\Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) &\leq \sum_{i=1}^{n} \Pr(||X - X_i||_2 \leq \varepsilon) = \sum_{i=1}^{n} \int_{||x-X_i||_2 \leq \varepsilon} dx\, f_X(x) \\ &\leq \sum_{i=1}^{n} \int_{||x-X_i||_2 \leq \varepsilon} dx\, M = M \sum_{i=1}^{n} \int_{||x-X_i||_2 \leq \varepsilon} dx \\ &= Mn \cdot \text{Vol}(B_d(\varepsilon)) = \frac{M\pi^{d/2} n\varepsilon^d}{\Gamma(1 + d/2)}\end{aligned}$$

where $\text{Vol}(B_d(\varepsilon))$ is the volume of a d-dimensional sphere with radius ε. As in our proof for **Theorem 1**, the denominator scales like $\mathcal{O}(d!)$ while the numerator is only exponential in d, meaning

$$\lim_{d \to +\infty} \Pr(\exists i : ||X - X_i||_2 \leq \varepsilon) \leq \lim_{d \to +\infty} \frac{M\pi^{d/2} n\varepsilon^d}{\Gamma(1 + d/2)} = 0 \tag{2.3}$$

for any $\varepsilon > 0$ finite. □

Theorem 1 and **Remark 1** argue that extrapolation is not only common, but *unavoidable* in high-dimensional spaces. Furthermore, the dominant terms in both (2.2) and (2.3) are factorial in d (with only linear dependence on the dataset size n), meaning the lower bound on the respective probabilities of extrapolation *rapidly* approaches 1. This result builds on the well-known result from Bárány and Füredi (1988), which we include below for completeness.

Theorem 2 (Vanishing Convex Hull of Finite Datasets (Bárány and Füredi, 1988))**.** *Suppose that we have a dataset of size n of d-dimensional points $\{X_i\}_{i=1}^n$ sampled i.i.d. according to $X_i \sim \mathcal{N}(0, I_d)$, and let $X \sim \mathcal{N}(0, I_d)$ be an independent test point. Define* Hull$(\cdot)$ *to be the convex hull of its arguments. Then*

$$\lim_{d \to +\infty} Pr(X \in \textit{Hull}(X_1, X_2, \ldots X_n)) = 0 \tag{2.4}$$

Proof. Define the random variable $u := X/||X||$. By construction, observe that

$$u^T X = \frac{X^T X}{||X||} = ||X||$$

Because the linear functional of a Gaussian is also Gaussian and since u has unit norm, note that $u^T X_i \sim \mathcal{N}(0,1)$ for any dataset member $X_i \sim \mathcal{N}(0, I_d)$. Define $M := \max_{1 \leq i \leq n} u^T X_i$. We have

$$\Pr(M \leq \varepsilon) = \Pr\left(\bigcap_{i=1}^{n} (u^T X_i \leq \varepsilon)\right) = \left[\Pr(u^T X_i \leq \varepsilon)\right]^n = \left[\frac{1}{2}\left(1 + \text{erf}\left(\frac{\varepsilon}{\sqrt{2}}\right)\right)\right]^n$$

Choosing $\varepsilon = \varepsilon_d := d^{1/4}$, note that

$$\lim_{d \to +\infty} \Pr(M \leq \varepsilon_d) = \lim_{d \to +\infty} \left[\frac{1}{2}\left(1 + \text{erf}\left(\frac{\varepsilon_d}{\sqrt{2}}\right)\right)\right]^n = \left[\frac{1}{2}(1+1)\right]^n = 1 \tag{2.5}$$

since $\lim_{z \to +\infty} \text{erf}(z) = 1$. Separately, a corollary of our proof to **Theorem 1** is that for any random variable $X \sim \mathcal{N}(0, I_d)$,

$$\lim_{d \to +\infty} \Pr(||X|| > \varepsilon_d) = 1 - \lim_{d \to +\infty} \frac{\varepsilon_d^d}{d \cdot 2^{d-1}\Gamma(d/2)} = 1 - \lim_{d \to +\infty} \frac{d^{d/4}}{d \cdot 2^{d-1}\Gamma(d/2)}$$

Using Stirling's formula in the limit of large $d \gg 1$,

$$\begin{aligned}
\lim_{d \to +\infty} \Pr(||X|| > \varepsilon_d) &= 1 - \lim_{d \to +\infty} \frac{d^{d/4}}{d \cdot 2^{d-1} \cdot \sqrt{2\pi} \cdot (d/2)^{(d/2)-(1/2)} \exp(-d/2)(1 + \mathcal{O}(1))} \\
&= 1 - \lim_{d \to +\infty} \frac{1}{d \cdot (\sqrt{2})^{d-3} \cdot \sqrt{2\pi} \cdot d^{(d/4)-(1/2)} \exp(-d/2)(1 + \mathcal{O}(1))} \\
&= 1 - 0 = 1
\end{aligned} \tag{2.6}$$

since $e^d = o(d^d)$ in the above limit. From independence,

$$\lim_{d \to +\infty} \Pr\left((||X|| > \varepsilon_d) \bigcap (M \leq \varepsilon_d)\right) = \lim_{d \to +\infty} \Pr(||X|| > \varepsilon) \cdot \Pr(M \leq \varepsilon_d) = 1 \tag{2.7}$$

combining (2.5) and (2.6). By definition of the convex hull, we know $z \in \textsc{Hull}(X_1, X_2, \ldots X_n))$ iff $z = \sum_{i=1}^n \alpha_i X_i$ for some $(\alpha_1, \alpha_2, \ldots, \alpha_n) \in \Delta(n)$ probability simplex. Observe that

$$u^T z = \sum_{i=1}^{n} \alpha_i u^T X_i \leq \sum_{i=1}^{n} \alpha_i (\max_{1 \leq j \leq n} u^T X_j) = M \sum_{i=1}^{n} \alpha_i = M \tag{2.8}$$

for any member z of the convex hull. In particular, note that if $u^T z \leq M \leq \varepsilon_d < ||X||$, then X

cannot be contained in the convex hull since it does not satisfy (2.8). Therefore,

$$\lim_{d\to+\infty} \Pr(X \in \textsc{Hull}(X_1, X_2, \ldots, X_n)) \leq 1 - \lim_{d\to+\infty} \Pr\left((||X|| > \varepsilon_d) \bigcap (M \leq \varepsilon_d)\right) = 1 - 1 = 0$$

using (2.8). □

A notable corollary to **Theorem 2** is presented by Balestriero et al. (2021): the size of training dataset n must increase exponentially with the intrinsic dimension of the data manifold to avoid the vanishing hull problem presented in this theorem. While this may be feasible in training modern generalist ML systems on Internet-scale data, the vast majority of problems in science and medicine are inherently limited by the availability of high-quality data. Finally, we extend beyond the notion of ε-neighborhoods from **Theorem 1** to consider how nearest neighbor distances scale in high dimensions:

Theorem 3 (Scaling Law of Nearest Neighbor Distances)**.** *Suppose we have a dataset of size n of d-dimensional points $\{X_i\}_{i=1}^n$ sampled i.i.d according to $X_i \sim \mathcal{N}(0, I_d)$, and let $X \sim \mathcal{N}(0, I_d)$ be an independent test point. Define $W_{min} = \inf_{i\in\{1,2,\ldots,n\}} ||X - X_i||$. Then W_{min} scales with $\mathcal{O}(\sqrt{d})$.*

Proof. Recall from Park (1961) that the random variable $||X - X_i||_2$ is distributed according to the χ-distribution. From Laurent and Massart (2000),

$$\Pr\left(\left|\frac{1}{2}||X - X_i||_2^2 - d\right| \leq 2\sqrt{\varepsilon d} + 2\varepsilon\right) = \Pr\left(\left|||X - X_i||_2^2 - 2d\right| \leq 4\sqrt{\varepsilon d} + 4\varepsilon\right) \geq 1 - 2e^{-\varepsilon}$$

for any i. Since $||X - X_i||_2, \sqrt{2d} > 0$, we know $\left|||X - X_i||_2 - \sqrt{2d}\right|^2 \leq \left|||X - X_i||_2^2 - 2d\right|$, meaning

$$\Pr\left(\left|||X - X_i||_2 - \sqrt{2d}\right|^2 \leq 4\sqrt{\varepsilon d} + 4\varepsilon\right) \geq 1 - 2e^{-\varepsilon}$$

In the limit $\varepsilon/d \ll 1$, we know $4\sqrt{\varepsilon d} + 4\varepsilon \to 4\sqrt{\varepsilon d}$, meaning

$$\lim_{d\to+\infty} \Pr\left(\left|||X - X_i||_2 - \sqrt{2d}\right|^2 \leq 4\sqrt{\varepsilon d}\right) \geq 1 - 2e^{-\varepsilon}$$

so long as $\varepsilon = o(\sqrt{d})$. Equivalently,

$$\lim_{d\to+\infty} \Pr\left(\exists i : \left|||X - X_i||_2 - \sqrt{2d}\right|^2 > 4\sqrt{\varepsilon d}\right) \leq n \cdot \lim_{d\to+\infty} \Pr\left(\left|||X - X_1||_2 - \sqrt{2d}\right|^2 > 4\sqrt{\varepsilon d}\right)$$
$$\leq 2ne^{-\varepsilon}$$

using the standard union bound. We can choose $\varepsilon = \frac{1}{2}\log d = o(\sqrt{d})$, giving

$$\lim_{d\to+\infty} \Pr\left(\exists i : \left|||X - X_i||_2 - \sqrt{2d}\right| > 2\left(\frac{1}{2}d\log d\right)^{1/4}\right) \leq \lim_{d\to+\infty} \frac{2n}{\sqrt{d}}$$

As we take $d \to +\infty$, the right hand side of the inequality approaches zero. Defining $W_{\min} := \min_{1\leq i\leq n} ||X - X_i||_2$, we have

$$\begin{aligned}
&\lim_{d\to+\infty} \Pr\left(\left|W_{\min} - \sqrt{2d}\right| \leq 2^{3/4}(d\log d)^{1/4} \leq 2^{3/4}\sqrt{d}\right) \\
&\quad = 1 - \lim_{d\to+\infty} \Pr\left(\exists i : \left|||X - X_i||_2 - \sqrt{2d}\right| > 2\left(\frac{1}{2}d\log d\right)^{1/4}\right) \\
&\quad \geq 1 - \lim_{d\to+\infty} \frac{2n}{\sqrt{d}} = 1
\end{aligned}$$

And so $|W_{\min} - \sqrt{2d}| = o(\sqrt{d})$ in probability. The claim follows. □

The aforementioned results establish that in high-dimensional settings—particularly those of practical interest—out-of-distribution (OOD) evaluation of learned algorithms is not only common, but *inevitable*. This thesis investigates two complementary strategies to address this challenge. Firstly in **Section 2.1**, we show how to reduce the effective dimensionality of a problem to circumvent the limitations imposed by bounds that become critical in the asymptotic regime where $d \gg 1$. In **Section 2.2**, we next develop a principled measure of 'out-of-distribution-ness' based on adversarial supervision, penalizing model predictions that exceed a tolerated error bound.

2.1. Interpretability as a Means to Generalizability

In this dissertation, we define a model as **interpretable-by-design** when its internal representations of the input space correspond to '*concepts*' that can be understood by humans. For example,

an interpretable-by-design regressor $f : \mathbb{R}^d \to \mathbb{R}$ might have the form

$$x \mapsto \phi(x) := [\phi_1(x), \phi_2(x), \ldots, \phi_k(x)] \mapsto \tilde{f}(\phi(x)) \tag{2.9}$$

where $\tilde{f}$ is the regressor that acts over the k-dimensional feature space $\phi(x)$ instead, and where each ϕ_j corresponds to a semantically meaningful concept. In practice, we often choose to restrict our hypothesis class such that $\tilde{f}$ is an easily inspected mapping—this allows human practitioners to audit and verify the learned relationships between features $\phi_j(x)$ and the dependent variable. Our core assumption is that interpretable-by-design models are aligned with human cognitive inductive biases—for example, categories and causal relations—that humans naturally use to generalize internal representations under covariate shift.

A common assumption made in the field of mechanistic interpretability is that the individual features ϕ_j are **monosemantic** and **semantically disentangled**, meaning that each learned feature corresponds to *exactly one distinct* semantic concept. Under this assumption, one can show that the learned features are better separated in the model latent representation space, yielding greater robustness and out-of-distribution performance Zhang et al. (2025a). For example, a well-known result from Bartlett and Mendelson (2003) places a bound on the empirical test error as a function of the decision boundary margin:

Theorem 4 (Misclassification Risk Bound (Theorem 21 from Bartlett and Mendelson (2003)))**.** *Suppose that the minimum decision margin bound is given by γ, and define $\hat{R}_{train}$ to be the empirical training classification error over a set of n points, and R_{test} to be the true test error. Then for every linear classifier*

$$R_{test} \leq \hat{R}_{train} + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n\gamma^2}}\right)$$

with probability $1 - \delta$.

We cite this result primarily for discussion—the proof of this result is provided by the original authors in Bartlett and Mendelson (2003). Informally, greater separability of classes in the feature

space (i.e., larger γ) can allow us to place a tighter bound on the true test error. While **Theorem 4** assumes a family of linear classifiers, prior works have extended similar results to model deep networks and regression models (Zhang et al., 2025a; Scherlis et al., 2022; Alain and Bengio, 2017; Lyu et al., 2022). Building off these prior work, we hypothesize that

Interpretable-by-design ML systems enable human-like compositionality in predictions, enabling better out-of-distribution generalization.

In **Chapters 3-4**, we show that interpretability allows us to construct ML systems with principled inductive biases aligned with clinician reasoning to improve their generalizability.

2.2. Adversarial Supervision of Black-Box Models

A key limitation in building interpretable models as described in the previous section is that it assumes we as machine learning practitioners have the agency to design and train deployed predictive models from scratch. However, in many real-world applications this is not a feasible assumption—many ML systems interact with sensitive data, are available only via remote procedure calls, or implemented in workflows where interpretability is unfeasible. In the most general setting, we can only think of ML models as static *black-box* systems where we only have access to a single prediction y given an input x. Under this framework, it is challenging (and in many cases impossible) to ensure robust generalizability of an ML model without any additional assumptions.

In this thesis, we therefore specialize to the setting where there is pre-existing data to learn from. This is often true in many real-world settings; for example, we might have observational data of patient outcomes from one hospital site, or a database of molecular sequences and their corresponding properties reported in prior scientific literature. In these cases, we will show how it is possible to leverage these static (and often imperfect) datasets as a form of additional supervision.

An important observation is that by having access to such a dataset $\mathcal{D}$, we can bound the empirical test error as a function of the error on the available dataset and the 1-Wasserstein distance between $\mathcal{D}$ and the test samples. Firstly, recall that the p-Wasserstein distance is defined as follows:

Definition 1 (p-Wasserstein Distance). *Define $(\mathcal{X}, \mathrm{d})$ to be a metric space and fix $p \in [1, \infty)$. For any two probability measures μ, ν over $\mathcal{X}$ with finite p-th moment, define the set of all couplings of μ, ν as*

$$\Gamma(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \ : \ \pi(\mathcal{A} \times \mathcal{X}) = \mu(\mathcal{A}),\ \pi(\mathcal{X} \times \mathcal{B}) = \nu(\mathcal{B})\}$$

*where $\mathcal{A}, \mathcal{B}$ are any Borel subsets of $\mathcal{X}$. Then the p-**Wasserstein distance** is defined by*

$$W_p(\mu, \nu) = \left[\inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d\pi(x, x') \, \mathrm{d}(x, x')^p \right]^{1/p}$$

For the remainder of this dissertation, we specialize to the $p = 1$ Wasserstein distance in a Euclidean metric space:

$$W_1(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d\pi(x, x') \, ||x - x'||_2$$

Given this definition, we obtain the following bound on the empirical test error:

Theorem 5 (Bound on Empirical Test Risk). *Define a real-valued, Borel-measurable function $f : \mathcal{X} \to \mathbb{R}$ defined over a domain $\mathcal{X} \subseteq \mathbb{R}^d$, and define $K := ||f(x)||_L$ to be the corresponding Lipschitz constant of f. Given a finite dataset of n observations $\mathcal{D} := \{(x_i, f(x_i))\}_{i=1}^n$, suppose we train a predictive model $\hat{f}$ on $\mathcal{D}$ with Lipschitz constant $K_{\hat{f}}$ finite such that the empirical training risk $\varepsilon := \mathbb{E}_{(x,y)\sim\mathcal{D}}|y - \hat{f}(x)|$ is finite. Then, the test risk on a new sample of T test inputs $\mathcal{T} = \{x_j\}_{j=1}^T$ is bounded from above by*

$$\mathbb{E}_{x\sim\mathcal{T}}|f(x) - \hat{f}(x)| \leq \varepsilon + (K + K_{\hat{f}}) W_1(\mu_{\mathcal{D}}, \mu_{\mathcal{T}}) \tag{2.10}$$

where $W_1(\mu_{\mathcal{D}}, \mu_{\mathcal{T}})$ is the 1-Wasserstein distance associated with $|| \cdot ||_2$.

Proof. Define $\gamma \in \Gamma(\mu_{\mathcal{D}}, \mu_{\mathcal{T}})$ as the optimal coupling between input observations x' and x in $\mathcal{D}$ and

$\mathcal{T}$, respectively. For pairs $(x', x) \sim \gamma$, note that

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{T}}|f(x) - \hat{f}(x)| &= \mathbb{E}_{(x',x)\sim\gamma}|f(x) - \hat{f}(x)| && \mu_{\mathcal{T}} \text{ is the } x\text{-marginal of } \gamma \\
&= \mathbb{E}_{(x',x)\sim\gamma}\left|\Big(f(x) - \hat{f}(x)\Big) - \Big(f(x') - \hat{f}(x')\Big) + \Big(f(x') - \hat{f}(x')\Big)\right| \\
&\leq \mathbb{E}_{(x',x)\sim\gamma}\left|\Big(f(x) - \hat{f}(x)\Big) - \Big(f(x') - \hat{f}(x')\Big)\right| && \text{Triangle inequality} \\
&\qquad + \mathbb{E}_{(x',x)\sim\gamma}\left|f(x') - \hat{f}(x')\right| \\
&= \mathbb{E}_{(x',x)\sim\gamma}\left|\Big(f(x) - \hat{f}(x)\Big) - \Big(f(x') - \hat{f}(x')\Big)\right| \\
&\qquad + \mathbb{E}_{x'\sim\mathcal{D}}\left|f(x') - \hat{f}(x')\right| && \mu_{\mathcal{D}} \text{ is the } x'\text{-marginal of } \gamma \\
&= \mathbb{E}_{(x',x)\sim\gamma}\left|\Big(f(x) - \hat{f}(x)\Big) - \Big(f(x') - \hat{f}(x')\Big)\right| + \varepsilon && \text{Definition of } \varepsilon \\
&= (K + K_{\hat{f}})\mathbb{E}_{(x',x)\sim\gamma}||x - x'||_2 + \varepsilon && \text{Definition of Lipschitz constants} \\
&= (K + K_{\hat{f}})W_1(\mu_{\mathcal{D}}, \mu_{\mathcal{T}}) + \varepsilon && \text{Definition of 1-Wasserstein bound}
\end{aligned}
$$

The claim follows. □

We remark that deriving the global Lipschitz bounds $K, K_{\hat{f}}$ is $\mathcal{NP}$-hard and infeasible in practice (Scaman and Virmaux, 2018; Hu et al., 2024). However, **Theorem 5** still holds if $K, K_{\hat{f}}$ only hold locally over a finite subset of $\mathcal{X}$ that contains $\mathcal{D} \bigcup \mathcal{T}$, which is much easier to derive. Furthermore, note that the constants ε and Lipschitz constants $K, K_{\hat{f}}$ in (2.10) are irreducible, since we assume that we do not have control over $\mathcal{D}$ or the functions $f, \hat{f}$. However, (2.10) also shows that bounding $W_1(\mu_{\mathcal{D}}, \mu_{\mathcal{T}})$ will yield a corresponding finite bound on the empirical test risk. **This is a key observation**—by intelligently choosing the test points in $\mathcal{T}$ that we use to evaluate with $\hat{f}$, we can guarantee a bound on the mean test error over $\mathcal{T}$.

In practice however, computing $W_1(\mu_{\mathcal{D}}, \mu_{\mathcal{T}})$ for real-world instances of $\mathcal{D}, \mathcal{T}$ is nontrivial. The challenge is in how the set of couplings Γ scales with the number of observations n in the dataset; classical algorithms that naïvely compute W_1 have a time complexity of $\mathcal{O}(n^3)$ (Pele and Werman, 2009). To overcome this limitation, we look to prior work (Arjovsky et al., 2017):

Lemma 1 (Kantorovich-Rubinstein Duality (Kantorovich and Rubinstein, 1958)). *Recall that the 1-Wasserstein distance between probability measures μ, ν over $\mathcal{X}$ is given by*

$$W_1(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d\pi(x, x') \, ||x - x'||_2$$

Then $W_1(\mu, \nu)$ can be equivalently written as

$$W_1(\mu, \nu) = \frac{1}{K} \sup_{||c||_L \leq K} \left\{ \mathbb{E}_{x\sim\mu}[c(x)] - \mathbb{E}_{x'\sim\nu}[c(x')] \right\}$$

where $|| \cdot ||_L$ is the Lipschitz norm.

The proof of this result is given in Kantorovich and Rubinstein (1958). Informally, the function $c : \mathcal{X} \to \mathbb{R}$ is a *source critic* function that learns to discriminate points sampled from μ and ν. If $c(x)$ is large (resp., small), then the point x is likely to have been sampled from μ (resp., ν). Arjovsky et al. (2017); Yao et al. (2024) demonstrate how to implement source critic functions as machine learning models—by training a neural network as an adversarial model that learns to discriminate between two distributions, we can compute (and therefore bound) the 1-Wasserstein distance between the aforementioned probability measures. In the Wasserstein GAN (WGAN) model proposed by Arjovsky et al. (2017), a generative network and source critic are co-trained in a minimax game where the generator (critic) seeks to minimize (maximize) the Wasserstein distance W_1 between the training and generated distributions. In this way, the generator can learn the distribution of training samples from nature—in our work, we extend this framework to the problem of generative *optimization* under distribution shift. Put simply, we hypothesize that

Adversarial source critic models can help us implement meaningful and computationally tractable bounds on the 1-Wasserstein distance, and therefore the empirical test risk.

In **Chapters 5-6**, we will show how such an approach can be used in **generative optimization problems**, where we have explicit control over the test set $\mathcal{T}$.

2.3. Constrained Optimization via Lagrangian Duality

In the latter half of this dissertation, we consider constrained optimization problems of the form

$$\begin{aligned} \text{minimize}_{x \in \mathcal{X}} \quad & f(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \quad \forall i \in \{1, \ldots, m\} \end{aligned} \tag{2.11}$$

given a set of m constraints. In general, satisfying any arbitrary set of (potentially nonlinear) constraints is challenging if not intractable, and it is often desirable instead to solve a related *unconstrained* optimization problem. One common mechanism to perform such a problem transformation is to define the *Lagrangian* of (2.11) as $\mathcal{L}(x; \vec{\lambda}) = f(x) + \left\langle \vec{\lambda}, \begin{bmatrix} f_1(x) & f_2(x) & \cdots & f_m(x) \end{bmatrix} \right\rangle$ where $\vec{\lambda} \in \mathbb{R}^m_+$ and $\mathcal{L} : \mathcal{X} \times \mathbb{R}^m_+ \to \mathbb{R}$. It can be shown (Boyd and Vandenberghe, 2004) that the constrained optimization problem in (2.11) is equivalent to the unconstrained problem

$$\text{minimize}_{x \in \mathcal{X}} \quad \text{maximize}_{\vec{\lambda} \in \mathbb{R}^m_+} \mathcal{L}(x; \vec{\lambda}) \tag{2.12}$$

in terms of the Lagrangian. The *dual problem* of (2.12) is constructed by reversing the order of the minimization and maximization problems:

$$\text{maximize}_{\vec{\lambda} \in \mathbb{R}^m_+} \quad \text{minimize}_{x \in \mathcal{X}} \mathcal{L}(x; \vec{\lambda}) = \text{maximize}_{\vec{\lambda} \in \mathbb{R}^m_+} g(\vec{\lambda}) \tag{2.13}$$

where we implicitly define the *dual function* $g(\vec{\lambda}) := \min_{x \in \mathcal{X}} \mathcal{L}(x; \vec{\lambda})$. In general, it is guaranteed that the optimal solution to the dual problem in (2.13) is a lower bound on the optimal solution of the original problem in (2.11) from weak duality; if $f(x)$ and $f_i(x)$ are convex and bounded from below such that Slater's condition applies, then strong duality guarantees that the optimal solutions to the dual and original problems are equal. Solving the dual optimization problem in (2.13) requires us to first solve for the dual function $g(\vec{\lambda})$, which is a challenging (often intractable) task in the most general case. In Yao et al. (2024), we approximate $g(\vec{\lambda})$ under specific assumptions on the search space $\mathcal{X}$. In Yao et al. (2025b), we show how our problem formulation admits an *exact* solution for the dual function $g(\vec{\lambda})$ (see **Lemma 6**).

CHAPTER 3

CLINICAL DECISION SUPPORT VIA GENERALIST LANGUAGE MODELS

Portions of this chapter are adapted from the following published first-author manuscript:

(Yao et al., 2025a) Michael S Yao, Allison Chae, Piya Saraiya, Charles E Kahn Jr., Walter R Witschey, James C Gee, Hersh Sagreiya†, and Osbert Bastani†. Evaluating acute image ordering for real-world patient cases via language model alignment with radiological guidelines. *Commun Med*, 5(332), 2025. doi: 10.1038/s43856-025-01061-9

Here, † denotes co-senior authorship. I helped conceive the study, planned and performed experiments, analyzed experimental data, and drafted the manuscript with input from all other authors.

3.1. Introduction

In this chapter, we demonstrate a method to build an interpretable computational clinical assistant that requires no additional model training of generalist large language models (LLMs). By constructing an interpretable-by-design predictive pipeline, we adapt consumer-grade LLMs for a challenging, real-world clinical task.

Ordering diagnostic imaging studies is an increasingly common task in the emergency department (ED) and other acute-care settings, and is associated with high cognitive burden for clinicians (Baloescu, 2018; Kwee et al., 2024; Litkowski et al., 2016; Salerno et al., 2019). While diagnostic imaging can play a crucial role in the acute workup of patients, ordering imaging studies with limited clinical utility are associated with increasing concerns regarding resource utilization, radiation exposure, and financial burden to both patients and healthcare systems (Francisco et al., 2024; Sadigh et al., 2022; Tung et al., 2017). Recent estimates suggest that up to 30% of diagnostic imaging studies ordered in the ED setting could be replaced with more appropriate alternatives at the time the order was placed (Francisco et al., 2024; Venkatesh et al., 2012).

Multiple factors contribute to the challenge of ordering appropriate and clinically indicated imag-

ing studies. Emergency medicine physicians often need to make rapid diagnostic decisions with limited clinical context while simultaneously managing high patient volumes and complex patient presentations (Lee et al., 2013; Pinto et al., 2016). Furthermore, there is significant variability between healthcare providers in imaging ordering patterns: recent studies have documented significant inter-physician differences in the utilization rates of different imaging studies, suggesting that factors beyond pure clinical necessity influence imaging decisions (Jameson et al., 2024; Valtchinov et al., 2019; Wintermark et al., 2016; Quinn et al., 2023; Young et al., 2020).

To help clinicians make more informed, evidence-based decisions in image ordering and simultaneously address inter-provider variability in imaging practices, the American College of Radiology (ACR) released the ACR Appropriateness Criteria®(ACR AC), which are a set of evidence-based guidelines that assist referring physicians in ordering the most appropriate diagnostic imaging studies for specific clinical conditions (of Radiology, 2024). As of June 2024, the ACR AC contains 224 unique imaging topics (i.e., patient scenarios).

However, despite the widespread availability of the ACR AC, low utilization of these guidelines remains a challenge in many emergency departments and inpatient settings (Bresnahan, 2010; Salerno et al., 2019). Bautista et al. showed that there is low utilization of the ACR AC by clinicians in practice: less than 1% of physicians interviewed in their study use the ACR AC as a first-line resource when ordering diagnostic imaging studies (Bautista et al., 2009; Taragin et al., 2003). The limited usage of the ACR AC may be partly due to how the Appropriateness Criteria are made accessible to clinicians; the evidence-based criteria are dense and can be difficult to parse through even for physician experts—especially in acute healthcare settings such as the emergency department where decision making is both time-sensitive and critical.

To address this problem, recent work has investigated the potential utility of generative artificial intelligence (AI) tools to synthesize dense passages of evidence-based guidelines to offer clinical decision support (CDS) in physician workflows (Barabucci et al., 2024; Goh et al., 2025, 2024; Labkoff et al., 2024; Nazario-Johnson et al., 2023; Zaki et al., 2024). In particular, large language models (LLMs) are generative AI models trained on large corpora of textual data to achieve im-

pressive performance on tasks such as language translation, summarization, and text generation (Chambon et al., 2023; Clusmann et al., 2023; Tay et al., 2024; Yan et al., 2022). However, LLMs can nonetheless struggle in challenging *domain-specific* tasks requiring human expertise and specialized training, such as in medicine, law, and engineering (Evans and Snead, 2024; Hager et al., 2024; Kawamura et al., 2022; Ong et al., 2024; Omiye et al., 2023; Malaviya et al., 2025). As a result, accessing the potential benefits of LLMs in these domains—such as for recommending appropriate imaging studies for patients—continues to be an ongoing challenge, deterring widespread adaptation of generative AI models in clinical medicine (Allen et al., 2021; Spotnitz et al., 2024).

Prior work has examined the ability of LLMs to rapidly process and contextualize large volumes of information could help transform the ACR Appropriateness Criteria into a more accessible, real-time clinical decision support tool. For example, Nazario-Johnson et al. (2023) and Zaki et al. (2024) evaluate the alignment of LLMs with the ACR Appropriateness Criteria; however, both studies leverage inputs that are not representative of the vernacular used in real-world clinical workflows. Other studies (Savage et al., 2024; Kim et al., 2024; Krithara et al., 2023; Jin et al., 2021; Rau et al., 2023) work with more realistic examples of real-world patient descriptions; however, these LLM inputs either (1) assume that all relevant medical information is provided to make a diagnostic decision; or (2) are phrased as multiple choice questions. Neither of these characteristics are representative of how clinicians might use LLMs for clinical decision support in practice, especially in acute emergency medicine settings that are notably characterized by the lack of a complete patient information. Finally, Liu et al. (2025a); Zhang et al. (2024, 2025b,c); and Singhal et al. (2023) introduce a number of performant models for medical tasks; however, these models again assume access to a relatively complete picture of the patient's clinical status and past medical history, which is rarely the case for acutely presenting patients in the emergency room.

In this work, we investigate how state-of-the-art LLMs can be used as CDS tools to help clinicians order guideline-recommended imaging studies according to the ACR AC. In **Figure 3.1**, we first motivate this problem by demonstrating how a state-of-the-art language model, such as Claude Sonnet-3.5, fails to accurately recommend diagnostic imaging studies that align with the ACR AC

for a variety of input patient descriptions. Given these initial findings, we hypothesize that while LLMs may struggle to directly recommend imaging studies for patients (a domain-specific task), they may be able to accurately describe patient conditions and presentations (phrased as ACR AC Topics) (Singhal et al., 2023; Williams et al., 2024). In this light, we apply LLMs to analyze patient case summaries and map them to topic categories from the ACR AC (**Fig. 1.2**). These ACR AC topics are our interpretable representations of the input patient case summaries as in (2.9). We represent these case summaries as "patient one-liners," which are concise summaries of patient presentations commonly used by clinicians to communicate relevant details quickly to other healthcare providers (Arman, 2023; Zussman et al., 2024). Importantly, our dataset of patient one-liners is representative of both the vernacular and limited patient context available in real-world text written by clinicians. Given a patient one-liner, we can then programmatically query the ACR AC based on the LLM-recommended topic category (without any explicit LLM usage) to determine the optimal imaging study for a patient. In this fashion, LLMs can be used to recommend diagnostic imaging studies according to recommendations from the guidelines.

We first introduce **RadCases**, a publicly available dataset of one-liners labeled by the most relevant ACR AC Panel and Topic. Second, we evaluate publicly available LLMs on our RadCases dataset to characterize how existing tools may be used out-of-the-box for diagnostic imaging support in inpatient settings, and show that generalist models such as Claude Sonnet-3.5 and Meta Llama 3 can accurately predict ACR AC Topic labels given patient one-liners. We then assess how popular techniques such as model fine-tuning (MFT), retrieval-augmented generation (RAG), in-context learning (ICL), and chain-of-thought prompting (COT) may be effectively leveraged to improve the alignment of existing LLMs with ACR AC, and also enable LLMs to outperform clinicians in the accuracy of ordered imaging studies in a retrospective study (Kresevic et al., 2024; Sivarajkumar et al., 2024). Finally, we conduct a prospective clinical study to show that LLM clinical assistants can improve clinician image ordering accuracy in simulated acute care environments.

3.2. Materials and Methods

3.2.1. RadCases Dataset Construction

Prior work in medical natural language processing has primarily focused on tasks such as documentation writing, medical question answering, and chatbot-clinician alignment. In each of these tasks, a relatively complete picture including hospital course, lab values, and advanced image studies of a patient presentation is often available. This is *not* representative of the limited patient history to guide acute image ordering in the emergency room. To best simulate decision-making contexts with limited patient information available, we first needed to curate a dataset of patient scenario descriptions—or "one-liners"—and corresponding ground-truth labels.

To build the **RadCases Dataset**, we leveraged five publicly available, retrospective sources of textual data. Firstly, we prompted the GPT-3.5 (`gpt-3.5-turbo-0125`) LLM from OpenAI to generate 16 **Synthetic** patient cases with a chief complaint related to each of the 11 particular ACR AC Panels related to diagnostic radiology. We also introduced the Medbullets patient cases consisting of challenging United States Medical Licensing Examination (**USMLE**) Step 2- and 3- style cases introduced by Chen et al. (2025a). The original Medbullets dataset consisted of paragraph-form patient cases accompanied by a multiple-choice question; to convert each question to a patient one-liner, we used the first sentence of each patient case.

Similarly, we leveraged the **JAMA** Clinical Challenge and **NEJM** Case Record datasets that include challenging, real-world cases published in the Journal of the American Medical Association (JAMA) and the New England Journal of Medicine (NEJM), respectively. These patient cases are often complex enough to be published as resources for the broader medical community. The JAMA Clinical Challenge (resp., NEJM Case Record) dataset was initially introduced by Chen et al. (2025a) (resp., Savage et al. (2024)); we follow the same protocol as for the Medbullets dataset described above to convert these document-form patient cases into patient one-liners.

Finally, we sought to evaluate LLMs on patient summaries written by clinicians in a real-world emergency department. We constructed the **BIDMC** dataset from anonymized, de-identified pa-

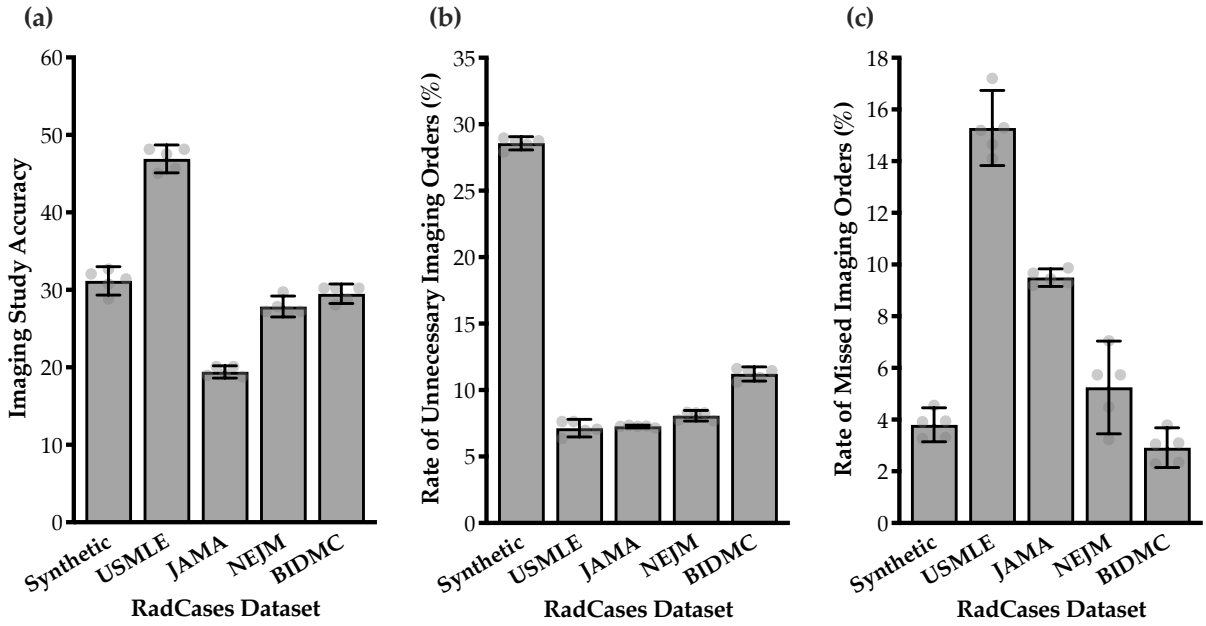

Figure 3.1: **LLMs struggle with diagnostic imaging ordering.** We evaluate Claude Sonnet-3.5, a state-of-the-art language model, on its ability to order imaging studies given an input patient case description, or "one-liner." The LLM is evaluated on five representative subsets of the RadCases dataset introduced in our work. To demonstrate the difficulty of ordering diagnostic imaging studies in practice, we show that (**a**) Claude Sonnet-3.5 frequently orders imaging studies that are not aligned with the ACR Appropriateness Criteria. (**b**) The language model also frequently orders unnecessary imaging studies, and (**c**) can incorrectly forego imaging even when it is clinically warranted. In our work, we introduce an LLM inference strategy to significantly improve the performance of language models according to these important clinical metrics. Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

tient admission notes introduced by Johnson et al. (2023) in the MIMIC-IV dataset by taking the first sentence of each clinical note as the patient one-liner. Briefly, the original MIMIC-IV dataset includes electronic health record data of patients from the Beth Israel Deaconess Medical Center (BIDMC) admitted to either the emergency department or an intensive care unit (ICU) between 2008 and 2019 (Johnson et al., 2023). We restrict our constructed one-liner dataset to those from the discharge summaries of a subset of 100 representative patients.

A patient one-liner was excluded if any of the following exclusionary criteria applied: (1) the ACR AC did not provide any guidance for the chief complaint (e.g., a primary dermatologic condition); (2) an appropriate imaging study was performed and/or a diagnosis was already made; (3) the

one-liner did not include sufficient information about the patient; or (4) the one-liner did not refer to a specific patient presentation (e.g., one-liners extracted from epidemiology-related USMLE practice questions). Of the original 2,513 patient cases, a total of 914 (36.3%) cases were excluded due to the above criteria (719 excluded cases due to criteria (1); 90 due to criteria (2); 49 due to criteria (3); and 56 due to criteria (4)). All 1,599 remaining one-liners in our final dataset were individually reviewed to be representative of true clinical one-liners in practice by one U.S. attending radiologist and two U.S. medical students.

3.2.2. Model Evaluation

Our curated RadCases dataset consists of 1,599 patient one-liner scenarios constructed from five different sources representing a diverse panel of patient presentations and clinical scenarios: (1) RadCases-Synthetic (156 out of the 1,599 total patient cases); (2) RadCases-USMLE (170 patient cases); (3) RadCases-JAMA (965 patient cases); (4) RadCases-NEJM (163 patient cases); and (5) RadCases-BIDMC (145 patient cases).

Our next task was to annotate ground-truth labels to each of the patient scenarios in the RadCases dataset. An intuitive ground-truth label might be to assign a single "best" imaging study (or lack thereof) to order for each patient scenario. However, such a singular ground-truth label is often non-existent: imaging studies can vary largely by clinician preference—even amongst expert physicians (Derbas et al., 2021; Guenette et al., 2024; Hughes et al., 2015)—and available hospital resources. Furthermore, the ultimate goal of the RadCases dataset is to align LLMs with evidence-based guidelines for image ordering; labels of imaging studies alone arguably contain weak, implicit signals on the underlying guidelines that dictate the "best" imaging study. We therefore labelled the patient one-liners according to the most relevant ACR Appropriateness Criteria Topic. As of June 2024, there are 224 possible diagnostic radiology Topic labels.

To characterize the out-of-the-box alignment of language models with the ACR Appropriateness Criteria, we evaluated 6 state-of-the-art LLM models on their ability to predict the most relevant ACR AC Topic given an input patient case summary (**Fig. 3.2**): (1) **DBRX** Instruct (`databricks/dbrx-instruct`) from Databricks is an open-source mixture-of-experts (MoE) model with a total of

132B parameters (Team, 2024); (2) **Llama 3** 70B Instruct (`meta-llama/Meta-Llama-3-70B-Instruct`) from Meta AI is an open-source LLM with 70B total parameters (AI, 2024b); (3) **Mistral** 8x7B Instruct (`mistralai/Mixtral-8x7B-Instru ct-v0.1`) from Mistral AI is an open-source sparse MoE model with 47B parameters (Jiang et al., 2024); (4) **Command R+** (`CohereForAI/c4ai-command-r-plus`) from Cohere for AI is an open-source retrieval-optimized model with 104B total parameters (AI, 2024a); (5) **GPT-4** Turbo (`gpt-4-turbo-2024-04-09`) from (OpenAI et al., 2024); and (6) **Claude Sonnet-3.5** (`anthropic.claude-3-5-sonnet-20240620-v1:0`) from (Anthropic, 2024) are proprietary LLMs with confidential model sizes.

Labelling one-liners by ACR Appropriateness Criteria Topics. In alignment with this plan, two fourth-year U.S. medical students—supervised by two attending radiologists—manually annotated all RadCases scenarios according to the ACR AC Topic that best describes the patient case. The two medical students and radiologists discussed cases where there was disagreement between proposed annotations, and the attending radiologists' decision was final. In scenarios where multiple ACR AC Topics might apply to a single patient case, the more acute, life-threatening scenario was used as the ground-truth label. Patient cases that were not well-described by any of the available ACR AC Topics were excluded from the dataset.

Evaluation metrics of language models according to the ACR Appropriateness Criteria. In our experiments, we are interested in evaluation metrics that help us elucidate the performance of LLMs as clinical decision support tools. We detail these metrics and how they are calculated below.

An LLM's *accuracy* is a score between 0 and 1. We evaluate two accuracy metrics in our experiments: Topic Accuracy (**Figs. 3.2-3.3**) and Imaging Accuracy (**Figs. 3.4-3.6**). For a given input patient case with ground truth ACR AC Topic y and model prediction y_{pred}, the Topic Accuracy is defined as the binary indicator variable equal to 1 if $y = y_{\text{pred}}$ and 0 otherwise. Separately, suppose that according to the ACR AC, the ground truth Topic y is associated with the set of clinically appropriate studies $\mathcal{K}$, and the model-predicted Topic y_{pred} is associated with the set of clinically appropriate studies

$\mathcal{K}_{\text{pred}}$. The Imaging Accuracy is then defined as

$$\text{Imaging Accuracy}\left(\mathcal{K}_{\text{pred}}, \mathcal{K}\right) = \frac{\left|\mathcal{K}_{\text{pred}} \cap \mathcal{K}\right|}{\left|\mathcal{K}_{\text{pred}}\right|}$$

Using the same notation as above, the *rate of unnecessary imaging studies* (i.e., false positive rate) ordered by an LLM is a score between 0 and 1 defined as the frequency of evaluated patient cases where (1) the ground truth set of appropriate studies $\mathcal{K}$ is identically equal to {`No Imaging`}; and (2) `No Imaging` is not a member of $\mathcal{K}_{\text{pred}}$. Similarly, the *rate of missed imaging studies* (i.e., false negative rate) ordered by an LLM is a score between 0 and 1 defined as the frequency of evaluated patient cases where (1) `No Imaging` is not a member of $\mathcal{K}$; and (2) $\mathcal{K}_{\text{pred}} =$ {`No Imaging`}. Finally, the *F_1 score* of an LLM is defined as $F_1 = \frac{2 \cdot TP}{2 \cdot TP + (FP + FN)}$, where TP is the number of patient cases where the LLM orders an imaging study that is clinically indicated, FP is the number of patient cases where the LLM orders an unnecessary study, and FN is the number of patient cases where the LLM incorrectly fails to order an imaging study according to the guidelines.

Importantly, we highlight that the construction of sets $\mathcal{K}$ and $\mathcal{K}_{\text{pred}}$ from the Topic labels y and y_{pred} are deterministically constructed and do not involve *any* LLM queries; instead, we use a custom Python (Python Software Foundation) web-scraping script with the Beautiful Soup (Leonard Richardson) open-source library to define each set of appropriate imaging studies for all Topics in the ACR AC from the URL `https://gravitas.acr.org/acportal`.

3.2.3. Optimization of Zero-Shot Prompt Engineering and Fine-Tuning Methods

In **Figure 3.3**, we explore 4 distinct LLM optimization strategies—retrieval-augmented generation (RAG), in-context learning (ICL), chain-of-thought (COT) prompting, and model fine-tuning (MFT)—to improve the ability of LLMs like Claude Sonnet-3.5 and Llama 3 to accurately predict relevant ACR AC Topics from input patient one-liner scenarios.

In our RAG approach, we first constructed the relevant reference corpus of guidelines made publicly available by the American College of Radiology (ACR). A link to our custom script implementation is made publicly available at this URL. Using a custom Python script included in our publicly

available code, we first used a web scraper, in compliance with the ACR Terms and Conditions, to download relevant Portable Document Format (PDF) narrative files from acsearch.acr.org/list on July 17, 2024. Each ACR AC Topic includes exactly one accompanying narrative document, resulting in a total of 224 narrative files extracted. We then used the Unstructured IO open-source library to extract the PDF content into raw text, and chunked the text into 3,380 disjoint corpus documents with sizes ranging between 1,119 and 2,048 characters per document. Our strategy for constructing the retrieval corpus is identical to that used by Xiong et al. (2024).

Using this corpus of relevant guidelines written by the ACR, we explored 8 different retriever algorithms to use for RAG: (1) **Random**, which randomly retrieves k corpus documents over a uniform probability distribution; (2) Okapi **BM25** bag-of-words retriever (Robertson and Zaragoza, 2009); (3) **BERT** (Devlin et al., 2019) and (4) **MPNet** (Song et al., 2020) trained on unlabeled, natural language text; (5) **RadBERT** (Yan et al., 2022) from fine-tuning BERT on radiology text reports; (6) MedCPT (Jin et al., 2023) leveraging a transformer trained on PubMed search logs; and (7) **OpenAI** (`text-embedding-3-large`) and (8) **Cohere** (`cohere.embed-english-v3`) embedding models from OpenAI and Cohere for AI, respectively. Retrievers (3) - (8) are embeddings-based retrievers that leverage cosine similarity as the ranking function. These 8 retrievers represent a diverse array of novel, well-studied, domain-agnostic, and domain-specific retrievers for RAG applications. In **Figure 3.3b-c**, we report the results using the best retriever specific to each language model and RadCases dataset subset, fixing the number of retrieved documents to $k = 8$ for each retriever. We include the experimental results for each individual retriever in **Supp. Figure A.3**.

Separately in our ICL approach, we use the RadCases-Synthetic dataset partition as the corpus of examples to retrieve from, and experimentally validate the same 8 retrievers used in RAG for retrieving relevant one-liner/ACR AC Topic pairs to provide as context to the language model. In **Figure 3.3b-c**, we report the results using the best retriever specific to each language model and RadCases dataset subset, fixing the number of retrieved examples to $k = 4$ for each retriever. To evaluate language models on the RadCases-Synthetic dataset using ICL, we constructed a separate corpus of synthetically generated, annotated one-liners to retrieve from that was created using the

identical prompting strategy as that for the RadCases-Synthetic dataset, except we used the Meta Llama 2 (7B) model (`meta-llama/Llama-2-7b-chat-hf`). We leveraged this separate corpus for in-context learning to avoid data leakage in our RadCases-Synthetic ICL evaluation experiments. We include the experimental results for each individual retriever in **Supp. Figure A.4**, and ablate the number of retrieved examples in **Supp. Figure A.5**.

In COT prompting, we explore four different reasoning strategies identical to those employed by Savage et al. (2024): (1) **Default** reasoning, which does not specify any particular reasoning strategy for the LLM to use; (2) **Differential** diagnosis reasoning, which encourages the model to reason through a differential diagnosis to arrive at a final prediction; (3) **Bayesian** reasoning, which encourages the model to approximate Bayesian posterior updates over the space of ACR AC Topics based on the clinical patient presentation; and (4) **Analytic** reasoning, which encourages the model to reason through the pathophysiology of the underlying disease process. We include the experimental results for each individual reasoning strategy in **Supp. Figure A.6**. In **Figure 3.2b-c**, we report the results using the best COT reasoning strategy specific to each language model and RadCases dataset subset. In **Figures 3.3-3.6**, we report results using the **Default** reasoning strategy when COT is leveraged together with Claude Sonnet-3.5.

For MFT, we explore three different fine-tuning strategies using the Meta Llama 3 base model: (1) **Full** fine-tuning where all the parameters of the LLM are updated; and (2) Low-Rank Adaptation (**LoRA**) (Hu et al., 2022) and (3) Quantized Low-Rank Adaptation (**QLoRA**) (Dettmers et al., 2024) fine-tuning where only the subset of linear LLM parameters are updated. We fix the number of training epochs to 3 and the learning rate to 0.0001. For LoRA (resp., QLoRA), we use a rank of 64 (resp., 512) and an α scaling value of 8 (resp., 8). We chose these particular values according to a hyperparameter grid search over the rank and α hyperparameters, logarithmically ranging from 8 to 512 (resp., 1 to 512), that maximize the accuracy of the fine-tuned model on a synthetic validation dataset. Due to limitations on local compute availability, we were only able to run the QLoRA fine-tuning experiments on the internal experimental cluster; LoRA and Full fine-tuning experiments were performed using a third-party platform (Together AI). Finally, we also investigate two

different fine-tuning datasets for each of the three strategies: (1) fine-tuning on the RadCases-Synthetic dataset; and (2) fine-tuning on 250 cases where 50 random cases come from each of the five RadCases dataset subsets. To prevent data leakage, we use the Llama 2-generated Synthetic dataset (constructed for a similar purpose in our ICL experiments above) to fine-tune the base Llama 3 model for evaluation on the RadCases-Synthetic dataset in strategy (1), and avoid evaluation on any cases from the individual patients represented in the fine-tuning dataset in strategy (2). In **Figure 3.3c**, we report the results using the LoRA fine-tuning strategy and "mixed" fine-tuning dataset of 250 cases described above, as this led to consistently superior fine-tuning results across all datasets and language models that were evaluated. We report additional experimental results in **Supp. Figures A.8-A.9**.

3.2.4. Translating ACR AC Topics Into Imaging Study Recommendations

In **Figure 3.4a**, we overview our Evidence-Based inference pipeline where we leverage LLMs to assign ACR AC Topics to input patient one-liner scenarios, and then deterministically map Topics to appropriate imaging studies based on the Appropriateness Criteria guidelines. These LLM-generated recommendations were used as the basis of our retrospective and prospective studies described in our work. Determining this mapping of Topics to imaging studies is a non-trivial task: for any particular Topic, there are often multiple, nuanced clinical variants that are described the ACR AC. For example, for the "Suspected Pulmonary Embolism" Topic, there are 4 variants in the guidelines as of June 2024:

1. Suspected pulmonary embolism. Low or intermediate pretest probability with a negative D-dimer. Initial imaging.

2. Suspected pulmonary embolism. Low or intermediate pretest probability with a positive D-dimer. Initial imaging.

3. Suspected pulmonary embolism. High pretest probability. Initial imaging.

4. Suspected pulmonary embolism. Pregnant patient. Initial imaging.

Each of these variants have different imaging recommendations: for example, variant (1) does not warrant any imaging study according to the ACR AC, whereas both computed tomography angiography (CTA) pulmonary arteries with intravenous (IV) contrast and a ventilation-perfusion (V/Q) scan lung are appropriate studies for variant (3). To define a deterministic mapping of topics to imaging studies, we therefore needed to isolate a single variant for each topic.

Our research team manually parsed through each of the 224 Topics to determine this single variant. In general, the process involved reverse engineering a "typical" patient presentation that would be described by a given Topic. In the above example, we reasoned that an acutely presenting patient where the most relevant Topic is "Suspected Pulmonary Embolism" would likely have a high pretest probability for a pulmonary embolism. Furthermore, pregnant patients are less common than non-pregnant patients in the emergency room, and the appropriate imaging studies for variant (3) are also appropriate for variant (4). For this reason, variant (3) was kept and the rest were discarded. As a result, a predicted imaging study of either CTA pulmonary arteries with IV contrast or V/Q scan lung were both considered correct answers in this example. If no imaging study was considered appropriate according to a guideline, then the ground-truth label was 'None.'

3.2.5. Retrospective Study on Autonomous Image Ordering Using LLMs

To power our retrospective study comparing language models with clinicians, we extracted a diverse sample of 242 de-identified admission notes derived from the MIMIC-IV dataset made available by Johnson et al. (2023). These notes were extracted from the medical records of 100 real patient admissions between 2008-2019 from the Beth Israel Deaconess Medical Center (Boston, MA). To account for the limited patient information available in acute presentations, we manually truncated the admission notes to only include relevant patient history and vitals. Admission notes were excluded from our analysis if either (1) the ACR Appropriateness Criteria contained no evidence-based guidance relevant to the patient scenario; or (2) the scenario described a patient admission that was not made in the emergency department (e.g., ICU downgrade to hospital floor). A total of 141 final patient scenarios were included in our analysis.

Using these patient scenarios, we prompted language models to predict up to m ACR AC Topics

that may be relevant for a given patient, and programmatically referenced the ACR AC guidelines to determine the recommended imaging studies based on the LLM-recommended Topic(s). We set $m = 1$ Topic in our experiments and evaluated two LLMs from our original RadCases evaluation suite: Claude Sonnet-3.5 from Anthropic AI using chain-of-thought (COT) prompting, and Llama-3 70B Instruct from Meta AI using no special prompt engineering (we further ablate the value of m in **Supp. Figure A.10**). We chose to evaluate these two models because they were the best performing proprietary and open-source models on the RadCases benchmarks, respectively (**Fig. 3.2b**). Simultaneously, we manually parsed through each of the full, original discharge summaries to determine what imaging study(s) where ordered by the patient's physician. The imaging studies ordered by both clinicians and language models were compared against the ground-truth best imaging study(s) as determined by consensus between two expert radiologists and two fourth-year U.S. medical students at the University of Pennsylvania.

3.2.6. Prospective Study on LLMs as Clinical Assistants

Constructing the patient cases for prospective user evaluation study. To enable our prospective evaluation of LLMs as clinical decision support tools for clinicians, we first constructed a separate dataset of 50 patient one-liners derived from the RadCases BIDMC one-liners. The initially redacted details such as patient name, age, or gender were manually replaced with fictitious name, age, and/or gender values. The cases were then reviewed and edited by three separate attending physicians to ensure that the cases were representative of typical real-world patient cases.

Participant recruitment and compensation. In our work, we conducted a clinical study with U.S. senior medical students and emergency medicine physicians to evaluate whether LLMs can serve as helpful assistants in deciding what imaging studies to order. Participants for this prospective study were recruited from the Perelman School of Medicine and the Hospital of the University of Pennsylvania where this study was conducted. We provided a monetary incentive of $50 USD to each opt-in, volunteer study participant, and the top 50% most accurate medical students and physicians (scored separately) within each treatment arm were compensated with an additional $10 USD to incentivize participants to perform to the best of their ability. Following prior work (Bickman et al., 2021; Dutz et al., 2023; Garland et al., 2021; Halpern et al., 2021), we chose to offer

this monetary compensation to improve recruitment rates and increase the diversity of opt-in participants, especially given the fact that our study posed minimal risk to the participants. A total of 23 medical students and 7 resident physicians participated in our experiment; all participating medical students were required to have passed and completed the emergency medicine clinical rotation at the University of Pennsylvania to participate in this study.

Participant task. Study participants were each tasked with ordering up to 1 diagnostic imaging study for a standardized set of 50 simulated patient case descriptions derived from the MIMIC-IV dataset (Johnson et al., 2023). Each case was presented on a custom-built website interface to display one patient case at a time; a visual of the custom interface is shown in **Supp. Figure A.11**. For each case, participants selected an imaging study from a dropdown menu containing an alphabetized list of all 1150 diagnostic imaging studies officially recognized in the ACR Appropriateness Criteria. Of the 50 simulated cases, a random subset of 25 cases was chosen at the per-participant level that also showed LLM-generated recommendations for the participant to consult. Study participants were allowed to consult any online resources that they would typically use in evaluating patients in the emergency department, but were not allowed to consult any other individuals for assistance. In some simulated patient cases, more than one correct answer may be possible—participants were instructed to select just one of those possible answers in these cases.

Separately, study participants were also asked to complete a 5-question multiple-choice survey asking questions about their prior experience with AI tools, and overall sentiment about the use of AI in medicine (**Supp. Table A.10**). All study participant answers to this short survey and the overall prospective study were anonymized and aggregated before analysis; participants were informed of this anonymization strategy in the informed consent.

3.2.7. Experimental Evaluation and Statistical Analysis

All models and prompting techniques were evaluated on a single internal cluster with 8 NVIDIA RTX A6000 GPUs. The temperature of all models was set to 0 to minimize variability in the model outputs. Each experiment was run using 5 random seeds, and we computed the mean accuracy of each method with 95% confidence intervals (CIs) against the human-annotated ground truth la-

bels. A p-value of $p < 0.05$ was used as the threshold for statistical significance. In all figures, "n.s." represents not significant (i.e., $p \geq 0.05$); a single asterisk $p < 0.05$; double asterisks $p < 0.01$, and triple asterisks $p < 0.001$. All statistical analyses were performed using Python software, version 3.10.13 (Python Software Foundation), the SciPy package, version 1.14.0 (Enthought) (Virtanen et al., 2020), and the PyFixest package, version 0.24.2 (Bergé, 2018).

3.3. Results

3.3.1. RadCases: A Dataset for Evaluating LLM Alignment with the ACR Appropriateness Criteria

Prior work evaluating LLMs for medical use cases have primarily relied on datasets that either contain complete pictures of patient presentations and outcomes (Kim et al., 2024; Savage et al., 2024; Williams et al., 2024; Xiong et al., 2024) or are not representative of how clinicians discuss acute patient presentations in practice (Nazario-Johnson et al., 2023; Zaki et al., 2024). As a result, such existing datasets cannot help us adequately interrogate the ability of LLMs to take "natural medical text" written by clinicians as input, and produce imaging recommendations that are aligned with the ACR Appropriateness Criteria. To address this limitation, we constructed the RadCases dataset, a labelled dataset of approximately 1,500 patient case descriptions that mimic the structure of one-liner patient scenarios contained in medical documentation written by clinicians. The RadCases dataset is partitioned into 5 subsets: (1) Synthetic; (2) USMLE; (3) JAMA; (4) NEJM; and (5) BIDMC—the source and construction of each subset is detailed in the **Section 3.2**. Each textual description is labelled by the most appropriate ACR Appropriateness Criteria guideline Topic that is most relevant to the patient case as determined by a consensus panel between U.S. attending radiologists and medical students. As an example, the input one-liner "49M with HTN, IDDM, HLD, and 20 pack-year smoking hx p/w 4 mo hx SOB and non-productive cough" is labelled with the ACR AC Topic "Chronic cough."

Neurologic topics were the most common label in all 5 RadCases subsets, followed by cardiac and gastrointestinal conditions (**Supp. Fig. A.1**, **Supp. Table A.1**). We also found that our RadCases patient case descriptions were representative of real-world patient one-liners previously written by physicians in acute clinical workflows (**Supp. Table A.2**). Of note, while there were 224 unique

diagnostic imaging topics in the ACR AC (as of June 2024 when this study was conducted), only 161 (71.9%) of all topics had nonzero support in the dataset. Furthermore, 73 (32.6%) unique topics are represented in the Synthetic dataset; 61 (27.2%) in the USMLE dataset; 119 (53.1%) in the JAMA dataset; 70 (31.3%) in the NEJM dataset; and 47 (21.0%) in the BIDMC dataset.

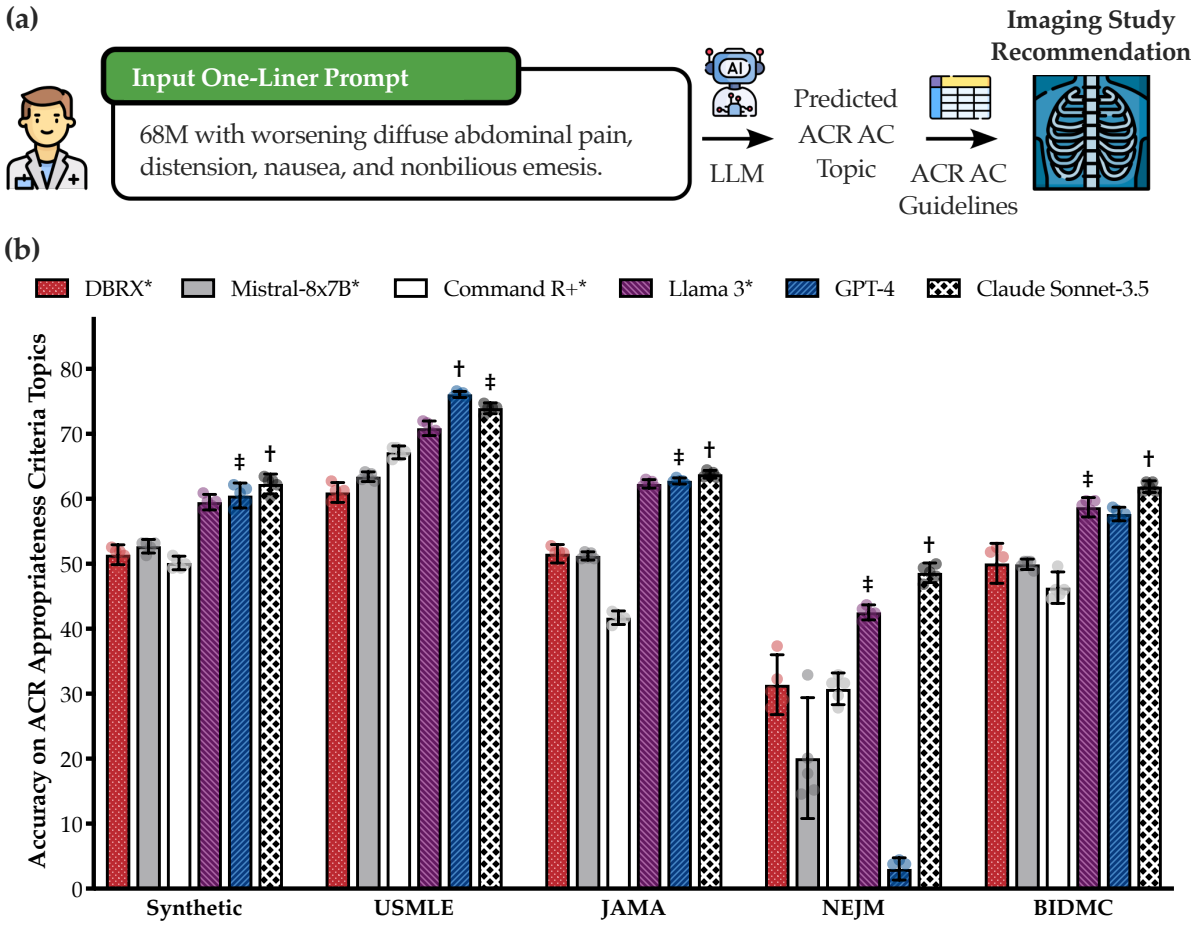

Figure 3.2: **Baseline LLM performance on the RadCases dataset.** (**a**) We query a language model to return the most relevant diagnostic radiology ACR AC Topic given an input patient one-liner description. We then query the ACR AC to return the most appropriate diagnostic imaging study (or lack thereof) given the predicted topic. (**b**) We evaluate six language models on their ability to correctly identify the ACR AC Topic most relevant to a patient one-liner. Open-source models are identified by an asterisk, and the best (second best) performing model for a RadCases dataset partition is identified by a dagger (double dagger). Error bars are $\pm$95% CI over $n = 5$ runs.

Using our RadCases dataset, we sought to evaluate if LLMs could yield better imaging study predictions if evidence-based guidelines were included as an explicit module in the patient scenario-

imaging study inference pipeline. If a language model classified patient scenarios to a specific guideline (i.e., a Topic of the ACR AC), then the best imaging study would be deterministically identified by the content of the guideline itself. More concretely, we looked to query LLMs to map input one-liners to output ACR Appropriateness Criteria Topics, and then programmatically map these Topics to their corresponding evidence-based imaging recommendations (**Fig. 3.2a**).

3.3.2. Evaluating Large Language Models on the RadCases Dataset

Figure 3.2b shows the performance of the LLMs evaluated on each of the RadCases dataset subsets. Of the language models evaluated, Claude Sonnet-3.5 performs the best on 4 out of the 5 subsets (i.e., Synthetic, JAMA, NEJM, and BIDMC) and the second best on the remaining subset (i.e., USMLE). Furthermore, Claude Sonnet-3.5 outperformed all open-source models with statistical significance (two-sample, two-tailed homoscedastic t-test; Synthetic $p = 0.0037$; USMLE $p = 0.0003$; JAMA $p = 0.0016$; NEJM $p < 0.0001$; BIDMC $p = 0.0010$). Separately, Llama 3 outperformed all other evaluated open-source models across all 5 RadCases subsets (two-sample, two-tailed homoscedastic t-test; $p < 0.0002$ for all 5 subsets). Based on these results, we chose to further optimize Claude Sonnet-3.5 and Llama 3 in subsequent experiments as the most promising overall and open-source large language models, respectively.

3.3.3. Optimizing Large Language Models for Imaging Ordering in Acute Clinical Workflows

While Claude Sonnet-3.5 and Llama 3 demonstrated impressive baseline accuracy on the RadCases dataset, recent work have introduced techniques to improve the performance of generative language models. For example, retrieval-augmented generation (RAG) provides relevant context to language models retrieved from an information corpus (i.e., the ACR AC narrative medical guidelines written by expert radiologists) to help improve the generative process. In-context learning (ICL) provides relevant examples of patient one-liners and their corresponding topic labels (i.e., examples from the RadCases-Synthetic dataset) as relevant context to improve the zero-shot performance of language models. Chain-of-thought (COT) prompting improves the complex reasoning abilities of language models by encouraging sequential, logical steps to arrive at a final answer. Finally, model fine-tuning (MFT) directly updates the parameters of a language model to improve

its performance on a specific task. We assess these strategies using Llama 3, and the zero-shot strategies RAG, ICL, and COT using Claude Sonnet-3.5 as there is no publicly available application programming interface (API) to fine-tune the proprietary model as of June 2024 (**Fig. 3.3a**).

Figure 3.3b shows that COT (chain-of-thought prompting) is the most effective strategy for Claude Sonnet-3.5, resulting in improvements of up to 17% in ACR AC Topic classification accuracy and consistent improvements across all five RadCases dataset subsets (two-sample, one-tailed homoscedastic t-test; $p < 0.0001$ for all subsets). Interestingly, this same strategy does not translate well to Llama 3 (**Fig. 3.3c**); COT marginally improves upon baseline prompting for Llama 3 only on the USMLE RadCases dataset. Instead, ICL (in-context learning) was the most effective prompt engineering strategy for Llama 3, resulting in improvements of up to 9% on ACR AC Topic classification accuracy compared with naïve prompting (two-sample, one-tailed homoscedastic t-test; $p < 0.0001$ for USMLE and NEJM datasets). Additional fine-grained optimization results are included in **Supp. Figures A.3-A.9**.

Our results show that while prompt engineering and other optimization techniques can indeed be effective in improving the performance of different language models on this task, the trends in improvements can be LLM-specific and fail to generalize across different language models. This finding highlights the inherent challenge in optimizing such models for challenging tasks such as diagnostic image ordering via alignment with the ACR AC.

3.3.4. Validating the LLM Prediction Pipeline

In **Figure 3.2b**, we demonstrated that LLMs could achieve promising accuracy on the ACR AC Topic classification task; in **Figure 3.3b-c**, we further optimized two state-of-the-art language models using prompt engineering techniques and model fine-tuning. Based on these results, we sought to validate our original hypothesis and evaluate whether assigning ACR AC Topic predictions to patient one-liners can meaningfully improve LLM diagnostic image study ordering.

We first mapped the ground-truth ACR AC Topic labels in the RadCases dataset to the ground-truth imaging study recommended by the relevant Topic guidelines. We evaluated 3 pipelines

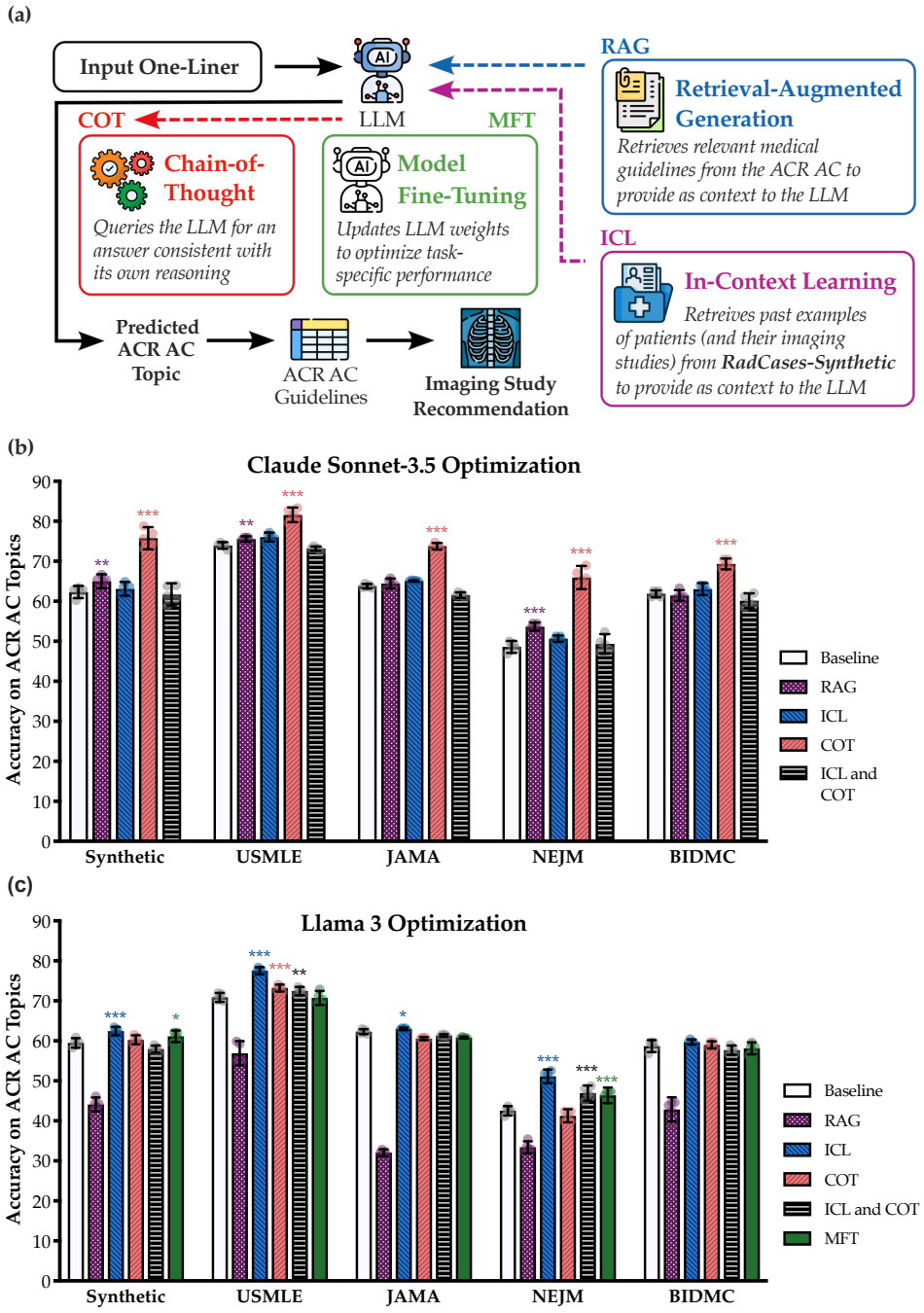

Figure 3.3: **Optimizing LLM performance on the RadCases dataset.** (**a**) We explore 4 strategies to further improve LLM alignment with the ACR AC: RAG and ICL provide additional context to an LLM as input, COT encourages deductive reasoning, and MFT optimizes the weights of the LLM itself. Each optimization strategy is independently implemented and compared against the baseline prompting results in **Figure 3.2** for (**b**) Claude Sonnet-3.5 and (**c**) Llama 3. Error bars represent ±95% CI over $n = 5$ independent experimental runs.

using Claude Sonnet-3.5: (1) **Baseline**, which queries an LLM to directly recommend a diagnostic imaging study; (2) **Evidence-Based Baseline**, which queries an LLM to recommend an ACR AC Topic that is then mapped to the imaging study; and (3) **Evidence-Based Optimized**, which is the same as (2) but uses optimized COT prompting from **Figure 3.3b** for Claude Sonnet-3.5 (**Fig. 3.4a**).

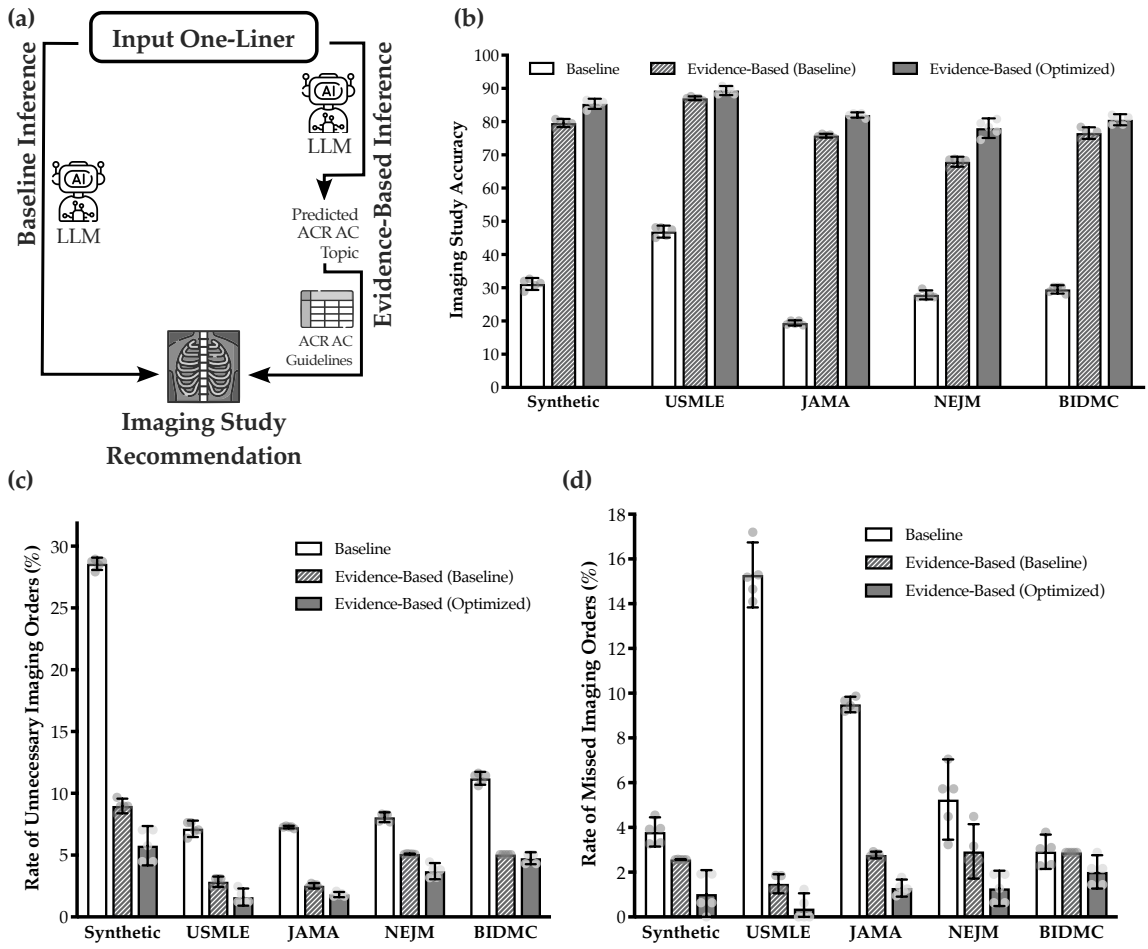

Figure 3.4: **Comparison of baseline and evidence-based inference pipelines with Claude Sonnet-3.5.** **(a)** Using our evidence-based inference pipeline, we query the LLM to predict the single ACR AC Topic most relevant to an input patient one-liner, and then refer to the ACR AC guidelines to make the final recommendation for diagnostic imaging. An alternative approach is the baseline inference pipeline where we query the LLM to recommend a diagnostic imaging study directly without the use of the ACR AC. **(b)** Our evidence-based pipelines (both using baseline prompting and optimized using chain-of-thought (COT) prompting) significantly outperform the baseline pipeline by up to 62.6% (two-sample, one-tailed, homoscedastic t-test; $p < 0.0001$ for all RadCases datasets). At the same time, they also reduce the rates of both **(c)** unnecessary imaging orders and **(d)** missed imaging orders (two-sample, one-tailed, homoscedastic t-test; $p < 0.05$ for all RadCases datasets). Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

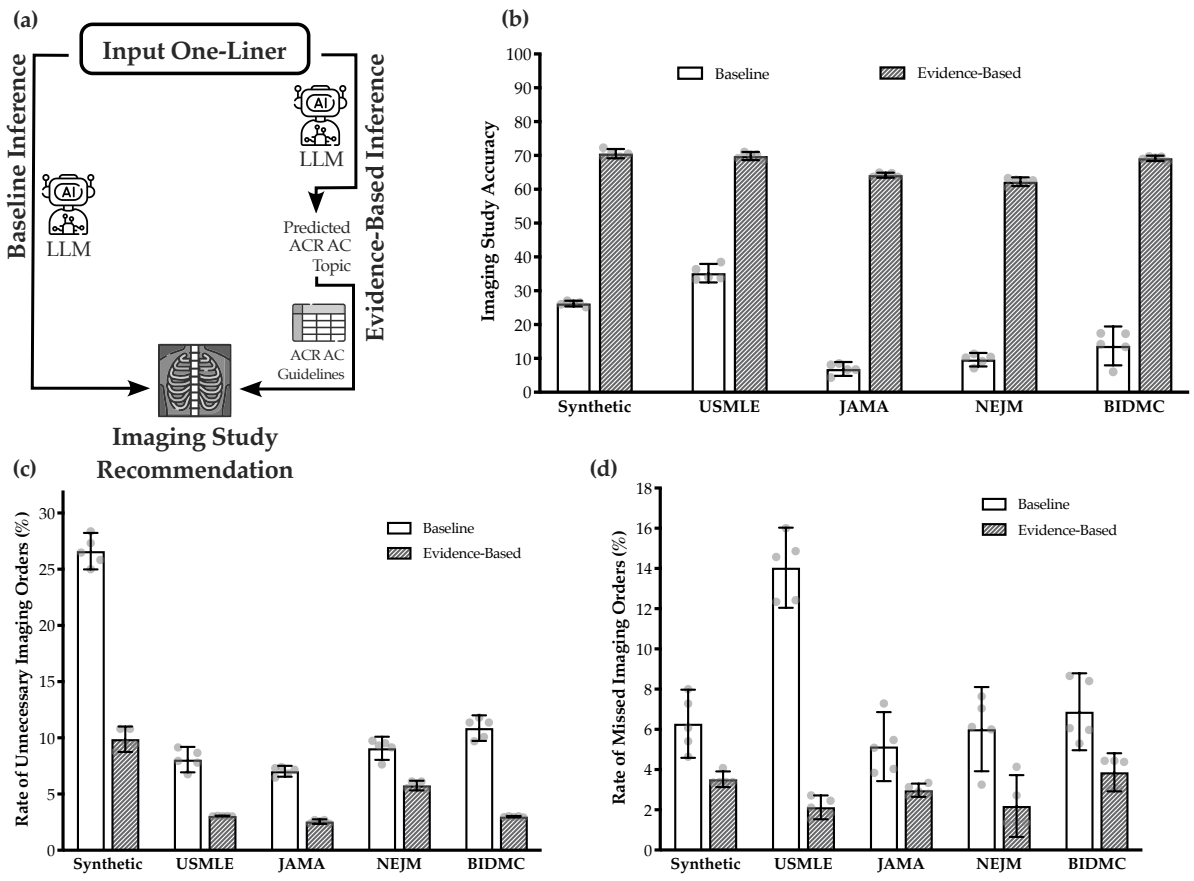

Figure 3.5: **Comparison of Llama 3 baseline and evidence-based inference pipelines.** (**a**) Using our evidence-based inference pipeline identical to that shown in Fig. 3a in the main text, we query the Llama 3 to predict the ACR AC Topic most relevant to an input patient one-liner, and programmatically refer to the evidence-based ACR AC guidelines to make the final recommendation for diagnostic imaging (Evidence-Based). An alternative approach is the baseline inference pipeline where we query the LLM to recommend a diagnostic imaging study directly without the use of the ACR AC (Baseline). Because there was no consistently optimal prompting or fine-tuning strategy that outperformed baseline prompting in **Fig. 3.3c**, we only empirically evaluated the baseline Evidence-Based inference strategy here. (**b**) Our evidence-based pipeline significantly outperforms the baseline pipeline by up to 57.3% (two-sample, one-tailed, homoscedastic t-test; $p < 0.0001$ for all RadCases datasets). At the same time, the also reduce the rates of both (**c**) unnecessary imaging orders and (**d**) missed imaging orders (two-sample, one-tailed, homoscedastic t-test; $p < 0.002$ for all RadCases datasets). Error bars are $\pm$95% CI over $n = 5$ experimental runs.

Our results demonstrate that our evidence-based algorithm of leveraging LLMs to map to ACR AC Topics provides significant improvements in the overall imaging accuracy achieved by the model. Across all 5 RadCases dataset subsets, our Evidence-Based Baseline (resp., Evidence-Based Opti-

mized) pipeline outperforms the Baseline pipeline by at least 40% (resp., 42%) on imaging accuracy (two-sample, one-tailed homoscedastic t-test; $p < 0.0001$ (**Fig. 3.4b**).

Interestingly, while the Evidence-Based Optimized pipeline significantly outperformed the Baseline pipeline on ACR AC Topic classification accuracy (**Fig. 3.3b**), we did not observe a statistically significant improvement in the optimized versus baseline Evidence-Based pipelines on the imaging classification accuracy (two-sample, one-tailed homoscedastic t-test; $p \geq 0.346$ for each of the 4 RadCases dataset subsets). Qualitatively, we found that although the Evidence-Based Baseline pipeline achieved a lower ACR AC Topic classification accuracy compared to the Evidence-Based Optimized inference strategy, its incorrect predictions were still closely related to the correct answer and underlying patient pathology. For example, a ground truth ACR AC Topic might be "Major Blunt Trauma" and the LLM prediction "Penetrating Torso Trauma;" although the LLM identified the incorrect ACR AC Topic label, both Topics warrant a "Radiography trauma series." As a result, both the optimized and baseline Evidence-Based pipelines achieve comparable imaging accuracy and significantly improve upon the Baseline pipeline.

We also evaluated the false positive and false negative rates in image ordering. Formally, false positives are cases where an imaging study is unnecessarily ordered, and false negatives are cases where a diagnostic imaging study was warranted but not ordered. Both Evidence-Based pipelines again outperform the Baseline pipeline according to both metrics, significantly reducing the rates of false positives and false negatives (two-sample, one-tailed homoscedastic t-test; $p < 0.0001$ for the Synthetic, USMLE, JAMA, and NEJM subsets).

3.3.5. Investigating Autonomous Image Ordering using LLMs versus Standard of Care

Based on the initial results in **Figures 3.4-3.5**, we next looked to assess if state-of-the-art, optimized language models could be used to accurately order imaging studies for acutely presenting patients without clinician intervention. Using a set of anonymized, de-identified admission notes derived from the medical records of 100 real patient admissions between 2008-2019 from the Beth Israel Deaconess Medical Center (Boston, MA) (Johnson et al., 2023) we compared the accuracy of diagnostic image ordering of the prompt-optimized versions of Claude Sonnet-3.5 and Llama

3 with that of clinicians. In **Figure 3.6**, our results suggest that autonomous LLMs can be effective tools in ordering diagnostic imaging: Claude Sonnet-3.5 achieved a higher accuracy score of 58.0% and F_1 score of 94.4% compared with clinicians (accuracy 39.3%; F_1 score 92.5%) (McNemar test; $p = 0.044$). Similarly, there was no statistically significant difference between Llama 3 (accuracy 61.5%; F_1 score 92.9%) and clinicians (McNemar test; $p = 0.099$). Across the patient cases assessed, clinicians ordered an average of 1.41 (95% CI: $[1.28 - 1.53]$) imaging studies per case; similarly, Claude Sonnet-3.5 ordered an average of 1.83 (95% CI: $[1.60 - 2.06]$) and Llama 3 an average of 1.54 (95% CI: $[1.33 - 1.76]$) studies per case. There was no statistically significant difference between the number of imaging studies ordered by Llama 3 and its clinician counterparts (two-sample paired t-test; Llama 3: $p = 0.269$).

We also evaluated the rates of both unnecessary and missed imaging studies: both Claude Sonnet-3.5 and Llama-3 were non-inferior to clinicians according to both metrics, achieving a false positive rates (FPR) of 6.90% and 6.90% (clinician FPR = 7.76%) (McNemar test; $p = 1.00$) and false negative rates (FNR) of 3.45% and 6.03% (clinician FNR = 6.03%) (McNemar test; Llama 3: $p = 1.00$; Claude Sonnet-3.5: $p = 0.549$), respectively (**Fig. 3.6b-c**). Altogether, these results suggest that LLMs can be promising tools for image ordering in clinical workflows.

Finally, to gauge the similarity between recommendations made by different language models and clinicians, we computed the pairwise Dice-Sørensen coefficient (DSC) between imaging recommendations made by different decision makers (**Fig. 3.6f**). According to this metric, we found that recommendations made by different language models consistently aligned significantly more closely than those made by language models and clinicians.

3.3.6. Evaluating LLMs As Support Tools for Clinician Diagnostic Image Ordering

We assessed LLMs as autonomous agents for clinical decision making above. Such retrospective studies help clarify the technical capabilities and limitations of these models compared with standard of care. However, LLMs can also act as *assistants* for clinicians in diagnostic image ordering.

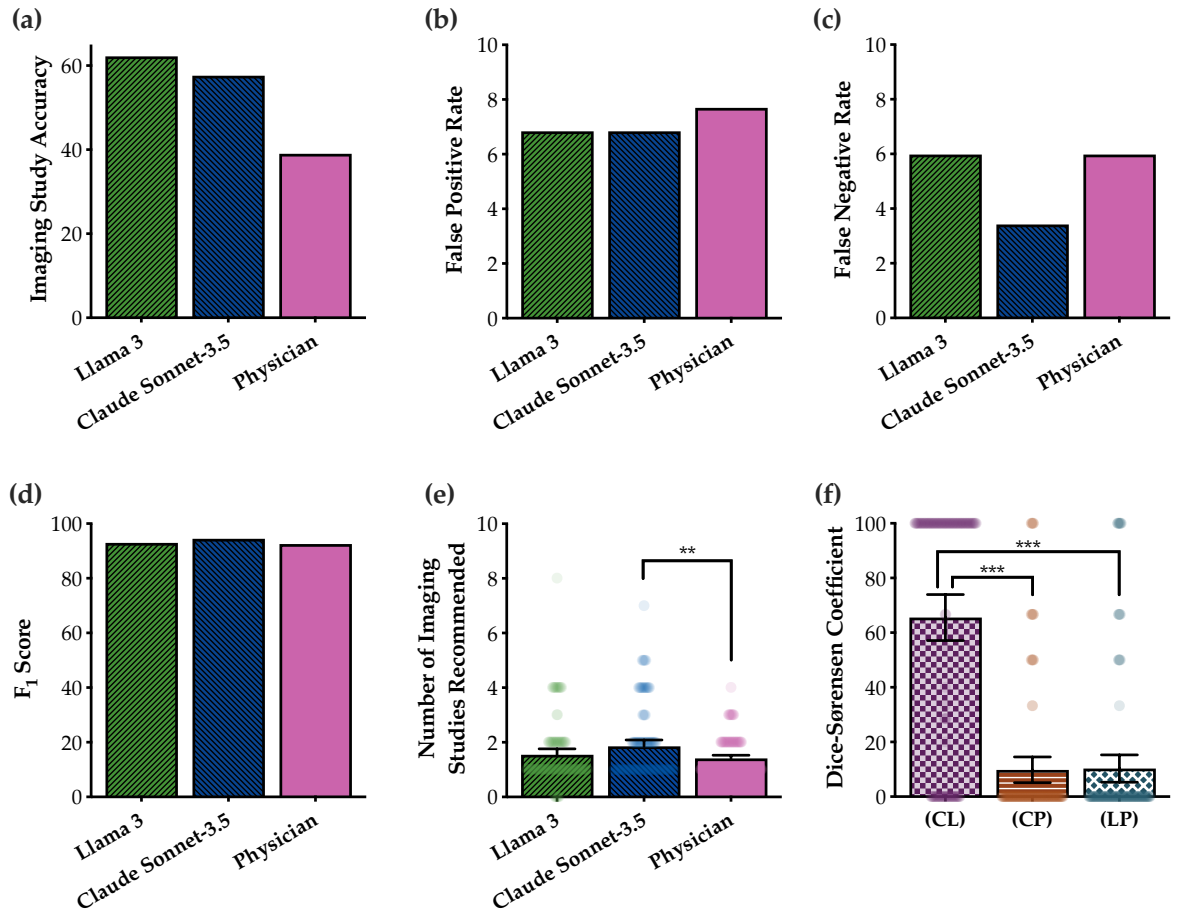

Figure 3.6: **Retrospective study of clinician- and LLM- ordered imaging studies.** We compare the diagnostic imaging studies ordered by the prompt-optimized LLMs Claude Sonnet-3.5 and Llama 3 against those ordered by clinicians in a retrospective study. Compared with clinicians, Claude Sonnet-3.5 and Llama 3 achieve the same or better (**a**) accuracy scores; and (**b**) false positive rates (i.e., the rate at which a patient received at least one unnecessary imaging recommendation); (**c**) false negative rates (i.e., the rate at which a patient should have received an imaging workup but did not); and (**d**) F_1 scores. (**e**) We observe that Claude Sonnet-3.5 orders a greater number of recommended imaging studies compared to clinicians. (**f**) Claude Sonnet-3.5 and Llama 3 order imaging studies that are more similar to one another than to clinicians (two-sample, two-tailed homoscedastic t-test; $p < 0.0001$).

To evaluate the utility of our evidence-based LLMs as clinical assistants, we conducted a prospective randomized control trial[4] asking volunteer clinician participants to order diagnostic imaging studies for simulated patient scenarios in an online testing environment. Participants were U.S.

[4]Pre-registration on AsPredicted, #185312. Link: https://aspredicted.org/x6b9-rcgh.pdf

medical students and emergency medicine resident physicians recruited from the Perelman School of Medicine and the Hospital of the University of Pennsylvania. This study was exempted by the University of Pennsylvania Institutional Review Board (Protocol #856530).

Each study participant was asked to order a single imaging study (or forego imaging if not indicated) for 50 simulated patient cases. For each participant, a random 50% of the patient cases included recommendations generated by Claude Sonnet-3.5 using the evidence-based optimized inference strategy in **Fig. 3.4**. To simulate the acuity and high-pressure of many emergency room environments, participants were required to complete the study at an average rate of 1 case per minute in a single setting. We then fitted a regression model according to

$$y_{s,q} = \beta_0 + (\beta_1 \cdot \text{WithLLMGuidance}_{s,q}) + \theta_q + \chi_s + \varepsilon_{s,q} \tag{3.1}$$

where s indexes study participants and q study questions, and $y_{s,q}$ is a binary variable indicating whether participant s answered study question q correctly. Here, θ is a q-vector of study question fixed effects, χ_s are control variables specific to the study participant (i.e., whether the study participant is a physician or medical student, the participant's personal experience with AI, and the participants' sentiment regarding AI), and $\varepsilon_{s,q}$ is the error term. We estimate (3.1) using standard errors clustered at the study participant level and question level. Furthermore, $\text{WithLLMGuidance}_{s,q}$ is a binary indicator that indicates whether LLM-generated guidance was provided for question q for participant s, respectively. Study participants generally found the study task challenging, with an average accuracy of 15.8% (95% CI: [12.2% - 19.3%]) without LLM guidance and 25.0% (95% CI: [20.7% - 29.3%]) with guidance. Offering LLM-based recommendations using our evidence-based optimized pipeline improved image ordering accuracy with statistical significance ($\beta_1 = 0.081$; 95% CI: [0.022 - 0.140]; $p = 0.011$) for both medical students and resident physicians.

To verify that participants were indeed taking advantage of LLM-generated recommendations when made available, we fitted a separate regression model analogous to that in (3.1) that instead measures that binary agreement between LLM recommendations and participant answers as the dependent variable. As expected, the agreement between answers and assistant recom-

mendations increases when the recommendations are made available to the clinician ($\beta_1 = 0.141$; 95% CI: [0.050 - 0.233]; $p = 0.005$). These results suggest that language models can act as clinical assistants to help clinicians order imaging studies more aligned with evidence-based guidelines.

Similarly, we did not observe statistically significant differences in either the false positive rate ($\beta_1 = 0.008$; 95% CI: [-0.012 - 0.027]; $p = 0.418$) or false negative rate ($\beta_1 = -0.019$; 95% CI: [-0.068 - 0.030]; $p = 0.431$). This ensures that the improvements in accuracy scores with LLM guidance were not at the cost of significantly increasing the number of unnecessary or missed imaging studies ordered by clinicians. Additional analysis is included in **Supp. Tables A.4-A.10**, and discussion of experimental results in **Section A.1**.

3.4. Discussion

Our study investigates the potential of LLMs in the domain of diagnostic image ordering—a task critical to the timely and appropriate management of acute patient presentations. Our results demonstrate how state-of-the-art language models can be used in the context of diagnostic image ordering in acute clinical settings, such as the emergency department. Firstly, we observed that generalist language models—such as Claude Sonnet-3.5 and Meta Llama 3—can accurately predict relevant ACR AC Topic labels to describe patient one-liner descriptions without any domain-specific fine-tuning. By leveraging LLMs to predict Topic labels instead of imaging studies directly, we achieved significant improvements in the quality of final imaging recommendations made by LLMs. Comparing language models with clinicians in a retrospective study, we show that LLMs achieve better accuracy with regard to image ordering in the ED without significant changes to the rate of missed imaging (FNR), rate of unnecessary imaging (FPR), or number of recommended imaging studies. Finally, we demonstrate that LLMs can be leveraged by clinicians as a CDS assistant to improve the accuracy of ordered imaging studies without significant changes to the FPR or FNR in a simulated acute care environment.

Importantly, we demonstrate how integrating evidence-based guidelines (i.e., the ACR AC) directly into the LLM-based inference pipeline can significantly improve the accuracy of clinical recommendations. This approach not only aligns model predictions with established guidelines, but

also provides a robust framework for reducing the rates of both unnecessary and missed imaging orders. In theory, such a framework could be readily translatable for other clinical use cases that make use of available guidelines, such as the American College of Gastroenterology guidelines to determine clinical indications for endoscopy (Gabelloni et al., 2020; Park et al., 2015; Patel et al., 2022), or the American Society of Addiction Medicine guidelines for the management of alcohol withdrawal syndrome (of Addiction Medicine, 2020). We leave these potential future applications of LLM-based CDS tools for future work.

We also highlight the challenges associated with integrating novel LLM toolkits into existing clinical workflows. Our prospective study demonstrated preliminary evidence that the utility of LLM clinical assistants can be largely dependent on factors such as user expertise, acuity of care, and existing user attitudes on AI, consistent with prior work (Agarwal et al., 2023; Prinster et al., 2024; Rau et al., 2023). Nonetheless, we observed that the accuracy of imaging studies ordered by clinicians increased by approximately 10 percentage points on average, which can potentially translate to hundreds of dollars saved per patient in reducing low-value and unnecessary imaging studies according to recent work (Brady et al., 2020; Kjelle et al., 2022; Yan et al., 2024). That being said, we highlight that our study was limited by a relatively small sample size of only 23 medical students and 7 junior emergency resident physicians. Furthermore, our study participants voluntarily opted in to participate in our study, and may not reflect the attitudes and behaviors of clinicians that may have a more conservative predisposition to the use of AI tools in healthcare. Finally, we highlight that ED residents at large academic institutions (such as the University of Pennsylvania where this study was conducted) may not currently be trained to order imaging studies in alignment with the ACR AC, as the benefit-to-cost ratio of obtaining more extensive imaging studies may be different institutionally than as dictated by national guidelines. Given these considerations, future work is warranted to better characterize the impact of these factors across diverse populations of healthcare workers as they affect real-world clinical workflows and physician thinking, ultimately ensuring that LLMs are used responsibly and can improve patient care.

Of note, our results consistently demonstrate that proprietary language models, such as Claude

Sonnet-3.5, consistently and significantly outperform open-source models. While the performance of Claude Sonnet-3.5 is impressive, it is unlikely that current publicly available inference APIs for the model are sufficient for widespread clinical deployment, as many hospitals understandably express concerns over patient privacy and unknown data handling practices from third-party vendors. In these settings, our experimental results suggest that open-source language models, such as Llama 3, can be potentially viable alternatives. Future work might investigate other strategies that better leverage open-source models in our proposed pipeline.

This study also has its limitations. Firstly, because most of the patient descriptions in our RadCases dataset are derived from real medical sources, they also reflect inherent biases with respect to patient demographics and medical conditions. For example, we found that our dataset most commonly included ground-truth ACR AC Topic labels related to gastrointestinal, cardiac, and neurologic pathologies—while these cases may reflect real-world ED visit patterns, it remains to be seen how LLMs perform on other patient cases sampled from different underlying distributions, such as in rare disease diagnostics and low-resourced patient populations.

Furthermore, closely related topics (e.g., "Major Blunt Trauma" and "Penetrating Torso Trauma") often share clinical indications for the same set of imaging studies (e.g., "Radiography trauma series"). As a result, our choice of imaging accuracy evaluation metrics in **Figures 3.4-3.6** still permits an LLM to predict the correct imaging study even if an incorrect ACR AC Topic was identified. This potential limitation of models achieving the "right answer through the wrong reasoning" is well-documented in prior and concurrent work examining discrepancies between model reasoning traces and final predictions (Agarwal et al., 2024a; Braun et al., 2019; Chen et al., 2025b; Hager et al., 2024; Lin et al., 2022; Turpin et al., 2023). However, even if we require our language models to obtain the "right answer through the *right* reasoning," our inference strategy in **Figure 3.4a** still outperforms baseline reasoning strategies (i.e., the ACR AC Topic Prediction accuracy in **Figure 3.3b-c** is greater than the imaging accuracy of baseline LLMs in **Figure 3.4b**).

We also highlight that our main experimental results are reported on the subset of patient cases with at least one ACR AC Topic label. In general, patient one-liners with no matching ground-

truth label were excluded from our analysis; see our **Section 3.2** for additional details. In **Supp. Table A.3**, we investigate the ability of language models to predict whether there exists at least one ACR AC Topic that is relevant for a given patient as a binary classification task. Future work might explore multi-step pipelines involving sequential LLM queries that first determine if a set of ACR AC Topics is relevant before predicting the most relevant Topic within the set.

Importantly, we also emphasize even though we leverage the ACR AC as a ground truth symbol in our experiments (**Fig. 3.2-3.3**), evidence-based guidelines like the ACR AC are ultimately *recommendations* that should be used in conjunction with clinical expertise to help physicians make the most appropriate decisions regarding the role of diagnostic imaging. Such recommendations may therefore fall short in more challenging patient cases not considered in this work, such as those with multiple medical conditions, complex admissions, and/or prior imaging studies that can drastically affect the appropriateness of different diagnostic methods. For these reasons, we argue that *any* LLM-based clinical decision support tool should ultimately be used in the same fashion, where LLM-generated recommendations are used by clinicians together with their individual expertise to best contextualize the role of diagnostic imaging in specific patient scenarios.

Finally, we emphasize that our experiments, while promising, are no substitute for true prospective evaluation of language models as clinical decision support tools in real-world clinical workflows, such as those in the emergency department (Chen et al., 2025b). We particularly highlight that practical applications of our work might focus on targeting clinical decision making for costly imaging studies (e.g., magnetic resonance imaging) and those associated with high radiation doses (e.g., computed tomography). Future work is needed in close collaboration with clinicians across a variety of clinical environments to truly validate the clinical utility of our LLM-based pipelines.

In conclusion, our study highlights the potential of LLMs to enhance the process of diagnostic image ordering by leveraging evidence-based guidelines. By simply mapping patient scenarios to interpretable ACR AC Topics, we show that LLMs can improve the accuracy of imaging decisions in simulated acute healthcare environments. Our findings suggest that our approach can better adapt LLMs for clinical tasks, such as improving patient care in acute diagnostic workflows.

CHAPTER 4

CLINICALLY DERIVED PRIORS FOR MEDICAL IMAGING ANALYSIS

This chapter is based on work published in the following co-first-author manuscripts:

(Chae et al., 2024) Allison Chae*, Michael S Yao*, Hersh Sagreiya, Ari D Goldberg, Neil Chatterjee, Matthew T MacLean, Jeffrey Duda, Ameena Elahi, Arijitt Borthakur, Marylyn D Ritchie, Daniel Rader, Charles E Kahn, Walter R Witschey†, and James C Gee†. Strategies for implementing machine learning algorithms in the clinical practice of radiology. *Radiology*, 310(1):e223170, 2024. doi: 10.1148/radiol.223170

(Yao et al., 2023) Michael S Yao*, Allison Chae*, Matthew T MacLean, Anurag Verma, Jeffrey Duda, James Gee, Drew Torigian, Daniel Rader, Charles Kahn, Walter R Witschey†, and Hersh Sagreiya†. SynthA1c: Towards clinically interpretable patient representations for diabetes risk stratification. In *Predictive Intelligence in Medicine*, pages 46–57, 2023. doi: 10.1007/978-3-031-46005-0_5

Here, * denotes the co-first authorship and † denotes co-senior authorship. In both works, I led the experimental design and analysis of results, developed and validated the technical methodology, and co-led the writing and submission of the manuscript texts. The contents of this chapter are not described in detail in any past, present, or future dissertation(s) by any co-author listed above.

4.1. Introduction

In **Chapter 3**, we explored how evidence-based medical guidelines could be used to define interpretable representations in ML prediction pipelines. Here, I discuss how clinical knowledge made available by human domain experts can be used to define interpretable representations, too.

Type 2 Diabetes Mellitus (T2DM) affects over 30 million patients in the United States, and is most commonly characterized by elevated serum hemoglobin A1c (HbA1c) levels measured through a blood sample (Khan et al., 2020; Xu et al., 2018). A patient is considered diabetic if their HbA1c is greater than 6.5% A1c. Patients diagnosed with T2DM are at an increased risk of many comorbidi-

ties, but early diagnosis and interventions can improve outcomes (Albarakat and Guzu, 2019).

However, delayed diagnosis of T2DM is frequent due to a low rate of screening. Up to a third of patients are not screened for T2DM as recommended by current national guidelines (Kaul et al., 2022; Polubriaginof et al., 2019), and Porter et al. (2022) estimate that it would take over 24 hours per day for primary care physicians to follow national screening recommendations for every adult visit. Furthermore, T2DM screening using patient bloodwork is not routinely performed in acute urgent care settings or emergency department (ED) visits. Given these obstacles, machine learning (ML) is a promising tool to predict patient risk for T2DM and other diseases (Farran et al., 2013).

Simultaneously, the usage of radiologic imaging in clinical medicine continues to increase every year (Dowhanik et al., 2022; Hong et al., 2020; Chae et al., 2024). Considering our own institution as an example, the number of imaging studies deposited in the Penn Medicine BioBank grows steadily year-over-year (**Fig. 4.1b**). Similarly, a large proportion of patients also have multiple imaging studies deposited (**Fig. 4.1c-d**), collectively amounting to hundreds of thousands of real-world patient data that can be used to learn clinical predictive algorithms (**Fig. 4.1a**). Of note, over 70 million computed tomography (CT) scans are performed annually, and their utilization has become increasingly common in both primary care and ED visits (Dowhanik et al., 2022). Consequently, the wealth of CT radiographic data can potentially be used to estimate patient risk for T2DM as an incidental finding in these clinical settings. For example, T2DM risk factors include central adiposity and the buildup of excess fat in the liver that can be readily estimated from clinical CT scans. Liver fat excess can be estimated by calculating the spleen-hepatic attenuation difference (SHAD), which is the difference between liver and spleen CT attenuation (MacLean et al., 2021). These metrics are examples of **image-derived phenotypes** (IDPs) derived from patient CT scans and other imaging modalities. Other IDPs, such as volume estimation of subcutaneous fat and visceral fat, can also be used to quantify central adiposity. In **Fig. 4.2**, we plot how the distribution of (the principal component of) the IDPs changes as a function of disease, demonstrating a statistically meaningful association between IDP values and clinical manifestations of diabetes and metabolic syndrome (e.g., obesity, obstructive sleep apnea, hypertension, and non-alcoholic fatty

liver disease). Using these metrics derived from both imaging data and expert knowledge of the clinical manifestations of diabetes, a prediction model could report estimated T2DM risk as an incidental finding during an unrelated outpatient imaging study or ED visit workup as a means of opportunistic risk stratification from analysis of CT scans and patient information, with automated referral of high-risk patients for downstream screening for diabetes.

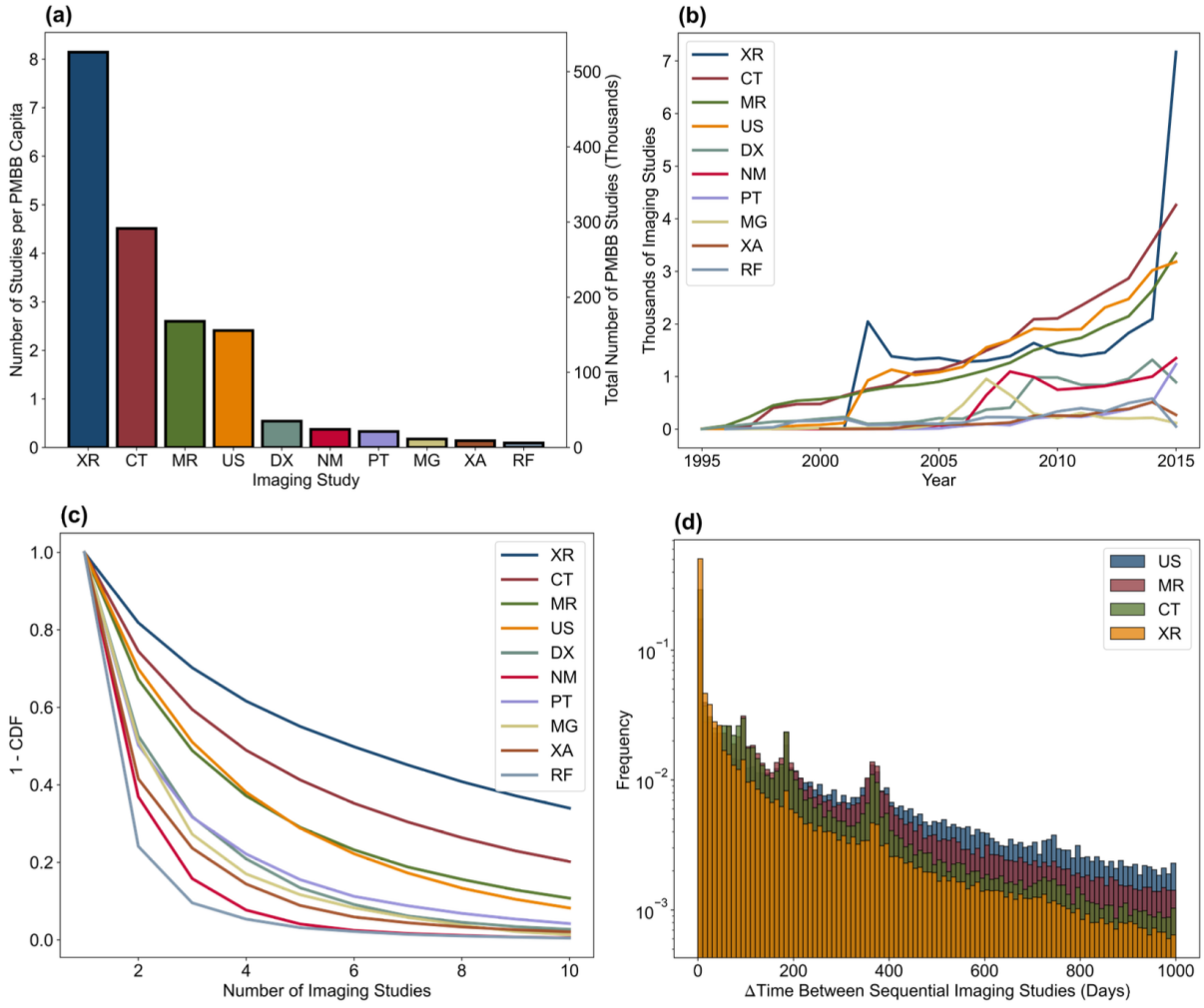

Figure 4.1: **Overview of Penn Medicine BioBank (PMBB) imaging data.** (**a**) Bar graph shows the number of studies within the Penn Medicine BioBank by imaging modality. The number of studies per Penn Medicine BioBank capita is the average number of studies per patient within the Penn Medicine BioBank. (**b**) Line graph shows the number of imaging studies acquired per year contained within the Penn Medicine BioBank by imaging modality. (**c**) Line graph of 1 – CDF, where CDF is the cumulative distribution function. 1 – CDF corresponds to the proportion of patients (by modality) according to number of examinations. (**d**) Histogram shows the time between sequential repeat imaging studies by patient for the four most common imaging modalities.

Existing machine learning methods for disease prediction have largely focused on developing classification models that output probability values for different physiologic states (Uddin et al., 2019; Kopitar et al., 2020; Deberneh and Kim, 2021). However, these metrics are difficult for clinicians to interpret at face value and cannot be intelligently integrated into existing clinician workflows, such as diagnostic pathways based on clinical lab findings (Sivaraman et al., 2023).

In this study, we hypothesized that radiomic metrics derived from CT scans could be used in conjunction with physical examination data to predict patient T2DM risk using SynthA1c, a novel synthetic *in silico* measurement approximating patient blood hemoglobin A1c (HbA1c) (**Fig. 1.1**). To predict model generalizability, we also propose a generalizable data augmentation-based model smoothness metric that predicts SynthA1c accuracy on previously unseen OOD patient datasets.

4.2. Materials and Methods

4.2.1. Patient Cohort and Data Declaration

The data used for our retrospective study were made available by the Penn Medicine BioBank (PMBB), an academic biobank established by the University of Pennsylvania (Chae et al., 2024). All patients provided informed consent to utilization of de-identified patient data, which was approved by the Institutional Review Board of the University of Pennsylvania (IRB protocol 813913). From the PMBB outpatient dataset, we obtained patient ages, genders, ethnicities, heights, weights, blood pressures, abdominal CT scans, and blood HbA1c measurements. Notably, the only laboratory value used was HbA1c as a ground truth metric in model training and evaluation—no blood biomarkers were used as model inputs. Patients with any missing features were excluded.

Using the pre-trained abdominal CT segmentation network trained and reported by MacLean et al. (2021), we estimated four IDPs from any given CT of the abdomen and pelvis study (either with or without contrast) to be used as model inputs. Our four IDPs of interest were mean liver CT attenuation, mean spleen CT attenuation, and estimated volume of subcutaneous fat and visceral fat. Briefly, their segmentation network achieved mean Sørenson-Dice coefficients $\geq$98% for all IDP extraction tasks assessed (including our four IDPs of interest).

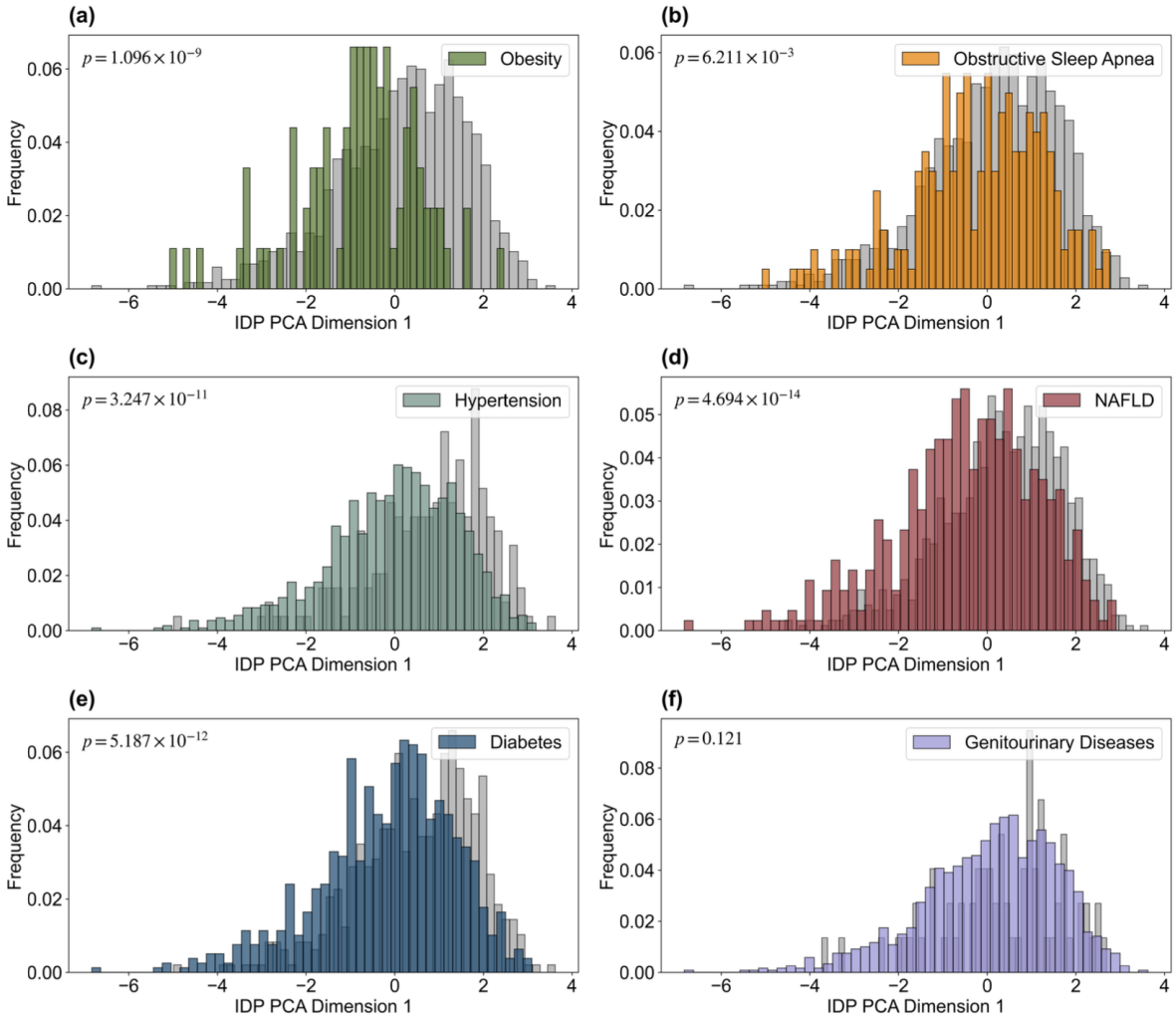

Figure 4.2: **Comparing principal component distributions of six image-derived phenotype (IDP) metrics computed from abdominal CT scans from 1276 anonymized patients in the Penn Medicine BioBank.** These image-derived phenotypes included liver CT attenuation, spleen CT attenuation, liver volume, spleen volume, visceral fat volume, and subcutaneous fat volume. Using principal component analysis (PCA), the principal component of these image-derived phenotypes was extracted and its distribution was plotted as a histogram for patients stratified by different clinical diagnoses. Bar graphs show different image-derived phenotype principal component distributions in patients without diagnoses (gray bars) versus in patients diagnosed with **(a)** obesity ($n = 91$), **(b)** obstructive sleep apnea ($n = 201$), and **(c)** hypertension ($n = 1082$). Image-derived phenotype principal component distributions in patients without diagnoses (gray bars) versus in patients diagnosed with **(d)** nonalcoholic fatty liver disease (NAFLD; $n = 429$) and **(e)** diabetes ($n = 790$). **(f)** Genitourinary diseases ($n = 1202$), which are not clinically associated with these image-derived phenotype metrics, were not associated with a statistically significant different principal component distribution compared with healthy patients. p values were calculated by comparing distributions of patients with and without the disease according to a two-sample Kolmogorov-Smirnov test for goodness of fit.

Any patient i has a set of measured values of any particular feature within the dataset. To construct a feature vector $\mathbf{x}$ associated with an HbA1c measurement y_i, we selected the patient's measurements that minimized the time between the dates the feature and y_i were measured.

4.2.2. Machine Learning Models: GBDT, NODE, and FT-Transformer

Current supervised methods for disease detection work with feature vectors derived from patient physical examinations and clinical laboratory values (Farran et al., 2013; Uddin et al., 2019; Kopitar et al., 2020). Our work builds on these prior advances by incorporating IDPs as additional input vector dimensions. Previously, Chen and Guestrin (2016) introduced gradient-boosted decision trees (**GBDTs**) that incorporate scalable gradient boosting with forest classifiers for state-of-the-art prediction accuracy across tasks. A separate class of machine learning models is deep neural networks (DNNs). Recently, neural oblivious decision ensemble (**NODE**) DNNs achieved classification performance on par with decision tree models on certain tasks (Popov et al., 2019) and the Feature Tokenizer + Transformer (Gorishniy et al., 2021) effectively adopts transformer architectures to tabular data. Here, we assessed NODE, FT-Transformer, and GBDT architectures as backbones for our SynthA1c encoders.

We sought to compare our proposed SynthA1c models against a number of baselines. We looked at Ordinary Least Squares (OLS) encoders and traditional diabetes *binary classifier* models with the same three architectures as proposed above, in addition to a zero-rule classifier and a multi-rule questionnaire-based classifier currently recommended for clinical practice by the American Diabetes Association and Centers for Disease Control and Prevention (Bang et al., 2009).

4.2.3. Model Training and Evaluation Strategy

Our model inputs can be divided into two disjoint sets: clinically derived phenotypes (CDPs), which are derived from physical examination, and image-derived phenotypes (IDPs) that are estimated from abdominal CT scans herein. The specific CDPs and IDPs used depended on the model class—broadly, we explored two categories of models, which we refer to as r-type and p-type. r-type models were trained on 'raw' data types (CDPs: height, weight, race, gender, age, systolic blood pressure [SBP], diastolic blood pressure [DBP]; IDPs: liver CT attenuation, spleen CT at-

tenuation, subcutaneous fat [SubQ Fat], visceral fat [Visc Fat]), while p-type models were trained on 'processed' data types (CDPs: BMI, race, gender, age, SBP, DBP; IDPs: SHAD, SubQ Fat, Visc Fat). Comparing the performance of r- and p- type models could help us better understand if using derivative processed metrics clinically correlated with T2DM risk yields better model performance.

SynthA1c encoders were trained to minimize the L_2 distance from the ground truth HbA1c laboratory measurement, and evaluated using the root mean square error (RMSE) and Pearson correlation coefficient (PCC). We then compared the predicted SynthA1c values with the traditional HbA1c $\geq$ 6.5% A1c diabetes cutoff to assess the utility of SynthA1c outputs in diagnosing T2DM. A p value of $p < 0.05$ was used to indicate statistical significance.

4.2.4. Implementation Details

NODE models were trained with a batch size of 16 and a learning rate of $\eta = 0.03$, which decayed by half every 40 epochs for a total of 100 epochs. FT-Transformer models were trained with a batch size of 128 and a learning rate of $\eta = 0.001$, which decayed by half every 50 epochs for a total of 100 epochs. GBDT models were trained using 32 boosted trees with a maximum tree depth of 8 with a learning rate of $\eta = 0.1$.

4.2.5. Assessing Out-of-Domain Performance

An important consideration in high-stakes clinical applications of machine learning is the generalizability of T2DM classifiers to members of previously unseen patient groups. Generalizability is traditionally difficult to quantify and can be affected by training data heterogeneity and the geographic, environmental, and socioeconomic variables unique to the PMBB dataset.

Prior work has shown that model smoothness can be used to predict out-of-domain generalization of neural networks (Ng et al., 2023; Jiang et al., 2021). However, these works largely limit their analysis to classifier networks. To evaluate SynthA1c encoder robustness, we develop an estimation of model manifold smoothness $\mathbb{M}$ for our encoder models. Under the mild assumption that our SynthA1c encoder function $y : \mathbb{R}^{|\mathbf{x}|} \rightarrow \mathbb{R}$ is Lipschitz continuous, we can define a local manifold

smoothness metric μ at $\mathbf{x} = \tilde{\mathbf{x}}$ given by

$$\mu(\tilde{\mathbf{x}}) = \mathbb{E}_{\mathcal{N}(\tilde{\mathbf{x}})}\left[\frac{\sigma_y^{-1}||y(\mathbf{x}) - y(\tilde{\mathbf{x}})||_1}{||\delta\mathbf{x} \oslash \sigma_{\mathbf{x}}||_2}\right] = \mathcal{V}[\mathcal{N}(\tilde{\mathbf{x}})]^{-1} \cdot \oint_{\mathcal{N}(\tilde{\mathbf{x}}) \in \mathcal{D}} d\mathbf{x} \, \frac{\sigma_y^{-1}|y(\mathbf{x}) - y(\tilde{\mathbf{x}})|}{\sqrt{(\delta\mathbf{x} \oslash \sigma_{\mathbf{x}})^T(\delta\mathbf{x} \oslash \sigma_{\mathbf{x}})}} \tag{4.1}$$

where we have a feature vector $\mathbf{x}$ in domain $\mathcal{D}$ and a neighborhood $\mathcal{N}(\tilde{\mathbf{x}}) \in \mathcal{D}$ around $\mathbf{x}$ with an associated volume of $\mathcal{V}[\mathcal{N}(\tilde{\mathbf{x}})]$. We also define $\delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}}$, $\oslash$ as the Hadamard division operator, and $\sigma_{\mathbf{x}}$ as the vector of the estimated standard deviations of each feature over $\mathcal{D}$. The exact expectation value over a given neighborhood $\mathcal{N}(\tilde{\mathbf{x}})$ is computationally intractable, but we can approximate it with a Monte Carlo integration through an empirical sampling of Q random points $\mathbf{x}_k$ from $\mathcal{N}(\tilde{\mathbf{x}})$:

$$\mu(\tilde{\mathbf{x}}) = \frac{1}{Q}\sum_{k=1}^{Q} \frac{\sigma_y^{-1}|y(\mathbf{x}_k) - y(\tilde{\mathbf{x}})|}{\sqrt{(\delta\mathbf{x}_k \oslash \sigma_{\mathbf{x}})^T(\delta\mathbf{x}_k \oslash \sigma_{\mathbf{x}})}} \tag{4.2}$$

We can now define a metric $\mathbb{M}$ for the global encoder manifold smoothness over a domain $\mathcal{D}$ as the expectation value of $\mu(\tilde{\mathbf{x}})$ over $\mathcal{D}$, which can similarly be approximated by an empirical sampling of N feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathcal{D}$. We hypothesized that this global smoothness metric $\mathbb{M}$ inversely correlates with model performance on out-of-domain datasets. To evaluate this experimentally, we assessed model performance on two previously unseen T2DM datasets: (1) the Iraqi Medical City Hospital dataset (Rashid, 2020); and (2) the PMBB inpatient dataset. The Iraqi dataset contains 1,000 sets of patient age, gender, BMI, and HbA1c measurements. Because of this limited feature set, we trained additional SynthA1c encoders (referred to as p'-type models) on the PMBB outpatient dataset using only these features. The PMBB inpatient dataset consists of 2,066 measurements of the same datatypes as the outpatient dataset (**Section 4.3.1**).

4.3. Results

4.3.1. Summary Statistics

Our model-building dataset from the PMBB consisted of 2,077 unique HbA1c measurements (1,159 diabetic, 619 prediabetic, 299 nondiabetic) derived from 389 patients (**Table 4.1**). 208 (10%) samples were set aside as a holdout test set partition disjoint by patient identity. We used each individual HbA1c measurement to construct an associated feature vector from that patient's data collected

Table 4.1: **PMBB outpatient dataset characteristics.** To reduce the effects of selection bias, all patients presenting to the University of Pennsylvania Health System were given the opportunity to enroll in the PMBB so as to best capture the population of patients seeking medical care and avoid overrepresentation of healthy patients as in traditional office visit patient recruitment strategies. However, the PMBB is still affected by hesitancies of patient sub-populations in study enrollment and the unique socioeconomic factors affecting different groups of patients. *HTN*: Hypertension.

Self-Reported Ethnicity	**Count (%)**
White	720 (34.7)
Hispanic	40 (1.9)
Black	1248 (60.1)
Asian	36 (1.7)
Pacific Islander	6 (0.3)
Native American	5 (0.2)
Other/Unknown	22 (1.1)
Self-Reported Gender	**Count (%)**
Male	880 (42.4)
Female	1197 (57.6)
Age Decade	**Count (%)**
20-29	31 (1.5)
30-39	89 (4.3)
40-49	362 (17.4)
50-59	593 (28.6)
60-69	680 (32.7)
70-79	299 (14.4)
80-89	23 (1.1)
Blood Pressure	**Count (%)**
Normal (SBP $<$ 120 mmHg and DBP $<$ 80 mmHg)	421 (20.2)
Elevated (120 $\leq$ SBP $<$ 130 mmHg and DBP $<$ 80 mmHg)	398 (19.2)
Stage 1 HTN (130 $\leq$ SBP $<$ 140 mmHg or 80 $\leq$ DBP $<$ 90 mmHg)	652 (31.4)
Stage 2 HTN (SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg)	606 (29.2)
BMI	**Count (%)**
Underweight or Healthy Weight (BMI $<$ 25.0)	275 (13.2)
Overweight (25.0 $\leq$ BMI $<$ 30.0)	443 (21.3)
Class 1 Obesity (30.0 $\leq$ BMI $<$ 35.0)	556 (26.8)
Class 2 Obesity (35.0 $\leq$ BMI $<$ 40.0)	389 (18.7)
Class 3 Obesity (BMI $\geq$ 40.0)	414 (20.0)
HbA1c	**Count (%)**
Not Diabetic (HbA1c $<$ 6.5% A1c)	918 (44.2)
Diabetic (HbA1c $\geq$ 6.5% A1c)	1159 (55.8)
CT Abdomen and Pelvis Enhancement	**Count (%)**
With Contrast	1570 (75.6)
Without Contrast	507 (24.4)
Image Derived Phenotypes (IDPs) Statistics	**Mean $\pm$ SD**
Spleen CT Attenuation (HU)	36.2 $\pm$ 16.7
Liver CT Attenuation (HU)	42.8 $\pm$ 20.2
Subcutaneous Fat Area (cm^2)	321.3 $\pm$ 170.1
Visceral Fat Area (cm^2)	172.4 $\pm$ 104.9
Total Count	**2077**

closest in time to each HbA1c measurement. To quantify the temporal association between a given patient's measurements, we defined the daterange of an observation vector $\mathbf{x}$ as the maximum length of time between any two features/imaging studies. The median daterange in our dataset was 18 days.

Table 4.2: **SynthA1c prediction results using different encoder models.** r- (resp., p-) prefixed models are fed raw (resp., processed) inputs as outlined in Section 4.2.3. RMSE in units of % A1c. For the SynthA1c encoder models, recall, precision, specificity, and accuracy metrics are reported based on the traditional T2DM cutoff of 6.5% A1c. The Multi-Rule binary classifier is the current risk stratification tool recommended by the American Diabetes Association (Bang et al., 2009).

SynthA1c Encoder	RMSE	PCC	Recall	Precision	Specificity	Accuracy
r-OLS	1.67	0.206	85.3	56.0	26.3	57.2
p-OLS	1.73	0.159	80.7	57.5	34.3	58.6
r-FT-Transformer	1.44	0.517	87.6	63.4	55.9	70.7
p-FT-Transformer	1.51	0.441	83.5	61.4	54.1	67.8
r-NODE	1.60	0.378	85.6	55.0	38.7	60.6
p-NODE	1.57	0.649	77.3	59.5	54.1	64.9
r-GBDT	1.36	0.567	87.2	66.4	51.5	70.2
p-GBDT	1.36	0.591	77.1	72.4	67.7	72.6
Binary Classifier	AUROC (%)		Recall	Precision	Specificity	Accuracy
Zero-Rule	—		100	52.4	0.0	52.4
Multi-Rule	56.3		67.0	54.9	39.4	53.8
r-FT-Transformer	82.1		85.3	73.8	66.7	76.4
r-NODE	83.5		82.6	76.9	72.7	77.9
r-GBDT	83.1		87.2	76.6	70.7	79.3

4.3.2. SynthA1c Encoder Experimental Results

Our results suggest that the GBDT encoder predicted SynthA1c values closest to ground truth HbA1c values, followed by both the NODE and FT-Transformer DNN models (**Table 4.2**). All the learning-based architectures assessed outperformed the baseline OLS encoder. When comparing SynthA1c outputs against the clinical HbA1c cutoff of 6.5% A1c for the diagnosis of diabetes, the r-GBDT SynthA1c model demonstrated the highest sensitivity of the assessed models at 87.6% on par with the best-performing binary classifier assessed. In terms of an opportunistic screening modality for T2DM, a high sensitivity ensures that a large proportion of patients with diabetes can be identified for additional lab-based diagnostic work-up with their primary care physicians.

Although the accuracy of SynthA1c encoders was lower than the corresponding binary classifier models assessed, this may be partially explained by the fact that the latter's threshold value for classification was empirically tuned to maximize the model's accuracy. In contrast, our SynthA1c encoders used the fixed clinical HbA1c cutoff of 6.5% A1c for diabetes classification. When comparing r- and p- type SynthA1c models, we did not observe a consistently superior data representation.

To further interrogate our SynthA1c encoders, we investigated whether model performance varied as a function of demographic features. Defining the difference between the model prediction and ground truth HbA1c values as a proxy for model performance, all SynthA1c encoders showed no statistically significant difference in performance when stratified by gender or BMI (**Fig. 4.3**).

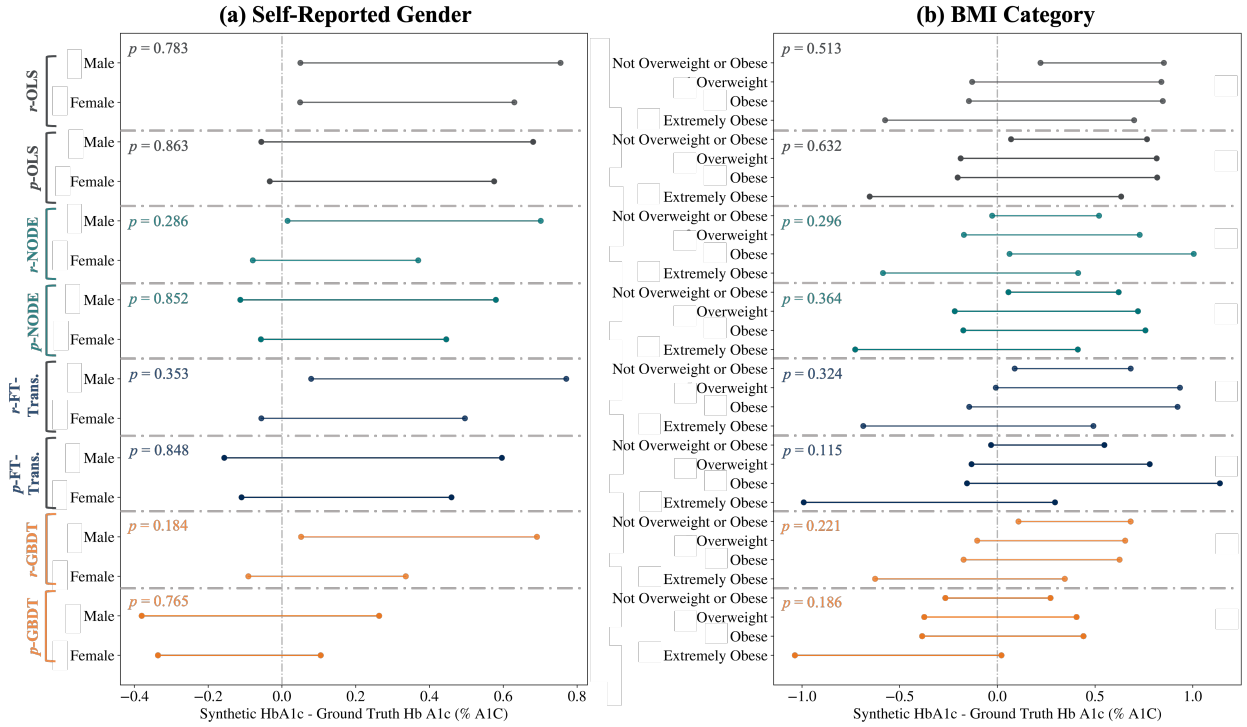

Figure 4.3: **Assessing for algorithmic bias in SynthA1c encoders.** We plotted the 95% confidence interval of the mean difference between the SynthA1c model output and ground truth HbA1c as a function of self-reported **(a)** gender and **(b)** BMI category. p values comparing the differences in SynthA1c model performance when stratified by gender (two-sample t-test) and BMI category (one-way ANOVA) are shown.

4.3.3. Ablation Studies: Relative Importance of CDPs and IDPs

Until now, prior T2DM classifiers have used only blood lab measurements and physical examination data to predict T2DM. In contrast, our models presented herein are the first to incorporate IDPs as input model features for the task of diabetes risk stratification. To better understand the benefit and value-add of using IDPs in conjunction with CDPs, we evaluated classifier performance on models trained using either only CDPs or only IDPs and compared them to corresponding models trained using both input types. Our results suggest that while classifier models trained only on CDPs generally outperform those trained only on IDPs, the best performance is achieved when combining CDPs and IDPs together (**Table 4.3**). This further validates the clinical utility of IDPs for patient diagnosis and disease risk stratification proposed by MacLean et al. (2021).

Table 4.3: **Patient feature ablation study.** We evaluate model performance as a function of whether clinically derived phenotypes (CDPs), image-derived phenotypes (IDPs), or both were used as input into the SynthA1c predictive model.

r**-NODE**	Recall	Precision	Specificity	Accuracy
CDPs Only	77.1	73.7	69.7	73.5
IDPs Only	73.4	76.9	75.8	74.5
CDPs + IDPs	82.6	76.9	72.7	77.9
r**-FT-Transformer**	Recall	Precision	Specificity	Accuracy
CDPs Only	78.0	76.6	73.7	75.9
IDPs Only	71.6	60.5	48.5	60.6
CDPs + IDPs	85.3	73.8	66.7	76.4
r**-GBDT**	Recall	Precision	Specificity	Accuracy
CDPs Only	80.7	68.6	59.6	70.7
IDPs Only	73.4	75.5	73.7	73.6
CDPs + IDPs	87.2	76.6	70.7	79.3

4.3.4. Characterizing Out-of-Domain Model Performance

As our metric $\mathbb{M}$ decreases across the three evaluated models, the RMSE in SynthA1c prediction decreases and the PCC increases, corresponding to better predictive performance on the OOD Iraqi Medical Center Dataset (**Table 4.4**). This supports our initial hypothesis that smoother models may generalize better to unseen datasets. We also noted larger RMSE values using the Iraqi Medical Center Dataset when compared to the PMBB outpatient test dataset results from **Table 4.2**.

Table 4.4: **SynthA1c model sensitivity and out-of-distribution (OOD) generalization results.** Smoothness metric values $\mathbb{M}$ were evaluated on the PMBB outpatient dataset. r-type models could not be evaluated on the Iraqi dataset because IDPs and medical imaging data were not available.

		Iraqi Dataset		PMBB Inpatient	
SynthA1c Encoder	$\mathbb{M}$	RMSE (% A1c)	PCC	RMSE (% A1c)	PCC
p'-/r- NODE	1.43	3.62 / —	0.154 / —	1.76 / 1.23	0.512 / 0.795
p'-/r- FT-Transformer	1.07	3.04 / —	0.246 / —	1.90 / 1.58	0.331 / 0.617
p'-/r- GBDT	3.28	6.25 / —	0.021 / —	1.54 / 1.12	0.674 / 0.823

Interestingly, we found that this relationship did not hold when considering the PMBB inpatient dataset; in fact, model predictive performance was *inversely* correlated with global smoothness. This suggested that the PMBB inpatient and outpatient dataset distributions were more similar than initially predicted. To validate this hypothesis, we computed the Kullback-Leibler (KL) divergence between each of the test dataset distributions and the training dataset distribution with respect to the features available in all datasets: ethnicity, gender, age, BMI, and HbA1c. We assumed the PMBB-derived outpatient training dataset was sampled from a distribution $\mathcal{Q}$ and each of the PMBB outpatient test, PMBB inpatient, and Iraqi medical center datasets were sampled from $\mathcal{P}_{\text{Outpatient}}$, $\mathcal{P}_{\text{Inpatient}}$, and $\mathcal{P}_{\text{Iraqi}}$, respectively. The greatest KL divergence was between the Iraqi medical center and training dataset distributions, as expected ($D_{KL}[\mathcal{P}_{\text{Iraqi}}||\mathcal{Q}] = 31.2$). Despite the fact that our training set included outpatient data alone, we found the KL divergence between the inpatient test and training datasets ($D_{KL}[\mathcal{P}_{\text{Inpatient}}||\mathcal{Q}] = 0.227$) was lower than that between the outpatient test and training dataset ($D_{KL}[\mathcal{P}_{\text{Outpatient}}||\mathcal{Q}] = 1.84$). To further characterize the feature distributions within our datasets, we analyzed the pairwise relationships between BMI, age, and HbA1c. Individual feature distributions were statistically significant between either of the PMBB datasets and the Iraqi Medical Center dataset (two-sample Kolmogorov-Smirnov (KS) test; $p < 0.0001$ between [PMBB inpatient dataset, Iraqi Medical Center dataset] and [PMBB outpatient dataset, Iraqi Medical Center dataset] pairs for individual age, HbA1c, and BMI quantitative features), but not between the PMBB inpatient and outpatient datasets (two-sample KS test; age: $p = 0.315$, HbA1c: $p = 0.463$, BMI: $p = 0.345$). These results help explain our initial findings regarding the relationship between $\mathbb{M}$ and model generalization.

4.4. Conclusion

Our work highlights the value of using CT-derived IDPs and CDPs for opportunistic screening of T2DM. We show that tabular learning architectures can act as novel SynthA1c encoders to predict HbA1c measurements noninvasively. Furthermore, we demonstrate that model manifold smoothness may be correlated with prediction performance on previously unseen data sampled from out-of-domain patient populations, although additional validation studies on separate tasks are needed. Ultimately, we hope that our proposed work may be used for opportunistic screening of type 2 diabetes—our proposed SynthA1c methodology will by no means replace existing diagnostic laboratory workups, but rather identify those at-risk patients who should consider consulting their physician for downstream clinical evaluation in an efficient and automated manner.

A few important limitations of our method warrant discussion. First, the mapping from the interpretable features (i.e. CDPs and IDPs) to the final SynthA1c prediction is modeled using a non-linear multilayer perceptron (MLP) in our method. We chose this strategy to capture more complex relationships between intermediate features and final predictions, and consistently yielded superior predictive performance across all evaluated datasets in our work. However, the non-linear nature of the MLP obfuscates direct attribution of individual feature representations to the output prediction, making the mapping less interpretable than alternative approaches (e.g., a linear mapping). Nevertheless, we argue that achieving strong, generalizable performance is more important than interpretability for its own sake in our application. Second, the current implementation does not explicitly address data missingness. For simplicity, we chose to exclude patients with *any* missing input features, although this strategy can potentially introduce systematic selection biases that may limit the external validity of our findings. This is especially relevant for generalizing our results to patient populations in resource-limited settings, where deriving image-derived phenotypes may not always be computationally feasible. Future iterations of this work may incorporate more principled imputation strategies or robust modeling techniques to better handle incomplete data. Lastly, real-world clinical validation of our method remains an essential direction for future work to assess the utility and generalizability of our method across diverse clinical settings.

CHAPTER 5

ADVERSARIAL SUPERVISION IN OFFLINE MODEL-BASED OPTIMIZATION

The following chapter is based on the first-author work:

(Yao et al., 2024) Michael S Yao, Yimeng Zeng, Hamsa Bastani, Jacob R Gardner, James C Gee, and Osbert Bastani. Generative adversarial model-based optimization via source critic regularization. In *Proc NeurIPS*, 2024. doi: 10.48550/arXiv.2402.06532.

I planned and performed experiments, analyzed the experimental data, and drafted the manuscript with input from all other authors.

5.1. Introduction

In the preceding chapters, we introduced a series of methods to build machine learning models that are more robust to common instances of distribution shift in the real-world. While it would be ideal for all current and future ML systems to adopt the algorithms we explored in this thesis, the current state-of-the-art models deployed in real-world user pipelines are most commonly black-box by design. Acknowledging this practical reality, a natural question arises: how can we prevent the out-of-distribution evaluation of any arbitrary machine learning model 'in-the-wild,' **including black-box models which may not have been originally designed with robustness in mind**?

In general, this is a challenging problem to solve. For instance, consider the example problem formulation shown in **Figure 1.3**. For a given function $f : \mathcal{X} \to \mathbb{R}$ mapping inputs from a domain $\mathcal{X}$ to real-valued outputs, we might only be able to train a machine learning model $f_\theta : \mathcal{X} \to \mathbb{R}$ on data spanning only a (potentially small) region of the entirety of the input domain. While the model may agree with f well for inputs similar to members of the training dataset of f_θ at inference time, there is no guarantee that the model's end users will not leverage f_θ to make predictions for new inputs that are wildly different from training examples. In these instances of *model extrapolation*, it is challenging to provide any meaningful bound on the error of f_θ compared to the true function f. If unrecognized, such prediction errors could have disastrous consequences in critical applications.

In this chapter, we consider a single such critical application—namely, **offline generative design**. To avoid instances of model extrapolation against f_θ in this setting, we propose a novel algorithm **GAMBO** (**G**enerative **A**dversarial **M**odel-**B**ased **O**ptimization) that achieves state-of-the-art performance across offline generative design problems spanning a wide variety of scientific domains.

5.2. Introduction to Generative Design

In many real-world tasks, we often seek to optimize the value of an objective function over some search space of inputs $\mathcal{X}$. Such **generative design** optimization problems span across a wide variety of domains, including molecule and protein design (Guimaraes et al., 2017; Brown et al., 2019; Maus et al., 2022), patient treatment effect estimation (Berrevoets et al., 2022), and resource allocation (Bastani et al., 2021). However, in many situations it may prove difficult or costly to estimate the objective function for any arbitrary input configuration. Evaluating newly proposed molecules requires expensive experimental laboratory setups, and testing multiple drug doses for a single patient can potentially be dangerous. In these scenarios, the allowable budget for objective function queries is prohibitive, limiting the utility of out-of-the-box online policy optimization methods.

To overcome this limitation, recent work has investigated the utility of optimization methods in the *offline* setting, where we are unable to query the objective function during the optimization process and instead only have access to a set of prior observations of inputs and associated objective values; this problem can often be referred to as offline *model-based optimization* (**MBO**) (Trabucco et al., 2021; Krishnamoorthy et al., 2023b). While one may naïvely attempt to learn a surrogate black-box model from the prior observations that approximates the true oracle objective function, such models can suffer from overestimation errors, yielding falsely promising objective estimates for inputs not contained in the offline dataset. As a result, offline optimization against the surrogate objective may yield low-scoring candidate designs according to the true oracle objective function—a key limitation of traditional policy optimization techniques in the offline setting.

To address this problem, we proposed a novel offline MBO algorithm (i.e., **GAMBO**) that leverages source critic models to optimize a surrogate objective while simultaneously remaining in-distribution when compared against a reference offline dataset. In this setting, an optimizer is re-

warded for proposing optima that are "similar" to reference data points, thereby minimizing overestimation error and allowing for more robust oracle function optimization in the offline setting. Inspired by recent work on generative adversarial networks (Goodfellow et al., 2014), we quantify design similarity by proposing a novel method that regularizes a surrogate objective model using a source critic actor, which we call *adaptive source critic regularization* (**aSCR**). We show how GAMBO and aSCR can be readily leveraged with common optimization methods, such as Bayesian optimization (BO) and first-order methods. Our contributions are as follows: first, we propose a novel approach for MBO that formulates the task as a constrained primal optimization problem, and we show how this framework can be used to solve for the optimal tradeoff between naïvely optimizing against the surrogate model and staying in-distribution relative to the offline dataset. Second, we introduce a computationally tractable method—which we call adaptive source critic regularization (aSCR)—to implement this framework with two popular optimization methods: Bayesian optimization and gradient ascent. Finally, we show that compared to prior methods, our proposed algorithm with Bayesian optimization empirically achieves the highest rank of **3.8** (second best is 5.5) on top-1 design evaluation, and highest rank of **3.0** (second best is 4.6) on top-128 design evaluation across a variety of tasks spanning multiple scientific domains.

5.3. Background

5.3.1. Offline Model-Based Optimization

In many real-world domains, we often seek to optimize an *oracle* objective function $f(\mathbf{x})$ over a space of design candidates $\mathcal{X}$ to solve for $\mathbf{x}^* = \text{argmax}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Examples of such problems include optimizing certain desirable properties of molecules in molecular design (Guimaraes et al., 2017; Brown et al., 2019; Maus et al., 2022), and estimating the optimal therapeutic intervention for patient care in clinical medicine (Berrevoets et al., 2022). In practice, however, the true objective function f may be costly to compute or even entirely unknown, making it difficult to query in optimizing $f(\mathbf{x})$. Instead, it is often more feasible to obtain access to a reference labeled dataset of observations from nature $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $y_i = f(\mathbf{x}_i)$. Optimization methods may use a variety of different strategies to leverage $\mathcal{D}_n$ in the offline setting (Krishnamoorthy et al.,

2023b,a; Chen et al., 2022); one common approach used by Trabucco et al. (2021) and others is to learn a regressor model f_θ parametrized by

$$\theta^* = \text{argmin}_\theta \ \mathbb{E}_{(\mathbf{x}_i, y_i) \sim \mathcal{D}_n} ||f_\theta(\mathbf{x}_i) - y_i||^2 \tag{5.1}$$

as a *surrogate model* for the true oracle objective $f(\mathbf{x})$. Rather than querying the oracle f as in the online setting, we can instead solve the related optimization problem

$$\mathbf{x}^* = \text{argmax}_{\mathbf{x} \in \mathcal{X}} f_\theta(\mathbf{x}) \tag{5.2}$$

with the hope that optimizing f_θ will also lead to desirable oracle values of f as well. Solving (5.2) is one instantiation of offline **model-based optimization** (**MBO**) for which a number of techniques have been developed, such as gradient ascent and Bayesian optimization (**BO**).

Of note, it is difficult to guarantee the reliability of the model's predictions for $\mathbf{x} \notin \mathcal{D}_n$ that are almost certainly encountered in the optimization trajectory. Thus, naïvely optimizing the surrogate objective f_θ can result in "optima" that are low-scoring according to the oracle objective f.

5.3.2. Optimization Over Latent Spaces

In certain cases, the search space $\mathcal{X}$ for an optimization task may be discretized over a finite set of structured inputs, such as amino acids for protein sequences or atomic building blocks for molecules. However, many historical optimization algorithms do not generalize well to these settings for a number of different reasons, such as the lack of gradients with respect to the input designs to guide the optimization trajectory. Instead of directly optimizing over $\mathcal{X}$, recent work leverages deep variational autoencoders (VAEs) to first map the input space into a continuous, (often) lower dimensional latent space $\mathcal{Z}$ and then performing optimization over $\mathcal{Z}$ instead (Tripp et al., 2020; Deshwal and Doppa, 2021; Maus et al., 2022). A VAE is composed of (1) an encoder with parameters ϕ that learns a posterior distribution $q_\phi(z|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}, z \in \mathcal{Z}$; and (2) a decoder with parameters φ that learns the conditional likelihood distribution $p_\varphi(\mathbf{x}|z)$ (Kingma and Welling, 2013).

The encoder and decoder are co-trained to maximize the evidence lower bound (ELBO)

$$\text{ELBO} = \mathbb{E}_{z \sim q_\phi} \left[\log p_\varphi(\mathbf{x}|z) \right] - D_{\text{KL}} \left[q_\phi(z|\mathbf{x}) \,||\, p_{\text{VAE}}(z) \right] \tag{5.3}$$

where D_{KL} is the Kullback-Leibler (KL) divergence and $p_{\text{VAE}}(z)$ is the prior distribution. A common choice is to set $p_{\text{VAE}} = \mathcal{N}(0, I)$ (i.e., the standard normal distribution). Optimization can then be performed over the continuous *latent space* $\mathcal{Z}$ of the VAE to propose 'latent space designs' that can be readily decoded using the decoder φ back into the original input space.

One such optimization method over VAE latent spaces is **Bayesian optimization (BO)**, a sample-efficient framework for solving optimization problems (Mockus, 1982; Snoek et al., 2012). While the utility of BO has primarily been explored for expensive-to-evaluate black-box functions in prior literature, recent work has shown that BO also outperforms baseline optimization methods in offline tasks involving models that are relatively inexpensive to evaluate, such as the neural network surrogates used in model-based optimization (MBO). Multiple prior works have shown that BO and related methods can outperform both gradient-based and stochastic evolutionary methods (Eriksson et al., 2019; Maus et al., 2022; Hvarfner et al., 2024; Eriksson and Jankowiak, 2021).

5.4. A Framework for Generative Adversarial Optimization

In this section we describe our proposed framework for generative adversarial model-based optimization using **adaptive source critic regularization** (**aSCR**). Our method uses a source critic model as in **Lemma 1** to dynamically regularize the optimization objective to avoid extrapolation against the proxy surrogate model f_θ in offline MBO.

5.4.1. Constrained Optimization Formulation

In offline generative optimization, we aim to optimize against a surrogate objective function f_θ. In order to ensure that we are achieving reliable estimates of the true, unknown oracle objective, we can add a regularization penalty to keep generated samples "similar" to those from the training dataset of f_θ according to an adversarial source critic trained to discriminate between generated and

offline samples. That is, in contrast to (5.2), aSCR considers a closely related *constrained* problem

$$\begin{aligned} \text{minimize}_{z \in \mathcal{Z}} \quad & -f_\theta(z) \\ \text{subject to} \quad & \mathbb{E}_{z' \in P}[c^*(z')] - c^*(z) \leq 0 \end{aligned} \tag{5.4}$$

over some configuration space $\mathcal{Z} \subseteq \mathbb{R}^d$, and where we define c^* as a source critic model that maximizes $\mathbb{E}_{z' \in P}[c^*(z')] - \mathbb{E}_{z \in Q}[c^*(z)]$ over all K-Lipschitz functions as in **Lemma 1**. We can think of $\mathbb{E}_{z' \in P}[c^*(z')] - c^*(z)$ as the contribution of a particular generated datum z to the overall $p = 1$ Wasserstein distance between the generated candidate (Q) and reference (P) distributions of designs. In practice, we model c^* as a fully connected neural net. Intuitively, the imposed constraint restricts the feasible search space to designs that score at least as in-distribution as the average sample in the offline dataset according to the source critic. Therefore, c^* acts as an adversarial model to regularize the optimization policy. Of note, our constraint in (5.4) may be highly non-convex, and so it is often impractical to directly apply (5.4) to any arbitrary MBO policy.

5.4.2. Dual Formulation

To solve this implementation problem, we instead look to reformulate (5.4) in its dual space by first considering the Lagrangian $\mathcal{L}$ of our constrained problem:

$$\mathcal{L}(z; \lambda) = -f_\theta(z) + \lambda \left[\mathbb{E}_{z' \in P}[c^*(z')] - c^*(z) \right] \tag{5.5}$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the constraint in (5.4). We can equivalently think of λ as a hyperparameter that controls the relative strength of the source critic-penalty term: $\lambda = 0$ equates to naïvely optimizing the surrogate objective, while $\lambda \gg 1$ asymptotically approaches a WGAN-like optimization policy. Minimizing $\mathcal{L}$ thus minimizes a relative sum of $-f_\theta$ and the Wasserstein distance contribution from any particular generated datum z with relative weighting dictated by the hyperparameter λ. From duality, minimizing $\mathcal{L}$ over z and simultaneously maximizing over $\lambda \in \mathbb{R}_+$ is equivalent to the original constrained problem in (5.4).

The challenge now is in determining this optimal value of λ: if λ is too small, then the objective

estimates may be unreliable; if λ is too large, then the optimization trajectory may be unable to adequately explore the input space. Prior work by Trabucco et al. (2021) has previously explored the idea of formulating offline optimization problems as a similarly regularized Lagrangian (albeit with a separate regularization constraint), although their method tunes a hyperparameter by hand. In contrast, aSCR treats λ as a dynamic parameter that adapts to the online optimization trajectory.

5.4.3. Computing the Lagrange Multiplier λ

Continuing with our dual formulation of (5.4), the Lagrange dual function $g(\lambda)$ is defined as $g(\lambda) = \inf_{z \in \mathbb{R}^n} \mathcal{L}(z; \lambda)$. The $z = \hat{z}$ that minimizes the Lagrangian in the definition of g is a function of λ. To show this, we use the first-order condition that $\nabla_z \mathcal{L} = 0$ at $z = \hat{z}$. Per (5.5), we have

$$\nabla_z \mathcal{L}(\hat{z}; \lambda) = -\nabla_z f_\theta(\hat{z}) - \lambda \nabla_z c^*(\hat{z}) = 0 \tag{5.6}$$

In general, solving (5.6) for $\hat{z}$ is computationally intractable—especially in high-dimensional problems. Instead, we can approximate $\hat{z}$ by relaxing the condition in (5.6) according to

$$\hat{z}(\lambda) = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \left|\left| -\nabla_z f_\theta(z) - \lambda \nabla_z c^*(z) \right|\right|^2 \tag{5.7}$$

Our key insight is that although minimizing the loss term in (5.7) is not practical when the feasible set is naïvely uniform over $\mathbb{R}^n$, we can instead choose to focus our attention on latent space coordinates with high associated probability according to the VAE prior distribution $p_{\text{VAE}}(z)$. This is because in optimization problems acting over the latent space of any variational autoencoder, the majority of the encoded information content is embedded according to $p_{\text{VAE}}(z)$ due to the Kullback-Leibler (KL) divergence contribution to VAE training. Put simply, the encoder distribution $q_\phi(z|\mathbf{x})$ is trained so that $D_{\text{KL}}[q_\phi(z|\mathbf{x})||p_{\text{VAE}}(z))]$ is optimized as a regularization term in (5.3). We argue that it is thus sufficient enough to approximate $\hat{z}(\lambda)$ using a Monte Carlo sampling schema with random samples $\mathcal{Z}_N = (z_1, z_2, \ldots, z_N) \sim p_{\text{VAE}}(z)$:

$$\hat{z}(\lambda) \approx \underset{\mathcal{Z}_N \sim p_{\text{VAE}}(z)}{\operatorname{argmin}} \frac{1}{2} \left|\left| -\nabla_z f_\theta(z) - \lambda \nabla_z c^*(z) \right|\right|^2 \tag{5.8}$$

We can now concretely write an approximation of the Lagrange dual problem of (5.4):

$$\begin{aligned} &\text{maximize} \quad g(\lambda) = -f_\theta(\hat{z}) + \lambda \left[\mathbb{E}_{z' \in P}[c^*(z')] - c^*(\hat{z})\right] \\ &\text{subject to} \quad \lambda \geq 0 \end{aligned} \tag{5.9}$$

where $\hat{z}$ is as in (5.8). Defining the surrogate variable α such that $\lambda = \frac{\alpha}{1-\alpha}$, we rewrite (5.9) as

$$\begin{aligned} &\text{maximize} \quad -(1-\alpha) f_\theta(\hat{z}) + \alpha \left[\mathbb{E}_{z' \in P}[c^*(z')] - c^*(\hat{z})\right] \\ &\text{subject to} \quad 0 \leq \alpha < 1 \end{aligned} \tag{5.10}$$

We discretize the search space for α to 200 evenly spaced points between 0 and 1 inclusive. From weak duality, finding the optimal solution to (5.9) provides a lower bound on the optimal solution to the primal problem in (5.4). **Algorithm 1** can now be used to choose the optimal α (and hence λ) adaptively during offline optimization: we refer to our method as **Adaptive SCR (aSCR)**.

Algorithm 1 Adaptive Source Critic Regularization (SCR)

Input: differentiable surrogate objective $f_\theta : \mathbb{R}^d \to \mathbb{R}$, differentiable source critic $c : \mathbb{R}^d \to \mathbb{R}$, reference dataset $\mathcal{D}_n = \{z'_j\}_{j=1}^n$, α step size $\Delta\alpha$, search budget $\mathcal{B}$, norm threshold τ
Sample candidates $\mathcal{Z}_\mathcal{B} \leftarrow \{z_i\}_{i=1}^\mathcal{B} \sim \mathcal{N}(0, I_d)$
Initialize $\alpha^* \leftarrow$ None and $g^* \leftarrow -\infty$
for α **in** range(start $= 0$, end $= 1$, stepsize $= \Delta\alpha$) **do**
  $z^* \leftarrow \text{argmin}_{z_i \in \mathcal{Z}_\mathcal{B}} ||(1-\alpha)\nabla f_\theta(z_i) + \alpha \nabla c(z_i)||_2$
  if $||(1-\alpha)\nabla f_\theta(z^*) + \alpha \nabla c(z^*)||_2 > \tau$ **then**
    continue // Discard α if best norm exceeds τ
  end if
  $g \leftarrow -(1-\alpha) f_\theta(z^*) + \alpha \left[\mathbb{E}_{\mathcal{D}_n}[c(z'_j)] - c(z^*)\right]$
  if $g > g^*$ **then**
    $\alpha^* \leftarrow \alpha$ and $g^* \leftarrow g$ // Implements (5.10)
  end if
end for
return α^*

5.4.4. Overall Algorithm

Using Adaptive SCR, we now have a proposed method for dynamically computing α (and hence the Lagrange multiplier λ) of the constrained optimization problem in (5.4). Importantly, aSCR can be integrated with any standard function optimization method by optimizing the Lagrangian objective in (5.5) over the candidate design space as opposed to the original unconstrained objec-

tive f_θ. We refer to this algorithm as *Generative Adversarial Model-Based Optimization* (GAMBO). To evaluate aSCR empirically, we instantiate two flavors of GAMBO: (1) **G**enerative **A**dversarial **B**ayesian **O**ptimization (**GABO**); and (2) **G**enerative **A**dversarial **G**radient **A**scent (**GAGA**)—see **Algorithm 2** for additional details.

Algorithm 2 Generative Adversarial Model-Based Optimization (GAMBO)

Input: surrogate objective $f_\theta : \mathcal{X} \to \mathbb{R}$, offline dataset $\mathcal{D}_n = \{z'_j\}_{j=1}^n$, acquisition function a, iterative sampling budget T, sampling batch size b, number of generator steps per source critic training $n_{\text{generator}}$, oracle query budget k, batched acquisition function $a^b : \mathcal{X} \times \mathbb{R} \to \mathcal{X}^b$
AdaptiveSCR Input: α step size $\Delta\alpha$, search budget $\mathcal{B}$, norm threshold τ
Define: Differentiable source critic $c : \mathbb{R}^d \to \mathbb{R}$
Define: Lagrangian $\mathcal{L}(z; \alpha) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ per (5.5): $\mathcal{L}(z; \alpha) = -f_\theta(z) + \frac{\alpha}{1-\alpha}[\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - c(z)]$
Sample candidates $\mathcal{Z}^1 \leftarrow \{z_i^1\}_{i=1}^b \sim$ SobolSequence
// Train the source critic per Lemma 1 to optimality:
$c \leftarrow \text{argmax}_{||c||_L \leq K} W_1(\mathcal{D}_n, \mathcal{Z}^1) = \text{argmax}_{||c||_L \leq K} [\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - \mathbb{E}_{z \sim \mathcal{Z}^1}[c(z)]]$
$\alpha \leftarrow$ **AdaptiveSCR**$(f_\theta, c, \mathcal{D}_n, \Delta\alpha, \mathcal{B}, \tau)$ // Alg. (1)
Evaluate candidates $\mathcal{Y}^1 \leftarrow \{y_i^1\}_{i=1}^b = \{-\mathcal{L}(z_i^1; \alpha)\}_{i=1}^b$
if backbone optimizer is Bayesian Optimization (i.e., GABO) **then**
  Place Gaussian Process (GP) prior on f_θ
end if
for t **in** $2, 3, \ldots, T$ **do**
  if backbone optimizer is Bayesian Optimization (i.e., GABO) **then**
    Update posterior on f_θ with $\mathcal{D}_{t-1} = \{(\mathcal{Z}^m, \mathcal{Y}^m)\}_{m=1}^{t-1}$
    Compute acquisition function a^b using fitted posterior
  end if
  Sample candidates $\mathcal{Z}^t \leftarrow \{z_i^t\}_{i=1}^b = a^b(\mathcal{D}_{t-1})$ (i.e., 4 gradient ascent steps for GAGA)
  $\alpha \leftarrow$ **AdaptiveSCR**$(f_\theta, c, \mathcal{D}_n, \Delta\alpha, \mathcal{B}, \tau)$
  Evaluate samples $\mathcal{Y}^t \leftarrow \{y_i^t\}_{i=1}^b = \{-\mathcal{L}(z_i^t; \alpha)\}_{i=1}^b$
  if t mod $n_{\text{generator}}$ equals 0 **then**
    // Train the source critic per Lemma 1 to optimality:
    $c \leftarrow \text{argmax}_{||c||_L \leq K} W_1(\mathcal{D}_n, \mathcal{Z}^t)$
      $= \text{argmax}_{||c||_L \leq K} [\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - \mathbb{E}_{z \sim \mathcal{Z}^t}[c(z)]]$
  end if
end for
return the top k samples from the $T \times b$ observations
    $\mathcal{D}_T = \{\{(z_i^m, y_i^m)\}_{i=1}^b\}_{m=1}^T$ according to y_i^m

5.5. Experimental Methods

We implement GABO using a quasi-expected improvement (qEI) acquisition function, iterative sampling budget of $T = 32$, sampling batch size of $b = 64$, and GAGA using a step size of $\eta = 0.05$, $T = 128$, and $b = 16$.

5.5.1. Datasets and Tasks

To evaluate our proposed algorithm, we focus on a set of eight tasks spanning multiple domains with publicly available datasets in the field of offline model-based optimization. (1) The **Branin** function is a well-known synthetic benchmark function where the task is to maximize the two-dimensional Branin function $f_{br} : [-5, 10] \times [0, 15] \rightarrow \mathbb{R}$ (Branin, 1972). (2) The **LogP** task is a well-studied optimization problem (Zhou et al., 2019; Chen et al., 2021; Flam-Shepherd et al., 2022) where we search over candidate molecules to maximize the penalized water-octanol partition coefficient (logP) score, which is an approximate measure of a molecule's hydrophobicity (Ertl and Schuffenhauer, 2009) that also rewards structures that can be synthesized easily and feature minimal ring structures. We use the publicly available Guacamol benchmarking dataset from Brown et al. (2019) to implement this task.

Tasks (3) - (7) are derived from Design-Bench, a publicly available set of MBO benchmarking tasks (Trabucco et al., 2022): (3) **TF-Bind-8** aims to maximize the transcription factor binding efficiency of an 8-base-pair DNA sequence (Barrera et al., 2016); (4) **GFP** the green fluorescence of a 237-amino-acid protein sequence (Brookes et al., 2019; Rao et al., 2019); (5) **UTR** the gene expression from a 50-base-pair 5'UTR DNA sequence (Sample et al., 2019; Angermueller et al., 2020a); (6) **ChEMBL** the mean corpuscular hemoglobin concentration (MCHC) biological response of a molecule using an offline dataset collected from the ChEMBL assay `CHEMBL3885882` (Gaulton et al., 2012); and (7) **D'Kitty** the morphological structure of the D'Kitty robot (Ahn et al., 2020).

Finally, (8) the **Warfarin** task uses the dataset of patients on warfarin medication from Consortium (2009) to estimate the optimal dose of warfarin given clinical and pharmacogenetic patient data. Of note, in contrast to tasks (1) - (7) and other traditional MBO tasks in prior work (Trabucco et al., 2022), the Warfarin task is novel in that only a subset of the input design dimensions may be optimized over (i.e., warfarin dose) while the others remain fixed as conditioning variables (i.e., patient covariates). Such a task can therefore be thought of as *conditional* model-based optimization.

5.5.2. Oracle Functions

All oracle functions for the tasks assessed are either exact functions or approximate oracles developed by domain experts. Specifically, the **Branin** and **TF-Bind-8** tasks utilize exact oracles described in detail by Branin (1972) and Barrera et al. (2016), respectively. The oracle for the Penalized **LogP** task is an approximate oracle from Wildman and Crippen (1999) that is the same oracle used by domain experts in the Guacamol benchmarking study (Brown et al., 2019). The **GFP**, **UTR**, and **ChEMBL** tasks feature approximate oracles from Angermueller et al. (2020a), Snoek et al. (2012), and Trabucco et al. (2022), respectively, that were trained on a larger, hidden datasets inaccessible to us for the respective tasks. The **D'Kitty** morphology task uses a MuJoCo (Todorov et al., 2012) simulation environment and learned control policy from Trabucco et al. (2022) to evaluate proposed designs. Finally, the **Warfarin** task uses a linear model (Consortium, 2009) to estimate a patient's optimal warfarin dose given their pharmacogenetic attributes.

5.5.3. Data Preprocessing

For the (1) **Branin** task, we sample 1000 points from the square input domain $[-5, 10] \times [0, 15]$ to construct the offline dataset, and remove the top 20%-ile according to the oracle function to make the task more challenging in line with prior work (Krishnamoorthy et al., 2023b). In this continuous task (along with the **D'Kitty** and **Warfarin** tasks), we treat input designs as their own latent space mappings, such that the VAE encoder and decoder for this task are both the identity function with zero trainable parameters. The offline dataset of the (2) Penalized **LogP** task is the validation partition of the Guacamol dataset from Brown et al. (2019), which consists of 79,564 unique molecules and their corresponding penalized LogP scores. The input molecules are represented as SMILES strings (Weininger, 1988), which is a molecule representation format shown to frequently yield invalid molecules in prior work (Krenn et al., 2020). Therefore, we encode the molecules instead as SELFIES strings—an alternative molecule representation from Krenn et al. (2020).

The (3) **TF-Bind-8**, (4) **GFP**, and (5) **UTR** tasks are assessed as-released by Design-Bench from Trabucco et al. (2022)—please refer to their work for task-specific descriptions. In the (6) **ChEMBL** and (7) **D'Kitty** tasks, we normalize all objective values y in the offline dataset to $\hat{y} = (y -$

$y_{\text{min}})/(y_{\text{max}} - y_{\text{min}})$ as done in prior work (Krishnamoorthy et al., 2023b), where $\hat{y}$ is the corresponding normalized objective value and y_{min} (y_{max}) is the minimum (maximum) observed objective value in the full, *unobserved* dataset. Because only the bottom 60%-ile (40%-ile) from the full dataset is used in the available offline dataset for the ChEMBL (D'Kitty) task, the respective maximum $\hat{y}$ values are less than 1.0 (**Table 5.1**). We also translate the original SMILES string representations in the ChEMBL task into SELFIES strings (Krenn et al., 2020) as in the LogP task.

Finally, the (8) **Warfarin** task uses the dataset of pharmacogenetic patient covariates published by Consortium (2009). We split the original dataset of 3,936 unique patient observations into training (validation) partitions with 3,736 (200) datums. The patient attributes in the Warfarin dataset consist of a combination of discrete and continuous values. All discrete attributes are one-hot encoded into binarized dimensions, and continuous values are normalized to zero mean and unit variance using the training dataset. Missing patient values were imputed following prior work (Truda and Marais, 2021). We define the cost $c(z|x)$ accrued by a patient with attributes $x \in \mathbb{R}^{32}$ as a function of the input dose $z \in \mathbb{R}$ is $c(z|x) = (z - d_{\text{oracle}}(x))^2$, where $d_{\text{oracle}} : \mathbb{R}^{32} \to \mathbb{R}$ is the domain-expert oracle warfarin dose estimator from Consortium (2009). The observed objective values y associated with each of the training datums is calculated as $y = [c(\bar{z}|x) - c(z|x)]/c(\bar{z}|x)$, where $\bar{z}$ is the mean warfarin dose over the training dataset and z is the true dose given to the patient. Using this constructed offline dataset, our task is then to assign optimal doses to the 200 validation patients to maximize y with *no* prior warfarin dosing observations.

5.5.4. Policy Optimization and Evaluation

For all experiments, the surrogate objective model f_θ is a fully connected net with two hidden layers of size 2048 and LeakyReLU activations. f_θ takes as input a VAE-encoded latent space datum and returns the predicted objective function value as output. The VAE encoder and decoder backbone architectures vary by MBO task and are detailed in **Table 5.1**.

Following Gómez-Bombarelli et al. (2018) and Maus et al. (2022), we co-train the VAE and surrogate objective models together using an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 3×10^{-4} for all tasks. For the optimization tasks over continuous design spaces (i.e., Branin,

Warfarin, and D'Kitty), we fix the VAE encoder and decoders as the identity functions, such that the latent and input spaces are equivalent.

Table 5.1: **MBO datasets and tasks.** Implementation details for each of the eight MBO tasks assessed in our work. *Denotes the life sciences-related discrete MBO tasks offered by the Design-Bench benchmarking repository (Trabucco et al., 2022).

Property	**Branin**	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**
Dataset Size	800	79,564	32,898	5,000	140,000	441	10,004	200
Input Shape	2	108	8	237	50	32	56	1 (33)
Vocabulary Size	—	97	4	20	4	40	—	—
VAE Backbone	Identity	Transformer	ResNet	ResNet	ResNet	Transformer	Identity	Identity
VAE Latent Shape	2	256	16	32	32	128	56	33
Oracle	Exact	Linear	Exact	Transformer	ResNet	Random Forest	Exact	Linear
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96

The source critic agent c introduced in **Lemma 5.4** is implemented as a fully connected net with two hidden layers with sizes equal to four (one) times the number of input dimensions for the first (second) layer. To constrain the Lipschitz norm of c as in **Lemma 1**, we clamp the weights of the model between [-0.01, 0.01] after each optimization step as done by Arjovsky et al. (2017). The model is trained using gradient descent with a learning rate of 0.001 to maximize the Wasserstein distance between the dataset and generated candidates in the VAE latent space.

During optimization, both GABO and GAGA alternate between sampling new designs and training the source critic actor $c(z)$ until there is no improvement to the Wasserstein distance W_1 according to c after 100 consecutive weight updates. We find that training c every $n_{\text{generator}} = 4$ sampling steps is a good choice across all tasks assessed, similar to prior work Arjovsky et al. (2017).

All MBO methods were evaluated using a fixed surrogate query budget of 2048. We focus on two evaluation metrics: 100th percentile (1) top $k = 1$; and (2) top $k = 128$ oracle score. The top $k = 128$ evaluation metric is commonly reported in prior offline MBO literature (Krishnamoorthy et al., 2023b; Trabucco et al., 2021; Yu et al., 2021); the top $k = 1$ metric better accounts for the limited oracle query budget of the real-world tasks in which offline MBO would be of use. In both settings, an optimizer selects the top k design that minimize the Lagrangian function value in (5.5) from the 2048 assessed designs to evaluate using the true oracle function, and the maximum score of those k designs is reported across 10 random seeds.

We evaluate both GABO and GAGA against a number of pre-existing baseline algorithms on one internal cluster with 8 NVIDIA RTX A6000 GPUs. We include vanilla Bayesian Optimization (**BO**-qEI) and gradient ascent (**Grad.**) in our evaluation to assess the utility of our proposed aSCR algorithm. Furthermore, we evaluate limited-memory BFGS (**L-BFGS**) Liu and Nocedal (1989), **CMA-ES** Hansen and Ostermeier (1996), and simulated annealing (**Anneal**) Kirkpatrick et al. (1983). We also compare our method against **TuRBO**-qEI (Eriksson et al., 2019), **COM** (Trabucco et al., 2021), **RoMA** (Yu et al., 2021), **BDI** (Chen et al., 2022), **DDOM** (Krishnamoorthy et al., 2023a), **BONET** (Krishnamoorthy et al., 2023b), **ExPT** (Nguyen et al., 2023), **BootGen** (Kim et al., 2023), and **ROMO** (Chen et al., 2023c). Of note, because BootGen is proposed by Kim et al. (2023) as an optimization method specifically for biological sequence design, we only assess this baseline method on the five relevant tasks in our evaluation suite.

Conditional MBO Tasks. To our knowledge, prior work in conditional model-based optimization is limited, and so previously reported algorithms are not equipped to solve such tasks out-of-the-box. Chen et al. (2023c) explore such tasks in their work, but primarily focus on conditional tasks that are built by arbitrarily fixing certain design dimensions from unconstrained problems, which are not representative of true conditional optimization problems in the real world. In our work, we introduce the Warfarin task to assess methods on their ability to design an optimal therapeutic drug regiment *conditioned* on a fixed patient state and lab values. To assess existing methods on this task, we implement conditional proxies of all baselines employing a first-order optimization schema via *partial* gradient ascent to only update the warfarin dose dimension while leaving the patient attribute conditional dimensions unchanged. Conditional BO-based methods are implemented by fitting separate Gaussian processes for each patient. In conditional DDOM, we exchange the algorithm's diffusion model with a *conditional* diffusion model (Gu et al., 2023).

Of note, the BONET algorithm (Krishnamoorthy et al., 2023b) requires multiple observations for any given patient to construct synthetic optimization trajectories. However, the key challenge in conditional MBO is that each condition (i.e., patient) has *no* past observations (i.e., warfarin doses), and instead relies on learning from offline datasets constructed from different permutations of

condition values. As a result, we could not evaluate BONET on conditional MBO tasks.

5.6. Results

Scoring of one-shot optimization candidates is shown in **Table 5.2**. Across all eight assessed tasks spanning a wide range of scientific domains, GABO with our aSCR algorithm achieved the best average rank of **3.8** when compared to other existing methods (next best is 5.5). Furthermore, GABO was able to propose top $k = 1$ candidate designs that outperform the best design in the pre-existing offline dataset for 6 of the 8 tasks–greater than any of the other methods assessed. If a larger oracle evaluation budget is available (i.e., $k = 128$), GABO with aSCR performs even better, achieving the best average rank of **3.0** (next best is 4.6). GABO is also the best algorithm on 3 of the 8 tasks and second best on 2 tasks according to this evaluation metric. Altogether, our results suggest that GABO is a promising method for proposing optimal design candidates in offline MBO.

Importantly, our aSCR algorithm improves upon both the naïve BO-qEI and Grad. Ascent parent optimizers assessed. GABO outperforms both baseline BO-based optimization methods in our evaluation suite: BO (TuRBO) only achieves a rank of 8.8 (9.0) on the top $k = 1$ evaluation metric and a rank of 6.6 (7.4) on the top $k = 128$ metric. Similarly, GAGA scores an average rank of 7.4 (7.6) on the top $k = 1$ ($k = 128$) evaluation metric; by leveraging aSCR, GAGA outperforms its base parent optimizer (Grad. Ascent), which only achieves an average rank of 9.0 and 11.0 on the same two evaluation metrics, respectively. Our results show that using aSCR to adaptively penalize the objective of two popular optimization methods can improve their offline performance.

5.6.1. Qualitative Evaluation: Penalized LogP Task

We evaluate GABO against naïve BO-qEI for the **LogP** task by inspecting the three-dimensional chemical structures of the top-scoring candidate molecules. As a general principle, molecules that are associated with high Penalized LogP scores are hydrophobic with minimal ring structures and therefore often feature long hydrocarbon backbones (Ertl and Schuffenhauer, 2009). In Figure 5.1, we see that BO-qEI using the unconstrained surrogate objective generates a candidate molecule of hydrogen and carbon atoms. However, the proposed candidate includes two rings in its structure, resulting in a suboptimal oracle Penalized LogP score.

Table 5.2: **Constrained budget** ($k = 1$) **oracle evaluation.** Each method proposes a single design that is evaluated using the oracle function to report the final score (higher is better) across 10 random seeds reported as mean ± standard deviation. $\mathcal{D}$ (best) reports the top oracle value in the task dataset. Each of the MBO methods are ranked by their mean one-shot oracle score, and the average rank (lower is better) across all eight tasks is reported in the final table column. **Bold** (resp., Underlined) entries indicate the best (resp., second best) entry in the column. *Denotes the MBO tasks from Trabucco et al. (2022).

Method	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin	Rank
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96	—
Grad.	-245.1 ± 81.3	-5.37 ± 1.44	0.429 ± 0.023	3.18 ± 0.88	6.82 ± 0.21	-1.95 ± 0.00	0.57 ± 0.19	0.86 ± 1.09	9.0
L-BFGS	-29.6 ± 0.0	3.82 ± 32.6	0.527 ± 0.140	3.51 ± 0.70	6.48 ± 1.20	-1.95 ± 0.00	0.31 ± 0.00	0.73 ± 1.83	8.5
CMA-ES	-8.6 ± 3.6	5.04 ± 6.83	0.438 ± 0.131	1.43 ± 0.00	6.39 ± 0.11	-1.95 ± 0.00	0.31 ± 0.00	-25.0 ± 150	10.6
Anneal	-9.6 ± 1.5	8.76 ± 0.15	0.807 ± 0.094	3.64 ± 0.03	5.01 ± 0.31	-1.95 ± 0.00	0.55 ± 0.18	**0.91 ± 0.08**	6.8
BO	-11.0 ± 7.8	-52.5 ± 88.8	0.586 ± 0.193	1.43 ± 0.00	5.65 ± 1.30	0.59 ± 0.10	0.61 ± 0.15	0.16 ± 1.67	8.8
TuRBO	-21.0 ± 5.1	-45.1 ± 93.8	0.564 ± 0.194	1.43 ± 0.00	6.53 ± 1.19	**0.65 ± 0.00**	0.44 ± 0.18	0.05 ± 0.11	9.0
BONET	-26.1 ± 0.9	10.8 ± 0.33	0.282 ± 0.000	**3.74 ± 0.00**	**9.12 ± 0.07**	0.55 ± 0.13	0.78 ± 0.00	—	5.7
DDOM	-6677 ± 6360	-4.23 ± 1.28	0.460 ± 0.030	1.43 ± 0.00	5.56 ± 0.02	0.54 ± 0.15	0.51 ± 0.20	-0.32 ± 0.40	11.1
COM	-3099 ± 32.6	**30.8 ± 19.5**	0.439 ± 0.000	3.62 ± 0.00	6.65 ± 0.43	0.63 ± 0.01	**0.90 ± 0.02**	0.72 ± 0.97	5.5
RoMA	-32.7 ± 18.4	6.37 ± 1.39	0.433 ± 0.040	3.37 ± 0.27	6.66 ± 0.98	0.50 ± 0.14	0.30 ± 0.27	-0.70 ± 0.02	9.4
BDI	-1050 ± 0.0	-0.20 ± 0.00	0.311 ± 0.000	3.26 ± 0.82	5.61 ± 0.00	0.48 ± 0.00	0.67 ± 0.00	-24.8 ± 233	10.8
ExPT	-57.2 ± 38.6	-15.9 ± 24.1	0.571 ± 0.076	1.43 ± 0.00	6.77 ± 1.38	0.56 ± 0.06	0.66 ± 0.20	-34.6 ± 61.4	9.1
BootGen	—	-13.0 ± 15.1	**0.942 ± 0.022**	3.10 ± 0.73	8.30 ± 0.93	0.59 ± 0.07	—	—	6.2
ROMO	-2614 ± 739.9	-20.5 ± 19.2	0.382 ± 0.203	3.55 ± 0.13	5.73 ± 1.42	**0.65 ± 0.00**	0.64 ± 0.27	-0.71 ± 2.10	9.6
GAGA	-2.9 ± 2.2	-68.6 ± 109.8	0.571 ± 0.120	**3.74 ± 0.00**	5.89 ± 1.42	-1.95 ± 0.00	0.89 ± 0.00	0.01 ± 0.14	7.4
GABO	**-2.6 ± 1.1**	21.3 ± 33.2	0.570 ± 0.131	3.60 ± 0.40	7.51 ± 0.39	0.60 ± 0.07	0.71 ± 0.01	0.60 ± 1.80	**3.8**

Table 5.3: **Relaxed budget** ($k = 128$) **oracle evaluation.** Each method now proposes 128 designs that are evaluated using the oracle function, and the maximum score out of these 128 designs is reported below (averaged across 10 random seeds and reported as mean ± standard deviation). $\mathcal{D}$ (best) reports the top oracle value in the task dataset. Each of the MBO methods are ranked by their mean $k = 128$-shot oracle score, and the average rank (lower is better) across all eight tasks is reported in the final table column. **Bold** (Underlined) entries indicate the best (second best) entry in the column. *Denotes the life sciences-related tasks from Design-Bench (Trabucco et al., 2022).

Method	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin	Rank
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96	—
Grad.	-115.3 ± 20.8	-5.14 ± 1.70	0.977 ± 0.025	3.49 ± 0.69	7.38 ± 0.15	-1.95 ± 0.00	0.87 ± 0.02	0.86 ± 1.08	11.0
L-BFGS	-4.0 ± 0.0	42.8 ± 9.44	0.633 ± 0.140	**3.74 ± 0.00**	7.51 ± 0.39	-1.95 ± 0.00	0.31 ± 0.00	0.75 ± 1.67	10.1
CMA-ES	-4.3 ± 1.7	47.6 ± 5.46	0.810 ± 0.235	**3.74 ± 0.00**	7.40 ± 0.32	-1.95 ± 0.00	0.74 ± 0.00	-8.62 ± 63.8	9.8
Anneal	-7.4 ± 2.8	11.3 ± 0.00	0.890 ± 0.035	3.72 ± 0.00	7.96 ± 0.22	-1.95 ± 0.00	0.88 ± 0.00	0.97 ± 0.08	9.3
BO	**-0.4 ± 0.0**	**135.3 ± 16.0**	0.942 ± 0.025	2.26 ± 1.03	8.26 ± 0.09	0.67 ± 0.00	0.72 ± 0.00	0.93 ± 0.11	6.6
TuRBO	-0.7 ± 0.4	59.7 ± 51.3	0.895 ± 0.049	1.89 ± 0.92	8.26 ± 0.11	0.67 ± 0.01	0.72 ± 0.00	0.99 ± 0.01	7.4
BONET	-26.0 ± 0.9	11.7 ± 0.38	0.951 ± 0.035	**3.74 ± 0.00**	9.13 ± 0.08	0.67 ± 0.01	0.95 ± 0.01	—	5.6
DDOM	-18.4 ± 29.8	-2.16 ± 0.60	0.936 ± 0.051	1.44 ± 0.00	8.30 ± 0.33	0.66 ± 0.01	0.89 ± 0.01	**1.00 ± 0.00**	8.4
COM	-1981 ± 224.5	42.0 ± 16.9	0.902 ± 0.056	3.62 ± 0.00	8.18 ± 0.00	0.64 ± 0.01	0.95 ± 0.02	0.77 ± 0.86	8.5
RoMA	-4.8 ± 3.0	10.8 ± 0.78	0.760 ± 0.113	**3.74 ± 0.00**	8.12 ± 0.09	0.69 ± 0.03	**1.02 ± 0.04**	0.67 ± 0.05	7.8
BDI	-65.0 ± 51.3	1.52 ± 5.79	0.735 ± 0.086	3.61 ± 0.05	6.31 ± 0.00	0.50 ± 0.12	0.94 ± 0.01	-5.07 ± 21.0	11.8
ExPT	-1.7 ± 1.0	-6.48 ± 4.58	0.927 ± 0.095	**3.74 ± 0.00**	8.13 ± 0.09	0.68 ± 0.04	0.97 ± 0.01	0.96 ± 0.05	6.5
BootGen	—	8.10 ± 3.31	**0.979 ± 0.002**	**3.74 ± 0.00**	**10.5 ± 0.95**	0.68 ± 0.00	—	—	4.6
ROMO	-2367 ± 787.5	-6.05 ± 14.5	0.572 ± 0.202	3.67 ± 0.03	6.94 ± 1.07	0.65 ± 0.00	0.90 ± 0.02	0.76 ± 1.91	12.1
GAGA	-1.0 ± 0.2	14.1 ± 25.0	0.722 ± 0.091	**3.74 ± 0.00**	7.98 ± 0.36	-1.95 ± 0.00	0.90 ± 0.01	0.95 ± 0.07	7.6
GABO	-0.5 ± 0.1	122.1 ± 20.6	0.954 ± 0.025	**3.74 ± 0.00**	8.36 ± 0.08	**0.70 ± 0.01**	0.72 ± 0.00	**1.00 ± 0.03**	**3.0**

We hypothesize that this may be due to a lack of ring-containing example molecules in the offline dataset, as only 6.7% (2.7%) of observed molecules contain at least one (two) carbon ring(s). As a result, the surrogate objective model estimator returns more inaccurate Penalized LogP estimates for input ring-containing structures (surrogate model root mean squared error (RMSE) = 25.5 for offline dataset molecules with at least 2 rings; RMSE = 16.5 for those with at least 1 ring; and RMSE = 4.6 for those with at least 0 rings), leading to sub-par BO-qEI optimization performance as the unconstrained algorithm extrapolates against the surrogate to find "optimal" molecules that are out-of-distribution. In contrast, GABO generates a candidate molecule with a long hydrocarbon backbone and *no* rings, resulting in a penalized logP score of 22.1—greater than the best observed value in the offline dataset for the task.

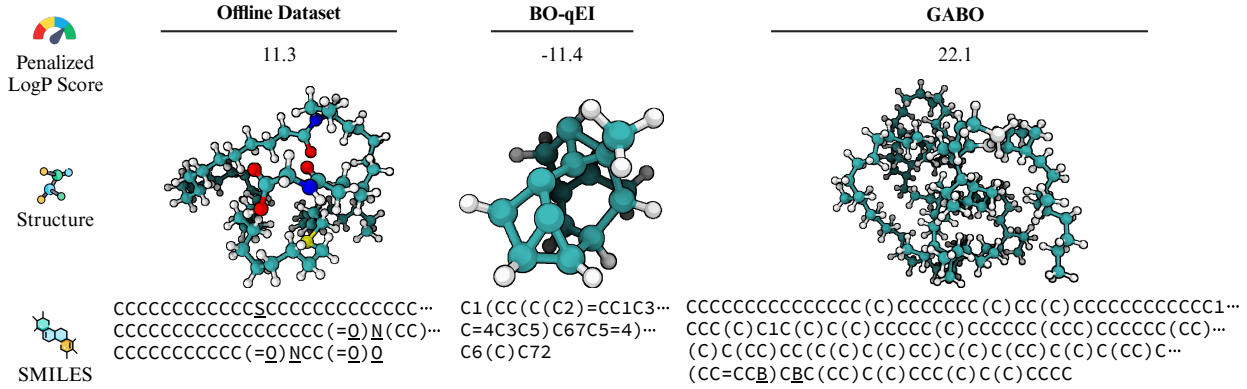

Figure 5.1: **Penalized LogP score maximization sample candidate designs.** (**Left**) The molecule with the highest penalized LogP score of 11.3 in the offline dataset. Here, we show the 100th percentile candidate molecules according to the surrogate objective generated from (**Middle**) vanilla BO-qEI and (**Right**) GABO. Teal- (white-) colored atoms are carbon (hydrogen). Non-hydrocarbon atoms are underlined in the SMILES (Weininger, 1988) string representations.

5.6.2. Ablation Studies

Adaptive SCR Algorithm Ablation. Taking inspiration from (Trabucco et al., 2021), it is possible to utilize our SCR algorithm in GABO *without* dynamically computing α (and hence the Lagrange multiplier λ). To better characterize the utility of aSCR, we ablate **Algorithm 1** by treating λ instead as a hand-tunable constant hyperparameter, and test our method using different values of $\lambda = \alpha/(1-\alpha)$ (**Table 5.4**). Setting $\alpha = 0$ (i.e., $\lambda = 0$) corresponds to naïve BO against the unconstrained surrogate model, while $\alpha = 1$ (i.e., $\lambda \to \infty$) is equivalent to a WGAN-like policy. Evaluating

constant values of α ranging from 0 to 1, we find that there is no consistently optimal constant value for all eight optimization tasks. In contrast, our method achieves an average rank of **1.9** (**2.4**) on the top-1 (top-128) evaluation metric, and is one of the top two methods when compared to the ablations for at least five of the eight tasks. These results suggest that the 'adaptive' nature of aSCR is an important component in solving the constrained optimization problem in (5.4).

Table 5.4: **GABO Adaptive SCR ablation study.** One-shot ($k = 1$) and few-shot ($k = 128$) oracle evaluations averaged across 10 random seeds reported as mean $\pm$ standard deviation. $\mathcal{D}$ (best) reports the top oracle value in the task dataset.

Top-1	**Branin**	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**	**Rank**
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96	—
$\alpha = 0.0$	-11.0 $\pm$ 7.8	-52.5 $\pm$ 88.8	0.586 $\pm$ 0.193	1.43 $\pm$ 0.00	5.65 $\pm$ 1.30	0.59 $\pm$ 0.10	0.61 $\pm$ 0.15	0.16 $\pm$ 1.67	4.5
$\alpha = 0.2$	-9.8 $\pm$ 3.9	-4.39 $\pm$ 60.7	0.535 $\pm$ 0.110	1.43 $\pm$ 0.00	4.69 $\pm$ 1.44	0.63 $\pm$ 0.03	0.61 $\pm$ 0.15	0.16 $\pm$ 1.79	3.9
$\alpha = 0.5$	-7.9 $\pm$ 6.6	-83.9 $\pm$ 166.3	0.601 $\pm$ 0.212	1.43 $\pm$ 0.00	5.69 $\pm$ 1.51	0.63 $\pm$ 0.04	0.66 $\pm$ 0.12	0.16 $\pm$ 1.79	3.6
$\alpha = 0.8$	-5.2 $\pm$ 3.1	-43.3 $\pm$ 170.0	**0.654 $\pm$ 0.218**	1.66 $\pm$ 0.69	6.49 $\pm$ 1.20	**0.64 $\pm$ 0.02**	**0.71 $\pm$ 0.01**	0.16 $\pm$ 1.80	2.4
$\alpha = 1.0$	-99.5 $\pm$ 61.2	-46.8 $\pm$ 114.3	0.454 $\pm$ 0.120	**3.74 $\pm$ 0.01**	5.26 $\pm$ 2.35	0.52 $\pm$ 0.16	0.62 $\pm$ 0.15	-9.04 $\pm$ 57.3	4.8
aSCR	**-2.6 $\pm$ 1.1**	**21.3 $\pm$ 33.2**	0.570 $\pm$ 0.131	3.60 $\pm$ 0.40	**7.51 $\pm$ 0.39**	0.60 $\pm$ 0.07	**0.71 $\pm$ 0.01**	**0.60 $\pm$ 1.80**	**1.9**

Top-128	**Branin**	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**	**Rank**
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96	—
$\alpha = 0.0$	**-0.4 $\pm$ 0.0**	135.3 $\pm$ 16.0	0.942 $\pm$ 0.025	2.26 $\pm$ 1.03	8.26 $\pm$ 0.09	0.67 $\pm$ 0.00	0.72 $\pm$ 0.00	0.93 $\pm$ 0.11	4.3
$\alpha = 0.2$	**-0.4 $\pm$ 0.1**	121.8 $\pm$ 20.6	0.925 $\pm$ 0.029	3.01 $\pm$ 1.04	8.20 $\pm$ 0.10	0.67 $\pm$ 0.01	0.72 $\pm$ 0.00	**1.00 $\pm$ 0.00**	4.8
$\alpha = 0.5$	**-0.4 $\pm$ 0.0**	127.7 $\pm$ 23.1	0.944 $\pm$ 0.040	3.49 $\pm$ 0.69	8.29 $\pm$ 0.08	0.67 $\pm$ 0.01	0.72 $\pm$ 0.00	**1.00 $\pm$ 0.00**	2.9
$\alpha = 0.8$	**-0.4 $\pm$ 0.0**	104.5 $\pm$ 31.8	0.933 $\pm$ 0.036	**3.74 $\pm$ 0.00**	8.38 $\pm$ 0.11	0.67 $\pm$ 0.02	0.72 $\pm$ 0.00	**1.00 $\pm$ 0.00**	3.4
$\alpha = 1.0$	-2.2 $\pm$ 1.4	**142.3 $\pm$ 2.41**	0.906 $\pm$ 0.061	**3.74 $\pm$ 0.00**	**8.54 $\pm$ 0.08**	0.68 $\pm$ 0.01	0.72 $\pm$ 0.00	0.99 $\pm$ 0.04	3.4
aSCR	-0.5 $\pm$ 0.1	122.1 $\pm$ 20.6	**0.954 $\pm$ 0.025**	**3.74 $\pm$ 0.00**	8.36 $\pm$ 0.08	**0.70 $\pm$ 0.01**	0.72 $\pm$ 0.00	**1.00 $\pm$ 0.03**	**2.4**

Of note, the top designs found across different constant values of α can be very similar for certain tasks. This reflects the inherent challenge in developing task-agnostic methods for policy regularization—if the magnitudes of the unconstrained objective and regularization function vastly differ, then constant values of α may over- or under- constrain the objective. Adaptive SCR overcomes this problem by dynamically setting α as an implicit function of prior observations.

Oracle Query Budget Ablation. We ablate the number of allowed k-shot oracle calls in the Penalized **LogP** task (**Fig. 5.2**). While the majority of first-order optimization methods we evaluated are able to reach local optima rapidly, the proposed designs from such approaches are suboptimal compared to those from GABO (and GAGA) with Adaptive SCR as the oracle query budget size increases. Separately, comparing the curves for GABO and vanilla BO-qEI, we see that GABO with Adaptive SCR is able to propose consistently superior design candidates in the small query budget

regime often encountered in real-world settings. This is due to the fact that GABO regularizes the surrogate function estimates such that the proposed candidates are both high-scoring according to the surrogate objective *and* relatively in-distribution. Our results demonstrate that especially for real-world tasks like molecule design with complex objective function landscapes, methods such as GABO with Adaptive SCR are able to explore diverse, high-performing design candidates effectively even in the setting of small oracle query budgets.

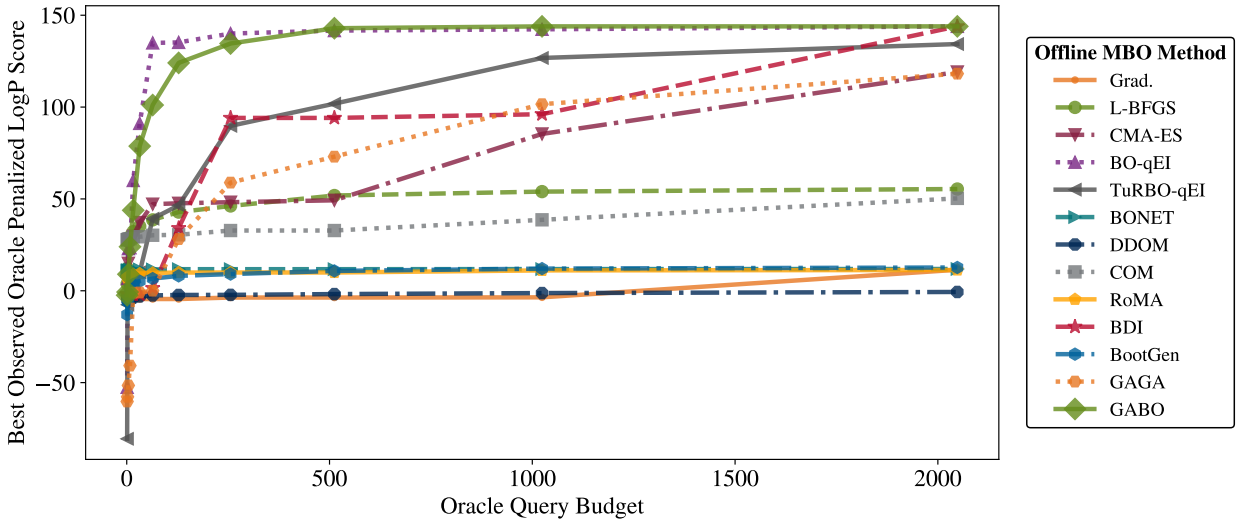

Figure 5.2: **100th percentile oracle scores versus k-shot oracle budget size.** We plot the 100th percentile oracle penalized LogP score averaged across 10 random seeds as a function of the number of allowed oracle calls k.

GP Initialization Ablation. Per **Algorithm 2**, GABO is based on the BO-qEI baseline optimization policy, which involves initializing the gaussian process (GP) to approximate the offline surrogate model. Consistent with prior work (Eriksson et al., 2019; Maus et al., 2022), we initialize the GP using the pseudo-random Sobol sequence (Sobol, 1967) at the beginning of the optimization procedure. However, an alternative approach is to instead initialize the GP using the top n_{init} samples from the offline dataset. In particular, this strategy is already employed in both related work describing the baseline first-order optimization methods assessed herein, with the idea that better designs can be generated by initializing from better designs. We compare these two GP initialization strategies in **Table 5.5**.

Interestingly, our results show that initializing the GABO GP from the Sobol sequence consistently

outperforms initialization from the top candidates in offline dataset. We hypothesize that this may be due to the fact that top-scoring candidates likely lie in similar regions of the input space, which significantly alters the ability of the optimizer to explore other regions of the design space over the course of the optimization process. Future work may help better interrogate the relationship between GP initialization and offline optimization, which is outside the scope of this work.

Table 5.5: **GABO GP initialization ablation study.** We investigate the effect of initializing the Gaussian process (GP) in GABO using the best n_{init} points from the offline dataset (i.e., **Best** initialization strategy) versus our method in **Algorithm 2** where the GP is initialized using the first n_{init} points from the Sobol sequence from (Sobol, 1967) (i.e., **Sobol** initialization strategy). In each evaluation setting, we rank all 2,048 proposed designs according to the penalized surrogate forward model in (5.5) and evaluate the top k designs using the oracle function, reporting the maximum out of the k oracle values. In the suboptimal evaluation setting, we report the oracle score of the single 90th percentile design according to the penalized surrogate ranking. **Bold** entries indicate the best entry in the column for the particular optimizer and evaluation metric. *Denotes the life sciences MBO tasks offered by Design-Bench (Trabucco et al., 2022).

Strategy	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96
Constrained Budget ($k = 1$) Oracle Evaluation								
Best	-3.6 ± 4.1	14.0 ± 18.4	0.504 ± 0.117	2.97 ± 1.02	5.36 ± 1.24	**0.61 ± 0.00**	0.50 ± 0.19	-2.97 ± 9.03
Sobol	**-2.6 ± 1.1**	**21.3 ± 33.2**	**0.570 ± 0.131**	**3.60 ± 0.40**	**7.51 ± 0.39**	0.60 ± 0.07	**0.71 ± 0.01**	**0.60 ± 1.80**
Relaxed Budget ($k = 128$) Oracle Evaluation								
Best	-0.5 ± 0.0	118.9 ± 19.5	0.918 ± 0.034	3.74 ± 0.00	**8.37 ± 0.09**	0.66 ± 0.01	**0.87 ± 0.05**	0.99 ± 0.09
Sobol	-0.5 ± 0.1	**122.1 ± 20.6**	**0.954 ± 0.025**	3.74 ± 0.00	8.36 ± 0.08	**0.70 ± 0.01**	0.72 ± 0.00	**1.00 ± 0.03**
Constrained Budget ($k = 1$) Suboptimal (90%-ile) Oracle Evaluation								
Best	**-11.8 ± 6.4**	-85.9 ± 124	0.382 ± 0.106	3.45 ± 0.77	**6.28 ± 1.70**	0.60 ± 0.03	**0.64 ± 0.23**	-0.65 ± 3.97
Sobol	-12.7 ± 10.0	**-12.2 ± 46.1**	**0.467 ± 0.066**	**3.56 ± 1.66**	6.12 ± 1.22	**0.61 ± 0.08**	0.57 ± 0.17	**0.02 ± 5.77**

Surrogate Model Ablation. In **Algorithm 2**, we leverage a surrogate forward model f_θ in model-based optimization and a separate GP to acquire samples in the Bayesian optimization framework. However, it may be possible to use the GP directly as the surrogate forward model. Our results in **Table 5.6** suggest that this is *not* an effective strategy with which to use GABO—using even a simple neural-network as the surrogate function (as done in our approach in **Algorithm 2**) outperforms the alternative GP-based approach in six of the eight tasks in the top-1 evaluation setting, and is non-inferior to the alternative GP-based approach in all eight tasks in the top-128 evaluation setting. These results suggest that using a more complex neural-network surrogate function for GABO leads to better optimization results than directly using the GP as the surrogate.

Table 5.6: **GABO neural network surrogate ablation study.** Instead of using a neural network (NN) as our surrogate forward model, we explore if the Gaussian process (GP) employed by the parent BO optimizer can directly be used as the surrogate model in GABO's framework. In each evaluation setting, we rank all 2,048 proposed designs according to the penalized surrogate forward model in (5.5) and evaluate the top k designs using the oracle function, reporting the maximum out of the k oracle values. In the suboptimal evaluation setting, we report the oracle score of the single 90th percentile design according to the penalized surrogate ranking. **Bold** entries indicate the best entry in the column for the particular optimizer and evaluation metric. *Denotes the life sciences MBO tasks offered by Design-Bench (Trabucco et al., 2022).

Surrogate	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96
Constrained Budget ($k = 1$) Oracle Evaluation								
GP	-37.4 ± 4.4	-57.9 ± 159.2	**0.576 ± 0.058**	3.51 ± 0.69	6.84 ± 1.24	**0.65 ± 0.01**	0.42 ± 0.17	-0.28 ± 2.13
NN	**-2.6 ± 1.1**	**21.3 ± 33.2**	0.570 ± 0.131	**3.60 ± 0.40**	**7.51 ± 0.39**	0.60 ± 0.07	**0.71 ± 0.01**	**0.60 ± 1.80**
Relaxed Budget ($k = 128$) Oracle Evaluation								
GP	-1.5 ± 0.5	119.9 ± 20.1	0.755 ± 0.071	3.74 ± 0.00	8.34 ± 0.07	0.67 ± 0.01	0.72 ± 0.00	-0.27 ± 2.13
NN	**-0.5 ± 0.1**	**122.1 ± 20.6**	**0.954 ± 0.025**	3.74 ± 0.00	**8.36 ± 0.08**	**0.70 ± 0.01**	0.72 ± 0.00	**1.00 ± 0.03**
Constrained Budget ($k = 1$) Suboptimal (90%-ile) Oracle Evaluation								
GP	**-10.1 ± 10.6**	-51.5 ± 108.8	**0.562 ± 0.091**	2.62 ± 1.13	**6.54 ± 1.56**	**0.65 ± 0.00**	0.50 ± 0.19	-0.27 ± 2.13
NN	-12.7 ± 10.0	**-12.2 ± 46.1**	0.467 ± 0.066	**3.56 ± 1.66**	6.12 ± 1.22	0.61 ± 0.08	**0.57 ± 0.17**	**0.02 ± 5.77**

5.7. Conclusion

We propose **adaptive source critic regularization (aSCR)** to solve the problem of off-distribution objective evaluation in offline MBO. When leveraged with vanilla Bayesian optimization, aSCR outperforms baseline methods to achieve an average rank of **3.8** (**3.0**) in one-shot $k = 1$ (few-shot $k = 128$) oracle evaluation, and most consistently proposes designs better than the offline dataset.

One important limitation of aSCR is that our algorithm requires preexisting knowledge of the prior distribution over the input space. While the tasks considered in this chapter are amenable to the imposed latent space priors, further work is needed to adapt aSCR to arbitrary configuration spaces. Future work may also extend aSCR to improve parent optimization methods more sophisticated than BO-qEI and Gradient Ascent. Finally, recent domain-specific foundation models (Lin et al., 2023; Ohana et al., 2025; Nguyen et al., 2024; Zeni et al., 2025) may also give rise to more sophisticated, accurate surrogate models that can be leveraged with GAMBO in future work.

CHAPTER 6

OBTAINING DIVERSE AND HIGH-QUALITY DESIGNS IN OFFLINE OPTIMIZATION

Portions of this chapter are adapted from the following published first-author manuscript:

(Yao et al., 2025b) Michael S Yao, James C Gee, and Osbert Bastani. Diversity by design: Leveraging distribution matching for offline model-based optimization. Proc ICML, 2025. doi: 10.48550/arXiv.2501.18768

I helped conceive the study, planned and performed experiments, analyzed the experimental data, and drafted the manuscript with input from all other authors.

6.0.1. Introduction

In the previous chapter, we considered the problem of *offline optimization* in generative design where our goal was to propose new materials, chemicals, proteins, and other scientific designs that optimize a desirable property of interest given only access to a static, offline dataset. Our experiments demonstrated that adversarial feedback via source critic networks could be a powerful tool to improve the quality of designs proposed by offline model-based optimization (MBO) frameworks; however, these same experiments also simultaneously revealed an important limitation of existing work in offline MBO: **how can we achieve designs that are high-quality *and* diverse?**

Prior work in offline MBO (Yu et al., 2021; Trabucco et al., 2021; Fu and Levine, 2021; Chen et al., 2022; Krishnamoorthy et al., 2023b,a; Nguyen et al., 2023; Kim et al., 2023) has almost exclusively focused on developing algorithms that propose high-quality candidate designs. A secondary metric often overlooked in these settings is ***candidate diversity*** (Jain et al., 2022; Kim et al., 2023): it is often ideal to include a diverse array of designs in the final samples proposed by an optimization procedure (**Fig. 1.4**). Different designs may achieve promising oracle rewards in different ways, and many real-world optimization tasks seek to capture as many of these 'modes of goodness' as possible (Mullis et al., 2019; Jain et al., 2022). Furthermore, there may be secondary optimization objective(s) (e.g., manufacturing cost or drug toxicity) that are better explored and evaluated in a

diverse sample set. In these settings, it may be more desirable to sample slightly suboptimal designs in addition to the most optimal design to achieve a greater diversity of proposed candidates.

To this end, we introduce **Diversity in Adversarial Model-based Optimization** (**DynAMO**) as a novel approach to explicitly control the trade-off between the reward-optimality and diversity of a proposed batch of designs in offline MBO. To motivate our contributions, we show how naïve optimization algorithms provably suffer from poor candidate diversity. To overcome this limitation, we will first propose a modified optimization objective in the offline setting that encourages discovery of designs that encapsulate the diversity of samples in the offline dataset—an approach inspired by recent advancements in imitation learning and offline reinforcement learning (Ho and Ermon, 2016; Kostrikov et al., 2020; Ke et al., 2021; Ma et al., 2022; Rafailov et al., 2023; Deka et al., 2023; Huang et al., 2024b). We will then derive DynAMO as a provably optimal solution to our modified optimization objective. Finally, we will empirically demonstrate how DynAMO can be used with a wide variety of different offline optimization methods to propose promising design candidates comparable to the state-of-the-art, while also achieving significantly better candidate diversity.

6.1. Related Work

Model-free offline optimization. In this chapter, we specifically look at *model-based optimization* methods that explicitly optimize against a forward surrogate model $r_\theta(x)$ that acts as a proxy for the hidden oracle function $r(x)$. However, related work have also proposed offline optimization methods that do not require access to a model $r_\theta(x)$ and instead impose constraints on the backbone optimization method—we refer to such work as *model-free* offline optimization. Krishnamoorthy et al. (2023b) frame generative design tasks as a 'next-sample' prediction problem and learn a transformer to roll out sample predictions; and Krishnamoorthy et al. (2023a); Yun et al. (2024) learn a diffusion model to sample candidate designs conditioned on reward values. Because DynAMO operates on MBO forward surrogate models $r_\theta(x)$, we cannot leverage DynAMO with these model-free methods. However, we compare them against DynAMO in **Appendix D.4**.

Active learning in optimization. In our work, we specifically consider the experimental setup of *one-shot, batched oracle evaluation*: that is, the final candidate designs that are scored by the oracle

function at the end of optimization are *not* used to subsequently update the prior over the design space to better inform subsequent optimization steps. In contrast, a separate body of recent work has investigated generative design in the setting of *active learning* where there can be multiple rounds of offline optimization to inform subsequent online acquisitions (Hernández-García et al., 2024; Li et al., 2022b,a; Wu et al., 2023; Palizhati et al., 2022). For example, Li et al. (2024) show how active learning can be formulated as a multi-fidelity optimization problem.

Reinforcement learning. Prior work has explored how to formulate offline generative design tasks as reinforcement learning (RL) problems. Trabucco et al. (2022) used REINFORCE-style methods similar to Williams (1992) to learn a myopic sampling policy, although do not use RL for offline generative design. Angermueller et al. (2020b); Korshunova et al. (2022); and Jang et al. (2022) leverage RL for offline optimization in the active learning setting outside the scope of our work.

6.2. Background and Preliminaries

6.2.1. Offline Model-Based Optimization

In the previous chapter, we formulated our model-based optimization (MBO) problem as a sampling task, where the goal was to solve $x^* = \arg\max_{x \in \mathcal{X}} r_\theta(x)$ as in (6.1). We primarily considered this problem formulation because in experimental settings where offline optimization methods are desirable, we are typically most concerned with the *single best* design we can obtain to evaluate with an expensive-to-evaluate oracle function. However, other experimental settings may offer a *relaxed* oracle evaluation budget, where it is feasible to evaluate *batches* of candidated designs as opposed to a single proposed candidate. In this setting, we can instead reformulate offline MBO as learning a generative policy π^* over a space of policies Π such that the admitted distribution $q^{\pi^*}(x) : \mathcal{X} \to [0, 1]$ of designs maximizes

$$\pi^* = \arg\max_{\pi \in \Pi} \; \mathbb{E}_{x \sim q^\pi(x)}[r_\theta(x)] \tag{6.1}$$

over a design space $\mathcal{X}$, where the hope again is that optimizing against $r_\theta(x)$ will learn a generative policy that also proposes optimal designs according to the hidden oracle function $r(x)$, too.

6.2.2. Optimization Algorithms

To solve (6.1) and similar problem formulations, a number of optimization algorithms have been reported in prior work. One of the most popular approaches is *first-order methods* such as gradient ascent, adaptive moment estimation (Adam) (Kingma and Ba, 2014), and derivative work (Duchi et al., 2011; Loshchilov and Hutter, 2019; Liu and Nocedal, 1989). Broadly, these optimizers leverage the gradient $\vec{\nabla}_x r_\theta$ of the forward surrogate to iteratively update a candidate design. However, such techniques have been shown to struggle in optimizing against highly non-convex functions typical of real-world offline optimization problems (Trabucco et al., 2021, 2022).

Evolutionary algorithms, such as covariance matrix adaptation evolution strategy (CMA-ES) from Hansen (2016, 2006) and cooperative synapse neuroevolution (CoSyNE) (Gomez et al., 2008), are an alternative approach to optimization. Inspired by biological evolution, such methods iteratively improve a population of candidate solutions using mechanisms like selection and mutation, and do not require gradient information from the forward model.

Separately, *Bayesian optimization* (BO) (Kushner, 1964) is another model-based optimization technique historically used to optimize reward functions that are non-convex, noisy, and/or lack a closed-form expression. Briefly, BO iteratively alternates between (1) fitting a probabilistic surrogate model (e.g., a Gaussian process) to the acquired data and their scores according to r_θ; and (2) acquiring new candidate designs according to an acquisition function, such as the expected improvement (EI) or upper confidence bound (UCB) (Ament et al., 2023; Wilson et al., 2018; Zhou et al., 2024a). While BO has traditionally been leveraged for optimization problems using expensive-to-evaluate black-box functions, recent work has shown that BO is also a powerful method for offline optimization tasks, too Maus et al. (2022); Yao et al. (2024); Eriksson et al. (2019); Hvarfner et al. (2024); Eriksson and Jankowiak (2021); Astudillo and Frazier (2019). Prior work from Maus et al. (2023); Jain et al. (2022); and others have investigated how to incorporate diversity in existing BO frameworks; however, such methods either (1) gate whether to sample candidate designs based on a diversity-based thresholding schema; or (2) have specifically been proposed for the BO optimization framework. In contrast, our method explicitly includes diversity

as an optimization objective, and is readily compatible with standard optimization algorithms.

6.2.3. Distribution Matching

Distribution matching is a technique leveraged in recent work on imitation learning and offline reinforcement learning (RL) (Kostrikov et al., 2020; Ke et al., 2021). The approach considers an experimental setup where RL agents cannot interact with the environment and instead must learn from static, offline expert demonstrations sampled from an unknown state-action-reward distribution. The Kullback-Leibler (KL)- divergence (de G Matthews et al., 2016) is commonly used to train an agent to minimize the discrepancy between state-action visitations made by the RL agent and the offline expert. Given a sufficiently large and diverse dataset of expert demonstrations, we can also think of the KL divergence as encouraging the agent to match the diversity of the non-zero support of $p(x)$. Distribution matching has been used in prior work to learn robotic control policies (Ho and Ermon, 2016; Wang et al., 2020; Kostrikov et al., 2020; Ke et al., 2021; Ma et al., 2022) and align language models (Rafailov et al., 2023; Huang et al., 2024b; Chakraborty et al., 2024); here, we demonstrate how distribution matching can also be leveraged in offline generative design (a non-RL application) by matching the distribution of designs learned by a generative policy with the distribution of designs from the offline dataset.

6.2.4. f-Divergence and Fenchel Conjugates

Definition 2 (f-Divergence)**.** *Suppose we are given two probability distributions $P(x), Q(x)$ defined over a common support $\mathcal{X}$. For any continuous, convex function $f : \mathbb{R}_+ \to \mathbb{R}$ that is finite over $\mathbb{R}_{++}$, we define the f-divergence between $P(x), Q(x)$ as*

$$D_f(Q(x)||P(x)) := \mathbb{E}_{x\sim P(x)}\left[f\left(\frac{Q(x)}{P(x)}\right)\right] \tag{6.2}$$

We refer to f as the generator of $D_f(\cdot||\cdot)$. Two commonly used f-divergences are the Kullback-Leibler (KL)-Divergence (defined by the generator $f_{KL}(u) = u\log u$) and the χ^2-Divergence (defined by the generator $f_{\chi^2}(u) = (u-1)^2/2$).

Definition 3 (Fenchel Conjugate)**.** *The Fenchel conjugate (i.e., Legendre-Fenchel transform) of a function*

$f : \mathcal{U} \to \mathbb{R}$ is defined as

$$f^\star(v) := -\inf \{-\langle u, v\rangle + f(u) \mid u \in \mathcal{U}\} \tag{6.3}$$

*where $\langle u, v\rangle$ is the inner product, and $f^\star : \mathcal{V} \to \mathbb{R}$ is the Fenchel conjugate defined over the dual space $\mathcal{V}$ of $\mathcal{U}$. Importantly, the Fenchel conjugate function is guaranteed to always be convex Borwein and Lewis (2006) regardless of the (non-)convexity of the original function f. This allows us to make important convergence guarantees in solving the Lagrangian dual problem in **Algorithm 3**. Fenchel conjugates are commonly used in optimization problems to rewrite difficult primal problems into more tractable dual formulations Ma et al. (2022); Borwein and Lewis (2006); Agrawal and Horel (2021)—we leverage a similar technique in our work in **Algorithm 3**.*

Lemma 2 (Fenchel Conjugate of the KL-Divergence Generator Function)**.** *Recall that the generator function of the KL-divergence is $f_{KL}(u) := u \log u$ for $u \in \mathbb{R}_{++}$. The Fenchel conjugate of this generator is $f^\star_{KL}(v) = e^{v-1}$.*

Proof. The proof follows immediately from the definition of the Fenchel conjugate in (6.3).

$$f^\star_{\text{KL}}(v) := \sup \{uv - u \log u \mid u \in \mathbb{R}_{++}\} \tag{6.4}$$

We differentiate the argument on the right hand side with respect to u to find the supremum given a particular $v \in \mathcal{V}$:

$$\frac{\partial}{\partial u}\left[uv - u \log u\right]\Big|_{u=u^*} = v - \log u^* - 1 = 0 \to u^* = e^{v-1} \tag{6.5}$$

It is easy to verify that u^* is a maxima. Plugging this result into (6.4),

$$f^\star_{\text{KL}}(v) = u^* v - u^* \log u^* = v e^{v-1} - (v-1)e^{v-1} = e^{v-1} \tag{6.6}$$

□

Lemma 3 (Fenchel Conjugate of the χ^2-Divergence Generator Function)**.** *Recall that the generator*

function of the χ^2-divergence is $f_{\chi^2}(u) := \frac{1}{2}(u-1)^2$ for $u \in \mathbb{R}_{++}$. The Fenchel conjugate of this generator is $f^\star_{\chi^2}(v) = \frac{v^2}{2} + v$.

Proof. The proof follows immediately from the definition of the Fenchel conjugate in (6.3).

$$f^\star_{\chi^2}(v) := \sup \left\{ uv - \frac{1}{2}(u-1)^2 \mid u \in \mathbb{R}_{++} \right\} \tag{6.7}$$

We differentiate the argument on the right hand side with respect to u to find the supremum given a particular $v \in \mathcal{V}$:

$$\frac{\partial}{\partial u}\left[uv - \frac{1}{2}(u-1)^2\right]\Big|_{u=u^*} = v - u^* + 1 = 0 \to u^* = v + 1 \tag{6.8}$$

It is easy to verify that u^* is a maxima. Plugging this result into (6.7),

$$f^\star_{\chi^2}(v) = u^* v - \frac{1}{2}(u^*-1)^2 = (v+1)v - \frac{1}{2}((v+1)-1)^2 = v^2 + v - \frac{1}{2}v^2 = \frac{v^2}{2} + v \tag{6.9}$$

□

Additional details and technical discussion are offered by Borwein and Lewis (2006); Ma et al. (2022); Nachum and Dai (2020); Amos (2023); and Terjék and González-Sánchez (2022).

6.3. Distribution Matching for Generative Offline Optimization

6.3.1. Motivating Limitation of Naïve MBO

Prior work from Mullis et al. (2019); Jain et al. (2022); Kim et al. (2023) have shown that an important challenge in offline optimization as in (6.1) is that of **reward hacking**: learned generative policies can exploit a small region of the design space, resulting in a low diversity of proposed designs. For example, consider the following lemma:

Lemma 4. (*Diversity Collapse in Reward Optimization*) *Suppose that there exists a finite set of globally optimal designs x^*_j such that $x^*_j := \arg\max_{x \in \mathcal{X}} r(x)$ and $r^* := r(x^*_j)$ is the optimal reward given a finite, non-uniform reward function $r(x)$. Given any distribution q^π, we can decompose it into the form $q^\pi(x) =$*

$\sum_j w_j \delta(x - x_j^*) + \sum_j w_j' \mathbb{1}(x = x_j^*) + \tilde{q}(x)$, *where* $w_j \geq 0$ *for all* j, *and* $\tilde{q}(x) \geq 0$ *and* $\tilde{q}(x_j^*) = 0$ *for all* j. *Then,* $\tilde{q}(x)$ *satisfies* $\int dx\, \tilde{q}(x) = 0$.

Proof. First, note that if $\int dx\, \tilde{q}(x) = 0$, we have

$$\begin{aligned}
\mathbb{E}_{x \sim q^\pi(x)}[r(x)] &= \int dx \sum_j w_j \delta(x - x_j^*) r(x) \\
&= \sum_j w_j \int dx\, \delta(x - x_j^*) r(x) = \sum_j w_j r^* = r^* \sum_j w_j \\
&= r^*,
\end{aligned} \tag{6.10}$$

which is optimal. Next, we show that if $\int dx\, \tilde{q}(x) > 0$, then $\mathbb{E}_{x \sim q^\pi}[r(x)] < r^*$. To this end, we define

$$\mathcal{X}_1 := \{x \in \mathcal{X} \,|\, 1 < r^* - r(x)\}, \quad \mathcal{X}_n := \left\{ x \in \mathcal{X} \,\middle|\, \frac{1}{n} < r^* - r(x) \leq \frac{1}{n-1} \right\} \subseteq \mathcal{X} \quad \forall n \geq 2 \tag{6.11}$$

for each $n \in \mathbb{N}$. Note that all $\mathcal{X}_n$ are disjoint by construction; also by construction, we have $\mathcal{X} \setminus \{x_j^*\} = \bigcup_{n=1}^\infty \mathcal{X}_n$. Furthermore, note that since $\tilde{q}(x) = 0$ for $x = x_j^*$ for some j, we have $0 < \int dx\, \tilde{q}(x) = \sum_{n=1}^\infty \int_{\mathcal{X}_n} dx\, \tilde{q}(x)$, so it must be that $\int_{\mathcal{X}_m} dx\, \tilde{q}(x) > 0$ for some m. Consequently,

$$\int_{\mathcal{X}_m} dx\, \tilde{q}(x)(r^* - r(x)) \geq \frac{1}{m} \int_{\mathcal{X}_m} dx\, \tilde{q}(x) > 0.$$

The expected reward would therefore be

$$\begin{aligned}
\mathbb{E}_{x \sim q^\pi}[r(x)] &= \int dx\, q^\pi(x) r(x) \\
&= \int dx \left(\sum_j w_j \delta(x - x_j^*) + \sum_j w_j' \mathbb{1}(x = x_j^*) \right) r(x) + \sum_{n=1}^\infty \int_{\mathcal{X}_n} dx\, \tilde{q}(x) r(x) \\
&< \int dx \left(\sum_j w_j \delta(x - x_j^*) + \sum_j w_j' \mathbb{1}(x = x_j^*) \right) r^* + \sum_{n=1}^\infty \int_{\mathcal{X}_n} dx\, \tilde{q}(x) r^* \\
&= \int dx\, q^\pi(x) r^* \\
&= r^*
\end{aligned} \tag{6.12}$$

so q^π is suboptimal. The claim follows. □

Note that this result holds for both the oracle function $r(x)$ and the forward surrogate $r_\theta(x)$. Intuitively, this lemma states that an optimal policy that maximizes (6.1) can only have non-zero support at the global optimizers over $\mathcal{X}$. However, many real-world reward functions do not have a large number of globally optimal designs (Trabucco et al., 2021), leading to a low diversity of generated designs seen in practice (Kim et al., 2023). Furthermore, there is no guarantee that the set of optimal x_j^* cover a large region of the design space; in practice, we might be interested in trading optimality of a subset of designs to achieve a greater diversity of candidate samples.

6.3.2. An Alternative MBO Problem Formulation

To reward generative policies in proposing diverse designs, we modify the original MBO objective in (6.1) according to

$$J(\pi) := \mathbb{E}_{x\sim q^\pi(x)}[r_\theta(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \tag{6.13}$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback–Leibler divergence (KL-divergence) and $\tau, \beta \in \mathbb{R}_+$ are hyperparameters. In subsequent steps, we abbreviate the expectation value over probability distributions $\mathbb{E}_{x\sim q^\pi(x)}[\cdot]$ as $\mathbb{E}_{q^\pi}[\cdot]$ for brevity.

The temperature hyperparameter τ. Equation (6.13) implicitly introduces a hyperparameter $\tau \in \mathbb{R}_+$ to control the trade-off between diversity and optimality. Note that the KL-divergence in (6.13) is computed with respect to a distribution $p_{\mathcal{D}}^\tau(x)$ defined as the *τ-weighted probability distribution*:

Definition 4 (τ-Weighted Probability Distribution)**.** *Suppose that we are given a reward function $r(x) : \mathcal{X} \to \mathbb{R}$ over a space of possible designs $\mathcal{X}$, and access to an static, offline dataset $\mathcal{D}$ of real designs. We define the τ-weighted probability distribution over $\mathcal{X}$ (for $\tau \geq 0$) as*

$$p^\tau(x) := \frac{\exp(\tau r(x))}{Z^\tau} \tag{6.14}$$

where the partition function $Z^\tau := \int_{\mathcal{X}} dx\ \exp(\tau r(x))$ is a normalizing constant. We use the dataset of prior observations $\mathcal{D} = \{(x_i, r(x_i))\}_{i=1}^n$ to empirically approximate $p^\tau(x)$, and refer to this approximation

as $p^{\tau}_{\mathcal{D}}(x) \approx p^{\tau}(x)$. For $\tau \gg 1$, near-optimal designs that are associated with high reward scores are weighted more heavily in $p^{\tau}_{\mathcal{D}}$; conversely, $\tau = 0$ weights all designs equally to achieve the greatest diversity in designs. The penalized objective in (6.13) thereby encourages the learned policy to capture the diversity of designs in the τ-weighted distribution $p^{\tau}_{\mathcal{D}}(x)$.

The KL-divergence strength hyperparameter β. Separately, the hyperparameter $\beta \geq 0$ controls the relative importance of the distribution matching objective. As $\beta \to \infty$, it becomes increasingly important for the generator to learn a distribution of designs that match $p^{\tau}_{\mathcal{D}}(x)$; setting $\beta = 0$ reduces $J(\pi)$ to the original MBO objective in (6.1).

6.3.3. Adversarial Source Critic as a Constraint

Separately, to address the problem of forward surrogate model overestimation of candidate design fitness according to $r_\theta(x)$, we constrain the optimization problem to ensure that expected source critic scores over $q^{\pi}(x)$ and $p^{\tau}_{\mathcal{D}}(x)$ differ by no more than a constant $W_0 \in \mathbb{R}_+$, similar to the approach to offline MBO used by Yao et al. (2024) introduced in the preceding chapter. That is,

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J(\pi) = \mathbb{E}_{q^{\pi}}[r_\theta(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^{\pi} || p^{\tau}_{\mathcal{D}}) \\ \text{s.t.} \quad & \mathbb{E}_{p^{\tau}_{\mathcal{D}}}[c^*(x)] - \mathbb{E}_{q^{\pi}}[c^*(x)] \leq W_0 \end{aligned} \tag{6.15}$$

where the source critic $c^* : \mathcal{X} \to \mathbb{R}$ is a neural network as in (**Lemma 1**) that maximizes $\mathbb{E}_{p^{\tau}_{\mathcal{D}}}[c^*(x)] - \mathbb{E}_{q^{\pi}}[c^*(x)]$ subject to the constraint $||c^*(x)||_L \leq 1$, where $|| \cdot ||_L$ is the Lipschitz norm. Intuitively, this constraint prevents the evaluation of the forward surrogate model $r_\theta(x)$ on wildly out-of-distribution inputs encountered in the offline setting.

We are now interested in finding a generative policy π^* that solves this optimization problem in (6.15); in our work below, we demonstrate how this approach can yield a policy that generates high-scoring candidate designs that also better capture the diversity of possible designs in $\mathcal{X}$.

6.3.4. Constrained Optimization via Lagrangian Duality

Our problem in (6.15) is ostensibly challenging to solve: both the objective $J(\pi)$ and the constraint imposed by the source critic can be arbitrarily non-convex, making traditional constrained opti-

mization techniques intractable in solving the optimization problem out-of-the-box. In this section, we derive an explicit solution to (6.15) to make the problem tractably solvable using any standard optimization algorithm.

Recall from Lagrangian duality that solving (6.15) is equivalent to the min-max problem

$$\min_{\pi \in \Pi} \max_{\lambda \in \mathbb{R}_+} \mathcal{L}(\pi; \lambda) \tag{6.16}$$

where the Lagrangian $\mathcal{L}(\pi; \lambda) : \Pi \times \mathbb{R}_+ \to \mathbb{R}$ is given by

$$\begin{aligned} \mathcal{L}(\pi; \lambda) = &- J(\pi) \\ &+ \beta\lambda \left[\mathbb{E}_{p^\tau_{\mathcal{D}}}[c^*(x)] - \mathbb{E}_{q^\pi}[c^*(x)] - W_0 \right] \end{aligned} \tag{6.17}$$

introducing $\lambda \in \mathbb{R}_+$ such that $\beta\lambda \in \mathbb{R}_+$ is the Lagrange multiplier associated with the constraint in (6.15). From weak duality, the *Lagrange dual problem* provides us with a tight lower bound on the primal problem in (6.15):

$$\max_{\lambda \in \mathbb{R}_+} \min_{\pi \in \Pi} \mathcal{L}(\pi; \lambda) := \max_{\lambda \in \mathbb{R}_+} g(\lambda) \leq \min_{\pi \in \Pi} \max_{\lambda \in \mathbb{R}_+} \mathcal{L}(\pi; \lambda) \tag{6.18}$$

where $g(\lambda) := \min_{\pi \in \Pi} \mathcal{L}(\pi; \lambda)$ is the *Lagrange dual function*. In general, computing $g(\lambda)$ is challenging for an arbitrary offline optimization problem; in prior work, Trabucco et al. (2021) bypassed this dual problem entirely by treating λ as a hyperparameter tuned by hand (albeit for a different constraint); and Yao et al. (2024) approximated the dual function under certain assumptions about the input space by performing a grid search over possible λ values. In our approach, we look to rewrite the problem into an equivalent representation that admits a closed-form, computationally tractable expression for $g(\lambda)$:

Lemma 5 (Entropy-Divergence Formulation)**.** *Define $J(\pi)$ as in (6.13). An equivalent representation of $J(\pi)$ is*

$$J(\pi) \simeq -\mathcal{H}(q^\pi(x)) - (1 + \beta) D_{KL}(q^\pi(x) || p^\tau_{\mathcal{D}}(x)) \tag{6.19}$$

where $\mathcal{H}(\cdot)$ is the Shannon entropy. Maximizing (6.19) is equivalent to maximizing (6.13) in the sense that both objectives admit the same optimal policy.

Proof. Firstly, note that

$$\begin{aligned} J(\pi) &= \mathbb{E}_{q^\pi}\left[r_\theta(x)\right] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \\ &\simeq \tau \cdot \mathbb{E}_{q^\pi}\left[\log e^{r_\theta(x)}\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \\ &= \mathbb{E}_{q^\pi}\left[\log e^{\tau r_\theta(x)}\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \end{aligned} \tag{6.20}$$

where $\simeq$ denotes an equivalent representation of the objective (i.e., scaling $J(\pi)$ by $\tau > 0$ does not change the optimal policy π^*). Further rewriting,

$$J(\pi) = \mathbb{E}_{q^\pi}\left[\log \frac{e^{\tau r_\theta(x)}}{Z^\tau}\right] + \mathbb{E}_{q^\pi} \log Z^\tau - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \simeq \mathbb{E}_{q^\pi}\left[\log \frac{e^{\tau r_\theta(x)}}{Z^\tau}\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \tag{6.21}$$

where we omit the constant $\mathbb{E}_{q^\pi} \log \mathcal{Z}_\theta^\tau$ because the expectation value argument is independent of the policy π. The remaining expectation value can be re-expressed via *importance weighting*:

$$J(\pi) = \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log \frac{e^{\tau r_\theta(x)}}{Z^\tau}\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \tag{6.22}$$

Assuming that the surrogate $r_\theta(x)$ is well-trained on the offline dataset $\mathcal{D}$ (i.e., $r(x) \approx r_\theta(x)\ \forall x \in \mathcal{D}$), we have

$$J(\pi) \approx \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log \frac{e^{\tau r(x)}}{Z^\tau}\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) = \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log p_{\mathcal{D}}^\tau\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \tag{6.23}$$

from **Definition 4**. Further rewriting, we have

$$\begin{aligned} J(\pi) &= \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log \left(p_{\mathcal{D}}^\tau \cdot \frac{q^\pi}{q^\pi}\right)\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \\ &= \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log \frac{p_{\mathcal{D}}^\tau}{q^\pi}\right] + \mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log q^\pi\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \\ &= -\mathbb{E}_{p_{\mathcal{D}}^\tau}\left[\frac{q^\pi}{p_{\mathcal{D}}^\tau} \log \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right] - \mathbb{E}_{q^\pi}\left[-\log q^\pi\right] - \beta D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) \end{aligned} \tag{6.24}$$

From the definition of KL-divergence,

$$\begin{aligned} J(\pi) &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log\frac{q^\pi}{p^\tau_{\mathcal{D}}}\right] - \mathbb{E}_{q^\pi}\left[-\log q^\pi\right] \\ &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[f_{\mathrm{KL}}\left(\frac{q^\pi}{p^\tau_{\mathcal{D}}}\right)\right] - \mathbb{E}_{q^\pi}\left[f_\ell\left(q^\pi\right)\right] \\ &= -(1+\beta)D_{\mathrm{KL}}(q^\pi||p^\tau_{\mathcal{D}}) - \mathcal{H}(q^\pi) \end{aligned} \tag{6.25}$$

up to a constant, where $f_{\mathrm{KL}}(x) := x\log x$ and $f_\ell(x) := -\log x$ are convex functions, $\mathcal{H}(\cdot)$ is the Shannon entropy, and $D_{\mathrm{KL}}(\cdot||\cdot)$ is the KL divergence. □

Remark 2 (Equivalence of **Lemma 5** and Canonical State-Matching)**.** *Of theoretical interest, we can also show that the entropy-divergence formulation of our penalized objective function in* ***Lemma 5*** *is equivalent to the canonical state-matching objective used in traditional imitation learning. Continuing from* (6.25)*, one might notice that* $J(\pi)$ *can be equivalently rewritten as*

$$\begin{aligned} J(\pi) &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log\frac{q^\pi}{p^\tau_{\mathcal{D}}}\right] - \mathbb{E}_{q^\pi}\left[-\log q^\pi\right] \\ &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log\frac{q^\pi}{p^\tau_{\mathcal{D}}}\right] - \mathbb{E}_{p^\tau_{\mathcal{D}}}\left[-\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log q^\pi\right] \\ &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log\frac{q^\pi}{p^\tau_{\mathcal{D}}}\right] - (1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log(q^\pi)^{-1/(1+\beta)}\right] \\ &= -(1+\beta)\mathbb{E}_{p^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{p^\tau_{\mathcal{D}}}\log\left(\frac{q^\pi}{p^\tau_{\mathcal{D}}}\cdot\frac{1}{(q^\pi)^{1/(1+\beta)}}\right)\right] \\ &= -(1+\beta)\mathbb{E}_{q^\pi}\left[\log\frac{(q^\pi)^{\beta/(1+\beta)}}{p^\tau_{\mathcal{D}}}\right] \end{aligned} \tag{6.26}$$

Assume that there exists a probability distribution $\hat{p}^\tau_{\mathcal{D}}(x)$ *such that* $\hat{p}^\tau_{\mathcal{D}}(x) \propto (p^\tau_{\mathcal{D}}(x))^{(1+\beta)/\beta}$*. Then*

$$J(\pi) \simeq -(1+\beta)\mathbb{E}_{q^\pi}\left[\log\left(\frac{q^\pi}{\hat{p}^\tau_{\mathcal{D}}(x)}\right)^{\beta/(1+\beta)}\right] = -\beta\mathbb{E}_{\hat{p}^\tau_{\mathcal{D}}}\left[\frac{q^\pi}{\hat{p}^\tau_{\mathcal{D}}}\log\frac{q^\pi}{\hat{p}^\tau_{\mathcal{D}}(x)}\right] = -\beta D_{\mathrm{KL}}(q^\pi||\hat{p}^\tau_{\mathcal{D}}) \tag{6.27}$$

In other words, the optimization objective considered in (6.13) *and in* ***Lemma 5*** *is equivalent to a pure state-matching objective* $-\beta D_{\mathrm{KL}}(q^\pi||\hat{p}^\tau_{\mathcal{D}})$ *predicated on the existence of a 'rescaled' probability distribution* $\hat{p}^\tau_{\mathcal{D}}(x)$ *as defined above.*

To build intuition about how (6.19) is equivalent to (6.13), we can consider the behavior of the objective in the limit of $\tau \to +\infty$: the reference distribution $p_{\mathcal{D}}^{\tau}$ approaches the sum-of-δ-distributions formulation in **Lemma 4**. In this setting, the entropy and KL-divergence terms are equivalent, and the optimal policy π^* admits a distribution q^{π} with nonzero support only at the globally optimal designs in $p_{\mathcal{D}}^{\tau}$. Alternatively in the limit that $\tau \to 0$ and $\beta \to +\infty$, $p_{\mathcal{D}}^{\tau}$ approaches a uniform distribution and both (6.13) and (6.19) simplify to a state-matching objective according to the KL-divergence loss term, without any explicit optimization against the surrogate model $r_{\theta}(x)$.

Lemma 5 enables us to write an exact formulation for the Lagrangian dual function $g(\lambda)$:

Lemma 6 (Explicit Dual Function of (6.15))**.** *Consider the primal problem*

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J(\pi) \simeq -\mathcal{H}(q^{\pi}) - (1+\beta) D_{KL}(q^{\pi} || p_{\mathcal{D}}^{\tau}) \\ s.t. \quad & \mathbb{E}_{p_{\mathcal{D}}^{\tau}}[c^*(x)] - \mathbb{E}_{q^{\pi}}[c^*(x)] \leq W_0 \end{aligned} \tag{6.28}$$

The Lagrangian dual function $g(\lambda)$ is bounded from below by the function $g_{\ell}(\lambda)$ given by

$$g_{\ell}(\lambda) := \beta \left[\lambda (\mathbb{E}_{p_{\mathcal{D}}^{\tau}}[c^*(x)] - W_0) - \mathbb{E}_{p_{\mathcal{D}}^{\tau}} e^{\lambda c^*(x) - 1} \right] \tag{6.29}$$

Proof. Define $f_{\text{KL}}(u) := u \log u$. From (6.18), the dual function $g(\lambda) : \mathbb{R}_+ \to \mathbb{R}$ is given by

$$\begin{aligned} g(\lambda) &:= \min_{\pi \in \Pi} \left[(1+\beta) \mathbb{E}_{p_{\mathcal{D}}^{\tau}} f_{\text{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}^{\tau}} \right) - \mathbb{E}_{q^{\pi}} \log (q^{\pi}) + \beta\lambda \left(\mathbb{E}_{p_{\mathcal{D}}^{\tau}} c^*(x) - \mathbb{E}_{q^{\pi}} c^*(x) - W_0 \right) \right] \\ &= \min_{\pi \in \Pi} \left[(1+\beta) \mathbb{E}_{p_{\mathcal{D}}^{\tau}} f_{\text{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}^{\tau}} \right) - \left(\mathbb{E}_{p_{\mathcal{D}}^{\tau}} f_{\text{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}^{\tau}} \right) + \mathbb{E}_{q^{\pi}} \log p_{\mathcal{D}}^{\tau} \right) + \beta\lambda \left(\mathbb{E}_{p_{\mathcal{D}}^{\tau}} c^*(x) - \mathbb{E}_{q^{\pi}} c^*(x) - W_0 \right) \right] \\ &= \min_{\pi \in \Pi} \left[\beta \mathbb{E}_{p_{\mathcal{D}}^{\tau}} f_{\text{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}^{\tau}} \right) - \mathbb{E}_{q^{\pi}} \log p_{\mathcal{D}}^{\tau} + \beta\lambda \left(\mathbb{E}_{p_{\mathcal{D}}^{\tau}} c^*(x) - \mathbb{E}_{q^{\pi}} c^*(x) - W_0 \right) \right] \end{aligned} \tag{6.30}$$

where we define $\beta\lambda \in \mathbb{R}_+$ as the Lagrangian multiplier associated with the constraint in (6.15). Rearranging terms,

$$g(\lambda) = \min_{\pi \in \Pi} \left[\beta \mathbb{E}_{p_{\mathcal{D}}^{\tau}} \left[- \left(\lambda c^* \cdot \frac{q^{\pi}}{p_{\mathcal{D}}^{\tau}} \right) + f_{\text{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}} \right) \right] - \mathbb{E}_{q^{\pi}} \log p_{\mathcal{D}}^{\tau} + \beta\lambda \mathbb{E}_{p_{\mathcal{D}}^{\tau}} c^*(x) - \beta\lambda W_0 \right] \tag{6.31}$$

Because the sum of function minima is a lower bound on the minima of the sum of the functions

themselves, we have

$$
\begin{aligned}
g(\lambda) &\geq \beta \mathbb{E}_{p^{\tau}_{\mathcal{D}}} \min_{\pi \in \Pi} \left[- \left(\lambda c^* \cdot \frac{q^{\pi}}{p^{\tau}_{\mathcal{D}}} \right) + f_{\mathrm{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}} \right) \right] - \max_{\pi \in \Pi} \left[\mathbb{E}_{q^{\pi}} \log p^{\tau}_{\mathcal{D}} \right] + \min_{\pi \in \Pi} \left[\beta \lambda \mathbb{E}_{p^{\tau}_{\mathcal{D}}} c^*(x) - \beta \lambda W_0 \right] \\
&\sim \beta \mathbb{E}_{p^{\tau}_{\mathcal{D}}} \min_{\pi \in \Pi} \left[- \left(\lambda c^* \cdot \frac{q^{\pi}}{p^{\tau}_{\mathcal{D}}} \right) + f_{\mathrm{KL}} \left(\frac{q^{\pi}}{p_{\mathcal{D}}} \right) \right] + \beta \lambda \mathbb{E}_{p^{\tau}_{\mathcal{D}}} c^*(x) - \beta \lambda W_0
\end{aligned}
\tag{6.32}
$$

ignoring the term $\max_{\pi \in \Pi} \left[\mathbb{E}_{q^{\pi}} \log p^{\tau}_{\mathcal{D}} \right]$ that is constant with respect to λ. In general, simplifying (6.32) is challenging if not intractable. Instead, we note that minimizing over the set of admissible policies Π achieves an optimum that is lower bounded by minimizing over the superset

$$
\begin{aligned}
g(\lambda) &\geq \beta \mathbb{E}_{p^{\tau}_{\mathcal{D}}} \min_{z \in \mathbb{R}_+} \left[- \left(\lambda c^*(x) \cdot z \right) + f_{\mathrm{KL}} \left(z \right) \right] + \beta \lambda \mathbb{E}_{p^{\tau}_{\mathcal{D}}} c^*(x) - \beta \lambda W_0 \\
&= \beta \left[- \mathbb{E}_{p^{\tau}_{\mathcal{D}}} f^{\star}_{\mathrm{KL}} (\lambda c^*(x)) + \lambda (\mathbb{E}_{p^{\tau}_{\mathcal{D}}} c^*(x) - W_0) \right]
\end{aligned}
\tag{6.33}
$$

where $f^{\star}(\cdot)$ is the Fenchel conjugate of a convex function $f(\cdot)$. The Fenchel conjugate of $f_{\mathrm{KL}}(u) = u \log u$ is $f^{\star}_{\mathrm{KL}}(v) = e^{v-1}$ following Borwein and Lewis (2006), and so

$$
g(\lambda) \geq \beta \left[- \mathbb{E}_{p^{\tau}_{\mathcal{D}}} e^{\lambda c^*(x) - 1} + \lambda (\mathbb{E}_{p^{\tau}_{\mathcal{D}}} c^*(x) - W_0) \right] \tag{6.34}
$$

Define the right hand side of this inequality as the function $g_{\ell}(\lambda)$ and the result is immediate. □

Lemma 6 admits an explicit concave function $g_{\ell}(\lambda)$ such that $g(\lambda) \geq g_{\ell}(\lambda)$ for all $\lambda \in \mathbb{R}_+$; because we are interested in maximizing the dual function in leveraging Lagrangian duality as in (6.18), it follows that maximizing $g_{\ell}(\lambda)$ bounds the maxima over $g(\lambda)$ from below. In subsequent steps, we therefore optimize over this explicit function $g_{\ell}(\lambda)$.

The utility of **Lemma 6** is in solving for the optimal λ that maximizes the dual function lower bound in (6.29). Prior work has explored approximating λ via a grid search (Yao et al., 2024) or using iterative implicit solvers; these methods cannot provide any formal guarantee in arriving at a reasonable solution for λ. In contrast, maximizing against $g_{\ell}(\lambda)$ is easy because the function is *guaranteed* to be concave for any β, τ, W_0 and source critic $c^*(x)$. We can therefore derive an *exact* solution for λ using any standard convex optimization problem solver (Agrawal et al., 2019; Diamond and Boyd,

2016). We now have a method to write an explicit expression for the Lagrangian $\mathcal{L}(\pi; \lambda)$ by exactly specifying the optimal λ, and then leverage any out-of-the-box policy optimization method to solve (6.15) via solving the easier, ostensibly unconstrained problem in (6.16).

6.3.5. Overall Algorithm

To summarize, our work aims to solve two separate but related problems in offline MBO in (6.1): traditional model-based optimization approaches can yield candidate designs that are [1] of low diversity; and [2] not optimal due to exploiting out-of-distribution errors of the forward surrogate $r_\theta(x)$. We introduce a KL-divergence-based distribution matching objective—with input hyperparameters τ and β—to solve the diversity problem; and build off prior work (Yao et al., 2024) to constrain the search space using source critic feedback to solve the out-of-distribution evaluation problem. We then show that there exists a provable, explicit solution to our modified offline MBO problem (i.e., **Lemma 6** and (6.16)). In contrast with prior work imposing specific constraints on the forward model (Trabucco et al., 2021; Yu et al., 2021) or design space (Yao et al., 2024), or requiring the use of model-free optimization methods (Krishnamoorthy et al., 2023a,b), *our approach only modifies the MBO objective and is therefore both optimizer- and task- agnostic*. We refer to our method as **Diversity in Adversarial Model-based Optimization (DynAMO)**.

6.4. Experimental Evaluation

6.4.1. Datasets and Offline Optimization Tasks

We evaluate DynAMO on a set of six real-world offline MBO tasks spanning multiple scientific domains and both discrete and continuous search spaces. Five of the tasks are from Design-Bench, a publicly available set of offline optimization benchmarking tasks from Trabucco et al. (2022): (1) **TFBind8** aims to maximize the transcription factor binding efficiency of a short DNA sequence (Barrera et al., 2016); (2) **UTR** the gene expression from a 5′ UTR DNA sequence (Sample et al., 2019; Angermueller et al., 2020a); (3) **ChEMBL** the mean corpuscular hemoglobin concentration (MCHC) biological response of a molecule using an offline dataset from the `CHEMBL3885882` public ChEMBL assay (Gaulton et al., 2012); (4) **Superconductor** the critical temperature of a superconductor material specified by its chemical formula design (Hamidieh, 2018); and (5) **D'Kitty** the

Algorithm 3 (**DynAMO**). **D**iversity **i**n **A**dversarial **M**odel-based **O**ptimization

Input: pre-trained forward surrogate model $r_\theta : \mathcal{X} \to \mathbb{R}$, initalized source critic model $c^* : \mathcal{X} \to \mathbb{R}$, reference dataset $\mathcal{D} = \{(x'_j, r(x'_j))\}_{j=1}^n$, regularization strength $\beta \geq 0$, temperature $\tau \geq 0$, batch size b, optimizer algorithm $a^b : \mathcal{X} \times \mathbb{R} \to \mathcal{X}^b$, source critic learning rate η_{critic}, λ dual step size η_λ, oracle evaluation budget k
Initialize sampled candidates $\mathcal{D}_{\text{gen}} = \varnothing \subset \mathcal{X} \times \mathbb{R}$
while a^b has not converged **do**
  // Solve for the globally optimal λ using (6.29)
  $\lambda \leftarrow \lambda_0 \quad (\lambda_0 = 1.0$ in our experiments)
  while λ has not converged **do**
    $\lambda \leftarrow \lambda + \eta_\lambda \frac{\partial g_\ell(\lambda)}{\partial \lambda}$
  end while
  // Sample new candidates using the optimizer
  $\{x_i^{\text{new}}\}_{i=1}^b \leftarrow a^b(\mathcal{D}_{\text{gen}})$
  // Re-train the source critic parameters θ_c
  $\delta W \leftarrow +\infty$
  while δW has not converged **do**
    $\delta W \leftarrow \vec{\nabla}_{\theta_c} \left[\mathbb{E}_{x' \sim \mathcal{D}}[c^*(x')] - \mathbb{E}_{x \sim \{x_i^{\text{new}}\}_{i=1}^b}[c^*(x)] \right]$
    $\theta_c \leftarrow \min(\max(\theta_c + \eta_{\text{critic}} \cdot \delta W, -0.01), 0.01)$
  end while
  // Evaluate and cache the candidates according to (6.17)
  $\mathcal{D}_{\text{gen}} \leftarrow \mathcal{D}_{\text{gen}} \cup \{(x_i^{\text{new}}, -\mathcal{L}(x_i^{\text{new}}; \lambda))\}_{i=1}^b$
end while
return top-k candidates from $\mathcal{D}_{\text{gen}}$ according to their penalized MBO objective values

morphological structure of a quadrupedal robot (Ahn et al., 2020). Tasks (1) - (3) (i.e., TFBind8, UTR, and ChEMBL) are discrete optimization tasks, where tasks (4) and (5) (i.e., Superconductor and D'Kitty) are continuous optimization tasks. We also evaluate our method on the discrete (6) **Molecule** task described in Brown et al. (2019); Maus et al. (2022); Yao et al. (2024), where the goal is to design a maximally hydrophobic molecule.

6.4.2. Experiment Implementation

All our optimization tasks include an offline, static dataset $\mathcal{D} = \{(x_i, r(x_i))\}_{i=1}^n$ of previously observed designs and their corresponding objective values. We first use $\mathcal{D}$ to train a task-specific forward surrogate model r_θ with parameters θ^* according to (5.1). We parameterize $r_\theta(x)$ as a fully connected neural network with two hidden layers of size 2048 and LeakyReLU activations,

trained using an Adam optimizer with a learning rate of $\eta = 0.0003$ for 100 epochs.

Importantly, optimization problems over *discrete* search spaces are generally NP-hard and often involve heuristic-based solutions (Papalexopoulos et al., 2022; Xiong, 2022). Instead, we use the standard approach of learning a variational autoencoder (VAE) (Kingma and Welling, 2013) to encode and decode discrete designs to and from a continuous latent space, and optimize over the continuous VAE latent space instead. Following prior work (Maus et al., 2022; Tripp et al., 2020; Yao et al., 2024), we co-train a Transformer-based VAE autoencoder (consisting of an encoder $e_\varphi : \hat{\mathcal{X}} \to \mathcal{X}$ parameterized by φ^* and decoder $d_\phi : \mathcal{X} : \mathcal{X} \to \hat{\mathcal{X}}$ parameterized by γ^*) with the surrogate model $r_\theta : \mathcal{X} \to \mathbb{R}$ (parameterized by θ^*) according to

$$\begin{aligned}\theta^*, \varphi^*, \phi^* = \underset{(\theta,\varphi,\phi)\in\Theta\times\Gamma\times\Phi}{\arg\min} \mathbb{E}_{(x,r(x))\sim\mathcal{D}}\Big[&- \log d_\phi(x|e_\varphi(x)) \\ &+ \beta D_{\text{KL}}(\mathcal{N}(0,I)||e_\varphi(x)) + \alpha||r_\theta(e_\varphi(x)) - r(x)||_2^2\Big]\end{aligned} \tag{6.35}$$

where $\mathcal{N}(0, I)$ is the standard multivariate normal prior and $\alpha = 1.0$, $\beta = 10^{-4}$ are constant hyperparameters. We can then perform optimization against r_θ trained on the 256-dimensional continuous latent space of the VAE, and then decode the candidate designs using $d_\phi(\cdot)$ to derive the corresponding discrete design following prior work from Maus et al. (2022); Gómez-Bombarelli et al. (2018). We again use an Adam optimizer with a learning rate of $\eta = 3 \times 10^{-4}$ for both the VAE and the forward surrogate. In this way, the search space for our discrete tasks becomes the $\mathcal{X} \subseteq \mathbb{R}^d$ for $d = 256$, the surrogate model is simply $r_\theta : \mathcal{X} \to \mathbb{R}$, and the reward function $r : \mathcal{X} \to \mathbb{R}$ is now

$$r(x) := \mathbb{E}_{\hat{x}\sim d_\phi(\hat{x}|x)}[\hat{r}(\hat{x})] \tag{6.36}$$

where $\hat{r} : \hat{\mathcal{X}} \to \mathbb{R}$ is the original expert oracle reward function over the discretized input space $\hat{\mathcal{X}}$, and $r(x)$ is the corresponding oracle reward function that accepts our continuous inputs from $\mathcal{X}$ as input. Note that for the MBO tasks over continuous search spaces (i.e., the **Superconductor** and **D'Kitty** tasks), we treat $\mathcal{X} = \hat{\mathcal{X}}$ and fix both the encoder e_φ and decoder d_ϕ to be the identity functions, as no transformation to a separate continuous search space is necessary.

DynAMO also involves training and implementing a source critic model $c^*(x)$ as in (**Lemma 1**); we implement c^* as a fully connected neural network with two hidden layers each with size 512. We implement the constraint on the model's Lipschitz norm by clamping the weights of the model such that the ℓ_∞-norm of the parameters is no greater than 0.01 after each optimization step, consistent with Arjovsky et al. (2017). We train the critic using gradient descent with a learning rate of $\eta = 0.01$ according to (**Lemma 1**). Separately to solve for the globally optimal λ using **Lemma 6**, we perform gradient ascent on λ until the algorithm converges. Finally, we fix the KL-divergence weighting $\beta = 1.0$, temperature hyperparameter $\tau = 1.0$, and constraint bound $W_0 = 0$ for all experiments to avoid overfitting DynAMO to any particular task or optimizer. All experiments were run for 10 random seeds on a single internal cluster with 8 NVIDIA RTX A6000 GPUs. Of note, all DynAMO experiments were run using only a single GPU.

Baseline Methods. Our proposed work, DynAMO, specifically looks to modify an offline MBO optimization problem as in (6.1) where we assume access to a forward surrogate model $r_\theta(x)$ to rank proposed design candidates and offer potential information about the design space. We compare DynAMO against other objective modifying MBO approaches: (1) Conservative Objective Models (**COMs**; Trabucco et al. (2021)) penalizes the objective at a 'look-ahead' gradient-ascent iterate to prevent falsely promising gradient ascent steps; (2) Robust Model Adaptation (**RoMA**; Yu et al. (2021)) modifies the objective $r_\theta(x)$ to enforce a local smoothness prior; (3) Retrieval-enhanced Offline Model-Based Optimization (**ROMO**; Chen et al. (2023c)) retrieves relevant samples from the offline dataset for more trustworthy gradient updates; and (4) Generative Adversarial Model-Based Optimization (**GAMBO**; Yao et al. (2024)) introduces a framework for initially leveraging source critic feedback to regularize an MBO objective. We evaluate each of these MBO objective transformation methods alongside **DynAMO** and naïve, unmodified **Baseline** MBO using representative first-order methods (1) **Grad.** (Gradient Ascent) and (2) **Adam** (Adaptive Moment Estimation (Kingma and Ba, 2014)); evolutionary algorithms (3) **CMA-ES** (Covariance Matrix Adaptation Evolution Strategy (Hansen, 2016)) and (4) **CoSyNE** (Cooperative Synapse Neuroevolution (Gomez et al., 2008)); and Bayesian optimization with (5) Expected Improvement (**BO-qEI**) and (6) Upper Confidence Bound (**BO-qUCB**) acquisition functions.

Notably, the baseline methods COMs and RoMA impose specific constraints on the training process for the forward surrogate model $r_\theta(x)$, and/or also assume that the forward model can be updated during the sampling process (Yu et al., 2021; Trabucco et al., 2021). These constraints are not generally satisfied for any arbitrary offline MBO problem; for example, r_θ may be a non-differentiable black-box simulator with fixed parameters. In contrast, both our method (DynAMO) and baseline methods GAMBO and ROMO are compatible with this more general experimental setting; to ensure a fair experimental comparison, we evaluate both RoMA and COMs using a baseline forward surrogate (i.e., RoMA$^-$, COMs$^-$) and using a specialized forward surrogate model trained and updated according to the methods described by the respective authors (i.e., RoMA$^+$, COMs$^+$).

Evaluating the Diversity of Candidate Designs. To empirically evaluate the diversity of a final set of $k = 128$ candidate designs $\{x_i^F\}_{i=1}^k$ proposed by an offline MBO experiment, we report the **Pairwise Diversity** (PD) of a batch of k candidate designs, defined by Jain et al. (2022); Kim et al. (2023); and Maus et al. (2023) as

$$\text{PD}(\{x_i^F\}_{i=1}^k) := \mathbb{E}_{x_i^F}\left[\mathbb{E}_{x_j^F \neq x_i^F}\left[d(x_i^F, x_j^F)\right]\right] \tag{6.37}$$

where $d(\cdot, \cdot)$ is the normalized Levenshtein edit distance Haldar and Mukhopadhyay (2011) (resp., Euclidean distance) for discrete (resp., continuous) tasks.

Evaluating the Quality of Candidate Designs. To ensure that diversity does not come at the expense of finding optimal design candidates, we report the **Best@**k oracle score obtained by evaluating $k = 128$ candidate designs $\{x_i^F\}_{i=1}^k$ proposed in an experiment. Consistent with prior work (Trabucco et al., 2021; Yao et al., 2024), we define

$$\text{Best@}k(\{x_i^F\}_{i=1}^k) := \max_{1 \leq i \leq k} r(x_i^F) \tag{6.38}$$

Crucially, the Best@k metric is computed with respect to the oracle function $r(x)$ that was hidden during optimization; we only use $r(x)$ in (6.38) to report the true reward associated with each candidate design.

Finally, we rank each method for a given optimizer and task and report the method's **Rank** averaged over the six tasks according to the Best@128 (6.38) and PD (6.37) metrics. We also report the **Optimality Gap** (Opt. Gap) (averaged over the six tasks), defined as as difference between the score achieved by an MBO optimization method and the score in the offline dataset, for both the Best@128 and PD metrics.

Oracle Functions for Optimization Tasks. The task-specific oracle functions $r(x)$ are developed by domain experts and assumed to return the exact, noiseless reward of all possible input designs in the search space $\mathcal{X}$. The oracle functions associated with tasks from the Design-Bench MBO evaluation suite are detailed by the original Design-Bench authors in Trabucco et al. (2022); briefly, the **TFBind8** task uses the oracle function from Barrera et al. (2016); the **UTR** task uses the oracle proposed by Angermueller et al. (2020a); the **ChEMBL** task uses the oracle function from Trabucco et al. (2022); the **Superconductor** task uses the oracle function from Hamidieh (2018); and the task uses a MuJoCo (Todorov et al., 2012) simulation environment and learned control policy from Trabucco et al. (2022) to evaluate input designs. The **Molecule** task uses the oracle function from Ertl and Schuffenhauer (2009).

Data Preprocessing. For all experiments, we follow Krishnamoorthy et al. (2023b) and normalize the objective values both in the offline dataset $\mathcal{D}$ and in those reported in **Section 6.5** according to:

$$y = \frac{\hat{y} - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \tag{6.39}$$

where $\hat{y} = r(x)$ is the original unnormalized oracle value for an input design x, and y_{max} (resp., y_{min}) is the maximum (resp., minimum) value in the full offline dataset. A reported value of $y > 1$ means that an offline optimization experiment proposed a candidate design better than the best design in the offline dataset. Note that in many of the MBO tasks, the publicly available offline dataset $\mathcal{D}$ is only a subset of the designs in the full offline dataset; it is therefore possible (and frequently the case) that $\max_{y \in \mathcal{D}} y < 1$ in our MBO tasks.

As introduced in the main text, we learn a VAE (Kingma and Welling, 2013) model to encode and

decode designs for discrete optimization tasks to and from a continuous latent space, and perform our optimization experiments over the continuous VAE latent space. Following prior work (Maus et al., 2022; Tripp et al., 2020; Yao et al., 2024), we co-train a Transformer-based VAE autoencoder (consisting of an encoder $e_\varphi : \hat{\mathcal{X}} \to \mathcal{X}$ parameterized by φ^* and decoder $d_\phi : \mathcal{X} : \mathcal{X} \to \hat{\mathcal{X}}$ parameterized by γ^*) with the surrogate model $r_\theta : \mathcal{X} \to \mathbb{R}$ (parameterized by θ^*) according to

$$\begin{aligned}\theta^*, \varphi^*, \phi^* = \underset{(\theta,\varphi,\phi)\in\Theta\times\Gamma\times\Phi}{\arg\min} \mathbb{E}_{(x,r(x))\sim\mathcal{D}}\Big[&- \log d_\phi(x|e_\varphi(x)) \\ &+ \beta D_{\mathrm{KL}}(\mathcal{N}(0,I)||e_\varphi(x)) + \alpha||r_\theta(e_\varphi(x)) - r(x)||_2^2\Big]\end{aligned} \tag{6.40}$$

where $\mathcal{N}(0, I)$ is the standard multivariate normal prior and $\alpha = 1$, $\beta = 10^{-4}$ are constant hyperparameters. We can then perform optimization against r_θ trained on the 256-dimensional continuous latent space of the VAE, and subsequently decode the candidate designs using $d_\phi(\cdot)$ to derive the corresponding discrete design following Maus et al. (2022). We again use an Adam optimizer with a learning rate of $\eta = 3 \times 10^{-4}$ for both the VAE and the forward surrogate. In this way, the search space for our discrete tasks becomes the $\mathcal{X} \subseteq \mathbb{R}^d$ for $d = 256$, the surrogate model is simply $r_\theta : \mathcal{X} \to \mathbb{R}$, and the reward function $r : \mathcal{X} \to \mathbb{R}$ is now

$$r(x) := \mathbb{E}_{\hat{x}\sim d_\phi(\hat{x}|x)}[\hat{r}(\hat{x})] \tag{6.41}$$

where $\hat{r} : \hat{\mathcal{X}} \to \mathbb{R}$ is the original expert oracle reward function over the discretized input space $\hat{\mathcal{X}}$, and $r(x)$ is the corresponding oracle reward function that accepts our continuous inputs from $\mathcal{X}$ as input. Note that for the MBO tasks over continuous search spaces (i.e., the **Superconductor** and **D'Kitty** tasks), we treat $\mathcal{X} = \hat{\mathcal{X}}$ and fix both the encoder e_φ and decoder d_ϕ to be the identity functions, as no transformation to a separate continuous search space is necessary.

Optimization Experiments. All baseline methods were evaluated using their official open-source implementations made publicly available by the respective authors. In DynAMO, we initialize all optimizers using the first b elements from a d-dimensional scrambled Sobol sequence (Sobol, 1967) using the official PyTorch quasi-random generator `SobolEngine` implementation, where b is the

sampling batch size and d is the dimensionality of the search space. Note that the Sobol sequence only returns points with dimensions between 0 and 1; for each task, we therefore un-normalize the sampled Sobol points $\tilde{x}_0$ according to $x_0 = x_{\text{min}} + (\tilde{x}_0 \cdot (x_{\text{max}} - x_{\text{min}}))$, where $x_{\text{max}}, x_{\text{min}}$ are the maximum and minimum bounds on the search space for our experiments, respectively. We fix $x_{\text{min}} = -4.0$ and $x_{\text{max}} = +4.0$ for all d dimensions across all tasks. In all experiments reported in **Table 6.1**, each optimizer continues to sample from the search space in batched acquisitions of b samples—we set $b = 64$ for all our experiments unless otherwise stated. After each acquisition, we score the sampled designs using the (penalized) forward surrogate model (i.e., the Lagrangian in (6.17) for DynAMO). If the maximum prediction from the recently sampled batch is not at least as optimal as the maximum prediction of the previously sampled designs, then we define the acquisition step as a *failure*; a sequence of 10 consecutive failures triggers a *restart* in the optimization process where the optimizer starts from the scratch beginning with sampling with the Sobol sequence to initialize the optimizer as described above. After 3 restarts, we consider the optimization process terminated, and all designs across all restarts are aggregated to choose the top $k = 128$ final candidate designs to be evaluated using the oracle reward function.

6.5. Results

6.5.1. Main Results

DynAMO consistently proposes the most diverse set of designs and achieves an Optimality Gap as high as **74.2** (DynAMO-BO-qUCB) and an average Rank as low as **1.2** (**Table 6.1**). We find that DynAMO offers the largest improvements in diversity for first-order methods, although also improves upon the evolutionary algorithms and Bayesian optimization methods. This makes sense, as both Grad. and Adam are only local optimizers that often end up exploring a much smaller region of the design space (without using DynAMO) compared to gradient-free methods. For example, DynAMO-Grad. (resp., DynAMO-CMA-ES; resp., DynAMO-BO-qEI) achieves a Pairwise Diversity Optimality Gap of 35.7 (resp., 55.2; resp., 74.2); in contrast, no other baseline method achieves a diversity score greater than -6.9 (resp., 16.8; resp., 51.4) within the same optimizer class.

These results do not come at the cost of the quality of designs; for example, for all 3 optimizers

Table 6.1: **Quality and diversity of designs under MBO objective transforms.** We evaluate DynAMO against other MBO objective-modifying methods using six different backbone optimizers. Each cell consists of '**Best@128 (Best)**/**Pairwise Diversity (PD)**' Rank and Optimality Gap scores separated by a forward slash. **Bolded** (resp., Underlined) entries indicate the best (resp., second best) performing algorithm for a given optimizer (i.e., within each column). See **Supp. Table D.1** for detailed results broken down by MBO task.

Best/PD	**Rank ↓**						**Optimality Gap ↑**					
	Grad.	Adam	CMA-ES	CoSyNE	BO-qEI	BO-qUCB	Grad.	Adam	CMA-ES	CoSyNE	BO-qEI	BO-qUCB
Baseline	5.0/5.5	4.5/6.0	3.7/3.8	5.3/4.5	5.8/5.2	3.7/3.0	6.8/-53.2	0.5/-52.4	14.4/9.5	-0.6/-52.1	18.7/47.1	19.4/43.5
COMs$^-$	7.3/6.5	6.0/5.3	5.7/5.2	5.7/5.7	4.3/4.0	4.5/4.5	-3.0/-53.5	-3.0/-52.4	7.6/-9.7	1.1/-51.6	19.2/51.4	19.0/45.5
COMs$^+$	**2.5**/2.8	3.2/3.0	7.3/7.8	5.7/3.3	6.0/5.7	5.2/5.7	12.3/-6.9	8.1/-12.3	7.1/-38.2	3.5/-40.4	17.9/40.3	18.6/51.3
RoMA$^-$	6.7/5.7	4.5/6.3	3.7/3.5	5.3/4.5	2.7/3.3	3.8/3.0	-1.2/-53.3	0.5/-52.4	14.4/9.5	-0.6/-52.2	**21.0**/48.4	19.2/43.6
RoMA$^+$	3.8/5.8	**2.8**/5.2	5.0/4.8	2.8/5.0	5.2/6.3	4.7/6.5	9.2/-46.8	**14.5**/-45.8	14.1/8.0	6.2/-52.0	18.3/32.9	18.5/39.9
ROMO	4.2/2.8	4.8/2.8	4.2/4.3	4.2/5.2	5.0/6.0	4.7/6.2	10.9/-12.7	6.4/-20.5	15.7/-3.1	3.1/-50.8	19.2/34.9	19.9/33.2
GAMBO	3.2/5.3	5.3/5.8	**2.2**/4.3	3.7/6.3	**2.2**/4.0	4.7/5.0	10.5/-52.1	8.6/-51.9	16.7/16.8	5.0/-53.6	20.8/30.0	20.2/30.3
DynAMO	2.8/**1.2**	**2.8**/**1.2**	3.3/**1.8**	**2.3**/**1.2**	3.0/**1.3**	**3.5**/**1.8**	**14.2**/**27.8**	**14.5**/**35.7**	**17.5**/**55.2**	**12.3**/**-20.7**	20.7/**74.2**	**20.5**/**59.4**

where DynAMO scores an average Rank of 1.2 (i.e., Grad., Adam, and CoSyNE backbone optimizers), DynAMO is also within the **top 2 methods** in proposing *high-quality* designs according to both Rank and Optimality Gap. In fact, DynAMO proposes the best designs for 5 out of the 6 backbone optimizers according to the Best@128 Optimality Gap. These results suggest that DynAMO can be used to improve both the quality *and* diversity of designs in a variety of experimental settings for both discrete and continuous search spaces.

6.5.2. Ablation Studies

In this section, we ablate key hyperparameters and algorithmic components of DynAMO in **Algorithm 3** to better interrogate and understand the utility of each algorithmic module.

Sampling Batch Size Ablation. Recall from (6.15) that a key component of our DynAMO algorithm is the estimation of the empirical KL-divergence between the τ-weighted probability distribution of real designs from the offline dataset and the distribution of sampled designs from the generative policy. The latter distribution of generated designs is fundamentally dependent on our sampling batch size b in **Algorithm 3**—the larger the batch size per sampling step, the better our empirical estimate of the KL divergence between our two distributions. However, as the batch size increases, there also exists a greater likelihood of significant regret in the sampling policy when compared to the optimal sequential policy (Gonzalez et al., 2016; Wilson et al., 2017). To better evaluate the impact of the sampling batch size parameter b, we experimentally evaluate sampling

batch size values logarithmically ranging between $2 \leq b \leq 512$. We use a BO-qEI sampling policy with the DynAMO-modified objective on the TFBind8 optimization task, and evaluate both the Best@128 oracle score and Pairwise Diversity of the 128 final proposed design candidates (**Fig. 6.1**). We find that the Best@128 design quality scores do not vary significantly as a function of the batch sizes that were evaluated; however, there exists an optimal batch size ($b = 64$ in our experiments) that maximizes the diversity of designs according to the pairwise diversity metric.

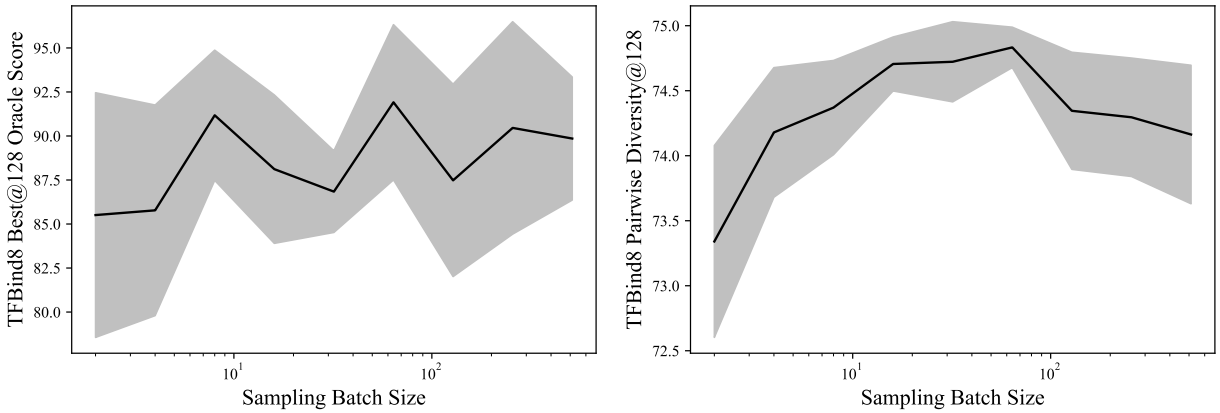

Figure 6.1: **Sampling batch size ablation.** We vary the sampling batch size b in **Algorithm 3** between 2 and 512, and report both the (**left**) Best@128 Oracle Score and (**right**) Pairwise Diversity score for 128 final designs proposed by a DynAMO-BO-qEI policy on the TFBind8 optimization task. We plot the mean $\pm$ 95% confidence interval over 10 random seeds.

Adversarial Critic Feedback and Distribution Matching Ablation. Recall that instead of solving the original MBO optimization problem in (6.1), DynAMO leverages weak Lagrangian duality to solve the constrained optimization problem in (6.15)—copied below for convenience:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J(\pi) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^\pi || p^\tau_{\mathcal{D}}) \\ \text{s.t.} \quad & \mathbb{E}_{x \sim p^\tau_{\mathcal{D}}(x)} c^*(x) - \mathbb{E}_{x \sim q^\pi(x)} c^*(x) \leq W_0 \end{aligned} \tag{6.42}$$

We can think of this problem formulation as as the fusion of two separable components: (1) Adversarial feedback via a source-critic model $c^*(x)$ to prevent out-of-distribution evaluation of $r_\theta(x)$; and (2) **Diversity** (via KL-divergence-based distribution matching with a diverse reference distribution $p^\tau_{\mathcal{D}}$) **in** Model-based Optimization. These two components together form the foundation of **DynAMO** presented in **Algorithm 3**. To better understand how each of these two com-

ponents affects the performance of DynAMO-augmented optimizers, we can separate these two components and study them individually.

AMO is our ablation method that solves the related optimization problem

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \mathbb{E}_{q^\pi}[r_\theta(x)] \\ \text{s.t.} \quad & \mathbb{E}_{x \sim p^\tau_{\mathcal{D}}(x)} c^*(x) - \mathbb{E}_{x \sim q^\pi(x)} c^*(x) \leq W_0 \end{aligned} \tag{6.43}$$

instead of (6.42). Note that AMO solves the same constrained optimization problem as DynAMO in the setting where $\beta = 0$, and is equivalent to the problem formulation considered in the previous chapter (see Yao et al. (2024) for additional details). We note that our derivation of the Lagrange dual function of (6.15) in **Lemma 5** is invalid when $\beta = 0$, and so we cannot exactly solve (6.43) using the same methodology presented in **Algorithm 3**. Instead, we leverage the **adaptive Source Critic Regularization (aSCR)** algorithm from Yao et al. (2024) to *approximate* a solution to (6.43) in the Lagrangian dual space—see **Chapter 5** for additional discussion regarding the specific implementation details of aSCR.

Separately, **Dyn**MO is our separate ablation method that solves the related (unconstrained) optimization problem

$$\max_{\pi \in \Pi} \quad J(\pi) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_{\text{KL}}(q^\pi || p^\tau_{\mathcal{D}}) \tag{6.44}$$

instead of (6.42). To implement DynMO empirically, we modify **Algorithm 3** by ignoring the subroutine to solve for the globally optimal Lagrange multiplier λ using (6.29), and instead fixing $\lambda = 0$ for the entire optimization process to effectively remove any contributions from the adversarial source critic $c^*(x)$. All other implementation details were kept constant.

We compare DynAMO with AMO and **Dyn**MO in **Tables 6.2-6.4**. Firstly, we note that DynAMO and AMO are competitive in proposing the high-quality designs according to the Best@128 oracle scores, alternating between having the highest and second best Rank and Optimality Gaps across all six tasks when compared with DynMO and the baseline optimizer for all optimizers evaluated. This makes sense, as the purpose of the adversarial source critic-dependent constraint in (6.42)

and (6.43) is to minimize out-of-distribution evaluation of $r_\theta(x)$ during optimization—as a result, the forward surrogate model $r_\theta(x)$ can provide a better estimate of the quality of sampled designs, leading to higher quality designs according to the true oracle function $r(x)$. Separately, we find that DynMO and DynAMO also perform similarly in terms of the all 3 diversity metrics evaluated. However, we find that DynMO (resp., AMO) struggles on proposing high-quality (resp., diverse) sets of final designs. These experimental results collectively allow us to conclude that both the adversarial source critic supervision and KL-divergence-based distribution matching are important for DynAMO to propose *both* high-quality and diverse sets of designs.

β and τ Hyperparameter Ablation. Fundamentally, DynAMO relies on two important hyperparameters that define the constrained optimization problem in (6.15): (1) the β hyperparameter dictates the relative weighting of the KL-divergence penalty relative to the original MBO objective; and (2) the τ temperature hyperparameter describes the distribution of reference designs weighted according to their oracle scores in the offline dataset. To better interrogate how these hyperparameters impact the performance of DynAMO-augmented MBO optimizers, we (independently) ablate the values of both β and τ logarithmically between $0.01 \leq \beta, \tau \leq 100$. We use a BO-qEI sampling policy with the DynAMO-modified objective on the TFBind8 optimization task, and evaluate both the Best@128 oracle score and Pairwise Diversity of the 128 final proposed design candidates.

Our results suggest that as the strength of the KL-divergence term β increases, the diversity of proposed designs (according to the Pairwise Diversity metric) increases roughly proportional to the logarithm of β (**Fig. 6.2**). This is expected: as the distribution matching objective becomes more important relative to the $r_\theta(x)$ forward surrogate model, the generative policy is rewarded for finding an increasingly diverse set of designs that matches the τ-weighted reference distribution. Similarly, we found that for sufficiently large values of β (i.e., $\beta \geq 0.03$ in our particular experimental setting), the *quality* of designs (according to the Best@128 oracle score) decreases due to the inherent trade-off between design quality (according to $r_\theta(x)$) and diversity (according to the KL-divergence in (6.15)). Interestingly, for small values of β (i.e., $\beta \leq 0.03$) the quality of designs actually *increases* with β. This is because in this regime, naïvely optimizing against pri-

Table 6.2: **Quality of design candidates in adversarial critic feedback (AMO) and diversity in (DynMO) model-based optimization.** We evaluate our method (1) with the KL-divergence penalized-MBO objective as in (6.19) only (**Dyn**MO); (2) with the adversarial source critic-dependent constraint as introduced by Yao et al. (2024) only (AMO); and (3) with both algorithmic components as in **Dyn**AMO described in **Algorithm 3**. We report the Best@128 (resp., Median@128) oracle score achieved by the 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Best@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	43.9	59.4	60.5	88.9	40.0	88.4	—	—
Grad.	**90.0**$^{(4.3)}$	**80.9**$^{(12.1)}$	**60.2**$^{(8.9)}$	**88.8**$^{(4.0)}$	**36.0**$^{(6.8)}$	**65.6**$^{(14.5)}$	3.2	6.8
AMO-Grad.	**73.1**$^{(12.8)}$	**77.1**$^{(9.6)}$	**64.4**$^{(1.5)}$	**92.8**$^{(8.0)}$	**46.0**$^{(6.8)}$	**90.6**$^{(14.5)}$	1.8	10.5
DynMO-Grad.	61.3$^{(9.7)}$	63.6$^{(11.6)}$	**59.8**$^{(8.6)}$	**89.3**$^{(5.6)}$	**36.3**$^{(6.9)}$	**70.4**$^{(12.0)}$	3.5	0.0
DynAMO-Grad.	**90.3**$^{(4.7)}$	**86.2**$^{(0.0)}$	**64.4**$^{(2.5)}$	**91.2**$^{(0.0)}$	**44.2**$^{(7.8)}$	**89.8**$^{(3.2)}$	**1.5**	**14.2**
Adam	62.9$^{(13.0)}$	69.7$^{(10.5)}$	**62.9**$^{(1.9)}$	**92.3**$^{(8.9)}$	**37.8**$^{(6.3)}$	**58.4**$^{(18.5)}$	2.8	0.5
AMO-Adam	**94.0**$^{(2.2)}$	60.0$^{(12.6)}$	**60.9**$^{(8.7)}$	**91.4**$^{(6.3)}$	**37.8**$^{(6.3)}$	**88.4**$^{(13.8)}$	2.8	8.6
DynMO-Adam	66.6$^{(12.9)}$	68.7$^{(10.1)}$	63.7$^{(0.4)}$	**92.0**$^{(8.3)}$	**38.6**$^{(5.7)}$	**66.5**$^{(14.6)}$	2.5	2.5
DynAMO-Adam	**95.2**$^{(1.7)}$	**86.2**$^{(0.0)}$	**65.2**$^{(1.1)}$	**91.2**$^{(0.0)}$	**45.5**$^{(5.7)}$	**84.9**$^{(12.0)}$	**1.7**	**14.5**
BO-qEI	**87.3**$^{(5.8)}$	**86.2**$^{(0.0)}$	**65.4**$^{(1.0)}$	117$^{(3.1)}$	**53.1**$^{(3.3)}$	84.4$^{(0.9)}$	3.5	18.7
AMO-BO-qEI	**94.1**$^{(1.9)}$	**86.3**$^{(0.2)}$	**66.8**$^{(0.7)}$	**121**$^{(0.0)}$	**50.8**$^{(3.3)}$	**86.7**$^{(1.1)}$	**1.5**	**20.8**
DynMO-BO-qEI	**93.2**$^{(3.3)}$	**86.2**$^{(0.0)}$	**66.0**$^{(0.8)}$	**121**$^{(0.0)}$	**49.6**$^{(2.6)}$	**85.9**$^{(1.0)}$	2.7	20.2
DynAMO-BO-qEI	**91.9**$^{(4.4)}$	**86.2**$^{(0.0)}$	**67.0**$^{(1.3)}$	**121**$^{(0.0)}$	**53.5**$^{(5.0)}$	**85.5**$^{(1.1)}$	1.8	20.7
BO-qUCB	88.1$^{(5.3)}$	**86.2**$^{(0.1)}$	**66.4**$^{(0.7)}$	**121**$^{(1.3)}$	**51.3**$^{(3.6)}$	**84.5**$^{(0.8)}$	**2.2**	19.4
AMO-BO-qUCB	**95.4**$^{(1.6)}$	**86.2**$^{(0.0)}$	**66.3**$^{(1.1)}$	**121**$^{(1.3)}$	**50.2**$^{(2.8)}$	**83.6**$^{(1.0)}$	2.7	20.2
DynMO-BO-qUCB	**93.6**$^{(3.0)}$	**86.2**$^{(0.1)}$	**66.0**$^{(0.9)}$	**121**$^{(0.0)}$	**49.9**$^{(3.0)}$	**83.9**$^{(1.1)}$	2.7	20.0
DynAMO-BO-qUCB	**95.1**$^{(1.9)}$	**86.2**$^{(0.0)}$	**66.7**$^{(1.5)}$	**121**$^{(0.0)}$	**48.1**$^{(4.0)}$	**86.9**$^{(4.5)}$	**2.2**	**20.5**
Baseline-CMA-ES	**87.6**$^{(8.3)}$	**86.2**$^{(0.0)}$	**66.1**$^{(1.0)}$	106$^{(5.9)}$	**49.0**$^{(1.0)}$	72.2$^{(0.1)}$	2.8	14.4
AMO-CMA-ES	**90.4**$^{(4.4)}$	**86.2**$^{(0.0)}$	**66.2**$^{(1.6)}$	**121**$^{(0.0)}$	45.2$^{(3.5)}$	72.2$^{(0.1)}$	**1.8**	16.7
DynMO-CMA-ES	**85.2**$^{(10.1)}$	**86.2**$^{(0.0)}$	**65.0**$^{(0.6)}$	104$^{(7.8)}$	**51.6**$^{(2.0)}$	**83.6**$^{(3.1)}$	2.7	15.8
DynAMO-CMA-ES	**89.8**$^{(3.6)}$	**85.7**$^{(5.8)}$	**63.9**$^{(0.9)}$	**117**$^{(6.7)}$	**50.6**$^{(4.8)}$	**78.5**$^{(5.5)}$	2.7	**17.5**
CoSyNE	61.7$^{(10.0)}$	**57.3**$^{(9.6)}$	**63.6**$^{(0.4)}$	94.8$^{(10.1)}$	**37.0**$^{(4.1)}$	**62.7**$^{(13.1)}$	3.5	-0.6
AMO-CoSyNE	**79.8**$^{(10.6)}$	**68.0**$^{(12.5)}$	**64.2**$^{(0.9)}$	**99.4**$^{(15.0)}$	**37.0**$^{(4.1)}$	**62.7**$^{(13.1)}$	2.2	5.0
DynMO-CoSyNE	63.6$^{(10.1)}$	**59.3**$^{(10.8)}$	**63.9**$^{(1.6)}$	90.1$^{(12.7)}$	**37.0**$^{(4.1)}$	**62.7**$^{(13.1)}$	3.0	-0.7
DynAMO-CoSyNE	**91.3**$^{(4.4)}$	**77.2**$^{(11.6)}$	**63.9**$^{(0.9)}$	**114**$^{(7.0)}$	**40.6**$^{(8.6)}$	**67.5**$^{(14.1)}$	**1.2**	**12.3**

Median@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	33.7	42.8	50.9	87.6	6.7	77.8	—	—
Grad.	**58.1**$^{(6.1)}$	**58.6**$^{(13.1)}$	**59.3**$^{(8.6)}$	**85.3**$^{(7.7)}$	**36.0**$^{(6.7)}$	**65.1**$^{(14.4)}$	3.2	10.5
AMO-Grad.	**63.8**$^{(13.7)}$	**75.3**$^{(9.9)}$	**60.1**$^{(3.3)}$	**91.6**$^{(11.2)}$	**46.0**$^{(6.7)}$	**90.1**$^{(14.4)}$	**1.2**	**21.2**
DynMO-Grad.	**50.5**$^{(6.5)}$	**58.6**$^{(13.1)}$	**59.7**$^{(8.7)}$	**85.1**$^{(8.1)}$	**36.3**$^{(6.9)}$	**70.0**$^{(12.0)}$	3.0	10.1
DynAMO-Grad.	47.0$^{(2.8)}$	**69.8**$^{(6.0)}$	**61.9**$^{(2.2)}$	**85.9**$^{(0.4)}$	23.4$^{(8.5)}$	**68.7**$^{(12.1)}$	2.7	9.5
Adam	**54.7**$^{(8.8)}$	**60.4**$^{(12.7)}$	**59.2**$^{(8.6)}$	**87.9**$^{(10.0)}$	**37.4**$^{(6.2)}$	56.8$^{(19.8)}$	2.3	9.5
AMO-Adam	**49.5**$^{(8.9)}$	**55.7**$^{(12.7)}$	**57.7**$^{(9.1)}$	**84.3**$^{(9.6)}$	**37.4**$^{(6.2)}$	**87.8**$^{(4.3)}$	3.0	**12.1**
DynMO-Adam	**54.0**$^{(9.9)}$	**60.5**$^{(12.6)}$	**59.3**$^{(8.6)}$	**85.9**$^{(10.8)}$	**37.7**$^{(6.4)}$	63.6$^{(15.6)}$	**2.2**	10.2
DynAMO-Adam	**47.7**$^{(3.0)}$	**69.0**$^{(5.2)}$	**62.4**$^{(1.9)}$	**86.4**$^{(0.6)}$	23.0$^{(6.0)}$	65.6$^{(14.1)}$	2.3	9.1
BO-qEI	48.5$^{(1.5)}$	59.9$^{(2.0)}$	**63.3**$^{(0.0)}$	**86.7**$^{(0.6)}$	**28.7**$^{(1.8)}$	72.4$^{(1.8)}$	2.5	10.0
AMO-BO-qEI	46.4$^{(1.8)}$	**63.4**$^{(3.3)}$	**63.3**$^{(0.0)}$	**86.3**$^{(0.5)}$	**28.9**$^{(1.1)}$	**79.1**$^{(0.7)}$	2.2	11.3
DynMO-BO-qEI	**50.5**$^{(1.5)}$	**61.1**$^{(2.8)}$	**63.3**$^{(0.0)}$	**86.4**$^{(0.7)}$	**28.4**$^{(0.7)}$	**79.1**$^{(0.9)}$	2.3	**11.5**
DynAMO-BO-qEI	**51.5**$^{(0.9)}$	**65.6**$^{(3.1)}$	**63.3**$^{(0.0)}$	**86.7**$^{(0.6)}$	23.5$^{(2.4)}$	77.0$^{(0.7)}$	**2.0**	11.3
BO-qUCB	**50.3**$^{(1.8)}$	**62.1**$^{(3.4)}$	**63.3**$^{(0.0)}$	**86.6**$^{(0.6)}$	**31.7**$^{(1.2)}$	**74.4**$^{(0.6)}$	**1.7**	**11.5**
AMO-BO-qUCB	**47.9**$^{(1.9)}$	59.8$^{(1.2)}$	**63.3**$^{(0.0)}$	**86.0**$^{(0.6)}$	**33.1**$^{(2.9)}$	**73.8**$^{(1.2)}$	2.8	10.7
DynMO-BO-qUCB	**50.3**$^{(1.7)}$	**60.1**$^{(2.2)}$	**63.3**$^{(0.0)}$	**86.4**$^{(0.6)}$	**32.4**$^{(2.7)}$	**74.3**$^{(0.8)}$	2.0	11.2
DynAMO-BO-qUCB	**48.8**$^{(1.8)}$	**65.9**$^{(3.7)}$	**63.3**$^{(0.0)}$	**86.5**$^{(0.5)}$	22.7$^{(2.0)}$	50.4$^{(14.6)}$	2.5	6.3
CMA-ES	**50.7**$^{(2.7)}$	**71.7**$^{(10.4)}$	**63.3**$^{(0.0)}$	83.9$^{(1.0)}$	**37.9**$^{(0.7)}$	**59.3**$^{(10.9)}$	2.2	11.2
AMO-CMA-ES	44.2$^{(0.8)}$	**72.7**$^{(3.8)}$	**62.7**$^{(1.1)}$	86.1$^{(0.5)}$	21.4$^{(2.0)}$	**54.9**$^{(9.6)}$	3.2	7.1
DynMO-CMA-ES	**50.7**$^{(2.7)}$	**75.2**$^{(9.1)}$	**63.3**$^{(0.0)}$	82.5$^{(1.2)}$	**38.7**$^{(3.9)}$	**65.8**$^{(9.1)}$	**1.7**	**12.8**
DynAMO-CMA-ES	45.3$^{(2.4)}$	**65.8**$^{(8.9)}$	59.3$^{(3.8)}$	**99.0**$^{(12.1)}$	22.5$^{(5.1)}$	**60.6**$^{(15.0)}$	2.8	8.8
CoSyNE	**55.3**$^{(8.0)}$	**53.6**$^{(10.2)}$	**60.8**$^{(3.2)}$	**87.4**$^{(16.6)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	2.5	8.9
AMO-CoSyNE	**59.5**$^{(12.0)}$	**63.5**$^{(11.2)}$	**55.4**$^{(9.6)}$	**84.2**$^{(17.2)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	**2.2**	9.8
DynMO-CoSyNE	**59.2**$^{(10.8)}$	**55.6**$^{(8.2)}$	**60.4**$^{(5.9)}$	**87.9**$^{(12.5)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	2.3	**9.9**
DynAMO-CoSyNE	**53.8**$^{(11.0)}$	**63.4**$^{(11.5)}$	**59.3**$^{(3.8)}$	**99.0**$^{(12.1)}$	20.5$^{(5.8)}$	**60.6**$^{(15.0)}$	2.5	9.5

Table 6.3: **Diversity of design candidates in adversarial critic feedback (AMO) and diversity in (DynMO) model-based optimization.** We evaluate our method (1) with the KL-divergence penalized-MBO objective as in (6.19) only (i.e., **Dyn**MO); (2) with the adversarial source critic-dependent constraint as introduced by Yao et al. (2024) only (AMO); and (3) with both algorithmic components as in **Dyn**AMO described in **Algorithm 3**. We report the pairwise diversity (resp., minimum novelty) oracle score achieved by the 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given optimizer.

Pairwise Diversity@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	65.9	57.3	60.0	36.7	66.0	85.7	—	—
Grad.	$12.5^{(8.0)}$	$7.8^{(8.8)}$	$7.9^{(7.8)}$	$24.1^{(13.3)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	3.0	-53.2
AMO-Grad.	$17.3^{(12.8)}$	$11.2^{(10.3)}$	$6.9^{(7.7)}$	$22.1^{(10.5)}$	$0.0^{(0.0)}$	$1.5^{(3.2)}$	2.7	-52.1
DynMO-Grad.	$20.9^{(15.1)}$	$3.0^{(3.2)}$	**$58.2^{(24.0)}$**	$13.5^{(8.6)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	3.0	-46.0
DynAMO-Grad.	**$66.9^{(6.9)}$**	**$68.2^{(10.8)}$**	**$77.2^{(21.5)}$**	**$93.0^{(1.2)}$**	**$129^{(55.3)}$**	**$104^{(56.1)}$**	**1.0**	**27.8**
Adam	$12.0^{(12.3)}$	$11.0^{(12.1)}$	$4.8^{(3.8)}$	$16.8^{(12.4)}$	$6.4^{(14.5)}$	$6.2^{(14.0)}$	3.0	-52.4
AMO-Adam	$15.1^{(11.2)}$	$10.3^{(11.5)}$	$12.1^{(11.3)}$	$19.6^{(15.2)}$	$0.3^{(0.8)}$	$2.6^{(3.9)}$	3.0	-51.9
DynMO-Adam	$13.1^{(11.4)}$	$10.3^{(9.6)}$	**$57.0^{(26.0)}$**	$23.8^{(15.1)}$	$6.4^{(14.5)}$	$0.0^{(0.0)}$	2.8	-43.5
DynAMO-Adam	**$54.8^{(8.9)}$**	**$72.3^{(3.4)}$**	**$84.8^{(9.2)}$**	**$89.9^{(5.3)}$**	**$158^{(37.3)}$**	**$126^{(57.3)}$**	**1.0**	**35.7**
BO-qEI	$73.7^{(0.6)}$	**$73.8^{(0.5)}$**	**$99.3^{(0.1)}$**	**$93.0^{(0.5)}$**	$190^{(0.8)}$	$124^{(7.4)}$	3.7	47.1
AMO-BO-qEI	$74.0^{(0.6)}$	**$74.3^{(0.4)}$**	**$99.3^{(0.1)}$**	**$93.3^{(0.4)}$**	$193^{(1.2)}$	$17.7^{(3.5)}$	2.8	30.0
DynMO-BO-qEI	**$74.5^{(0.3)}$**	**$74.3^{(0.6)}$**	**$99.3^{(0.1)}$**	**$93.3^{(0.7)}$**	**$200^{(3.0)}$**	$135^{(11.2)}$	2.3	50.8
DynAMO-BO-qEI	**$74.8^{(0.2)}$**	**$74.6^{(0.3)}$**	**$99.4^{(0.1)}$**	**$93.5^{(0.4)}$**	**$198^{(1.9)}$**	**$277^{(59.7)}$**	**1.2**	**74.2**
BO-qUCB	$73.9^{(0.5)}$	**$74.3^{(0.4)}$**	**$99.4^{(0.1)}$**	**$93.6^{(0.5)}$**	**$198^{(10.3)}$**	$94.1^{(3.9)}$	2.5	43.5
AMO-BO-qUCB	$74.0^{(0.5)}$	**$74.3^{(0.3)}$**	**$99.3^{(0.1)}$**	**$93.4^{(0.4)}$**	**$190^{(9.3)}$**	$22.0^{(2.1)}$	3.7	30.3
DynMO-BO-qUCB	**$74.7^{(0.2)}$**	**$74.3^{(0.4)}$**	$99.2^{(0.1)}$	**$93.6^{(0.5)}$**	**$198^{(12.0)}$**	$92.6^{(3.8)}$	2.2	43.5
DynAMO-BO-qUCB	**$74.3^{(0.5)}$**	**$74.4^{(0.6)}$**	**$99.3^{(0.1)}$**	**$93.5^{(0.6)}$**	**$211^{(22.8)}$**	**$175^{(44.7)}$**	**1.7**	**59.4**
CMA-ES	$47.2^{(11.2)}$	$44.6^{(15.9)}$	**$93.5^{(2.0)}$**	$66.2^{(9.4)}$	$12.8^{(0.6)}$	$164^{(10.6)}$	2.3	9.5
AMO-CMA-ES	$39.6^{(15.5)}$	$53.4^{(8.4)}$	$84.8^{(4.8)}$	$61.3^{(14.6)}$	**$173^{(19.4)}$**	$59.9^{(19.6)}$	2.3	16.8
DynMO-CMA-ES	$33.5^{(2.6)}$	$11.1^{(1.1)}$	$34.5^{(2.8)}$	$4.5^{(5.0)}$	$38.1^{(5.4)}$	$14.4^{(1.2)}$	3.8	-39.3
DynAMO-CMA-ES	**$73.6^{(0.6)}$**	**$73.1^{(3.1)}$**	$72.0^{(3.1)}$	**$94.0^{(0.5)}$**	$97.8^{(13.2)}$	**$292^{(83.5)}$**	**1.5**	**55.2**
CoSyNE	$5.6^{(5.0)}$	**$12.7^{(9.8)}$**	**$28.2^{(11.3)}$**	**$12.2^{(7.3)}$**	$0.0^{(0.0)}$	$0.0^{(0.0)}$	2.8	-52.1
AMO-CoSyNE	$5.2^{(5.7)}$	**$9.1^{(9.0)}$**	**$28.4^{(15.7)}$**	**$7.1^{(8.0)}$**	$0.0^{(0.0)}$	$0.0^{(0.0)}$	3.7	-53.6
DynMO-CoSyNE	**$27.4^{(10.3)}$**	**$13.0^{(11.8)}$**	**$53.3^{(24.2)}$**	**$18.1^{(13.8)}$**	$0.0^{(0.0)}$	$0.0^{(0.0)}$	2.2	-43.3
DynAMO-CoSyNE	**$18.1^{(13.0)}$**	**$20.3^{(2.3)}$**	**$35.0^{(17.9)}$**	**$22.8^{(11.9)}$**	**$74.4^{(46.3)}$**	**$77.0^{(35.9)}$**	**1.3**	**-20.7**

Minimum Novelty@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	0.0	0.0	0.0	0.0	0.0	0.0	—	—
Grad.	**$21.2^{(3.0)}$**	**$51.7^{(2.9)}$**	**$97.4^{(3.9)}$**	**$79.5^{(19.7)}$**	**$95.0^{(0.7)}$**	**$102^{(6.1)}$**	2.3	74.5
AMO-Grad.	$14.0^{(2.0)}$	$46.7^{(2.7)}$	**$96.8^{(3.9)}$**	**$76.8^{(19.7)}$**	$83.8^{(6.8)}$	$31.5^{(3.6)}$	3.8	58.3
DynMO-Grad.	**$21.9^{(3.0)}$**	**$53.6^{(2.1)}$**	**$93.1^{(6.6)}$**	**$86.4^{(10.2)}$**	**$95.0^{(0.7)}$**	**$102^{(6.1)}$**	1.8	75.4
DynAMO-Grad.	**$21.1^{(1.1)}$**	**$52.2^{(1.3)}$**	**$98.6^{(1.5)}$**	**$85.8^{(1.0)}$**	**$95.0^{(0.4)}$**	**$107^{(6.7)}$**	**1.7**	**76.7**
Adam	**$23.7^{(2.8)}$**	**$51.1^{(3.5)}$**	**$95.5^{(5.3)}$**	**$79.3^{(21.2)}$**	**$94.8^{(0.7)}$**	**$103^{(6.3)}$**	2.7	74.5
AMO-Adam	**$23.7^{(3.1)}$**	**$51.3^{(3.4)}$**	**$95.0^{(5.1)}$**	**$80.0^{(20.6)}$**	$84.8^{(6.4)}$	$27.3^{(3.4)}$	3.0	60.4
DynMO-Adam	**$22.9^{(2.3)}$**	**$51.6^{(2.9)}$**	**$99.2^{(0.6)}$**	**$87.3^{(9.7)}$**	**$94.7^{(0.7)}$**	**$103^{(6.3)}$**	**1.8**	**76.4**
DynAMO-Adam	$14.7^{(1.9)}$	$46.2^{(0.5)}$	**$98.7^{(1.2)}$**	**$85.9^{(1.8)}$**	**$94.9^{(0.4)}$**	**$108^{(7.2)}$**	2.3	74.7
BO-qEI	**$21.8^{(0.5)}$**	**$51.5^{(0.3)}$**	**$97.6^{(0.3)}$**	**$85.4^{(1.5)}$**	$94.6^{(0.1)}$	$106^{(2.9)}$	2.5	76.2
AMO-BO-qEI	$15.4^{(0.3)}$	**$51.8^{(0.2)}$**	**$97.8^{(0.3)}$**	**$84.9^{(0.9)}$**	$85.1^{(0.4)}$	$14.3^{(1.5)}$	3.3	58.2
DynMO-BO-qEI	$20.4^{(0.4)}$	**$51.8^{(0.1)}$**	**$97.7^{(0.3)}$**	**$85.7^{(1.3)}$**	**$94.4^{(0.5)}$**	$108^{(3.2)}$	2.2	76.4
DynAMO-BO-qEI	**$21.0^{(0.5)}$**	**$51.9^{(0.2)}$**	**$97.4^{(0.4)}$**	**$85.2^{(0.9)}$**	**$94.8^{(0.1)}$**	**$126^{(14.6)}$**	**2.0**	**79.4**
BO-qUCB	**$21.6^{(0.3)}$**	**$51.7^{(0.2)}$**	**$97.9^{(0.4)}$**	**$85.3^{(1.1)}$**	$93.8^{(0.6)}$	$98.8^{(1.1)}$	**2.0**	74.8
AMO-BO-qUCB	**$21.9^{(0.4)}$**	**$51.7^{(0.3)}$**	**$97.5^{(0.4)}$**	**$85.2^{(1.0)}$**	$81.9^{(1.9)}$	$25.9^{(1.4)}$	2.7	60.7
DynMO-BO-qUCB	$20.7^{(0.4)}$	**$51.8^{(0.2)}$**	**$97.1^{(0.5)}$**	**$84.9^{(0.6)}$**	**$93.2^{(1.4)}$**	$98.0^{(1.8)}$	3.2	74.3
DynMO-BO-qUCB	**$21.4^{(0.5)}$**	**$51.7^{(0.2)}$**	**$97.1^{(0.5)}$**	**$85.3^{(1.1)}$**	**$94.7^{(0.2)}$**	**$109^{(4.5)}$**	2.2	**76.6**
CMA-ES	$16.5^{(2.1)}$	$47.8^{(1.0)}$	$96.5^{(0.7)}$	**$73.0^{(18.0)}$**	**$100^{(0.0)}$**	$100^{(0.0)}$	2.3	72.3
AMO-CMA-ES	**$24.3^{(0.9)}$**	**$53.3^{(1.4)}$**	$95.0^{(1.5)}$	**$72.5^{(23.6)}$**	$85.6^{(3.0)}$	$41.5^{(2.0)}$	3.0	62.0
DynMO-CMA-ES	$14.3^{(0.3)}$	$46.1^{(0.4)}$	**$98.2^{(1.0)}$**	**$83.3^{(1.0)}$**	**$100^{(0.0)}$**	$100^{(0.0)}$	**2.0**	73.7
DynAMO-CMA-ES	$12.9^{(0.8)}$	$48.0^{(1.6)}$	**$96.7^{(3.5)}$**	**$81.8^{(13.4)}$**	$94.5^{(0.7)}$	**$112^{(7.8)}$**	2.3	**74.3**
CoSyNE	**$24.5^{(3.5)}$**	**$49.7^{(3.1)}$**	**$98.5^{(1.6)}$**	**$86.6^{(12.7)}$**	**$93.2^{(1.0)}$**	$91.9^{(2.0)}$	**1.8**	74.1
AMO-CoSyNE	**$22.8^{(2.8)}$**	**$50.8^{(1.5)}$**	**$90.8^{(14.2)}$**	**$91.9^{(3.4)}$**	$86.0^{(3.3)}$	$29.6^{(3.6)}$	2.5	62.0
DynMO-CoSyNE	**$19.3^{(5.5)}$**	$46.3^{(2.3)}$	**$93.4^{(3.7)}$**	**$88.3^{(5.2)}$**	$85.7^{(3.2)}$	$29.6^{(3.6)}$	3.2	60.4
DynAMO-CoSyNE	**$17.8^{(5.5)}$**	**$48.4^{(2.3)}$**	**$96.7^{(3.5)}$**	**$80.2^{(12.9)}$**	**$94.5^{(0.7)}$**	**$112^{(7.8)}$**	2.5	**75.0**

Table 6.4: **Diversity of design candidates in adversarial critic feedback (AMO) and diversity in (DynMO) model-based optimization (cont.).** We evaluate our method (1) with the KL-divergence penalized-MBO objective as in (6.19) only (**Dyn**MO); (2) with the adversarial source critic-dependent constraint as introduced by Yao et al. (2024) only (AMO); and (3) with both algorithmic components as in **DynA**MO described in **Algorithm 3**. We report the L_1 coverage score achieved by the 128 evaluated designs as mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

L_1 **Coverage@128**	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank** ↓	**Opt. Gap** ↑
Dataset $\mathcal{D}$	0.42	0.31	1.42	0.68	6.26	0.58	—	—
Grad.	$0.16^{(0.10)}$	$0.20^{(0.13)}$	$0.21^{(0.10)}$	$0.42^{(0.18)}$	$0.00^{(0.00)}$	$0.00^{(0.00)}$	3.3	-1.44
AMO-Grad.	$0.17^{(0.10)}$	$0.24^{(0.13)}$	$0.25^{(0.16)}$	$0.37^{(0.11)}$	$0.00^{(0.00)}$	$0.09^{(0.14)}$	2.5	-1.42
DynMO-Grad.	$\mathbf{0.20^{(0.13)}}$	$0.22^{(0.15)}$	$\mathbf{1.23^{(0.63)}}$	$0.28^{(0.17)}$	$0.00^{(0.00)}$	$0.00^{(0.00)}$	2.8	-1.29
DynAMO-Grad.	$\mathbf{0.36^{(0.04)}}$	$\mathbf{0.52^{(0.06)}}$	$\mathbf{1.46^{(0.38)}}$	$\mathbf{2.49^{(0.06)}}$	$\mathbf{6.47^{(1.24)}}$	$\mathbf{5.85^{(1.35)}}$	**1.0**	**1.25**
Adam	$0.11^{(0.06)}$	$0.22^{(0.09)}$	$0.23^{(0.15)}$	$0.48^{(0.31)}$	$0.27^{(0.55)}$	$0.24^{(0.49)}$	2.8	-1.35
AMO-Adam	$0.14^{(0.06)}$	$0.22^{(0.10)}$	$0.35^{(0.24)}$	$0.50^{(0.34)}$	$0.26^{(0.53)}$	$0.09^{(0.12)}$	3.0	-1.35
DynMO-Adam	$\mathbf{0.20^{(0.12)}}$	$0.21^{(0.15)}$	$\mathbf{1.10^{(0.60)}}$	$0.45^{(0.29)}$	$0.27^{(0.55)}$	$0.03^{(0.00)}$	3.0	-1.23
DynAMO-Adam	$\mathbf{0.33^{(0.05)}}$	$\mathbf{0.55^{(0.03)}}$	$\mathbf{1.44^{(0.39)}}$	$\mathbf{2.40^{(0.16)}}$	$\mathbf{7.06^{(0.73)}}$	$\mathbf{6.91^{(0.71)}}$	**1.0**	**1.50**
BO-qEI	$\mathbf{0.41^{(0.02)}}$	$\mathbf{0.55^{(0.01)}}$	$2.37^{(0.03)}$	$2.11^{(0.15)}$	$7.84^{(0.01)}$	$6.61^{(0.33)}$	2.8	1.70
AMO-BO-qEI	$\mathbf{0.40^{(0.03)}}$	$\mathbf{0.55^{(0.01)}}$	$\mathbf{2.38^{(0.10)}}$	$\mathbf{2.53^{(0.05)}}$	$7.45^{(0.01)}$	$1.29^{(0.08)}$	3.7	0.82
DynMO-BO-qEI	$\mathbf{0.40^{(0.02)}}$	$\mathbf{0.55^{(0.01)}}$	$\mathbf{2.42^{(0.05)}}$	$\mathbf{2.55^{(0.03)}}$	$7.83^{(0.02)}$	$6.72^{(0.48)}$	2.3	1.80
DynAMO-BO-qEI	$\mathbf{0.42^{(0.01)}}$	$\mathbf{0.56^{(0.01)}}$	$\mathbf{2.47^{(0.03)}}$	$\mathbf{2.54^{(0.03)}}$	$\mathbf{7.87^{(0.01)}}$	$\mathbf{7.92^{(0.04)}}$	**1.2**	**2.02**
BO-qUCB	$\mathbf{0.40^{(0.02)}}$	$0.54^{(0.01)}$	$\mathbf{2.40^{(0.05)}}$	$\mathbf{2.52^{(0.07)}}$	$7.78^{(0.04)}$	$6.64^{(0.09)}$	2.8	1.77
AMO-BO-qUCB	$\mathbf{0.40^{(0.01)}}$	$\mathbf{0.56^{(0.01)}}$	$\mathbf{2.39^{(0.05)}}$	$\mathbf{2.52^{(0.04)}}$	$7.37^{(0.09)}$	$1.34^{(0.04)}$	3.3	0.82
DynMO-BO-qUCB	$\mathbf{0.39^{(0.02)}}$	$0.55^{(0.00)}$	$\mathbf{2.40^{(0.08)}}$	$\mathbf{2.52^{(0.04)}}$	$7.76^{(0.07)}$	$6.64^{(0.13)}$	2.5	1.77
DynAMO-BO-qUCB	$\mathbf{0.40^{(0.02)}}$	$\mathbf{0.55^{(0.01)}}$	$\mathbf{2.47^{(0.07)}}$	$\mathbf{2.54^{(0.05)}}$	$\mathbf{7.88^{(0.03)}}$	$\mathbf{7.80^{(0.23)}}$	**1.3**	**2.00**
CMA-ES	$\mathbf{0.33^{(0.05)}}$	$0.48^{(0.04)}$	$\mathbf{2.18^{(0.04)}}$	$1.82^{(0.12)}$	$\mathbf{3.26^{(1.42)}}$	$\mathbf{3.77^{(1.36)}}$	2.3	0.36
AMO-CMA-ES	$0.31^{(0.04)}$	$0.51^{(0.01)}$	$\mathbf{2.17^{(0.06)}}$	$1.83^{(0.15)}$	$\mathbf{3.37^{(0.49)}}$	$\mathbf{3.12^{(0.40)}}$	2.5	0.27
DynMO-CMA-ES	$\mathbf{0.34^{(0.09)}}$	$0.39^{(0.12)}$	$0.66^{(0.43)}$	$0.60^{(0.42)}$	$\mathbf{1.85^{(2.28)}}$	$\mathbf{0.94^{(1.73)}}$	3.7	-0.81
DynAMO-CMA-ES	$\mathbf{0.40^{(0.03)}}$	$\mathbf{0.56^{(0.01)}}$	$\mathbf{1.82^{(0.72)}}$	$\mathbf{2.54^{(0.05)}}$	$\mathbf{4.75^{(2.16)}}$	$\mathbf{3.29^{(1.56)}}$	**1.5**	**0.62**
CoSyNE	$0.10^{(0.07)}$	$\mathbf{0.22^{(0.10)}}$	$\mathbf{0.39^{(0.20)}}$	$\mathbf{0.27^{(0.13)}}$	$0.10^{(0.00)}$	$0.10^{(0.00)}$	2.8	-1.41
AMO-CoSyNE	$\mathbf{0.12^{(0.08)}}$	$\mathbf{0.22^{(0.15)}}$	$\mathbf{0.53^{(0.18)}}$	$\mathbf{0.14^{(0.13)}}$	$0.10^{(0.00)}$	$0.02^{(0.00)}$	2.8	-1.42
DynMO-CoSyNE	$\mathbf{0.28^{(0.09)}}$	$\mathbf{0.32^{(0.13)}}$	$\mathbf{0.33^{(0.23)}}$	$\mathbf{0.79^{(0.65)}}$	$0.10^{(0.00)}$	$0.02^{(0.00)}$	2.3	-1.30
DynAMO-CoSyNE	$\mathbf{0.21^{(0.11)}}$	$\mathbf{0.18^{(0.09)}}$	$\mathbf{0.64^{(0.42)}}$	$\mathbf{0.43^{(0.34)}}$	$\mathbf{1.85^{(0.22)}}$	$\mathbf{0.94^{(0.17)}}$	**1.8**	**-0.90**

marily $r_\theta(x)$ leads to the policy exploiting suboptimal regions of the design space—penalizing the optimization objective with a 'small amount of' the diversity objective helps the policy explore new regions of the design space that can contain more optimal designs.

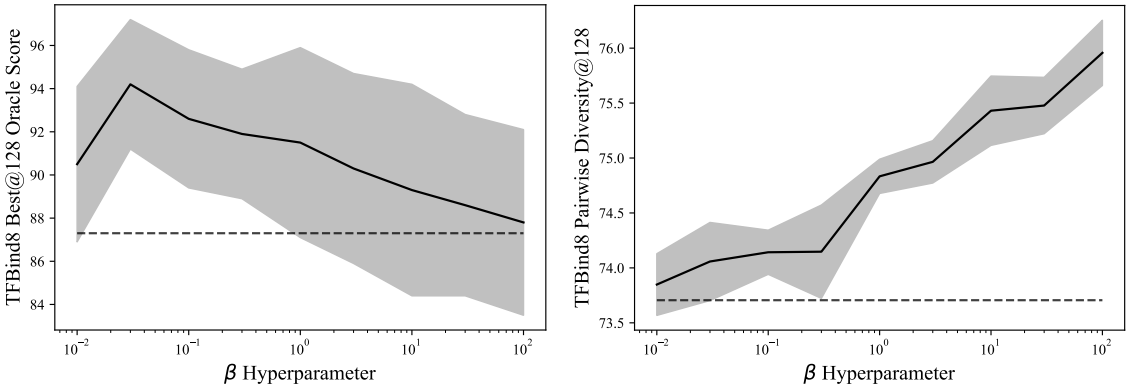

Figure 6.2: β **hyperparameter ablation.** We vary the value of the KL-divergence regularization strength hyperparameter β in **Algorithm 3** between 0.01 and 100, and report both the (**left**) Best@128 Oracle Score and (**right**) Pairwise Diversity score for 128 final design candidates proposed by a DynAMO-BO-qEI policy on the TFBind8 optimization task. We plot the mean $\pm$ 95% confidence interval over 10 random seeds. The dotted horizontal line corresponds to the $\beta = 0$ experimental mean score, which could not be plotted as a point on the logarithmic x-axis.

Separately, the experimental results for our τ ablation study are shown in **Figure 6.3**. (Note that in these experiments, we fix $\beta = \tau$ so that the ratio β/τ in **Algorithm 3** remains constant.) As the value of τ increases, the diversity of designs captured by the reference τ-weighted probability distribution decreases and approaches a (potential mixture of) Dirac delta functions with non-zero support at the optimal designs in the offline dataset. As a result, distribution matching via the KL-divergence objective no longer encourages the generative policy to find a diverse sample of designs, as the reference distribution is no longer diverse itself for $\tau \gg 1$. Similar to our β ablation study, we find that there is a unique exploration-exploitation trade-off phenomenon according to the Best@128 oracle score as a function of τ: in our particular experimental setting, we find that for $\tau \leq 1$, the Best@128 oracle score (modestly) increases, while for $\tau \geq 1$, the score decreases. For $\tau \approx 1$, we find that the generative policy is encouraged to match high-quality samples that are *diverse* enough together for the generative policy to explore new regions of the design space.

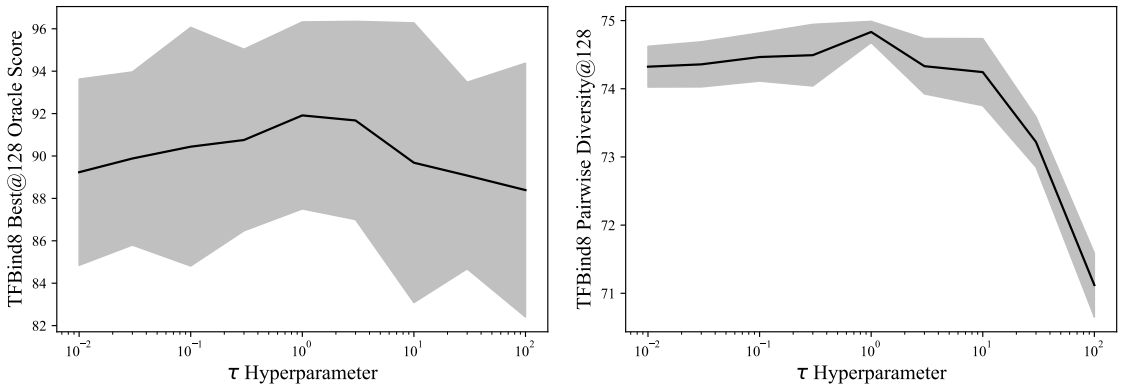

Figure 6.3: τ **temperature hyperparameter ablation.** We vary the temperature hyperparameter τ in **Algorithm 3** between 0.01 and 100, and report both the (**left**) Best@128 Oracle Score and (**right**) Pairwise Diversity score for 128 final designs proposed by a DynAMO-BO-qEI policy on the TFBind8 optimization task. We plot the mean $\pm$ 95% confidence interval over 10 random seeds.

Oracle Evaluation Budget Ablation. Recall that in our experiments, we evaluate DynAMO and baseline methods using an oracle evaluation budget of $k = 128$ samples consistent with prior work (Krishnamoorthy et al., 2023b; Yu et al., 2021; Trabucco et al., 2021; Chen et al., 2023c; Yao et al., 2024). More specifically, this means that any offline optimization method proposes exactly k design candidates that are evaluated by the hidden oracle function $r(x)$ as the final step for experimental evaluation. In **Table 6.1**, we reported both the Best@k and Pairwise Diversity@k metrics, where Best@k represents the maximum oracle score achieved by the k final design candidates; and Pairwise Diversity@k represents the pairwise diversity averaged over the k candidates.

However, in different experimental settings we might have a different evaluation budget available—larger values of k are more costly but enable us to evaluate more designs that are potentially promising, whereas smaller, more practical budgets may preclude the evaluation of optimal designs according to $r(x)$. In this section, we evaluate the performance of DynAMO as a function of the allowed evaluation budget $16 \leq k \leq 1024$. We compare DynAMO-augmented optimizers against the corresponding vanilla backbone optimization method on the TFBind8 task, and plot the Best@k and Pairwise Diversity@k metrics as a function of k in **Figure 6.4**.

As expected, the Best@k oracle score is monotonically non-decreasing as a function of k for all DynAMO-enhanced and baseline optimizers (**Fig. 6.4**). We also find that in the limit of $k \gg 1$,

the DynAMO optimizers are able to propose best designs that are more optimal than the designs by their baseline counterparts for first-order, evolutionary, and Bayesian optimization algorithms. Furthermore, DynAMO achieves a mean Best@k score non-inferior to that of the baseline method for all $k \geq 128$ across all the optimization methods evaluated on the TFBind8 task.

Separately, we find that the Pairwise Diversity of the k designs proposed by DynAMO-augmented first-order optimizers (i.e., **DynAMO-Grad.** and **DynAMO-Adam**) increases as a function of k. This makes sense, as first-order methods generally produce optimization trajectories that are simple curves in the design space as a function of the acquisition step. In contrast, we find that the Pairwise Diversity *decreases* after a certain optimizer-dependent threshold k for evolutionary and Bayesian optimization-based backbone optimizers. This is because as both classes of optimization methods do not necessarily sample repeatedly from any given region of the input space; as a result, the pairwise diversity between any two sampled points may decrease as more of the design space has been explored as a function of k. Finally, we found that leveraging DynAMO improves the Pairwise Diversity of designs compared to the baseline objective for almost all optimizers and values of k assessed. These results suggest that DynAMO helps optimization methods discover both high-quality and diverse sets of designs across a wide range of oracle evaluation budgets.

Optimization Initialization Ablation. In **Algorithm 3**, we initialize DynAMO by sampling the initial batch of $b = 64$ designs according to a pseudo-random Sobol sequence as described in **Section 6.3**. This initial batch of designs is used as the 'starting point' in our first-order optimization experiments. However, most first-order offline MBO algorithms reported in prior work (Trabucco et al., 2021; Yu et al., 2021) do not follow this same initialization schema. Instead, they perform a *top-k* initialization strategy where the top $k = b$ designs in the dataset with the highest associated reward score constitute the initial batch of designs. First-order optimization is then performed on these initial top-k designs. However, it is possible that for many MBO problems, these top-k initial designs constitute only a small 'area' of the overall search space, resulting in a lower diversity of final designs when compared to Sobol sequence initialization.

To interrogate whether the gains in diversity of designs obtained with DynAMO are due to our

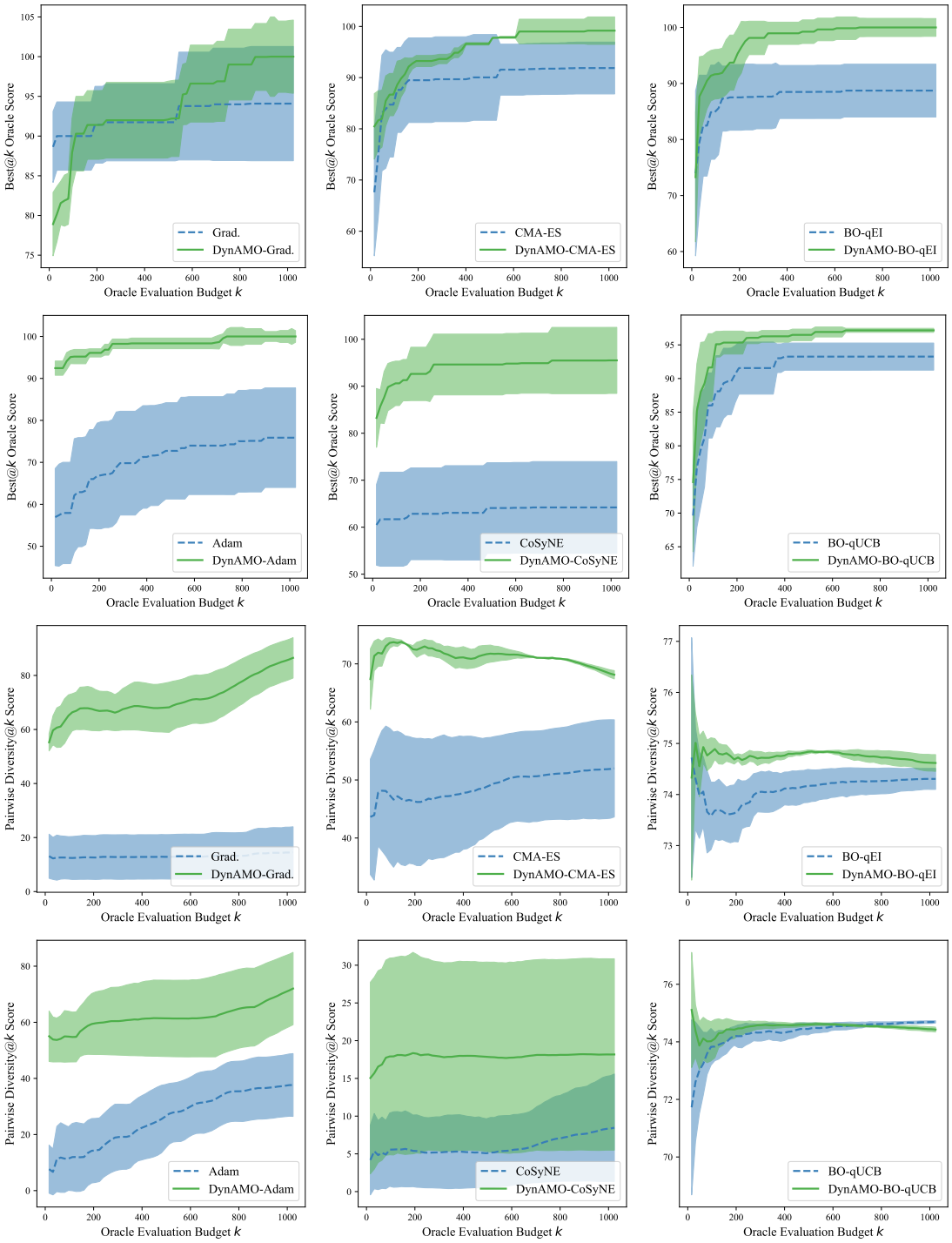

Figure 6.4: **Oracle evaluation budget ablation.** We vary the allowed oracle evaluation budget k in **Algorithm 3** between 16 and 1024, and report both the (**first two rows**) Best@128 Oracle Score and (**last two rows**) Pairwise Diversity score for k final designs proposed by both DynAMO-augmented and base optimizers on the TFBind8 task. We plot the mean $\pm$ 95% confidence interval over 10 random seeds.

Sobol sequence-based initialization strategy, we evaluated Gradient Ascent, COMs, RoMA, ROMO, GAMBO with Gradient Ascent, and DynAMO with Gradient Ascent using both Sobol sequence-based and top-k-based initialization strategies. All algorithms were initialized using $k = b = 64$ samples and used Gradient Ascent as the backbone optimizer (except for RoMA (Yu et al., 2021), which used Adam Ascent as in the original method proposed by the authors).

Our results are shown in **Table 6.5**. Empirically, we found that the relative performance of Sobol sequence-initialized and Top-k-initialized optimizer largely depends on the specific algorithm; for example, COMs and RoMA strongly benefit from using Top-k initialization in obtaining high-quality designs. This makes sense, as the original authors for both methods use Top-k initialization for all their experiments. In contrast, the quality of designs proposed by GAMBO and DynAMO is better with Sobol sequence initialization.

While DynAMO using Sobol sequence initialization does indeed outperform the Top-k-initialized counterpart across all tasks, both initialization strategies consistently propose batches of designs with competitive pairwise diversity scores when compared to other first-order optimization algorithms. This suggests that DynAMO is able to provide a significant advantage in proposing diverse designs that extend beyond the choice of initialization strategy alone. Separately for the other first-order optimization methods assessed, there is no clear advantage in obtaining diverse designs when using Sobol sequence initialization according to the pairwise diversity metric across all tasks. Our results suggest that DynAMO is able to propose both high-quality and diverse sets of designs with performance exceeding what is possible with using a Sobol initialization alone.

6.5.3. Theoretical Guarantees

We seek to place an upper bound on the difference between *true* diversity-penalized objective

$$J^\star(\pi) := \mathbb{E}_{x\sim q^\pi(x)}[r(x)] - \frac{\beta}{\tau} D_{\text{KL}}(q^\pi(x)||p^\tau(x)) \tag{6.45}$$

realized by the final generative policy $\hat{\pi} \in \Pi$ learned by DynAMO, and the true diversity-penalized objective realized by the true optimal policy $\pi^\star := \arg\max_{\pi\in\Pi} J^\star(\pi)$. Note that this objective

Table 6.5: **Optimization initialization ablation.** We evaluate both Sobol sequence-based and Top-k initialization strategies for DynAMO with Grad. Ascent and other first-order MBO methods. We report the maximum oracle score (resp., pairwise diversity score) achieved out of 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. $\max(\mathcal{D})$ reports the top oracle score in the offline dataset. All metrics are multiplied by 100 for easier legibility. **Bolded** entries indicate the higher average scores for a given optimization method.

Best@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Win Rate ↑**
Dataset $\mathcal{D}$	43.9	59.4	60.5	88.9	40.0	88.4	—
Grad. (Sobol)	**90.0**$^{(4.3)}$	**80.9**$^{(12.1)}$	60.2$^{(8.9)}$	88.8$^{(4.0)}$	**36.0**$^{(6.8)}$	65.6$^{(14.5)}$	3/6
Grad. (Top-k)	85.1$^{(3.1)}$	64.0$^{(0.7)}$	**63.3**$^{(0.0)}$	**90.1**$^{(0.3)}$	27.1$^{(1.0)}$	**67.8**$^{(0.0)}$	3/6
COMs (Sobol)	84.7$^{(5.3)}$	60.4$^{(2.2)}$	63.3$^{(0.0)}$	91.4$^{(0.4)}$	17.3$^{(0.5)}$	82.8$^{(2.9)}$	0/6
COMs (Top-k)	**93.1**$^{(3.4)}$	**67.0**$^{(0.9)}$	**64.6**$^{(1.0)}$	**97.1**$^{(1.6)}$	**41.2**$^{(4.8)}$	**91.8**$^{(0.9)}$	6/6
RoMA (Sobol)	**96.5**$^{(0.0)}$	**77.8**$^{(0.0)}$	**63.3**$^{(0.0)}$	**85.5**$^{(2.4)}$	46.5$^{(2.5)}$	93.9$^{(1.0)}$	4/6
RoMA (Top-k)	**96.5**$^{(0.0)}$	**77.8**$^{(0.0)}$	**63.3**$^{(0.0)}$	**84.7**$^{(0.0)}$	**49.8**$^{(1.4)}$	**95.7**$^{(1.6)}$	6/6
ROMO (Sobol)	97.7$^{(1.2)}$	**67.0**$^{(1.3)}$	**68.3**$^{(0.5)}$	90.8$^{(0.4)}$	**45.5**$^{(1.6)}$	86.1$^{(0.5)}$	3/6
ROMO (Top-k)	**98.1**$^{(0.7)}$	66.8$^{(1.0)}$	63.0$^{(0.8)}$	**91.8**$^{(0.9)}$	38.7$^{(2.5)}$	**87.8**$^{(0.9)}$	3/6
GAMBO (Sobol)	73.1$^{(12.8)}$	**77.1**$^{(9.6)}$	**64.4**$^{(1.5)}$	**92.8**$^{(8.0)}$	**46.0**$^{(6.8)}$	**90.6**$^{(14.5)}$	5/6
GAMBO (Top-k)	**78.5**$^{(9.3)}$	68.3$^{(0.5)}$	63.0$^{(0.0)}$	90.6$^{(0.3)}$	27.1$^{(1.0)}$	77.8$^{(0.0)}$	1/6
DynAMO (Sobol)	**90.3**$^{(4.7)}$	**86.2**$^{(0.0)}$	**64.4**$^{(2.5)}$	**91.2**$^{(0.0)}$	**44.2**$^{(7.8)}$	**89.8**$^{(3.2)}$	6/6
DynAMO (Top-k)	81.9$^{(8.4)}$	64.4$^{(1.2)}$	63.3$^{(0.0)}$	90.8$^{(0.3)}$	29.4$^{(4.4)}$	75.3$^{(11.6)}$	0/6

Pairwise Diversity@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Win Rate ↑**
Dataset $\mathcal{D}$	33.7	42.8	50.9	87.6	6.7	77.8	—
Grad. (Sobol)	**12.5**$^{(8.0)}$	7.8$^{(8.8)}$	7.9$^{(7.8)}$	24.1$^{(13.3)}$	**0.0**$^{(0.0)}$	**0.0**$^{(0.0)}$	3/6
Grad. (Top-k)	8.3$^{(4.7)}$	**40.3**$^{(3.6)}$	**63.1**$^{(8.3)}$	**28.4**$^{(6.0)}$	**0.0**$^{(0.0)}$	**0.0**$^{(0.0)}$	5/6
COMs (Sobol)	65.4$^{(0.5)}$	57.3$^{(0.1)}$	59.3$^{(1.1)}$	**72.6**$^{(0.7)}$	43.9$^{(16.5)}$	**33.8**$^{(1.7)}$	2/6
COMs (Top-k)	**66.6**$^{(1.0)}$	**57.4**$^{(0.2)}$	**81.6**$^{(4.9)}$	3.8$^{(0.9)}$	**99.5**$^{(25.8)}$	21.1$^{(23.5)}$	4/6
RoMA (Sobol)	21.0$^{(0.2)}$	**3.8**$^{(0.0)}$	**5.9**$^{(0.0)}$	**1.8**$^{(0.0)}$	**70.3**$^{(13.6)}$	8.1$^{(0.3}$	4/6
RoMA (Top-k)	**21.3**$^{(0.3)}$	**3.8**$^{(0.0)}$	**5.9**$^{(0.2)}$	**1.8**$^{(0.0)}$	49.4$^{(6.1)}$	**14.8**$^{(0.6)}$	5/6
ROMO (Sobol)	**64.4**$^{(1.3)}$	56.9$^{(0.2)}$	**59.3**$^{(0.9)}$	39.0$^{(0.9)}$	**58.3**$^{(12.9)}$	10.9$^{(0.5)}$	3/6
ROMO (Top-k)	62.1$^{(0.8)}$	**57.1**$^{(0.1)}$	53.9$^{(0.6)}$	**48.7**$^{(0.1)}$	51.7$^{(31.7)}$	**22.1**$^{(5.5)}$	3/6
GAMBO (Sobol)	15.1$^{(11.2)}$	10.3$^{(11.5)}$	12.1$^{(11.3)}$	19.6$^{(15.2)}$	**0.3**$^{(0.8)}$	**2.6**$^{(3.9)}$	2/6
GAMBO (Top-k)	**59.2**$^{(5.0)}$	**54.1**$^{(2.5)}$	**79.3**$^{3.4)}$	**33.4**$^{(1.9)}$	0.0$^{(0.0)}$	0.0$^{(0.0)}$	4/6
DynAMO (Sobol)	**66.9**$^{(6.9)}$	**68.2**$^{(10.8)}$	**77.2**$^{(21.5)}$	**93.0**$^{(1.2)}$	**129**$^{(55.3)}$	**104**$^{(56.1)}$	6/6
DynAMO (Top-k)	55.2$^{(10.5)}$	46.4$^{(5.2)}$	76.8$^{(4.5)}$	36.4$^{(3.1)}$	120$^{(30.0)}$	85.7$^{(50.0)}$	0/6

$J^\star(\pi)$ is *not* equivalent to the offline MBO objective $J(\pi)$ introduced in (6.13); importantly, the objective $J(\pi)$ is a function of the *true*, hidden oracle reward $r(x)$ as opposed to the forward surrogate model $r_\theta(x)$. Furthermore, the KL-divergence penalty is computed with respect to the *true* τ-weighted probability distribution $p^\tau(x)$, as opposed to its empirical estimate computed from the offline dataset $\mathcal{D}$ as in **Definition 4**. In principle, (6.45) captures the true trade-off between diversity and quality of designs that we hope to achieve by the theoretically optimal zero-regret generative policy $\pi^\star$ that maximizes (6.45) over the space of policies Π.

Our main result is in **Theorem 6** below, although we first step through the relevant assumptions and intermediate results necessary to arrive at (6.45). Firstly, we assume the following:

Assumption 1 (Surrogate Model Error Bound)**.** *There exists a finite $\varepsilon_0^2 \in \mathbb{R}_+$ such that*

$$\mathbb{E}_{x\sim p^\tau(x)}[r(x) - r_\theta(x)]^2 \leq \varepsilon_0^2/4 \tag{6.46}$$

for any choice in $\tau \geq 0$, where $p^\tau(x)$ is the true τ-weighted probability distribution over $\mathcal{X}$.

Assumption 2 (Policy Realizability)**.** *Both the true optimal sampling policy $\pi^\star$ according to (6.45) and optimal sampling policy $\hat{\pi}$ according to (6.15) are contained in the (finite) policy class Π.*

Assumption 3 (Bounded Importance Weights)**.** *Define the importance weight $w(x)$ as the ratio between probability distributions $q^\pi(x)$ and $p(x)$. There exists a finite $M \in \mathbb{R}_+$ such that for all possible permutations of $\pi \in \{\hat{\pi}, \pi^\star\}$ and $p(x) \in \{p^\tau(x), p^\tau_{\mathcal{D}}(x)\}$, we have $w(x) := q^\pi(x)/p(x) \leq M$ for all $x \in \mathcal{X}$.*

Under these assumptions, we first place a bound on the error of the forward surrogate model over the distribution of generated designs from the optimal policies according to both the offline objective $J(\pi)$ and true objective $J^\star(\pi)$:

Lemma 7 (Bounded Prediction Error)**.** *Assume there exists an $M \in \mathbb{R}_+$ finite satisfying **Assumption 3**. Then with probability at least $1 - \delta$ we have (for any $\delta > 0$ and for both $\pi = \pi^\star$ and $\pi = \hat{\pi}$)*

$$\mathbb{E}_{x\sim q^\pi(x)} |r(x) - r_\theta(x)| \leq \frac{\varepsilon_0}{2} + M\sqrt{\frac{2\log(2|\Pi|/\delta)}{n}} \tag{6.47}$$

where $n := |\mathcal{D}|$ is the number of datums in the offline dataset $\mathcal{D}$.

Proof. Under **Assumption 1**, Jensen's inequality gives us

$$\mathbb{E}_{x\sim p^\tau(x)} |r(x) - r_\theta(x)| \leq \sqrt{\mathbb{E}_{x\sim p^\tau(x)}[r(x) - r_\theta(x)]^2} \leq \sqrt{\frac{\varepsilon_0^2}{4}} =: \frac{\varepsilon_0}{2} \tag{6.48}$$

Furthermore, **Assumption 3** and Cortes et al. (2010) yield

$$\begin{aligned} &\Big|\mathbb{E}_{x\sim p^\tau(x)} |r(x) - r_\theta(x)| - \mathbb{E}_{x\sim q^\pi(x)} |r(x) - r_\theta(x)| \Big| \\ &\qquad = \left|\mathbb{E}_{x\sim p^\tau(x)} |r(x) - r_\theta(x)| - \mathbb{E}_{x\sim p^\tau_{\mathcal{D}}(x)} \left[\frac{q^\pi(x)}{p^\tau_{\mathcal{D}}(x)} \Big| r(x) - r_\theta(x)\Big|\right]\right| \\ &\qquad \leq M\sqrt{\frac{2\log(2|\Pi|/\delta)}{n}} \end{aligned} \tag{6.49}$$

with probability at least $1-\delta$. In the offline setting (as in our work) and assuming that the forward surrogate model $r_\theta(x)$ has been well-trained according to (5.1) or a similar learning paradigm (e.g., see Trabucco et al. (2021); Yu et al. (2021)), we can reasonably assume that $\varepsilon_0^2 \leq \mathbb{E}_{x\sim q^\pi(x)}[r(x) - r_\theta(x)]^2$. We therefore have an upper bound on the prediction error of the forward surrogate model over the distribution $q^\pi(x)$ over generated designs:

$$\begin{aligned} \mathbb{E}_{x\sim q^\pi(x)} |r(x) - r_\theta(x)| &\leq \mathbb{E}_{x\sim p^\tau(x)} |r(x) - r_\theta(x)| \\ &\quad + M\sqrt{\frac{2\log(2|\Pi|/\delta)}{n}} \leq \frac{\varepsilon_0}{2} + M\sqrt{\frac{2\log(2|\Pi|/\delta)}{n}} \end{aligned} \tag{6.50}$$

with probability at least $1-\delta$. □

Under **Assumption 3**, we can also place an upper bound on the true and realized KL-divergence penalties:

Lemma 8 (Bounded KL-Divergence)**.** *Assume there exists an $M \in \mathbb{R}_+$ finite satisfying **Assumption 3**. Then with probability at least $1-\delta$ we have (for any $\delta > 0$ and for both $\pi = \pi^\star$ and $\pi = \hat{\pi}$)*

$$|D_{KL}(q^\pi(x)||p^\tau(x)) - D_{KL}(q^\pi(x)||p^\tau_{\mathcal{D}}(x))| \leq M\sqrt{\log(|\Pi|/\delta)} \tag{6.51}$$

Proof. According to the definition of M,

$$D_{\mathrm{KL}}(q^\pi(x)||p^\tau(x)) = \mathbb{E}_{x\sim p^\tau(x)}\left[\frac{q^\pi(x)}{p^\tau(x)}\log\left(\frac{q^\pi(x)}{p^\tau(x)}\right)\right] \leq M\log M \tag{6.52}$$

From Hoeffding's inequality (Hoeffding, 1963),

$$\mathbb{P}\left(|D_{\mathrm{KL}}(q^\pi(x)||p^\tau(x)) - D_{\mathrm{KL}}(q^\pi(x)||p^\tau_{\mathcal{D}}(x))| \geq \varepsilon\right) \leq |\Pi|\cdot\exp\left(-\frac{2\varepsilon^2}{M\log M}\right) \tag{6.53}$$

for any $\varepsilon > 0$. We can choose to define $\varepsilon := \sqrt{(M\log M)\cdot\log(|\Pi|/\delta)/2}$ such that

$$\begin{aligned}&|D_{\mathrm{KL}}(q^\pi(x)||p^\tau(x)) - D_{\mathrm{KL}}(q^\pi(x)||p^\tau_{\mathcal{D}}(x))| \\ &\quad\leq \frac{\sqrt{(M\log M)\cdot\log(|\Pi|/\delta)}}{\sqrt{2}} \leq \frac{M\sqrt{\log(|\Pi|/\delta)}}{\sqrt{2}} \leq M\sqrt{\log(|\Pi|/\delta)}\end{aligned} \tag{6.54}$$

with probability at least $1-\delta$. □

We are now ready to prove our main result:

Theorem 6 (Bounded Diversity-Penalized Objective $J^\star(\pi)$)**.** *Assume that there exists an $M \in \mathbb{R}_+$ finite satisfying **Assumption 3**. Then with probability at least $1-\delta$, we have (for any $\delta > 0$)*

$$J^\star(\pi^\star) - J^\star(\hat{\pi}) \leq \varepsilon_0 + 2M\left(\frac{2}{\sqrt{n}} + \frac{\beta}{\tau}\right)\sqrt{\log\left(\frac{8|\Pi|}{\delta}\right)} \tag{6.55}$$

where $n := |\mathcal{D}|$ is the size of the offline dataset $\mathcal{D}$.

Proof. Firstly, we combine **Lemmas 7** and **8** using the triangle inequality to bound the difference between the true reward $J^\star(\hat{\pi})$ and the offline reward $J(\hat{\pi})$, where $\hat{\pi} \in \Pi$ maximizes $J(\pi)$ as defined

in (6.13).

$$
\begin{aligned}
J(\hat{\pi}) - J^{\star}(\hat{\pi}) &:= \left(\mathbb{E}_{x\sim q^{\hat{\pi}}(x)}[r_\theta(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^{\hat{\pi}}(x)||p^{\tau}_{\mathcal{D}})\right) - \left(\mathbb{E}_{x\sim q^{\hat{\pi}}(x)}[r(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^{\hat{\pi}}(x)||p^{\tau}(x))\right) \\
&\leq \mathbb{E}_{x\sim q^{\hat{\pi}}(x)}\left|r(x) - r_\theta(x)\right| + \frac{\beta}{\tau}\left|D_{\mathrm{KL}}(q^{\hat{\pi}}(x)||p^{\tau}(x)) - D_{\mathrm{KL}}(q^{\hat{\pi}}(x)||p^{\tau}(x))\right| \\
&\leq \left(\frac{\varepsilon_0}{2} + M\sqrt{\frac{2\log(8|\Pi|/\delta)}{n}}\right) + M\cdot\frac{\beta}{\tau}\sqrt{\log(4|\Pi|/\delta)} \\
&\leq \frac{\varepsilon_0}{2} + M\left(\frac{2}{\sqrt{n}} + \frac{\beta}{\tau}\right)\sqrt{\log\left(\frac{8|\Pi|}{\delta}\right)}
\end{aligned}
\tag{6.56}
$$

with probability $1-(\delta/2)$. Because $\hat{\pi} := \mathrm{argmax}_{\pi\in\Pi} J(\pi)$, we must have $J(\pi^\star) \leq J(\hat{\pi})$. Substituting this into the left hand side of (6.56) gives

$$
J(\pi^\star) - J^\star(\hat{\pi}) \leq \frac{\varepsilon_0}{2} + M\left(\frac{2}{\sqrt{n}} + \frac{\beta}{\tau}\right)\sqrt{\log\left(\frac{8|\Pi|}{\delta}\right)} \tag{6.57}
$$

Separately, we have (with probability $1-(\delta/2)$)

$$
\begin{aligned}
J^\star(\pi^\star) - J(\pi^\star) &:= \left(\mathbb{E}_{x\sim q^{\pi^\star}(x)}[r(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^{\pi^\star}(x)||p^{\tau}(x))\right) \\
&\quad - \left(\mathbb{E}_{x\sim q^{\pi^\star}}[r_\theta(x)] - \frac{\beta}{\tau} D_{\mathrm{KL}}(q^{\pi^\star}(x)||p^{\tau}_{\mathcal{D}}(x))\right) \\
&\leq \mathbb{E}_{x\sim q^{\pi^\star}(x)}\left|r(x) - r_\theta(x)\right| + \frac{\beta}{\tau}\left|D_{\mathrm{KL}}(q^{\pi^\star}(x)||p^{\tau}(x)) - D_{\mathrm{KL}}(q^{\pi^\star}(x)||p^{\tau}_{\mathcal{D}}(x))\right| \\
&\leq \left(\frac{\varepsilon_0}{2} + M\sqrt{\frac{2\log(8|\Pi|/\delta)}{n}}\right) + M\cdot\frac{\beta}{\tau}\sqrt{\log(|4\Pi|/\delta)} \\
&\leq \frac{\varepsilon_0}{2} + M\left(\frac{2}{\sqrt{n}} + \frac{\beta}{\tau}\right)\sqrt{\log\left(\frac{8|\Pi|}{\delta}\right)}
\end{aligned}
\tag{6.58}
$$

following the derivation in (6.56) except for $\pi^\star$ (as opposed to $\hat{\pi}$) that maximizes $J^\star(\pi)$ (as opposed to $J(\pi)$). Summing (6.57) and (6.58) gives

$$
J^\star(\pi^\star) - J^\star(\hat{\pi}) \leq \varepsilon_0 + 2M\left(\frac{2}{\sqrt{n}} + \frac{\beta}{\tau}\right)\sqrt{\log\left(\frac{8|\Pi|}{\delta}\right)} \tag{6.59}
$$

with probability $1-\delta$. □

We remark that we only prove **Theorem 6** in the unconstrained optimization setting; in principle, a tighter bound could exist in the adversarially constrained formulation introduced in (6.15), as a bound on the 1-Wasserstein distance between $q^{\hat{\pi}}(x)$ and $p^{\tau}_{\mathcal{D}}(x)$ will almost surely place a favorably *tighter* bound on the forward surrogate model prediction error than **Lemma 7**.

6.6. Discussion and Conclusion

We introduce **DynAMO**, a novel task- and optimizer- agnostic approach to MBO that improves the diversity of proposed designs in offline optimization tasks. By framing diversity as a distribution-matching problem, we show how DynAMO can enable generative policies to sample both high-quality and diverse sets of designs. Our experiments reveal that DynAMO significantly improves the diversity of proposed designs while also discovering high-quality candidates.

Limitations and Future Work. There are also important limitations of our method. Firstly, we note that while DynAMO can significantly improve the diversity of proposed designs in offline MBO while preserving Best@128 performance, our method is not as competitive with existing baselines according to the *median* score obtained by the 128 designs (**Supp. Fig. D.2**). The suboptimal performance of DynAMO according to this Median@128 metric is unsurprising given that our primary motivation of DynAMO is to obtain a diverse sample of designs while simultaneously ensuring that a nonzero subset of them are (near-) optimal. Furthermore, we empirically observe that no method is state-of-the-art on both Best@128 and Median@128 metrics. While it would be ideal for DynAMO (or any method) to be state-of-the-art for all quality and diversity metrics, we argue that obtaining a good Best@128 score is more important than a good Median@128 score, as the principle real-world goal of offline MBO is to find *a* design that maximizes the oracle function.

Secondly, we also limit our study of DynAMO to offline MBO tasks that are well-described and studied in prior work. In principle, real-world optimization problems may be complicated by noisy and/or sparse objective functions, ultra-high dimensional search spaces, small offline datasets, and other practical limitations. We leave a more rigorous interrogation of how such offline MBO methods perform in such settings for future work.

CHAPTER 7

CONCLUSION

This dissertation has examined the design of distributionally robust machine learning systems that operate reliably in real-world biomedical and scientific settings. Motivated by the potential of machine learning to be applied in high-stakes domains—such as clinical medicine and scientific discovery—this work has sought to address a central and pressing challenge: while modern learning systems demonstrate exceptional performance under idealized conditions, they often fail when deployed in the real world, where they are faced with environments that differ meaningfully from those represented in their training data. In **Chapter 2**, we introduced the theoretical underpinnings of distribution shift 'in-the-world,' although a key takeaway from this dissertation is that such failures are not merely theoretical curiosities: poor generalization of naïve ML systems is deeply consequential and can compromise safety, equity, and scientific validity. The results presented in this dissertation collectively advocate for more robust generalization under distribution shift. This dissertation was centered around two key hypotheses to achieve this goal:

1. Interpretable-by-design ML systems enable human-like compositionality in predictions, enabling better out-of-distribution generalization.

2. Adversarial source critic models can help us implement meaningful and computationally tractable bounds on the 1-Wasserstein distance, and therefore the empirical test risk.

The first part of this dissertation explored the hypothesis that interpretability can be leveraged as a means of achieving generalizability. Through the introduction of models that are interpretable-by-design, whose internal representations are constrained to be semantically meaningful and aligned with concepts that can be understood and verified by human experts, we showed that we can improve the out-of-distribution performance of ML systems. In **Chapter 3**, we showed how interpretable concepts derived from evidence-based medical guidelines can enable generalist language models to better generalize to domain-specific tasks, such as assisting with medical image ordering

in acute patient care. Separately in **Chapter 4**, we demonstrated how to integrate clinical knowledge from clinical experts to define the internal representation of multimodal patient data. In both settings, interpretable-by-design systems not only performed competitively with black-box counterparts, but also enabled direct alignment with medical knowledge for improved generalizability.

Our second line of inquiry considered scenarios in which full control over model architecture or training data is not possible—a common constraint in real-world deployments. In such cases, machine learning systems can be thought of as black-box predictors, and the challenge becomes one of constraining their post-training behavior to input regimes where their predictions are reliable. To address this, I developed a novel framework based on adversarial supervision to regularize the generalization behavior of black-box models used in offline optimization problems in **Chapter 5**. By leveraging source critic models trained to discriminate between in-distribution and out-of-distribution inputs, this computationally tractable approach enabled us to better solve generative design problems across a wide range of scientific domains. In **Chapter 6**, we then extended this method to accommodate the problem of diversity in offline generative design, where our objective is not only to propose optimal candidates, but also to generate a diverse set of high-quality solutions. This is particularly relevant in scientific discovery, where greater coverage of many (near-) optimal designs can enable better secondary downstream exploration (**Supp. Table D.10**).

Taken together, the contributions of this dissertation highlight a broader methodological insight: learning systems that incorporate structured priors—whether in the form of domain knowledge, semantically constrained representations, or adversarial model feedback—exhibit improved reliability and adaptability when compared to purely data-driven models. Ultimately, I hope that the methods proposed in this dissertation enable us to work towards deploying robust, verifiable, and generalizable machine learning systems in mission-critical settings. Moving forward, the methods presented here naturally open several promising research directions to explore in future work:

Human-AI Collaboration. A key qualitative observation made in the work described in **Chapters 3-4** and in **Appendix B** is how human experts can be affected by AI prediction models in applications to patient care. Preliminary findings reported by Yao et al. (2025a); Wu et al. (2025) sug-

gest that interpretable ML tools can improve the diagnostic performance of clinicians in simulated hospital environments, corroborating a growing body of evidence that human workflows can be effectively augmented with interpretable and generalizable AI tools (Nori et al., 2025; Korom et al., 2025; Perivolaris et al., 2024). However, Dell'Acqua et al. (2023); Dell'Acqua (2022) previously reported that human-AI interactions can be challenging to characterize properly, and humans are subject to over-reliance on AI recommendations that are plausible but non-verifiable, or even grossly incorrect (Agarwal et al., 2024a). Furthermore, Bastani et al. (2025) reported that generative AI tools can have detrimental impacts on education without appropriate usage guardrails in place. Given these challenges alongside the growing adoption of ML tools in high-stakes pipelines, the challenge of accurately characterizing human-AI interactions remains paramount and an important direction for future work. The algorithms proposed in this dissertation were designed to align with human reasoning, but actual deployment scenarios will often involve human users in the loop—whether it be clinicians, scientists, or engineers—whose individual, group-wise, and/or collective beliefs may evolve over time. Designing generalizable ML systems that can communicate uncertainty, justify their reasoning, and adapt to human feedback remains an open and urgent challenge.

Active Offline-Online Learning. In **Chapters 5-6**, we considered optimization problems in the offline setting, where the oracle objective function is assumed to be completely inaccessible during optimization. However, such a strict assumption may not necessarily be the case—for instance, we may be able to empirically test a small number of candidate designs in the laboratory before a subsequent round of optimization, or may even be able to run more computationally expensive simulations to arrive at better approximations of the true fitness of a proposed design (Swanson et al., 2025; Ghareeb et al., 2025; Narayanan et al., 2024; Laurent et al., 2024). Such experimental settings are common in **active learning** (Hernández-García et al., 2024; Li et al., 2022b,a; Wu et al., 2023; Palizhati et al., 2022; Li et al., 2024), where the challenge is not only sampling which designs to evaluate, but also *when* to sample and how to best allocate the limited oracle objective query budget. In these settings, more complex problem formulations—such as multi-fidelity optimization (Hernández-García et al., 2024; Li et al., 2022b), Markov decision processes (Fang et al., 2017), and

multi-armed bandits(Ganti and Gray, 2013)—may admit solutions that better generalize to this setting. Future work may consider adapting the methodology introduced in this dissertation to such temporally extended and partially observed environments.

Incorporating Prior Knowledge in Optimization. **Chapters 3-4** highlight the utility of prior knowledge in learning better predictive machine learning models. Given the problem formulations described in **Chapters 5-6**, a natural question is whether prior knowledge could improve the performance of backbone *optimization* methods in the offline setting. More explicitly, recent work has shown that large language models can act as effective black-box optimizers through iterative prompting (Yang et al., 2024a; Ma et al., 2024; Guo et al., 2024; Qiu et al., 2025; Hong et al., 2025). Given that LLMs also have the impressive ability to synthesize prior domain knowledge in the form of unstructured text, future work might explore how LLMs may be used to better guide offline optimization through contextualizing relevant domain knowledge (Liu et al., 2024).

Safety and Robustness of Digital Twins. Digital twins are virtual representations of a physical system or process that can be used to simulate how real-world interventions might affect the modeled system Gupta et al. (2024). In healthcare and biology, digital twins are computational models of a specific patient or biological system that integrate clinical, physiological, and behavioral data to mirror the state of the patient and how they may evolve over time (Wu and Koelzer, 2024; Barber et al., 2022; Laubenbacher et al., 2022). Recent work has explored how digital twins can be used for counterfactual treatment effect estimation (Holt et al., 2024; Das et al., 2023; Qian et al., 2021), prediction of immune system perturbations (Laubenbacher et al., 2022), clinical trial simulations (Wang et al., 2024b; Das et al., 2023), and pharmacokinetic modeling (Mujahid et al., 2024; Visentin et al., 2014). Such digital twins may therefore play a role in *in silico* biomedical optimization experiments analogous to those discussed in **Chapters 5-6**.

However, even state-of-the-art digital twins are rarely perfect models of the underlying physical system: historical patient observations and therapeutic interventions represent only a small fraction of the space of possible input perturbations to a digital twin. Furthermore, certain patient populations can be systemically underrepresented in clinical datasets, making it challenging to

build accurate and robust models of minority-specific pathophysiologic processes. As a result, it is possible (and frequently the case) that digital twins may fail to generalize in clinical use cases. For example, Zhu et al. (2025) and Guan et al. (2025) reported that covariate shifts across hospitals can cause patient models to generalize poorly in out-of-distribution clinical environments. Digital twins trained on pre-pandemic patient data often saw their performance collapse during the COVID-19 pandemic (Kagerbauer et al., 2024). Such miscalibration can misguide clinical decisions and may benefit from designing better digital twins that are aligned with clinical knowledge, and methods to constrain when and where digital twin-based predictions can be trusted.

In conclusion, this dissertation has argued for and demonstrated the value of machine learning approaches that prioritize robustness, interpretability, and distributional alignment in high-stakes applications. The methods and insights presented herein are offered as a step toward more trustworthy and actionable machine intelligence—one that is not only effective under ideal conditions but remains reliable, comprehensible, and adaptable in the complexity of the real world.

APPENDIX A

Clinical Decision Support via Generalist Language Models: Additional Experimental Results

The following appendix contains additional experimental results and discussion for the interested reader related to the work titled "Evaluating Image Ordering for Acute Patient Presentations via Language Model Alignment with the ACR Appropriateness Criteria."

A.1. Prospective Clinician-AI Study: Additional Discussion and Results

In this section, we offer additional experimental results for our prospective study with U.S. medical students and emergency medicine resident physicians. The main text of our work describes the results of our **Timed** experimental arm, where participants were required to complete the study at an average rate of no slower than 1 question per minute. We also ran a separate, **Untimed** experimental arm, where no constraints were imposed on the rate of completion so long as the study was completed in one sitting. Participants were randomized in exactly one of the two experimental arms: of the 30 participants who completed the study, 16 (14) were assigned to the Timed (Untimed) arm. On average, participants in the Timed experimental arm completed the study in 36.74 minutes (95% CI: $[29.88 - 43.60]$), while participants in the Untimed experimental arm completed the study in 46.80 minutes (95% CI: $[33.90 - 59.70]$). The overall accuracy of the Timed arm participants was 20.4% (95% CI: $[17.6\% - 23.2\%]$), the accuracy of the Untimed arm participants was 20.9% (95% CI: $[17.8\% - -23.9\%]$). The performance of language model on the study tasks was not made available to the study participants.

In **Supp. Table A.7** and in the main text, we describe a statistically significant improvement in the accuracy of ordered diagnostic imaging studies by study participants when LLM-generated guidance is offered in the Timed experimental arm. Interestingly, we found that the *inverse* is true in the Untimed arm: participant accuracy *decreased* with statistical significance when LLM-generated guidance was available ($\beta_1 = -0.089$; 95% CI: [-0.170 - -0.009]; $p = 0.032$). We hypothesize that this may be because participants have more time to carefully think through cases and

consult external resources in the Untimed arm. The absence of the "pressure" imposed by a time limit may also have psychological impacts in clinical decision making that are outside the scope of this work. Regardless, ostensibly paradoxical experimental findings—such as the results in the Untimed study arm—have been previously reported in related work studying human-computer interaction with generative AI systems; for example, Dell'Acqua (2022); Dell'Acqua et al. (2023) describe how AI systems can adversely impact expert performance on specialized tasks under certain conditions, and Bastani et al. (2025) characterize how generative AI tools deter student learning if key guardrails are not properly implemented. We believe these results emphasize the importance of carefully studying how different clinical workflows are affected by generative AI tools.

Supp. Tables A.8-A.10 describe the effect of LLM-generated recommendations on (1) LLM agreement; (2) false positive rate; and (3) false negative rate. For both the Timed and Untimed experimental arms, we observed that LLM-generated recommendations increased LLM agreement and did not affect the false positive or false negative rates of image ordering. Interestingly, a self-reported positive sentiment regarding AI in medicine was associated with *lower* LLM agreement scores in both the Timed ($\beta_3 = -0.219$; 95% CI: [-0.353 - -0.084]; $p = 0.004$) and Untimed ($\beta_3 = -0.086$; 95% CI: [-0.153 - -0.019]; $p = 0.016$) experimental arms.

A.2. Supplementary Figures

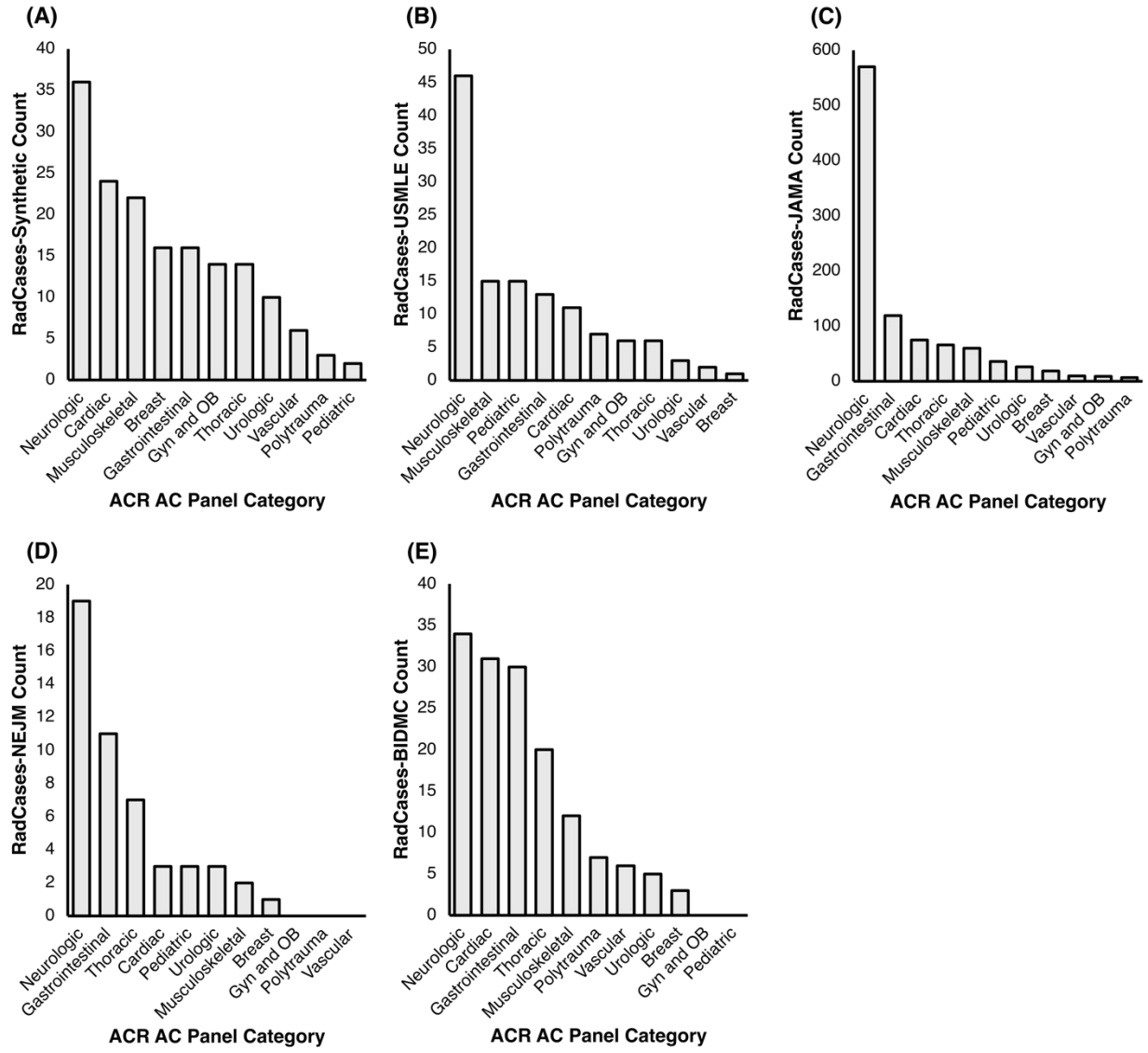

Figure A.1: **ACR AC Panel counts in the RadCases dataset.** As of June 2024, there are 224 ACR AC Topics that each have at least one assigned parent ACR AC Panel. Panels are more general categories for conditions, and there are 11 as of June 2024: Breast, Cardiac, Gastrointestinal, Gyn and OB, Musculoskeletal, Neurologic, Pediatric, Polytrauma, Thoracic, Urologic, and Vascular. To illustrate the distribution of conditions present in the RadCases dataset, we plot the counts of each of these 11 parent ACR AC Panels for the (**A**) **Synthetic**; (**B**) **USMLE**; (**C**) **JAMA**; (**D**) **NEJM**; and (**E**) **BIDMC** subsets of the RadCases dataset.

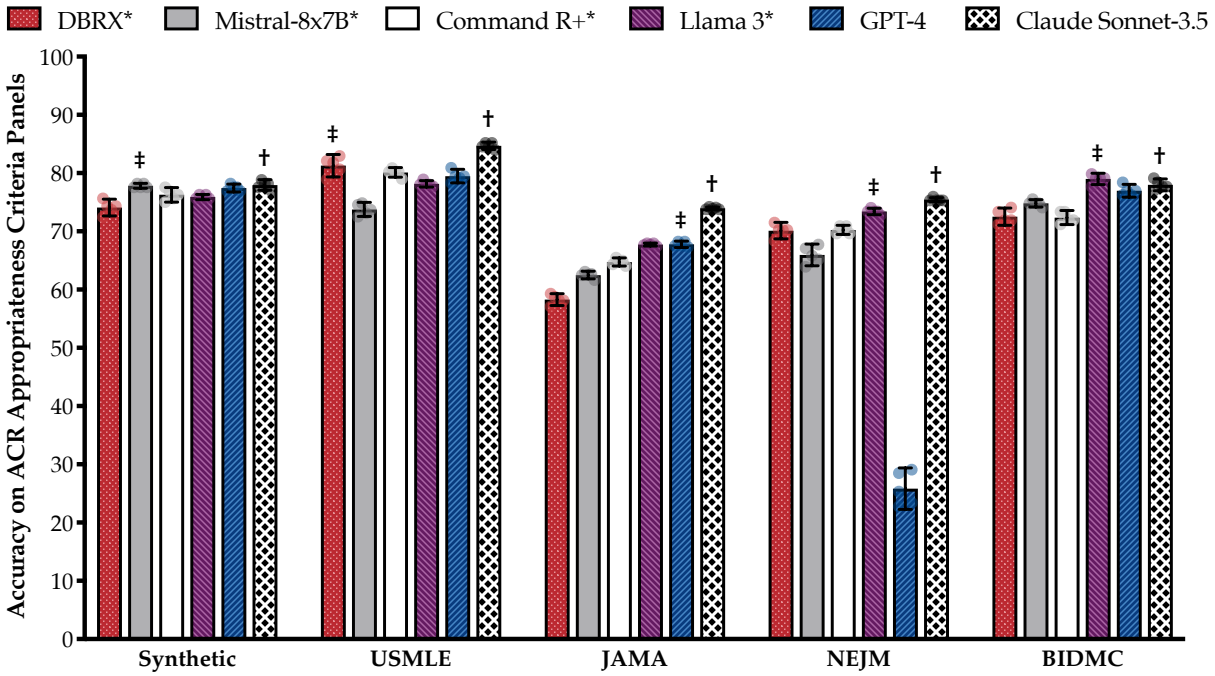

Figure A.2: **Baseline LLM performance on ACR AC Panel classification using the RadCases dataset.** In **Figure 3.2b**, we evaluate six state-of-the-art large language models (LLMs) on their ability to correctly assign 1 of 224 ACR AC Topics to an input one-liner. Here, we include analogous results on the related ACR AC *Panel* classification task, which queries an LLM to correctly assign 1 of 11 ACR AC Panels to an input one-liner. Because ACR AC Panels are much more coarse-grained when compared to Topics, a language model's accuracy on this task can help assess the model's ability to identify the general body part or organ system affected by pathophysiology. However, accuracy on this task is not helpful for ordering image studies, as there is no clear method for assigning a "correct" imaging study given only an ACR AC Panel. Open-source models are identified by an asterisk, and the best (second best) performing model for a RadCases dataset partition is identified by a dagger (double dagger). Error bars represent ±95% CI over $n = 5$ independent experimental runs.

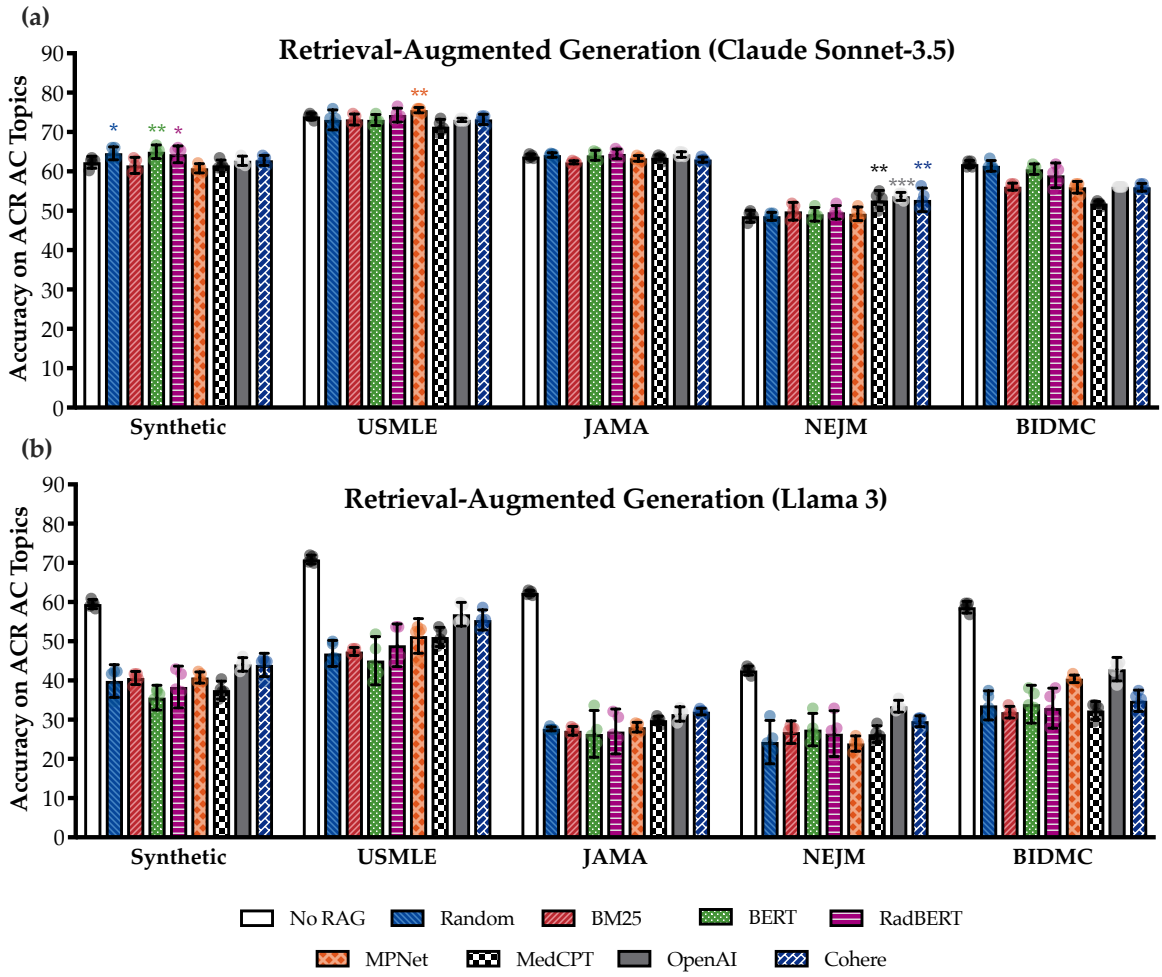

Figure A.3: **Retrieval-augmented generation (RAG) performance versus retriever algorithm.** To optimize RAG for LLM accuracy on the ACR AC Topic classification task, we investigated the use of 8 different retrieval algorithms to use in RAG: (1) **Random**, which randomly documents from the corpus over a uniform probability distribution; (2) Okapi **BM25** bag-of-words retriever; (3) **BERT** and (4) **MPNet** trained on unlabeled, natural language text; (5) **RadBERT** from fine-tuning BERT on radiology text reports; (6) **MedCPT** leveraging a transformer trained on PubMed search logs; and (7) **OpenAI** (`text-embedding-3-large`) and (8) **Cohere** (`cohere.embed-english-v3`) embedding models from OpenAI and Cohere for AI. Using (**a**) Claude Sonnet-3.5 and (**b**) Llama 3, we retrieve $k = 8$ documents from the ACR AC narrative guidelines corpus using each retriever, and compare each method against baseline ACR AC Topic accuracy achieved by each model. Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

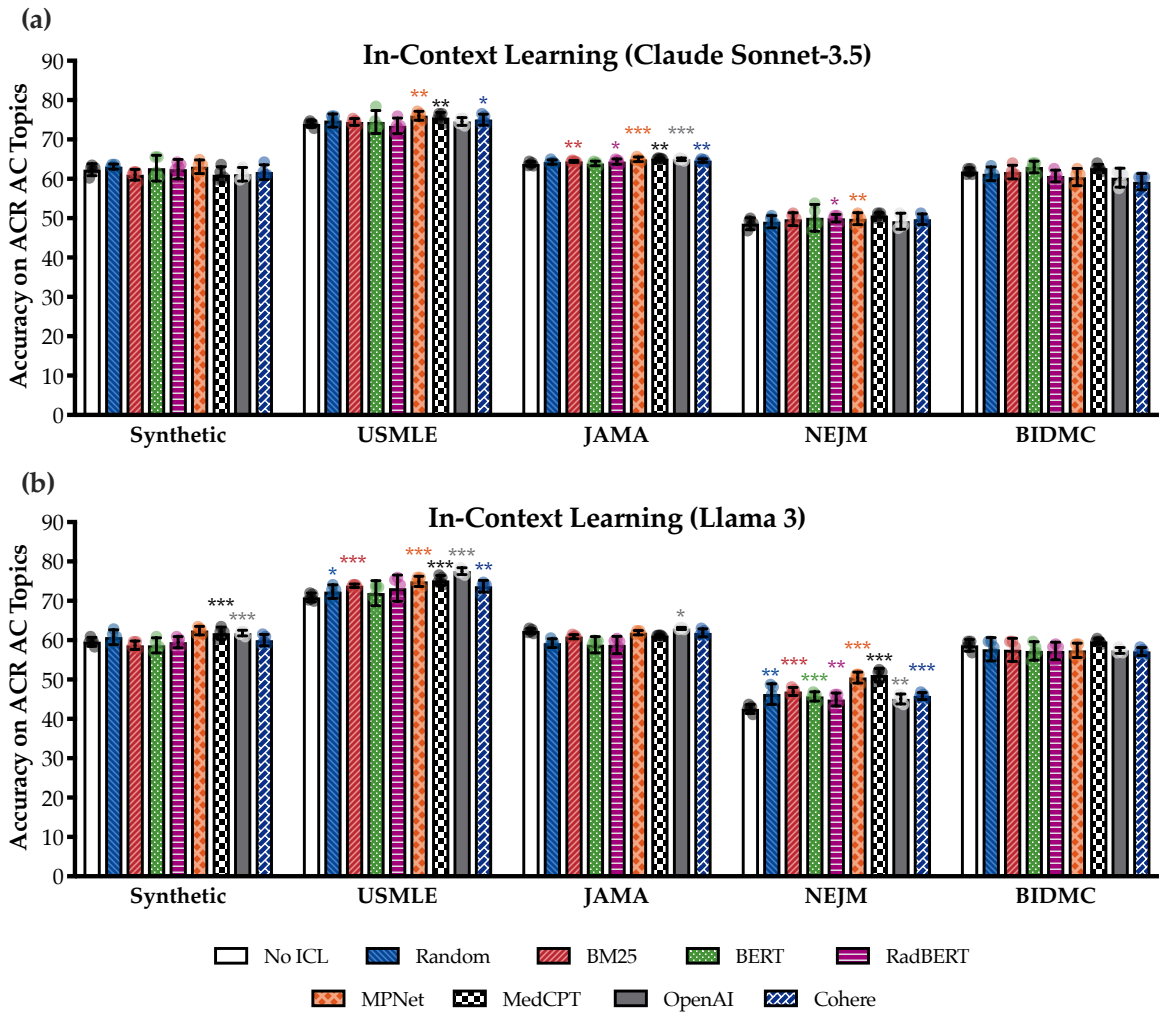

Figure A.4: **In-context learning (ICL) performance versus retriever algorithm.** To optimize ICL for LLM accuracy on the ACR AC Topic classification task, we investigated the use of 8 different retrieval algorithms to use in ICL identical to those explored in RAG (see caption of **Supp. Fig. A.3**). Using (**a**) Claude Sonnet-3.5 and (**b**) Llama 3, we retrieve $k = 4$ example one-liner/Topic pairs from the RadCases-Synthetic dataset corpus using each retriever, and compare each method against baseline ACR AC Topic accuracy achieved by each model. Note that a separate synthetically generated dataset (generated using Meta Llama 2 instead of OpenAI GPT-3.5) was used to evaluate ICL on the RadCases-Synthetic dataset to avoid data leakage. Error bars represent ±95% CI over $n = 5$ independent experimental runs.

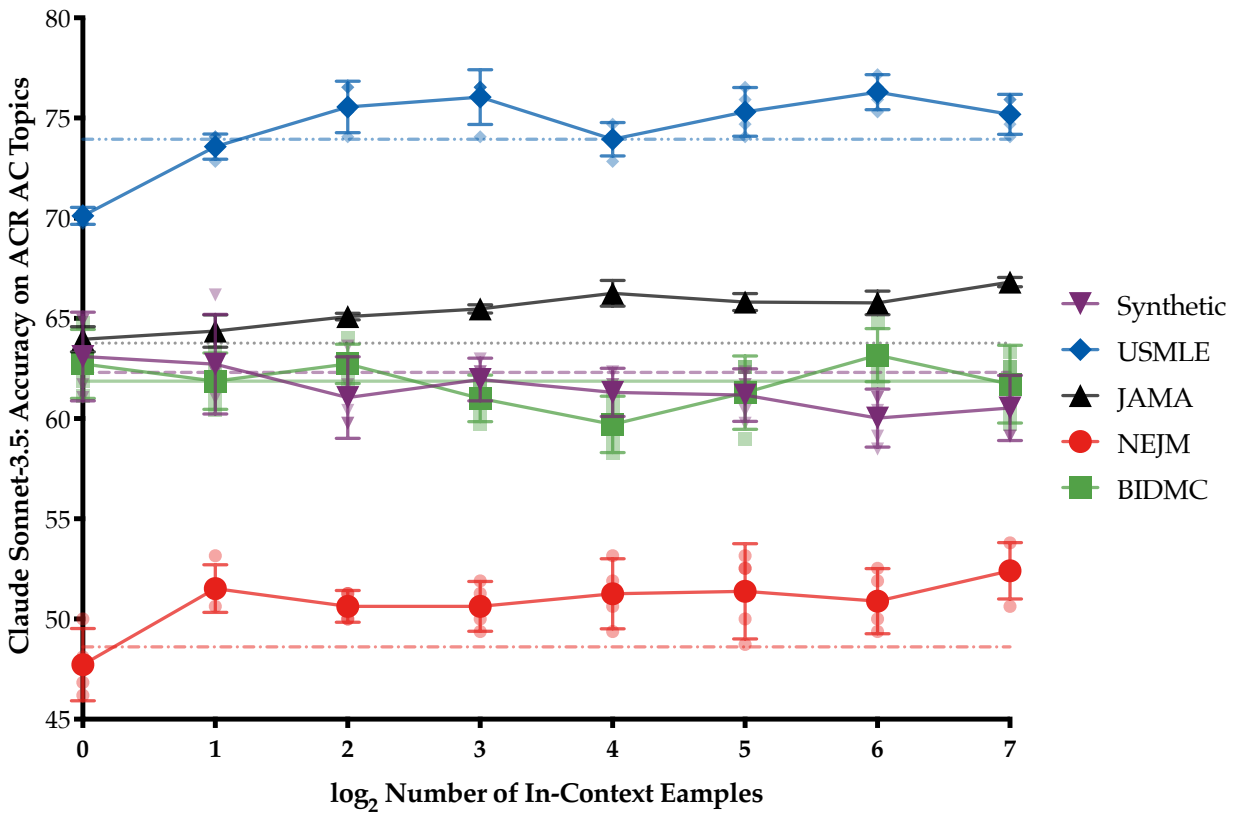

Figure A.5: **In-context learning (ICL) performance versus retriever budget.** Using the subjectively best retriever algorithm evaluated in **Supp. Fig. A.4** (i.e. the MedCPT retriever), we ablated the number of ICL examples retrieved by the retriever to pass as context to Claude Sonnet-3.5. Note that the purple solid, blue medium-dashed, black long-dashed, green dotted-dashed, and red dotted horizontal lines correspond to the baseline, no-ICL accuracy scores of Claude Sonnet-3.5 on the Synthetic, USMLE, JAMA, BIDMC, and NEJM subsets of the RadCases dataset, respectively. For the USMLE, JAMA, and NEJM subsets, we find that the performance of the model increases as the number of ICL examples increases from $k = 1$ to $k = 128$. Note that a separate synthetically generated dataset (generated from Meta Llama 2 instead of OpenAI GPT-3.5) was used to evaluate ICL on the RadCases-Synthetic dataset to avoid data leakage. Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

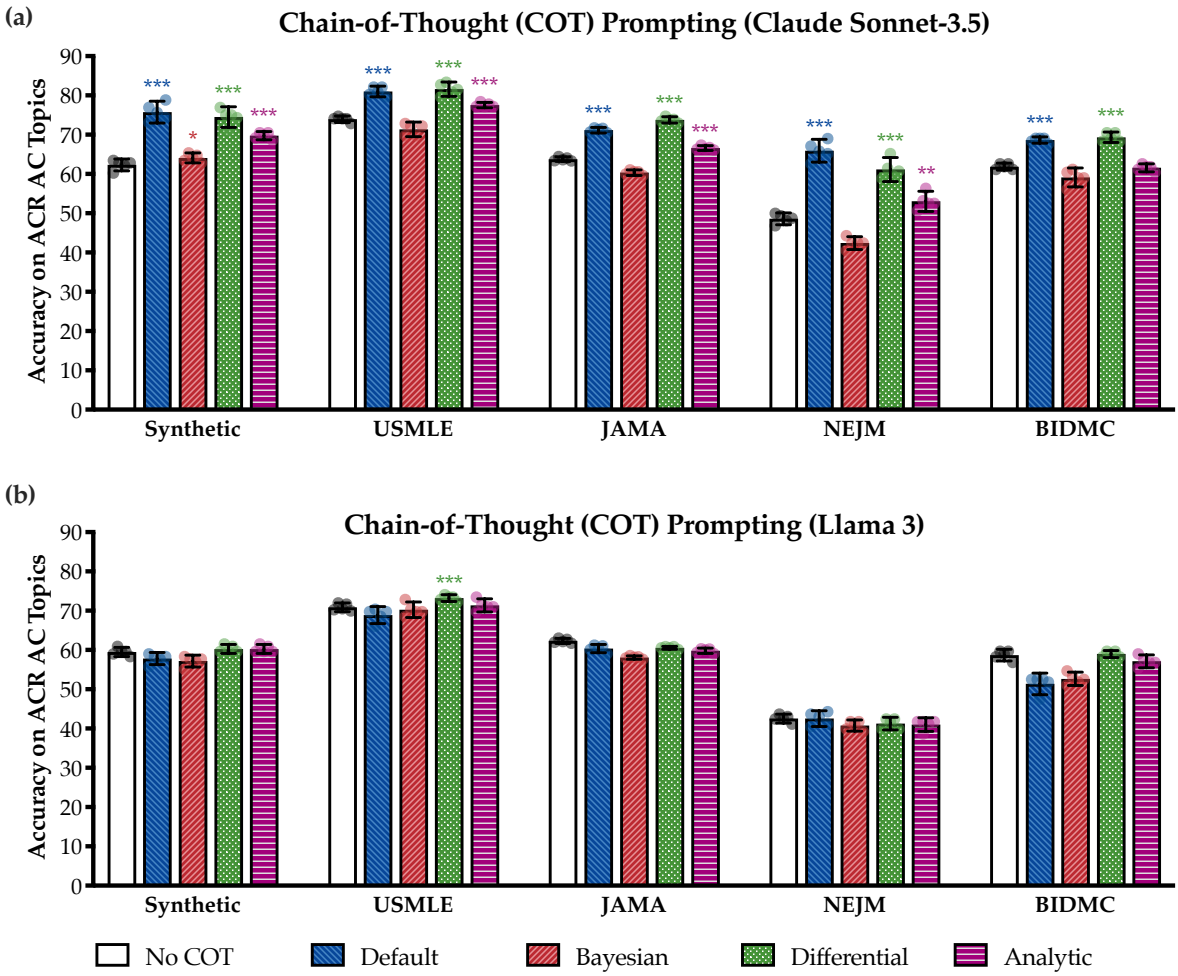

Figure A.6: **Chain-of-thought (COT) prompting performance versus reasoning algorithm.** To optimize COT for LLM accuracy on the ACR AC Topic classification task for both (**a**) Claude Sonnet-3.5 and (**b**) Llama 3, we investigated 4 different COT reasoning methods: (1) **Default** reasoning, which does not specify any particular reasoning strategy for the LLM to use; (2) **Differential** diagnosis reasoning, which encourages the model to reason through a differential diagnosis to arrive at a final prediction; (3) **Bayesian** reasoning, which encourages the model to approximate Bayesian posterior updates over the space of ACR AC Topics based on the clinical patient presentation; and (4) **Analytic** reasoning, which encourages the model to reason through the pathophysiology of the underlying disease process. We compare each method against the baseline ACR AC Topic accuracy achieved by each model. Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

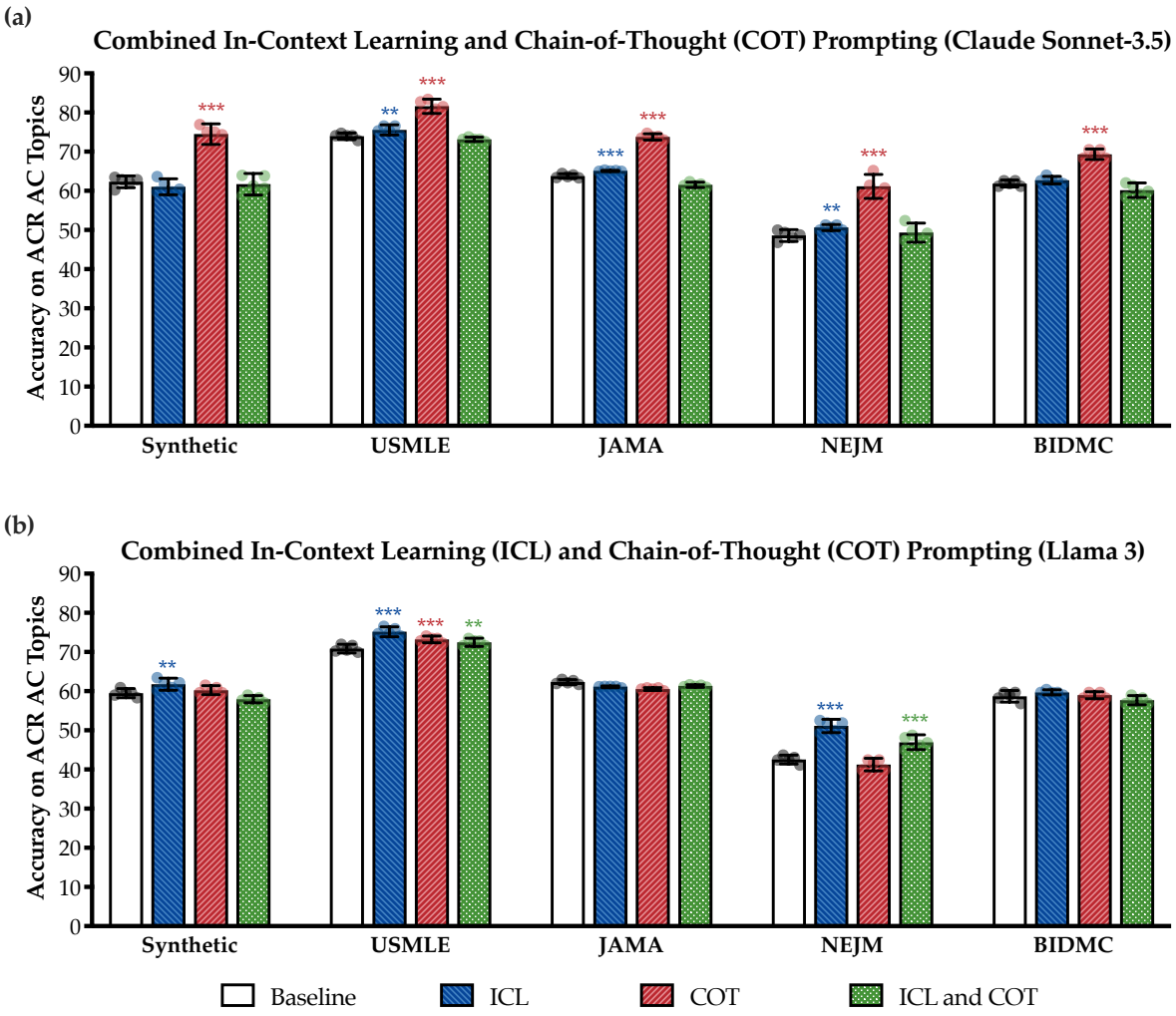

Figure A.7: **Combining in-context learning (ICL) and chain-of-thought (COT).** We observed that ICL (using the MedCPT retriever) and COT (using the Default reasoning strategy) were effective prompting strategies to improve the performance of Claude Sonnet-3.5 and/or Llama 3 in **Supp. Figures A.4** and **A.6**. We combine both of these strategies together to evaluate if the combination of these techniques together could further improve model performance of both (**a**) Claude Sonnet-3.5 and (**b**) Llama 3. We compare each method against the baseline, ICL-only, and COT-only ACR AC Topic accuracy achieved by each model. Error bars represent ±95% CI over $n = 5$ independent experimental runs.

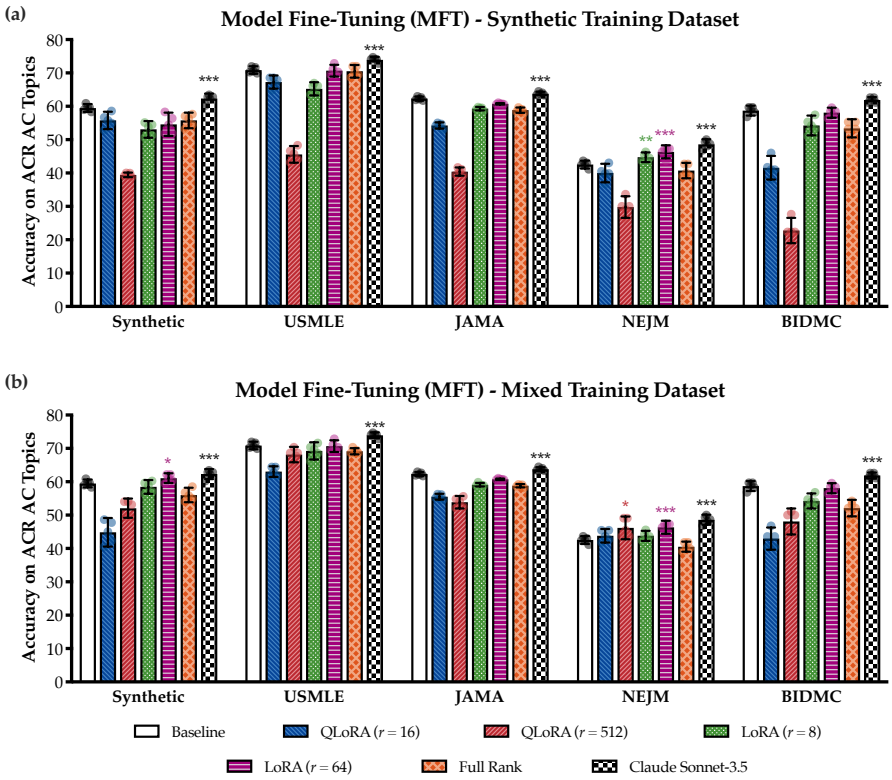

Figure A.8: **Model fine-tuning (MFT) algorithm evaluation with Llama 3.** We evaluate 5 different fine-tuning experimental setups in our MFT experiments: quantized low-rank adaptation (**QLoRA**) with a rank of (1) $r = 16$ and (2) $r = 512$; low-rank adaptation (**LoRA**) with a rank of (3) $r = 8$ and (4) $r = 64$; and (5) **Full Rank** model fine-tuning. We use an α scaling value of 8 for all QLoRA and LoRA experiments. To construct the MFT training dataset, we use either (**a**) all $n = 156$ labeled one-liners from the RadCases-Synthetic dataset; or (**b**) a Mixed dataset including 50 randomly selected cases from each of the 5 RadCases dataset subsets for a total of $n = 250$ labeled one-liners. The first scenario simulates a setting where we can only fine-tune models on synthetically generated data due to privacy concerns, and the latter scenario simulates a setting where we are able to train on real patient data sampled from the relevant distribution(s) of interest. Note that a separate synthetically generated dataset (generated from Meta Llama 2 instead of OpenAI GPT-3.5) was used to fine-tune the base model for evaluation on the RadCases-Synthetic dataset to avoid data leakage. Error bars represent $\pm$95% CI over $n = 5$ independent experiments.

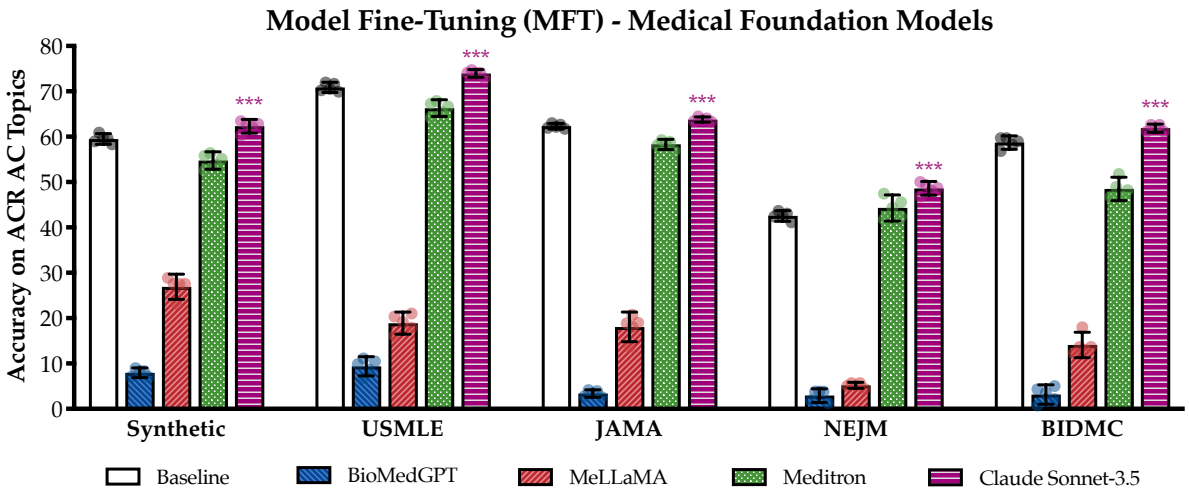

Figure A.9: **Evaluating medical foundation models fine-tuned on Llama LLMs.** Separate from the results presented in **Supp. Figure A.8**, an alternative approach to model fine-tuning is to instead leverage language models fine-tuned on large corpuses of domain-specific medical text. Such *foundation models* include BioMedGPT-7B (Zhang et al., 2024); MeLLaMA-70B (Xie et al., 2024); and Meditron-70B (Chen et al., 2023d). We evaluate their accuracies on predicting correct ACR AC Topic labels; none of the three medical foundation models evaluated outperformed the base Meta Llama 3 70B model with statistical significance on any of the RadCases datasets. Our results are consistent with findings reported by prior work (Jeong et al., 2024; Dorfner et al., 2024; Hager et al., 2024; Maharjan et al., 2024) and highlight the challenge in fine-tuning language models specifically for RadCases and other medical tasks. Error bars represent $\pm$95% CI over $n = 5$ independent experimental runs.

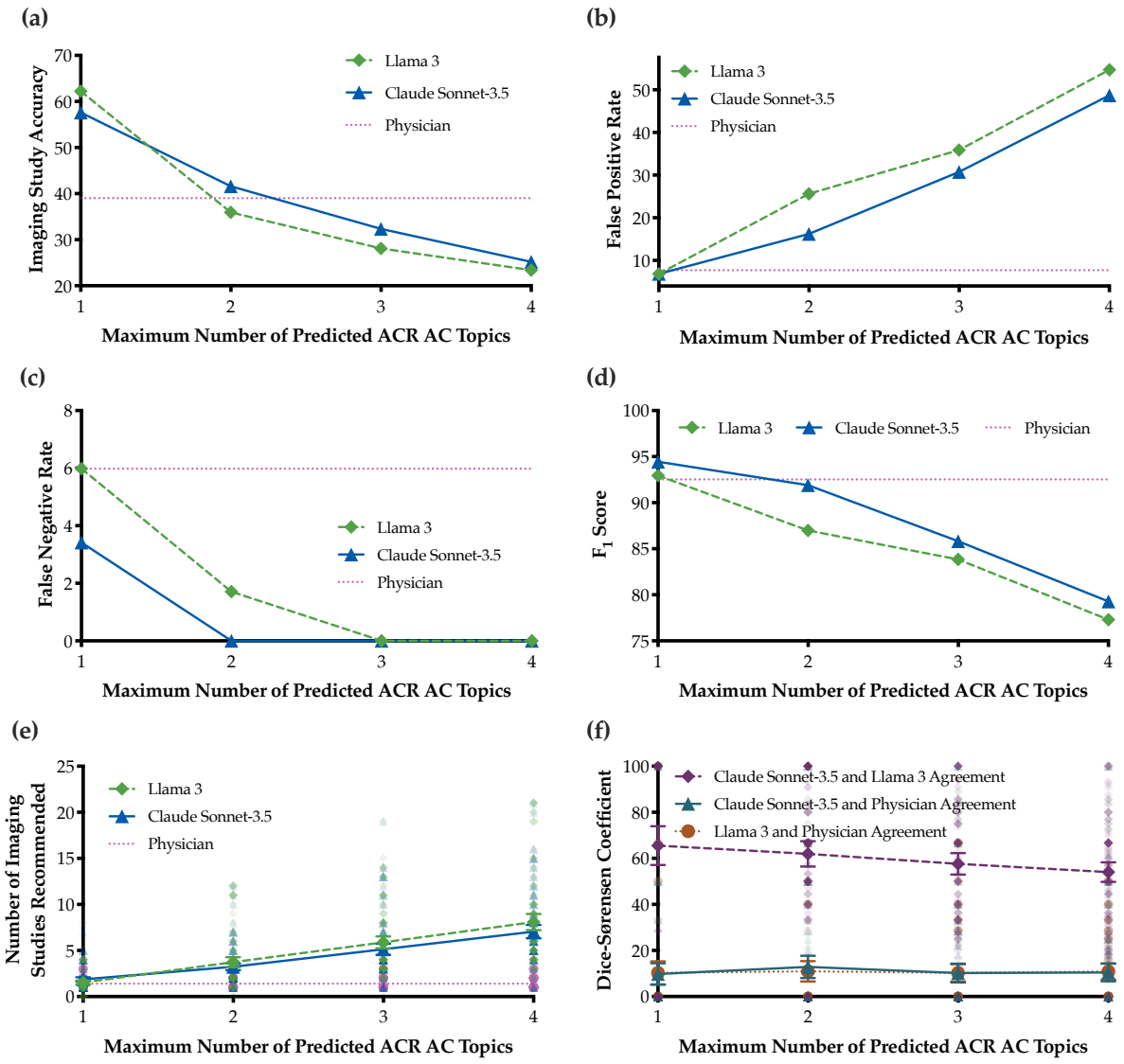

Figure A.10: **Ablating the number of ACR AC Topic predictions in retrospective study of clinician-ordered versus LLM-ordered imaging studies.** In **Figure 3.6**, we show the results of our retrospective study evaluating diagnostic imaging orders of both LLMs and clinicians—both Claude Sonnet-3.5 and Llama 3 were prompted to predict the single $m = 1$ best ACR AC Topic for an input patient description. Here, we vary the maximum number m of ACR AC Topic predictions requested from each language model on the x-axis. We compare the (**a**) accuracy scores; (**b**) false positive rates (i.e., the rate at which a patient received at least one unnecessary imaging recommendation); (**c**) false negative rates (i.e., the rate at which a patient should have received an imaging workup but did not); (**d**) F_1 scores; (**e**) number of recommended imaging studies; and (**f**) similarity of ordered imaging studies of Claude Sonnet-3.5 and Llama 3 versus m.

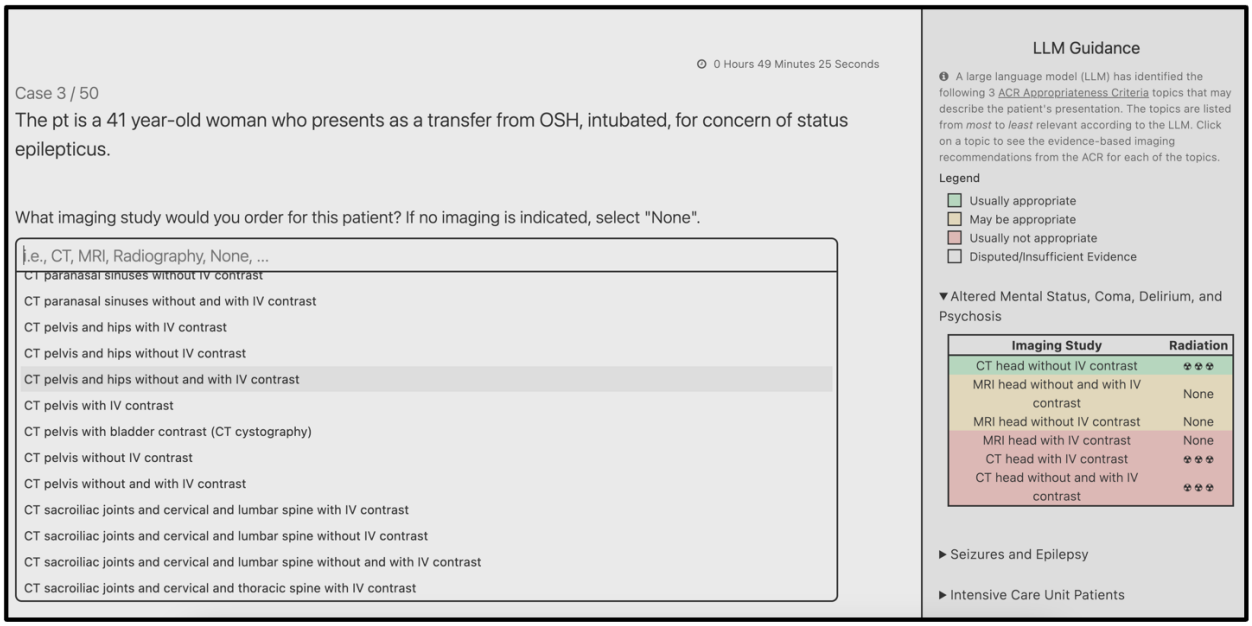

Figure A.11: **User interface for prospective study.** The LLM is asked to predict up to three ACR Appropriateness Criteria (AC) Topics that may be relevant for the patient case, and the table of corresponding ACR AC recommendations is displayed as reference to the user. In questions where LLM guidance is not made available, the right column does not show any recommendations and instead shows "LLM guidance is not available for this patient scenario."

A.3. Supplementary Tables

Table A.1: **Commonly appearing ACR AC Topics in the RadCases dataset.** We list the most commonly appearing ACR AC Topics for each of the RadCases datasets. The topics are listed as "Panel > Topic," where "Topic" is the ACR AC Topic and "Panel" is the parent ACR AC Panel.

Synthetic Subset ($n_{\textbf{total}} = 156$)	**Count (%)**
Cardiac > Chest Pain-Possible Acute Coronary Syndrome	10 (6.41)
Cardiac/Vascular > Suspected Pulmonary Embolism	8 (5.13)
Gyn and OB > Acute Pelvic Pain in the Reproductive Age Group	6 (3.85)
Neurologic > Low Back Pain	6 (3.85)
Breast > Breast Pain	5 (3.21)
USMLE Subset ($n_{\textbf{total}} = 164$)	**Count (%)**
Neurologic > Altered Mental Status, Coma, Delirium, and Psychosis	17 (10.4)
Neurologic > Headache	12 (7.32)
Cardiac > Chest Pain-Possible Acute Coronary Syndrome	9 (5.49)
Polytrauma > Major Blunt Trauma	9 (5.49)
Cardiac > Dyspnea-Suspected Cardiac Origin	8 (4.88)
JAMA Subset ($n_{\textbf{total}} = 971$)	**Count (%)**
Neurologic > Orbits, Vision, and Visual Loss	280 (28.8)
Neurologic > Neck Mass/Adenopathy	48 (4.94)
Neurologic > Staging and Post-Therapy Assessment of Head and Neck Cancer	42 (4.33)
Neurologic > Headache	31 (3.19)
Gastrointestinal > Acute Nonlocalized Abdominal Pain	28 (2.88)
NEJM Subset ($n_{\textbf{total}} = 159$)	**Count (%)**
Neurologic > Altered Mental Status, Coma, Delirium, and Psychosis	11 (6.92)
Neurologic > Orbits, Vision, and Visual Loss	8 (5.03)
Gastrointestinal > Acute Nonlocalized Abdominal Pain	7 (4.40)
Neurologic > Headache	7 (4.40)
Urologic > Renal Failure	7 (4.40)
BIDMC Subset ($n_{\textbf{total}} = 139$)	**Count (%)**
Cardiac > Chest Pain-Possible Acute Coronary Syndrome	13 (9.35)
Neurologic > Head Trauma	11 (7.91)
Gastrointestinal > Acute Nonlocalized Abdominal Pain	10 (7.19)
Neurologic > Altered Mental Status, Coma, Delirium, and Psychosis	7 (5.04)
Cardiac > Dyspnea-Suspected Cardiac Origin	6 (4.32)

Table A.2: **Comparing the RadCases dataset with real patient case summaries.** To validate our RadCases dataset, we first had 3 independent U.S. attending physicians review a set of 50 true one-liners and confirm that they are representative of real-world patient case summaries used in clinical practice. We then computed the (1) Maximum and (2) Mean Similarity Score using the NV-Embed-v2 Retriever (Lee et al., 2025; Moreira et al., 2024) between each of the RadCases datasets and a dataset of true one-liners derived from real patient cases. We also computed the average (3) Perplexity according to GPT-2 Large Medical (Radford et al., 2019; Gabarin, 2023; Jin et al., 2019); and (4) the average number of tokens per one-liner according to the GPT-4o tokenizer (OpenAI et al., 2024). We compare RadCases against other corpora such as arXiv computer science abstracts (arXiv NLP); Wikipedia articles (Wikitext); PubMed articles; and the MedQA dataset (Jin et al., 2021). Finally, we also compare against Random sentences admission notes in the MIMIC-IV dataset (Johnson et al., 2023); random sentences from Radiology imaging reports in the MIMIC-IV dataset, Full Admission Notes from the MIMIC-IV dataset; and a separate Test set of extracted patient one-liners from the MIMIC-IV dataset. Each metric is reported as $\text{Mean}^{95\%\text{ CI}}$, where [Mean] is the mean metric value, and [95% CI] is the 95% confidence interval. The best (resp., second best) values in each column—and all values with intersecting confidence intervals—are bolded (resp., underlined). Our results show that RadCases is a promising set of simulated patient one-liners compared with other domain-specific text corpora.

Text Source	Max Similarity ↑	Mean Similarity ↑	Perplexity	Token Count
True One-Liners	$100^{(100\text{-}100)}$	$40.0^{(38.7\text{-}41.3)}$	$115^{(81.1\text{-}148.5)}$	$51.3^{(44.2\text{-}58.3)}$
arXiv NLP	$13.8^{(13.7\text{-}13.9)}$	$6.09^{(6.04\text{-}6.13)}$	$30.0^{(29.7\text{-}30.2)}$	$171^{(170\text{-}172)}$
Wikitext	$14.3^{(14.1\text{-}14.4)}$	$6.57^{(6.48\text{-}6.66)}$	$\mathbf{134^{(127\text{-}141)}}$	$\underline{101}^{(99\text{-}103)}$
PubMed	$16.5^{(16.4\text{-}16.6)}$	$8.11^{(8.04\text{-}8.19)}$	$21.5^{(21.2\text{-}21.7)}$	$224^{(222\text{-}227)}$
MedQA	$38.4^{(38.2\text{-}38.6)}$	$25.1^{(24.9\text{-}25.2)}$	$14.6^{(14.5\text{-}14.7)}$	$161^{(160\text{-}163)}$
MIMIC-IV Random	$31.3^{(30.1\text{-}32.6)}$	$19.4^{(18.5\text{-}20.4)}$	$624^{(318\text{-}929)}$	$26.4^{(20.8\text{-}32.0)}$
MIMIC-IV Radiology	$26.7^{(25.3\text{-}28.0)}$	$16.1^{(15.0\text{-}17.2)}$	$437^{(323\text{-}552)}$	$19.3^{(16.1\text{-}22.6)}$
MIMIC-IV Full Note	$37.3^{(36.5\text{-}38.2)}$	$26.1^{(25.6\text{-}26.7)}$	$40.6^{(39.0\text{-}42.1)}$	$3220^{(3040\text{-}3390)}$
MIMIC-IV Test	$49.4^{(48.8\text{-}50.1)}$	$33.5^{(33.0\text{-}33.9)}$	$317^{(265\text{-}370)}$	$40.1^{(38.5\text{-}31.8)}$
RadCases Synthetic	$\underline{52.2}^{(51.0\text{-}53.4)}$	$\underline{35.3}^{(34.7\text{-}35.9)}$	$\underline{60.3}^{(47.8\text{-}72.8)}$	$23.6^{(22.7\text{-}24.4)}$
RadCases USMLE	$47.4^{(45.9\text{-}48.9)}$	$29.9^{(29.0\text{-}30.9)}$	$14.0^{(12.9\text{-}15.2)}$	$21.0^{(20.1\text{-}21.8)}$
RadCases JAMA	$43.1^{(42.6\text{-}43.7)}$	$27.9^{(27.5\text{-}28.2)}$	$20.0^{(19.2\text{-}20.8)}$	$33.0^{(32.3\text{-}33.8)}$
RadCases NEJM	$42.3^{(41.2\text{-}43.4)}$	$27.3^{(26.6\text{-}28.1)}$	$15.0^{(14.1\text{-}16.0)}$	$\mathbf{51.1^{(49.1\text{-}53.2)}}$
RadCases BIDMC	$\mathbf{65.1^{(62.1\text{-}68.2)}}$	$\mathbf{36.2^{(35.4\text{-}37.2)}}$	$\mathbf{173^{(140\text{-}207)}}$	$\mathbf{45.8^{(42.0\text{-}50.0)}}$

Table A.3: **Binary classification of ACR AC corpus relevancy for diagnostic image ordering.** In our main text, we limit our evaluation of language models to patient one-liners that can be $^{\text{(and are)}}$ assigned a ground-truth ACR AC Topic label. This implicitly assumes that we can filter out the patient one-liners where no ACR AC Topic is applicable. Here, we assess the ability of language models to perform this filtering task: we evaluate both Claude Sonnet-3.5 and Llama 3 on the binary classification task of determining whether the corpus of ACR AC Topics contains at least one ACR AC Topic that is applicable to an input patient one-liner. Each metric is reported as [Mean$^{\text{95\% CI}}$], where [Mean] is the mean metric value $^{\text{(averaged over 5 random seeds)}}$, and [95% CI] is the 95% confidence interval.

Claude Sonnet-3.5	**Synthetic**	**USMLE**	**JAMA**	**NEJM**	**BIDMC**
Balanced Accuracy ↑	$97.3^{(96.1\text{-}98.5)}$	$92.5^{(91.5\text{-}93.4)}$	$88.0^{(87.7\text{-}88.3)}$	$81.6^{(80.3\text{-}83.0)}$	$93.9^{(92.2\text{-}95.6)}$
F_1 Score ↑	$97.3^{(96.2\text{-}98.4)}$	$92.9^{(91.6\text{-}94.3)}$	$86.5^{(86.1\text{-}86.9)}$	$78.9^{(76.5\text{-}81.4)}$	$95.9^{(94.5\text{-}97.3)}$
False Positive Rate ↓	$0.1^{(0.0\text{-}0.2)}$	$6.4^{(5.2\text{-}7.6)}$	$1.1^{(0.2\text{-}1.9)}$	$8.1^{(6.5\text{-}9.7)}$	$8.4^{(7.0\text{-}9.7)}$
False Negative Rate ↓	$5.3^{(3.1\text{-}7.5)}$	$8.5^{(6.7\text{-}10.2)}$	$22.9^{(22.4\text{-}23.4)}$	$28.6^{(27.0\text{-}30.2)}$	$3.7^{(1.5\text{-}5.9)}$
Llama 3	**Synthetic**	**USMLE**	**JAMA**	**NEJM**	**BIDMC**
Balanced Accuracy ↑	$99.0^{(98.6\text{-}99.3)}$	$94.9^{(93.6\text{-}96.2)}$	$92.2^{(91.9\text{-}92.4)}$	$80.2^{(79.7\text{-}80.8)}$	$87.4^{(87.4\text{-}87.4)}$
F_1 Score ↑	$99.0^{(98.7\text{-}99.3)}$	$95.6^{(95.0\text{-}96.3)}$	$92.5^{(92.4\text{-}92.6)}$	$77.9^{(77.6\text{-}78.2)}$	$87.6^{(87.5\text{-}87.6)}$
False Positive Rate ↓	$0.0^{(0.0\text{-}0.0)}$	$7.3^{(4.7\text{-}9.9)}$	$3.8^{(3.3\text{-}4.3)}$	$11.0^{(10.1\text{-}11.8)}$	$9.1^{(9.1\text{-}9.1)}$
False Negative Rate ↓	$2.1^{(0.9\text{-}3.3)}$	$3.4^{(3.1\text{-}3.7)}$	$11.6^{(11.5\text{-}11.7)}$	$28.6^{(28.2\text{-}29.0)}$	$16.1^{(16.0\text{-}16.2)}$

Table A.4: **Simulated patient demographics for retrospective study assessing LLMs versus clinician performance.** In our retrospective study described in the main text, we analyzed the performance of autonomous LLM agents versus clinicians in ordering diagnostic imaging studies for simulated patient cases crafted from anonymized, de-identified discharge summaries from the MIMIC-IV dataset from Johnson et al. (2023). To better simulate actual patient cases, we manually annotated the patient cases to include simulated patient ages and genders if they were removed during the original de-identification process. The resulting distributions of these simulated patient variables are shown.

Gender	**Count (%)**
Male	63 (53.8)
Female	54 (46.2)
Age Decade (Years)	**Count (%)**
10-19	2 (1.71)
20-29	5 (4.27)
30-39	7 (5.98)
40-49	15 (12.8)
50-59	27 (23.1)
60-69	30 (25.6)
70-79	27 (23.1)
80-89	4 (3.42)
Total Number of Patient Cases	**117**

Table A.5: **Simulated patient demographics for prospective study assessing clinician performance with versus without LLM-based assistance.** In our prospective clinical study, we analyzed the performance of clinicians both with and without LLM-based imaging recommendations in ordering diagnostic imaging studies for simulated patient one-liners. These one-liners were crafted from anonymized, de-identified discharge summaries from the MIMIC-IV dataset from Johnson et al. (2023). To better simulate actual patient cases, we manually re-introduced simulated patient ages and/or genders if they were removed during the original de-identification process. The resulting distributions of these simulated patient variables are shown above.

Gender	**Count (%)**
Male	26 (52.0)
Female	24 (48.0)
Age Decade (Years)	**Count (%)**
10-19	2 (4.00)
20-29	3 (6.00)
30-39	2 (4.00)
40-49	7 (14.0)
50-59	8 (16.0)
60-69	12 (24.0)
70-79	13 (26.0)
80-89	3 (6.0)
Total Number of Patient Cases	**50**

Table A.6: **Study participant demographic information in prospective study assessing clinician performance with versus without LLM-based assistance.** Demographic and self-reported pre-study questionnaire information of the clinician study participants in our prospective study detailed in the main text are summarized here. Column (1) describes the participants randomized to the Timed study arm described in **Section A.1**, and column (2) describes the participants randomized to the Untimed study arm in **Section A.1**.

	(1)	**(2)**
Gender	**Count (%)**	**Count (%)**
Male	4 (25.0)	9 (64.3)
Female	12 (75.0)	5 (35.7)
Stage of Medical Training	**Count (%)**	**Count (%)**
Third- or Fourth- Year U.S. Medical Student	14 (87.5)	9 (64.3)
U.S. Emergency Medicine Resident Physician	2 (12.5)	5 (35.7)
Prior Experience Using AI in Everyday Life	**Count (%)**	**Count (%)**
No Experience or A Little Experience	8 (50.0)	9 (64.3)
Some Experience or A Lot of Experience	8 (50.0)	5 (35.7)
Overall Sentiment of the Use of AI in Healthcare	**Count (%)**	**Count (%)**
Negative	2 (12.5)	2 (14.3)
Neutral or Positive	14 (87.5)	12 (85.7)

Table A.7: **Accuracy scores of clinicians with and without LLM-generated recommendations.** The treatment effect of offering LLM-generated diagnostic imaging recommendations is analyzed according to (3.1). The accuracy score is a binary dependent variable equal to 1 if the clinician orders a ground-truth imaging study according to the ACR Appropriateness Criteria, and 0 otherwise. The regression coefficients are shown as mean (standard error). Columns (1) and (2) correspond to the timed experimental arm where participants are required to answer questions at an average rate no slower than 1 question per minute, while columns (3) and (4) correspond to the separate untimed experimental arm where participants can answer questions at their own pace in one sitting. Odd- (even-) numbered columns do not (do) factor in the fixed effects of participant self-reported personal experience with AI and personal sentiment on the use of AI in medicine, which are both modeled as binary variables, into the regression model. *Denotes $p < 0.05$.

	(1)	(2)	(3)	(4)
LLM Guidance Available	0.081*	0.081*	-0.089*	-0.089*
	(0.028)	(0.028)	(0.037)	(0.037)
p Value	0.011	0.011	0.032	0.032
R2	0.138	0.138	0.121	0.121
Number of Observations	800	800	700	700

Table A.8: **LLM agreement scores of clinicians with and without LLM-generated recommendations.** The treatment effect of offering LLM-generated diagnostic imaging recommendations is analyzed according to

$$z_{s,q} = \beta_0 + (\beta_1 * \text{WithLLMGuidance}_{s,q}) + \left(\beta_2 * \text{PriorExperienceUsingAI}_s\right) + (\beta_3 * \text{PositiveSentimentAboutAI}_s) + \theta_q + \chi_s + \varepsilon_{s,q}$$

where $z_{s,q}$ is the agreement score represented as a binary dependent variable equal to 1 if the clinician and LLM recommend the same imaging study and 0 otherwise; θ_q is the fixed effects of study question q; χ_s is the fixed effects of study participant s; and $\varepsilon_{s,q}$ is the error term. All the independent variables are binarized to take on values in $\{0, 1\}$. The regression coefficients are shown as mean (standard error). Column (1) corresponds to the timed experimental arm where participants are required to answer questions at an average rate no slower than 1 question per minute, while column (2) corresponds to the separate untimed experimental arm where participants can answer questions at their own pace in one sitting. *Denotes $p < 0.05$. **Denotes $p < 0.01$.

	(1)	(2)
LLM Guidance Available	0.141**	0.107**
	(0.043)	(0.031)
p Value	0.005	0.004
Prior Experience Using AI	0.043	-0.169**
	(0.051)	(0.055)
p Value	0.407	0.009
Positive Sentiment About AI	-0.219**	-0.086*
	(0.063)	(0.031)
p Value	0.004	0.016
R2	0.380	0.365
Number of Observations	800	700

Table A.9: **False positive rates of clinicians with and without LLM-generated recommendations.** The treatment effect of offering LLM-generated diagnostic imaging recommendations is analyzed according to (3.1). A false positive is a binary dependent variable equal to 1 if the clinician orders an unnecessary imaging study according to the ACR Appropriateness Criteria, and 0 otherwise. The regression coefficients are shown as mean (standard error). Columns (1) and (2) correspond to the timed experimental arm where participants are required to answer questions at an average rate no slower than 1 question per minute, while columns (3) and (4) correspond to the separate untimed experimental arm where participants can answer questions at their own pace in one sitting. Odd- (even-) numbered columns do not (do) factor in the fixed effects of participant self-reported personal experience with AI and personal sentiment on the use of AI in medicine, which are both modeled as binary variables, into the regression model.

	(1)	(2)	(3)	(4)
LLM Guidance Available	0.008	0.008	-0.005	-0.005
	(0.009)	(0.009)	(0.011)	(0.011)
p Value	0.412	0.418	0.633	0.630
R2	0.057	0.054	0.061	0.059
Number of Observations	800	800	700	700

Table A.10: **False negative rates of clinicians with and without LLM-generated recommendations.** The treatment effect of offering LLM-generated diagnostic imaging recommendations is analyzed according to (3.1). A false negative is a binary dependent variable equal to 1 if the clinician orders no imaging study even when diagnostic imaging is warranted according to the ACR Appropriateness Criteria, and 0 otherwise. The regression coefficients are shown as mean (standard error). Columns (1) and (2) correspond to the timed experimental arm where participants are required to answer questions at an average rate no slower than 1 question per minute, while columns (3) and (4) correspond to the separate untimed experimental arm where participants can answer questions at their own pace in one sitting. Odd- (even-) numbered columns do not (do) factor in the fixed effects of participant self-reported personal experience with AI and personal sentiment on the use of AI in medicine into the regression model.

	(1)	(2)	(3)	(4)
LLM Guidance Available	-0.019	-0.019	0.001	0.001
	(0.023)	(0.023)	(0.021)	(0.021)
p Value	0.440	0.431	0.952	0.951
R2	0.221	0.180	0.227	0.218
Number of Observations	800	800	700	700

Table A.11: **Study participant pre-study survey results.** Study participants were asked to complete an anonymized survey of multiple-choice questions (tabularized below) prior to beginning the study. The results of (Q1) and (Q5) were used to define the PriorExperienceUsingAI$_s$ and PositiveSentimentAboutAI$_s$ binary variables used in the regression models, respectively. For a subject s, PriorExperienceUsingAI$_s$ is equal to 1 if the subject answers "Some experience" or "A lot of prior experience" to (Q1) and 0 otherwise. Similarly, PositiveSentimentAboutAI$_s$ is equal to 1 if the subject answers "Neutral", "Somewhat positive", or "Very positive" to (Q5) and 0 otherwise.

(Q1) In your personal life, how much prior experience do you have with using machine learning models, such as ChatGPT or other AI tools?	**Count (%)**
No prior experience	2 (7.4)
A little prior experience	17 (63.0)
Some prior experience	8 (29.6)
A lot of prior experience	0 (0.0)
(Q2) In your personal role, how much prior experience do you have with using machine learning models, such as ChatGPT or other AI tools?	**Count (%)**
No prior experience	23 (85.2)
A little prior experience	4 (14.8)
Some prior experience	0 (0.0)
A lot of prior experience	0 (0.0)
(Q3) AI can help improve patient care and clinical workflows in the future.	**Count (%)**
I strongly disagree	0 (0.0)
I somewhat disagree	0 (0.0)
I am neutral	1 (3.7)
I somewhat agree	15 (55.6)
I strongly agree	11 (40.7)
(Q4) I am scared about the potential unknown impact of AI on healthcare.	**Count (%)**
I strongly disagree	1 (3.7)
I somewhat disagree	8 (29.6)
I am neutral	2 (7.4)
I somewhat agree	13 (48.2)
I strongly agree	3 (11.1)
(Q5) Overall, how positive or negative do you feel about the potential use of AI in medicine?	**Count (%)**
Very negative	0 (0.0)
Somewhat negative	3 (11.1)
Neutral	2 (7.4)
Somewhat positive	18 (66.7)
Very positive	4 (14.8)

APPENDIX B

Clinically Derived Priors for Medical Imaging Analysis: Additional Discussion

The following appendix discusses two co-authored works principally led by collaborators, titled (1) "A Textbook Remedy for Domain Shifts: Knowledge Priors for Medical Image Analysis" published in the Proceedings of the 2024 Conference of Neural Information Processing Systems as a spotlight work (Yang et al., 2024c); and (2) "A Concept-based Interpretable Model for the Diagnosis of Choroid Neoplasias using Multimodal Data" published in Nature Communications (Wu et al., 2025). Because these works have been discussed in detail in other dissertations by co-authors, the discussion based on these two manuscripts in this appendix will be brief, and primarily focus on how they relate to experimental findings reported in the main text. We refer the interested reader to Yang et al. (2024c) and Wu et al. (2025) for additional details.

B.1. Deriving Interpretable Features

Recall that a core focus of this dissertation is the construction of **interpretable-by-design** machine learning systems in order to improve the generalizability of predictions across different patient populations. In **Chapter 3**, we demonstrated how interpretable patient representations could be derived from consensus medical guidelines endorsed by expert clinicians; in **Chapter 4**, we separately used clinical knowledge from our understanding of disease processes to derive the interpretable representation space. However, other strategies exist on how to best derive interpretable feature spaces for representing input patient data.

B.2. Clinical Motivation

In Yang et al. (2024c), work led by co-author Yue Yang explores how concept bottleneck models (Koh et al., 2020) can be used to construct generalizable medical image classification models that are robust to distribution shifts. It is well-documented that ML models often fail in unexpected ways due to overfitting on confounding variables, such as patient sex, race, and environment that do not causally affect the underlying pathophysiology of disease. As a result, existing deep networks can be highly sensitive to domain shifts. To address this issue as it pertains to the diagnosis

of disease from chest X-rays and medical photographs of skin lesions, Yang et al. (2024c) investigates how to build more robust medical image classifiers.

Separately in Wu et al. (2025), work led by co-author Yifan Wu specializes to the problem of diagnosing rare ocular diseases. Making such diagnoses is a critical challenge in ophthalmology because there are a very limited number of clinical specialists with the expertise to diagnose ocular cancers and distinguish uveal melanoma, ocular hemangioma, and metastatic carcinoma. Clinically, it is important to distinguish these three diseases because they are individually associated with different prognoses and treatment strategies even though they can present similarly in practice. Furthermore, the scarcity of data on rare diseases makes it challenging to develop generalizable and accurate machine learning methods for diagnosing these diseases. To this end, Wu et al. (2025) explores how interpretable feature representations of multimodal ocular data can not only be used to build more accurate and generalizable predictive ML models, but also improve the predictive accuracy of non-specialist physicians in diagnosing rare eye diseases.

B.3. Methods

Yang et al. (2024c) introduces **Kno**ledge **Bo**ttlenecks (**Knobo**), a novel class of concept bottleneck models that incorporate knowledgeable priors derived from foundational medical knowledge to construct the space of concepts used to represent input imaging data. At a high-level, KnoBo takes as input a pretraining dataset $\mathcal{D}_{\text{pre}} = \{(I_k, t_k)\}_{k=1}^{n_{\text{pre}}}$ of image-text pairs and an annotated dataset $\mathcal{D}_{\text{train}} = \{(I_k, y_k)\}_{k=1}^{n_{\text{train}}}$ of images $I_k \in \mathcal{I}$ labeled by their corresponding disease $y \in \mathcal{Y}$. KnoBo first constructs a set of *concepts* $\mathcal{C}$ by iteratively querying a language model conditioned on the label space $\mathcal{Y}$ and a corpus of relevant medical documents to learn from and perform retrieval on. For each concept $c \in \mathcal{C}$, a corresponding feature encoder $f_c : \mathcal{I} \to [0, 1]$ is learned by first prompting a language model to map a text input t to a binary output indicating whether t contains c, and then learning f_c as a binary classifier using the mapped dataset $\mathcal{D}_{\text{pre}}^c = \{(I_k, \delta(c \in t_k))\}_{k=1}^{n_{\text{pre}}}$. After pretraining, each image I can then be mapped at inference time to an interpretable feature vector $[0, 1]^{|\mathcal{C}|}$ with dimensions representing the probability of the presence of each semantically meaningful concept in $\mathcal{C}$. A linear function can then be learned via the traditional cross entropy

loss function to map vectors in $[0,1]^{|\mathcal{C}|}$ to a vector of logits over the label space via training on $\mathcal{D}_{\text{train}}$. Altogether, the final pipeline can be written as

$$y_{\text{pred}} = \text{softmax}\left(f_{\mathcal{C}}(I) \cdot W^T\right)$$

where $f_{\mathcal{C}} : \mathcal{I} \to [0,1]^{|\mathcal{C}|}$ is a shorthand contraction of all learned f_c that collectively maps an input image to all $|\mathcal{C}|$ concept activations, and W is the $|\mathcal{Y}| \times |\mathcal{C}|$ weight matrix of the final linear function. In this way, we are able to construct a set of interpretable, semantically meaningful features $\mathcal{C}$ directly from unstructured, domain-specific natural language using LLM-based pipelines.

Wu et al. (2025) introduces **M**ultimodal **M**edical **C**oncept **B**ottleneck **M**odels (MMCBM), which share a similar methodology with KnoBo except for a few key modifications. First, the input image space $\mathcal{I}$ is inherently multimodal, meaning that the set of feature encoders f_c scales with not only the number of relevant concepts $|\mathcal{C}|$, but also the number of distinct imaging modalities available as input. Secondly, to overcome the fact that data for MMCBM pre-training is inherently limited in the rare-disease setting, the set of concepts $\mathcal{C}$ is manually refined by expert clinicians prior to learning the concept encoders f_c. This allows us to manually refine the space of concepts learned from natural language alone using the knowledge available from rare disease experts.

B.4. Main Results

In Yang et al. (2024c), we find that KnoBo significantly improves out-of-distribution (OOD) generalization across both chest X-ray and skin lesion multiclass image classification tasks. In comprehensive evaluations with explicitly constructed confounding factors, KnoBo outperforms fine-tuned baselines without significantly sacrificing in-distribution (ID) performance. KnoBo also outperforms other interpretable models, and its concept representations were found to be more robust and effective than standard CLIP-derived features. Altogether, we found that incorporating medically grounded priors directly into model design can help us build more generalizable medical image classification models.

In Wu et al. (2025), we show that the MMCBM method matches and sometimes even outperforms

traditional black-box baselines—especially when tested on a dataset with a known domain shift (i.e., patients from a different hospital setting). In this case, MMCBM was able to generalize to the new patient population in a zero-shot manner more effectively than existing methods. Furthermore, clinician users could interactively inspect and adjust concept activations in real-time to further boost model performance via human-in-the-loop inference. Our results suggest that MMCBM achieves strong performance in diagnosing rare ocular diseases compared to baseline methods.

B.5. Discussion

Here, we relate Yang et al. (2024c); Wu et al. (2025) to the main text of the dissertation (particularly **Chapters 3-4**). While the implementation details may vary between the works on building interpretable-by-design models discussed in this dissertation (i.e., Yao et al. (2025a, 2023); Wu et al. (2025); Yang et al. (2024c)), they all share a fundamental core hypothesis: **interpretable-by-design ML models aligned with human knowledge are more generalizable**. In Yao et al. (2025a), the concept set $\mathcal{C}$ is the space of American College of Radiology (ACR) Appropriateness Criteria (AC) Topics explicitly constructed from parsing the ACR AC medical guidelines; we use an LLM to map input patient descriptions to the concept space. In Yao et al. (2023), the concept set $\mathcal{C}$ is the set of clinically derived phenotypes (CDPs) and image-derived phenotypes (IDPs) extracted from multimodal patient data; we query tabular CDPs from health record databases and single-task image segmentation models to extract IDPs. In Yang et al. (2024c) and Wu et al. (2025), the concept space $\mathcal{C}$ is constructed using state-of-the-art language models given access to both the label space $\mathcal{Y}$ of the prediction task and a corpus of unstructured medical text; we use CLIP-based encoders (Radford et al., 2021; Wang et al., 2022; Eslami et al., 2023) to map input images to the concept space. Similarly, the set of concepts $\mathcal{C}$ is transformed into a label prediction via tabular lookup in Yao et al. (2025a); a simple non-linear multilayer perceptron in Yao et al. (2023); and a linear layer in Yang et al. (2024c) and Wu et al. (2025). In this way, the primary differences between the methodologies introduced in these works are (1) how the intermediate concept spaces $\mathcal{C}$ are constructed; (2) the mapping implementation from input space to $\mathcal{C}$; and (3) the mapping implementation from $\mathcal{C}$ to label space.

APPENDIX C

Adversarial Supervision in Offline Model-Based Optimization: Additional Experimental Results

The following appendix contains additional experimental results for the interested reader using **GAMBO**—our novel algorithm for offline optimization that leverages adversarial supervision to improve the quality of proposed designs. Specifically, we provide additional experimental results that help better characterize both the strengths and limitations of GABO and GAGA.

C.1. Additional Design Quality Results

To evaluate the robustness of optimization methods, we report one-shot 90th percentile oracle scores in **Tables C.1-C.2**. For each method, all proposed designs are ranked according to the surrogate forward model ((5.5) for Generative Adversarial Bayesian Optimization (GABO) and Generative Adversarial Gradient Ascent (GAGA)), and the single 90th percentile design according to this ranking is selected and evaluated using the oracle function. We report the oracle score of this suboptimal design averaged over 10 seeds.

We found that GABO and GAGA do not propose suboptimal designs that are better than those proposed by other methods, such as BONET (Krishnamoorthy et al., 2023b), Simulated Annealing (Kirkpatrick et al., 1983), L-BFGS (Liu and Nocedal, 1989), and ExPT (Nguyen et al., 2023). This is not surprising, as aSCR is not designed to target this metric (and it is not our primary metric of interest). Separately for GABO, we also hypothesize that the algorithm's performance according to this metric may partially be explained by the limitations of the underlying Bayesian optimization (BO) optimization algorithm. Because BO is not an iterative first-order algorithm, the designs proposed by any BO-based algorithm often have high variance in practice—this is indeed what we observe across all of our experiments, including in **Table 5.2** and **Supp. Tables C.1-C.2**.

Finally, we note that in most applications of offline optimization, the 90th percentile metric—or any metric that does not use the best proposed design(s)—is not as useful as the other metrics

assessed where GABO does perform well. This is because in offline optimization tasks with a restricted budget to query the hidden, expensive-to-evaluate oracle function, we are not interested in "wasting" this limited budget on subpar design candidates. While the 90th percentile and similar metrics can be helpful to understand the limitations of algorithms, we believe that the alternative evaluation metrics reported in the main text—namely, the 100th percentile top-1 and top-128 oracle score metrics—are more useful and practical in assessing each of the optimization algorithms.

Table C.1: **Constrained budget ($k = 1$) suboptimal (90%-ile) oracle evaluation.** The oracle score of the 90th percentile design candidate according to the surrogate across 10 random seeds is reported as mean ± standard deviation. $\mathcal{D}$ (best) reports the top oracle value in the task dataset. The average rank across all eight tasks is reported in the final table column. **Bolded** (Underlined) entries indicate the best (second best) entry in the column. *Denotes the life sciences-related discrete MBO tasks from Design-Bench (Trabucco et al., 2022).

Method	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin	Rank
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96	—
Grad.	-94.4 ± 20.9	-5.47 ± 1.32	0.429 ± 0.023	3.43 ± 0.67	7.16 ± 0.21	-1.95 ± 0.00	0.53 ± 0.20	0.87 ± 1.08	9.1
L-BFGS	**-4.0 ± 0.0**	4.96 ± 6.64	0.547 ± 0.163	3.50 ± 0.70	7.36 ± 0.92	-1.95 ± 0.00	0.31 ± 0.00	0.75 ± 1.66	6.9
CMA-ES	-10.4 ± 3.0	-4.35 ± 6.18	0.448 ± 0.068	**3.74 ± 0.00**	6.95 ± 1.13	-1.95 ± 0.00	0.60 ± 0.29	-4.02 ± 21.8	7.1
Anneal	-13.2 ± 0.0	9.57 ± 0.66	0.439 ± 0.000	3.65 ± 0.04	7.41 ± 0.22	-1.95 ± 0.00	0.56 ± 0.00	**0.96 ± 0.08**	**6.0**
BO	-11.5 ± 2.3	-56.2 ± 91.9	0.552 ± 0.152	1.42 ± 0.00	5.80 ± 1.71	0.64 ± 0.01	0.46 ± 0.18	-36.9 ± 205	9.6
TuRBO	-16.3 ± 10.2	-24.3 ± 66.3	**0.563 ± 0.087**	1.42 ± 0.00	6.79 ± 1.25	**0.65 ± 0.00**	0.71 ± 0.01	-32.3 ± 94.9	7.9
BONET	-29.2 ± 2.2	**10.8 ± 0.43**	0.324 ± 0.041	**3.74 ± 0.00**	**8.70 ± 0.32**	0.56 ± 0.11	0.78 ± 0.00	—	**6.0**
DDOM	-1870 ± 2693	-7.10 ± 1.42	0.386 ± 0.224	1.43 ± 0.00	7.91 ± 0.29	**0.65 ± 0.01**	0.50 ± 0.19	-56.6 ± 79.6	9.6
COM	-3468 ± 679	-37.4 ± 23.0	0.346 ± 0.093	3.62 ± 0.00	5.26 ± 1.01	0.60 ± 0.04	**0.90 ± 0.01**	0.80 ± 0.93	9.6
RoMA	-18.5 ± 8.2	5.21 ± 1.39	0.500 ± 0.153	3.58 ± 0.11	6.94 ± 1.11	0.43 ± 0.18	0.41 ± 0.21	-2.44 ± 2.16	8.1
BDI	-109 ± 0.0	0.93 ± 0.88	0.471 ± 0.000	3.58 ± 0.05	5.62 ± 0.00	0.49 ± 0.00	0.76 ± 0.00	-24.8 ± 233	9.0
ExPT	-23.1 ± 11.3	-16.7 ± 25.1	0.480 ± 0.091	**3.74 ± 0.00**	6.70 ± 0.39	0.62 ± 0.04	0.75 ± 0.07	-0.40 ± 1.61	6.9
BootGen	—	-116.8 ± 85.7	0.388 ± 0.007	3.60 ± 0.04	7.74 ± 0.56	0.61 ± 0.03	—	—	8.8
ROMO	-3142 ± 330	-25.6 ± 23.1	0.354 ± 0.247	3.59 ± 0.08	5.49 ± 1.38	0.62 ± 0.04	0.42 ± 0.17	-2.77 ± 5.21	11.1
GAGA	-14.2 ± 15.2	-16.7 ± 81.1	0.546 ± 0.148	3.22 ± 0.86	6.40 ± 1.13	-1.95 ± 0.00	0.89 ± 0.01	0.24 ± 0.20	8.5
GABO	-12.7 ± 10.0	-12.2 ± 46.1	0.467 ± 0.066	3.56 ± 1.66	6.12 ± 1.22	0.61 ± 0.08	0.57 ± 0.17	0.02 ± 5.77	7.9

Table C.2: **GABO Adaptive SCR ablation study—Constrained budget ($k = 1$) suboptimal (90%-ile) oracle evaluation.** The oracle score of the 90th percentile design candidate according to the surrogate across 10 random seeds reported as mean ± standard deviation. $\mathcal{D}$ (best) reports the top oracle value in the task dataset. Task-averaged method rank is reported in the final column. *Denotes the life sciences-related discrete MBO tasks from Design-Bench (Trabucco et al., 2022).

GABO α Value	Branin	LogP	TF-Bind-8*	GFP*	UTR*	ChEMBL*	D'Kitty	Warfarin	Rank
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 ± 1.96	—
$\alpha = 0.0$	-11.5 ± 2.3	-56.2 ± 91.9	0.552 ± 0.152	1.42 ± 0.00	5.80 ± 1.71	**0.64 ± 0.01**	0.46 ± 0.18	-36.9 ± 205	3.3
$\alpha = 0.2$	-9.0 ± 2.6	-40.2 ± 77.4	**0.612 ± 0.114**	1.42 ± 0.00	5.81 ± 1.83	0.59 ± 0.13	0.49 ± 0.18	-51.7 ± 265	2.9
$\alpha = 0.5$	**-8.6 ± 4.4**	-90.1 ± 107.2	0.501 ± 0.109	1.65 ± 0.69	**6.64 ± 1.42**	0.52 ± 0.15	0.41 ± 0.16	-63.5 ± 336	3.9
$\alpha = 0.8$	-10.9 ± 2.1	-41.9 ± 82.5	0.433 ± 0.158	1.97 ± 0.88	4.89 ± 1.23	0.56 ± 0.15	0.38 ± 0.15	-48.5 ± 265	4.4
$\alpha = 1.0$	-104.6 ± 68.9	-77.1 ± 146.1	0.452 ± 0.179	2.05 ± 0.98	5.15 ± 1.51	0.60 ± 0.08	0.41 ± 0.16	-82.1 ± 552	4.5
aSCR	-12.7 ± 10.0	**-12.2 ± 46.1**	0.467 ± 0.066	**3.56 ± 1.66**	6.12 ± 1.22	0.61 ± 0.08	**0.57 ± 0.17**	**0.02 ± 5.77**	**2.1**

Separately, to further characterize the distribution of designs and their associated oracle scores proposed by GABO, **Figure C.1** plots a histogram of the oracle scores of (1) all 2,048 oracle scores, and (2) the oracle scores of the top 256 designs according to the penalized surrogate objective in (5.5) for the **LogP** task. Compared with the other optimization methods assessed, we notice that the range of oracle scores is larger for BO-based optimization methods. This helps motivate our design choice to leverage aSCR and **Algorithm 1** with BO-qEI, as BO is able to explore a larger region of the design space and is an effective parent optimizer for complex design spaces. Secondly, we also find that the distribution of scores is similar between BO-qEI and GABO, even though the performance of these two methods is remarkably different in **Tables 5.2** and **5.3**. This is likely due to the fact that while BO enables us to explore a larger effective region of the design space (compared with first-order iterative methods), **aSCR more accurately ranks proposed designs using the penalized surrogate so that we can identify promising candidates even in the low-budget oracle evaluation regime**.

C.2. Correlation of Offline Objective and Oracle Function Values

A key component of GABO with Adapative SCR critical to the above discussion in **Section C.1** is that generated designs score similarly according to the hidden oracle function and the regularized Lagrangian objective as in (5.5) in order to solve the problem of surrogate objective overestimation encountered in traditional offline optimization settings (**Fig. 1.3**). To assess this quantitatively, we computed the distance covariance $\text{dCov}_n[\{\mathcal{L}(\mathbf{x}_k;\lambda^*)\}_{k=1}^n, \{f(\mathbf{x}_k)\}_{k=1}^n]$ between the oracle scores $f(\mathbf{x}_k)$ and the constrained Lagrangian scores $\mathcal{L}(\mathbf{x}_k;\lambda^*)$ with $\lambda = \lambda^*(t)$ computed using our Adaptive SCR algorithm. The empirical distance covariance metric is computed over the $n = 2048$ design candidates generated using our GABO algorithm. Briefly, the distance covariance is a nonnegative measure of dependence between two vectors which may be related nonlinearly; a greater distance covariance implies a greater degree of association between observations (Székely et al., 2007). We focus our subsequent discussion on the Penalized **LogP** task.

Across five random seeds, GABO with Adaptive SCR achieves a distance covariance score of 0.535 $\pm$ 0.067 (mean $\pm$ standard deviation). In contrast, naïve BO-qEI (i.e., $\lambda = 0$) only achieves a

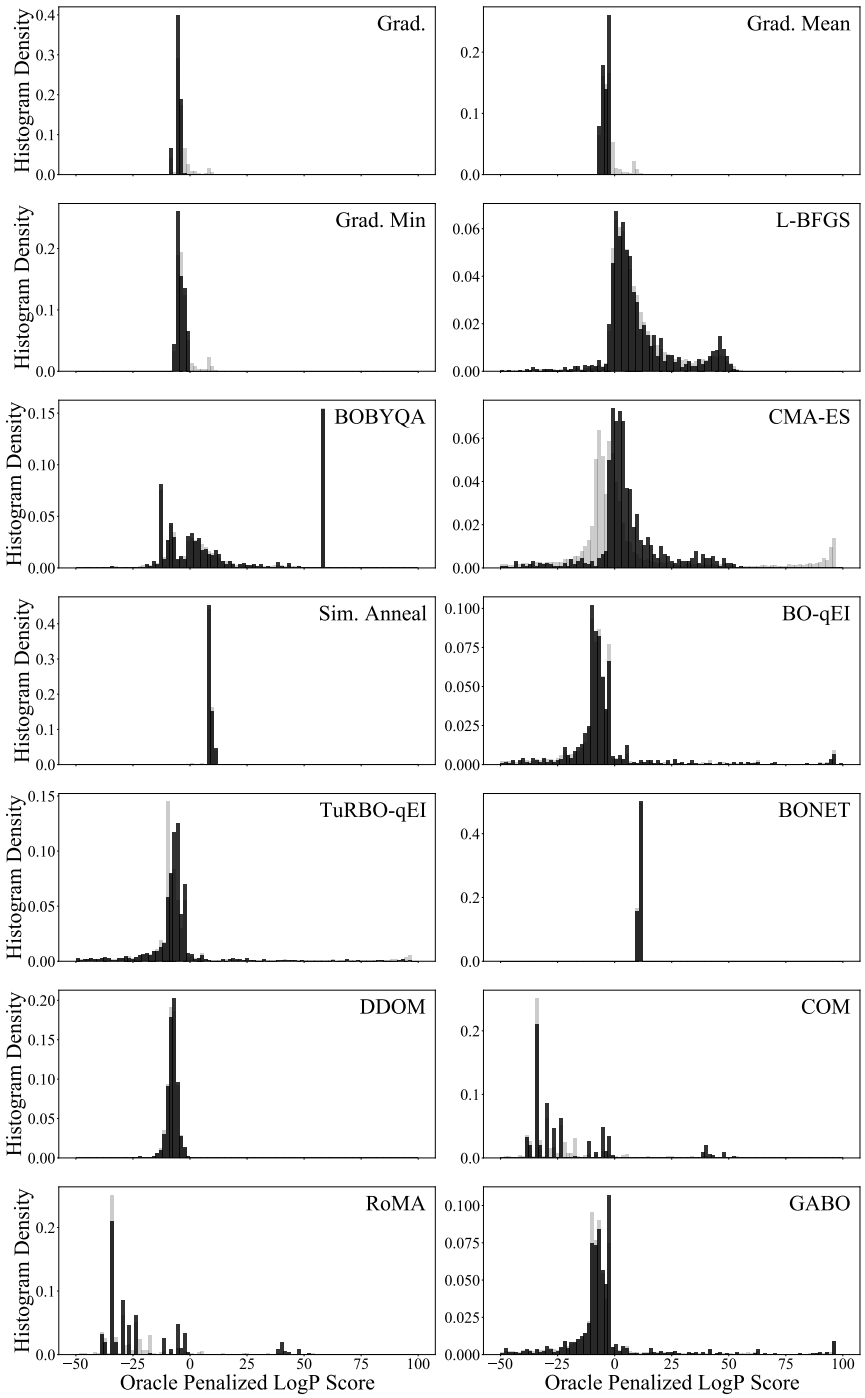

Figure C.1: **Distribution of oracle penalized LogP scores.** We plot the distribution of oracle scores for the top 128 surrogate model-ranked designs in black, and the distribution for all 2,048 generated designs in light gray for each of the offline model-based optimization methods assessed in our work across 10 random seeds. While GABO and BO-qEI have similar distributions, GABO is able to more reliably rank top-performing designs higher, such that these designs can be identified even under limited oracle query budgets.

distance covariance score of 0.392 ± 0.040. Using $p < 0.05$ as a cutoff for statistical significance, the distance covariance scores are significantly different between these two methods ($p \approx 0.006$, unpaired two-tailed t-test). These results help support our conclusion that GABO with Adaptive SCR is able to provide better estimates of design candidate performance according to the hidden oracle function when compared to the corresponding unconstrained BO policy.

C.3. GAGA Algorithm and Ablation Experimental Results

In our ablation experiments presented in **Table 5.4**, we showed how 'adaptive' nature of aSCR is an important component in solving the constrained optimization problem in (5.4) for GABO, and outperforms alternative approaches that manually hand-tune α (and hence λ) as a constant hyperparameter. We explore whether this conclusion also applies for GAGA as well here.

For clarity, we first offer the explicit formulation of GAGA in **Supp. Algorithm 4**. We ablate **Algorithm 1** in GAGA by instead evaluating our method using different values of $\lambda = \alpha/(1-\alpha)$. Setting $\alpha = 0$ (i.e., $\lambda = 0$) corresponds to naïvely performing gradient ascent against the unconstrained surrogate model; setting $\alpha = 1$ (i.e., $\lambda \to \infty$) is equivalent to a WGAN-like generative policy.

Algorithm 4 Generative Adversarial Gradient Ascent (GAGA)

Input: surrogate objective $f_\theta : \mathbb{R}^d \to \mathbb{R}$, offline dataset $\mathcal{D}_n = \{z'_j\}_{j=1}^n$, iterative sampling budget T, sampling batch size b, number of generator steps per source critic training $n_{\text{generator}}$, oracle query budget k, step size η
AdaptiveSCR Input: α step size $\Delta\alpha$, search budget $\mathcal{B}$, norm threshold τ
Define: Differentiable source critic $c : \mathbb{R}^d \to \mathbb{R}$
Define: Lagrangian $\mathcal{L}(z; \alpha) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} = -f_\theta(z) + \frac{\alpha}{1-\alpha}[\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - c(z)]$ // Eq. (5.5)
Sample $\mathcal{Z}^1 \leftarrow \{z_i^1\}_{i=1}^b$ as the top b designs in $\mathcal{D}_n$ according to their previously observed oracle scores
// Train the source critic per Lemma 1 to optimality:
$c \leftarrow \text{argmax}_{||c||_L \leq K} W_1(\mathcal{D}_n, \mathcal{Z}^1) = \text{argmax}_{||c||_L \leq K} [\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - \mathbb{E}_{z \sim \mathcal{Z}^1}[c(z)]]$
$\alpha \leftarrow$ **AdaptiveSCR**$(f_\theta, c, \mathcal{D}_n, \Delta\alpha, \mathcal{B}, \tau)$ // Alg. (1)
Evaluate candidates $\mathcal{Y}^1 \leftarrow \{y_i^1\}_{i=1}^b = \{-\mathcal{L}(z_i^1; \alpha)\}_{i=1}^b$
for t **in** $2, 3, \ldots, T$ **do**
  $\mathcal{Z}^t \leftarrow \{z_i^t\}_{i=1}^b = \{z_i^{t-1} - \eta \nabla_{z_i^{t-1}} \mathcal{L}(z_i^{t-1}; \alpha)\}_{i=1}^b$
  $\alpha \leftarrow$ **AdaptiveSCR**$(f_\theta, c, \mathcal{D}_n, \Delta\alpha, \mathcal{B}, \tau)$
  Evaluate samples $\mathcal{Y}^t \leftarrow \{y_i^t\}_{i=1}^b = \{-\mathcal{L}(z_i^t; \alpha)\}_{i=1}^b$
  if t mod $n_{\text{generator}}$ equals 0 **then**
    // Train the source critic per Lemma 1 to optimality:
    $c \leftarrow \text{argmax}_{||c||_L \leq K} W_1(\mathcal{D}_n, \mathcal{Z}^t) = \text{argmax}_{||c||_L \leq K} [\mathbb{E}_{z' \sim \mathcal{D}_n}[c(z')] - \mathbb{E}_{z \sim \mathcal{Z}^t}[c(z)]]$
  end if
end for
return the top k samples from the $T \times b$ observations $\mathcal{D}_T = \{\{(z_i^m, y_i^m)\}_{i=1}^b\}_{m=1}^T$ according to y_i^m

Our results are shown in **Supp. Table C.3**: similar to the analogous ablation results for GABO in **Table 5.4**, dynamically adjusting the strength of source critic regularization using our aSCR algorithm outperforms manually setting the value of α to a constant in both the one-shot $k = 1$ and few-shot $k = 128$ evaluation settings.

Table C.3: **GAGA Adaptive ACR ablation study.** We ablate the dynamic computation of α (and hence λ in (5.5)) by instead choosing to manually fix α to a constant value. A value of $\alpha = 0.0$ corresponds to naïve gradient ascent, and a value of $\alpha = 1.0$ corresponds to a WGAN-like generative policy. Oracle values are averaged across 10 random seeds and reported as mean $\pm$ standard deviation. In each evaluation setting, we rank all 2,048 proposed designs according to the penalized surrogate forward model in (5.5) and evaluate the top k designs using the oracle function, reporting the maximum out of the k oracle values. In the suboptimal evaluation setting, we report the oracle score of the single 90th percentile design according to the penalized surrogate ranking. **Bold** (resp., Underlined) entries indicate the best (resp., second best) entry in the column for the particular evaluation metric. *Denotes the life sciences MBO tasks from Design-Bench (Trabucco et al., 2022).

	Branin	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**	**Rank**
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96	—
Constrained Budget ($k = 1$) **Oracle Evaluation**									
$\alpha = 0.0$	-245.1 $\pm$ 81.3	**-5.37 $\pm$ 1.44**	0.429 $\pm$ 0.023	3.18 $\pm$ 0.88	6.82 $\pm$ 0.21	-1.95 $\pm$ 0.00	0.57 $\pm$ 0.19	**0.86 $\pm$ 1.09**	3.6
$\alpha = 0.2$	-13.7 $\pm$ 0.0	-70.3 $\pm$ 115.8	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	**7.73 $\pm$ 0.46**	-1.95 $\pm$ 0.00	0.88 $\pm$ 0.00	-0.17 $\pm$ 0.00	2.5
$\alpha = 0.5$	-13.7 $\pm$ 0.0	-70.3 $\pm$ 114.7	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	6.75 $\pm$ 0.72	-1.95 $\pm$ 0.00	0.88 $\pm$ 0.00	0.44 $\pm$ 0.00	2.4
$\alpha = 0.8$	-13.7 $\pm$ 0.0	-84.6 $\pm$ 115.8	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	6.75 $\pm$ 0.72	-1.95 $\pm$ 0.00	0.88 $\pm$ 0.00	0.44 $\pm$ 0.00	2.6
$\alpha = 1.0$	-14.4 $\pm$ 1.5	-27.8 $\pm$ 99.8	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	5.88 $\pm$ 1.04	-1.95 $\pm$ 0.00	**0.89 $\pm$ 0.00**	-8.61 $\pm$ 6.15	3.4
aSCR	**-2.9 $\pm$ 2.2**	-68.6 $\pm$ 109.8	**0.571 $\pm$ 0.120**	**3.74 $\pm$ 0.00**	5.89 $\pm$ 1.42	-1.95 $\pm$ 0.00	**0.89 $\pm$ 0.00**	0.01 $\pm$ 0.14	**2.3**
Relaxed Budget ($k = 128$) **Oracle Evaluation**									
$\alpha = 0.0$	-115.3 $\pm$ 20.8	-5.14 $\pm$ 1.70	**0.977 $\pm$ 0.025**	3.49 $\pm$ 0.69	7.38 $\pm$ 0.15	-1.95 $\pm$ 0.00	0.87 $\pm$ 0.02	0.86 $\pm$ 1.08	4.8
$\alpha = 0.2$	-13.2 $\pm$ 0.0	4.70 $\pm$ 10.3	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	7.92 $\pm$ 0.24	-1.95 $\pm$ 0.00	**0.95 $\pm$ 0.00**	**1.00 $\pm$ 0.00**	2.4
$\alpha = 0.5$	-13.2 $\pm$ 0.0	5.07 $\pm$ 4.56	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	7.77 $\pm$ 0.21	-1.95 $\pm$ 0.00	**0.95 $\pm$ 0.01**	**1.00 $\pm$ 0.00**	2.9
$\alpha = 0.8$	-13.2 $\pm$ 0.0	5.13 $\pm$ 4.28	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	7.44 $\pm$ 0.30	-1.95 $\pm$ 0.00	**0.95 $\pm$ 0.01**	**1.00 $\pm$ 0.00**	2.9
$\alpha = 1.0$	-13.1 $\pm$ 0.0	5.11 $\pm$ 4.11	0.445 $\pm$ 0.017	**3.74 $\pm$ 0.00**	7.40 $\pm$ 0.28	-1.95 $\pm$ 0.00	0.90 $\pm$ 0.01	0.96 $\pm$ 0.05	3.0
aSCR	**-1.0 $\pm$ 0.2**	**14.1 $\pm$ 25.0**	0.722 $\pm$ 0.091	**3.74 $\pm$ 0.00**	**7.98 $\pm$ 0.36**	-1.95 $\pm$ 0.00	0.90 $\pm$ 0.01	0.95 $\pm$ 0.07	**2.1**
Constrained Budget ($k = 1$) **Suboptimal (90%-ile) Oracle Evaluation**									
$\alpha = 0.0$	-94.4 $\pm$ 20.9	**-5.47 $\pm$ 1.32**	0.429 $\pm$ 0.023	3.43 $\pm$ 0.67	**7.16 $\pm$ 0.21**	-1.95 $\pm$ 0.00	0.53 $\pm$ 0.20	0.87 $\pm$ 1.08	3.6
$\alpha = 0.2$	-18.1 $\pm$ 0.5	-10.9 $\pm$ 14.9	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	6.57 $\pm$ 0.94	-1.95 $\pm$ 0.00	0.89 $\pm$ 0.02	**0.97 $\pm$ 0.04**	2.5
$\alpha = 0.5$	-16.2 $\pm$ 0.6	-15.2 $\pm$ 14.4	0.445 $\pm$ 0.017	**3.74 $\pm$ 0.00**	6.75 $\pm$ 1.18	-1.95 $\pm$ 0.00	**0.90 $\pm$ 0.02**	0.93 $\pm$ 0.18	**2.4**
$\alpha = 0.8$	-15.7 $\pm$ 1.0	-12.7 $\pm$ 13.8	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	6.84 $\pm$ 1.29	-1.95 $\pm$ 0.00	0.88 $\pm$ 0.01	-0.24 $\pm$ 2.89	3.1
$\alpha = 1.0$	-14.6 $\pm$ 1.4	-16.9 $\pm$ 13.1	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	6.82 $\pm$ 1.01	-1.95 $\pm$ 0.00	0.89 $\pm$ 0.01	-2.71 $\pm$ 7.71	3.3
aSCR	**-14.2 $\pm$ 15.2**	-16.7 $\pm$ 81.1	**0.546 $\pm$ 0.148**	3.22 $\pm$ 0.86	6.40 $\pm$ 1.13	-1.95 $\pm$ 0.00	0.89 $\pm$ 0.01	0.24 $\pm$ 0.20	3.5

C.4. Dynamic Re-Training of the Adversarial Source Critic

In **Algorithm 2** and **Supp. Algorithm 4**, we describe how generative adversarial optimization alternates between batched acquisition steps according to the optimizer and re-training the source

critic on the newly sampled trajectory points. To better interrogate the significance of dynamically re-training the source critic during optimization, we compare the performance of the default GABO and GAGA algorithms (with $n_{\text{generator}} = 4$ as the number of acquisition steps per critic re-training step) against the respective methods without source critic re-training (i.e., $n_{\text{generator}} = \infty$) in **Supp. Table C.4**. Across all three evaluation metrics and all eight tasks, dynamically retraining the source-critic improves upon the performance of the GABO when $n_{\text{generator}} = \infty$ by 67.4% in the top-1 evaluation metric; 0.0% in the top-128 evaluation metric; and 33.5% in the 90%-ile evaluation metric. Intuitively, these results align with the value of the source critic in being able to implicitly set the value of the regularization strength α in (5.5) according to the sampled trajectory points—especially in the constrained budget oracle evaluation setting.

Interestingly, we do not observe similar performance improvements with dynamic re-training of the source critic in GAGA. Qualitatively, we find that this is because of the iterative first-order nature of the parent gradient ascent algorithm—because the sampled designs are clustered in the same regions of the design space over the course of optimization, the energy landscape of the penalized surrogate (i.e., the negative of the Lagrangian expression in (5.5)) does not change significantly during source critic re-training. This reinforces the optimizer to stay in roughly the same regions of the design space.

C.5. Empirical Convergence Analysis

For all of our experimental results, we restrict the surrogate query budget to a total of 2048 allowed offline surrogate model queries in order to ensure a fair comparison between different optimization methods. To ensure that such a budget is sufficient for optimizer convergence across different methods evaluated, we plot the best achieved oracle Penalized LogP value (i.e., assuming an unlimited oracle evaluation budget) as a function of the number of optimizer surrogate queries (**Supp. Fig. C.2**) for the Penalized LogP task. These results show that our methods are indeed able to converge over the course of the optimization trajectory.

Table C.4: **Ablating dynamic updates to the source critic.** We study the effect of training the source critic model *exactly once* (i.e., setting $n_{\text{generator}} = \infty$ in **Algorithm 2** and **Supp. Algorithm 4**) as opposed to re-training the source critic model every $n_{\text{generator}} = 4$ acquisition steps on the newly sampled designs. Oracle values are averaged across 10 random seeds and reported as mean $\pm$ standard deviation. In each evaluation setting, we rank all 2,048 proposed designs according to the penalized surrogate forward model in (5.5) and evaluate the top k designs using the oracle function, reporting the maximum out of the k oracle values. In the suboptimal evaluation setting, we report the oracle score of the single 90th percentile design according to the penalized surrogate ranking. **Bold** entries indicate the best entry in the column for the particular optimizer and evaluation metric. *Denotes the life sciences MBO tasks from Design-Bench (Trabucco et al., 2022).

GABO	**Branin**	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96
Constrained Budget ($k = 1$) **Oracle Evaluation**								
$n_{\text{generator}} = \infty$	-3.5 $\pm$ 2.5	-55.6 $\pm$ 52.1	**0.577 $\pm$ 0.151**	**3.74 $\pm$ 0.00**	6.73 $\pm$ 1.10	**0.65 $\pm$ 0.00**	0.46 $\pm$ 0.18	-0.27 $\pm$ 13.7
$n_{\text{generator}} = 4$	**-2.6 $\pm$ 1.1**	**21.3 $\pm$ 33.2**	0.570 $\pm$ 0.131	3.60 $\pm$ 0.40	**7.51 $\pm$ 0.39**	0.60 $\pm$ 0.07	**0.71 $\pm$ 0.01**	**0.60 $\pm$ 1.80**
Relaxed Budget ($k = 128$) **Oracle Evaluation**								
$n_{\text{generator}} = \infty$	-0.5 $\pm$ 0.1	**128.0 $\pm$ 19.5**	0.946 $\pm$ 0.035	3.74 $\pm$ 0.00	**8.38 $\pm$ 0.11**	0.67 $\pm$ 0.01	0.72 $\pm$ 0.00	1.00 $\pm$ 0.00
$n_{\text{generator}} = 4$	-0.5 $\pm$ 0.1	122.1 $\pm$ 20.6	**0.954 $\pm$ 0.025**	3.74 $\pm$ 0.00	8.36 $\pm$ 0.08	**0.70 $\pm$ 0.01**	0.72 $\pm$ 0.00	1.00 $\pm$ 0.03
Constrained Budget ($k = 1$) **Suboptimal (90%-ile) Oracle Evaluation**								
$n_{\text{generator}} = \infty$	**-8.9 $\pm$ 6.6**	-54.1 $\pm$ 62.6	**0.471 $\pm$ 0.061**	3.06 $\pm$ 1.04	6.02 $\pm$ 1.41	**0.63 $\pm$ 0.07**	0.26 $\pm$ 0.62	-5.32 $\pm$ 4.59
$n_{\text{generator}} = 4$	-12.7 $\pm$ 10.0	**-12.2 $\pm$ 46.1**	0.467 $\pm$ 0.066	**3.56 $\pm$ 1.66**	**6.12 $\pm$ 1.22**	0.61 $\pm$ 0.08	**0.57 $\pm$ 0.17**	**0.02 $\pm$ 5.77**

GAGA	**Branin**	**LogP**	**TF-Bind-8***	**GFP***	**UTR***	**ChEMBL***	**D'Kitty**	**Warfarin**
$\mathcal{D}$ (best)	-13.0	11.3	0.439	3.53	7.12	0.61	0.88	-0.19 $\pm$ 1.96
Constrained Budget ($k = 1$) **Oracle Evaluation**								
$n_{\text{generator}} = \infty$	-14.6 $\pm$ 0.8	**-1.87 $\pm$ 14.9**	0.439 $\pm$ 0.000	3.74 $\pm$ 0.00	**6.45 $\pm$ 0.54**	-1.95 $\pm$ 0.00	0.88 $\pm$ 0.00	-0.17 $\pm$ 0.29
$n_{\text{generator}} = 4$	**-2.9 $\pm$ 2.2**	-68.6 $\pm$ 109.8	**0.571 $\pm$ 0.120**	3.74 $\pm$ 0.00	5.89 $\pm$ 1.42	-1.95 $\pm$ 0.00	**0.89 $\pm$ 0.00**	**0.01 $\pm$ 0.14**
Relaxed Budget ($k = 128$) **Oracle Evaluation**								
$n_{\text{generator}} = \infty$	-13.3 $\pm$ 0.2	**50.2 $\pm$ 2.48**	0.439 $\pm$ 0.000	3.74 $\pm$ 0.00	7.38 $\pm$ 0.31	-1.95 $\pm$ 0.00	0.90 $\pm$ 0.01	**0.99 $\pm$ 0.01**
$n_{\text{generator}} = 4$	**-1.0 $\pm$ 0.2**	14.1 $\pm$ 25.0	**0.722 $\pm$ 0.091**	3.74 $\pm$ 0.00	**7.98 $\pm$ 0.36**	-1.95 $\pm$ 0.00	0.90 $\pm$ 0.01	0.95 $\pm$ 0.07
Constrained Budget ($k = 1$) **Suboptimal (90%-ile) Oracle Evaluation**								
$n_{\text{generator}} = \infty$	-17.0 $\pm$ 1.6	**5.88 $\pm$ 4.88**	0.439 $\pm$ 0.000	**3.74 $\pm$ 0.00**	**7.08 $\pm$ 0.73**	-1.95 $\pm$ 0.00	0.89 $\pm$ 0.01	-1.38 $\pm$ 1.68
$n_{\text{generator}} = 4$	**-14.2 $\pm$ 15.2**	-16.7 $\pm$ 81.1	**0.546 $\pm$ 0.148**	3.22 $\pm$ 0.86	6.40 $\pm$ 1.13	-1.95 $\pm$ 0.00	0.89 $\pm$ 0.01	**0.24 $\pm$ 0.20**

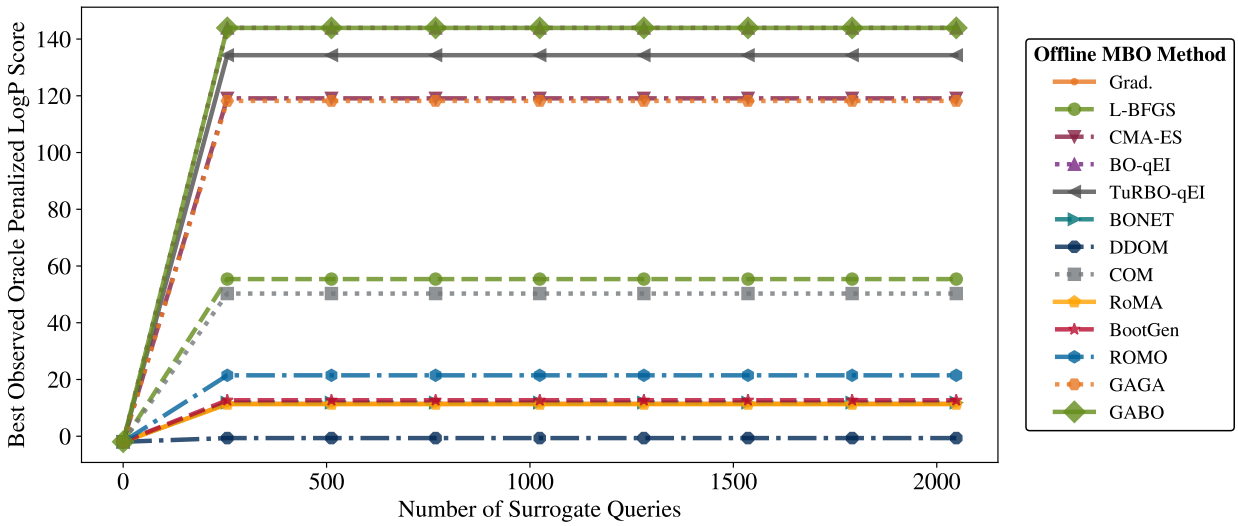

Figure C.2: **Best oracle penalized LogP value versus optimization step count.** We plot the best Penalized LogP score averaged across 10 random seeds as a function of the number of surrogate queries made over the optimization trajectory. All offline model-based optimization (MBO) methods assessed consistently converge within the allowed oracle query budget used in our experimental setup as described in **Section 5.5**.

APPENDIX D

Obtaining Diverse and High-Quality Designs in Offline Optimization: Additional Experimental Results

The following appendix contains additional experimental results for the interested reader using **DynAMO**—our novel algorithm to generate both high-quality and diverse sets of designs in offline optimization.

D.1. Additional Design Quality Results

We supplement the results shown in **Table 6.1** with the raw Best@128 oracle quality scores reported for each of the 6 tasks in our evaluation suite in **Supp. Table D.1**.

In **Section 6.4**, we define the **Best@**k score to evaluate the quality of observed designs according to a hidden oracle function used for evaluation of candidate fitness. Achieving a high Best@$B = k$ score ensures that a desirable design is found. Consistent with prior work on batched optimization methods (Trabucco et al., 2021; Krishnamoorthy et al., 2023b,a), we are also interested in the **Median@**k score defined as

$$\text{Median@}k(\{x_i^F\}_{i=1}^k) := \text{median}_{1 \leq i \leq k}\ r(x_i^F) \tag{D.1}$$

to evaluate whether a *batch* of candidate designs (as opposed to any singular design) is generally of high quality according to the oracle $r(x)$. We report the Median@k score for $k = 128$ in **Supp. Table D.2**; in general, we find that DynAMO does not perform as well as other objective-modifying baseline methods according to this metric. However, we note that in many applications of offline optimization, we are often not as interested in how the median design performs, but rather if we are able to discover optimal and near-optimal designs. For this reason, we chose to focus on the Best@128 oracle scores in **Table 6.1** to evaluate the quality of designs proposed by an optimizer in our main results. Nonetheless, future work may explore how to better tune DynAMO (e.g., the τ and β hyperparameters in **Algorithm 3**) to achieve more desirable Median@128 scores.

Table D.1: **Quality and diversity of designs under MBO objective transforms (full).** We evaluate DynAMO against other MBO objective-modifying methods using six different backbone optimizers. Each cell consists of '**Best@128**/**Pairwise Diversity**' oracle scores separated by a forward slash. Both metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Dataset** $\mathcal{D}$ reports the maximum oracle score and mean pairwise diversity in the offline dataset. **Bolded** entries indicate overlapping 95% confidence intervals with the best performing algorithm (according to the mean) per optimizer. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Grad.	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Dataset $\mathcal{D}$	43.9/65.9	59.4/57.3	60.5/60.0	88.9/36.7	40.0/66.0	88.4/85.7	—/—	—/—
Baseline	90.0$^{(4.3)}$/12.5$^{(8.0)}$	**80.9$^{(12.1)}$**/7.8$^{(8.8)}$	60.2$^{(8.9)}$/7.9$^{(7.8)}$	88.8$^{(4.0)}$/24.1$^{(13.3)}$	36.0$^{(6.8)}$/0.0$^{(0.0)}$	65.6$^{(14.5)}$/0.0$^{(0.0)}$	5.0/5.5	6.8/-53.2
COMs$^-$	60.4$^{(9.8)}$/10.4$^{(8.7)}$	60.2$^{(12.4)}$/7.5$^{(9.2)}$	60.2$^{(8.8)}$/7.9$^{(7.5)}$	88.4$^{(4.0)}$/24.8$^{(10.0)}$	22.5$^{(3.2)}$/0.0$^{(0.0)}$	71.2$^{(10.7)}$/0.0$^{(0.0)}$	7.3/6.5	-3.0/-53.5
COMs$^+$	93.1$^{(3.4)}$/**66.6$^{(1.0)}$**	67.0$^{(0.9)}$/57.4$^{(0.2)}$	**64.6$^{(1.0)}$**/**81.6$^{(4.9)}$**	**97.1$^{(1.6)}$**/3.8$^{(0.9)}$	41.2$^{(4.8)}$/99.5$^{(2.6)}$	91.8$^{(0.9)}$/21.1$^{(23.5)}$	**2.5**/2.8	12.3/-6.9
RoMA$^-$	62.0$^{(10.7)}$/12.3$^{(8.3)}$	60.9$^{(12.1)}$/7.9$^{(8.9)}$	60.2$^{(8.8)}$/7.7$^{(7.6)}$	88.8$^{(4.0)}$/24.2$^{(13.3)}$	36.0$^{(6.8)}$/0.0$^{(0.0)}$	65.6$^{(14.5)}$/0.0$^{(0.0)}$	6.7/5.7	-1.2/-53.3
RoMA$^+$	66.5$^{(0.0)}$/20.3$^{(0.7)}$	77.8$^{(0.0)}$/3.8$^{(0.0)}$	63.3$^{(0.0)}$/6.2$^{(0.0)}$	84.5$^{(0.0)}$/1.8$^{(0.0)}$	**49.0$^{(1.6)}$**/54.1$^{(1.4)}$	**95.2$^{(1.2)}$**/4.9$^{(0.0)}$	3.8/5.8	9.2/-46.8
ROMO	**98.1$^{(0.7)}$**/**62.1$^{(0.8)}$**	66.8$^{(1.0)}$/57.1$^{(0.1)}$	63.0$^{(0.8)}$/53.9$^{(0.6)}$	91.8$^{(0.9)}$/48.7$^{(0.1)}$	38.7$^{(2.5)}$/51.7$^{(3.2)}$	87.8$^{(0.9)}$/22.1$^{(5.5)}$	4.2/2.8	10.9/-12.7
GAMBO	73.1$^{(12.8)}$/17.3$^{(12.8)}$	**77.1$^{(9.6)}$**/11.2$^{(10.3)}$	**64.4$^{(1.5)}$**/6.9$^{(7.7)}$	**92.8$^{(8.0)}$**/22.1$^{(10.5)}$	**46.0$^{(6.8)}$**/0.0$^{(0.0)}$	**90.6$^{(14.5)}$**/1.5$^{(3.2)}$	3.2/5.3	10.5/-52.1
DynAMO	90.3$^{(4.7)}$/**66.9$^{(6.9)}$**	**86.2$^{(0.0)}$**/**68.2$^{(1.8)}$**	**64.4$^{(2.5)}$**/**77.2$^{(2.2)}$**	91.2$^{(0.0)}$/**93.0$^{(1.2)}$**	**44.2$^{(7.8)}$**/**129$^{(5.5)}$**	89.8$^{(3.2)}$/**104$^{(5.6)}$**	2.8/**1.2**	**14.2**/**27.8**

Adam	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	62.9$^{(13.0)}$/12.0$^{(12.3)}$	69.7$^{(10.5)}$/11.0$^{(12.1)}$	**62.9$^{(1.9)}$**/4.8$^{(3.8)}$	**92.3$^{(8.9)}$**/16.8$^{(12.4)}$	37.8$^{(6.3)}$/6.4$^{(14.5)}$	58.4$^{(18.5)}$/6.2$^{(14.0)}$	4.5/6.0	0.5/-52.4
COMs$^-$	62.9$^{(13.0)}$/13.6$^{(12.2)}$	65.1$^{(11.0)}$/11.0$^{(10.6)}$	**62.9$^{(1.9)}$**/5.0$^{(3.8)}$	92.4$^{(1.0)}$/21.2$^{(18.2)}$	22.5$^{(3.2)}$/0.0$^{(0.0)}$	57.3$^{(19.5)}$/6.3$^{(14.3)}$	6.0/5.3	-3.0/-52.4
COMs$^+$	**95.6$^{(2.6)}$**/44.2$^{(1.5)}$	67.1$^{(0.6)}$/57.4$^{(0.2)}$	**64.6$^{(0.9)}$**/**81.5$^{(5.7)}$**	**95.3$^{(1.9)}$**/3.7$^{(1.3)}$	39.6$^{(5.8)}$/79.3$^{(3.3)}$	67.1$^{(9.5)}$/31.8$^{(34.5)}$	3.2/3.0	8.1/-12.3
RoMA$^-$	62.9$^{(13.0)}$/12.3$^{(12.4)}$	69.7$^{(10.5)}$/10.9$^{(12.0)}$	**62.9$^{(1.9)}$**/4.7$^{(3.8)}$	84.7$^{(0.0)}$/16.8$^{(12.4)}$	37.8$^{(6.3)}$/6.4$^{(14.5)}$	58.4$^{(1.9)}$/6.2$^{(14.0)}$	4.5/6.3	0.5/-52.4
RoMA$^+$	**96.5$^{(0.0)}$**/21.3$^{(0.3)}$	77.8$^{(0.0)}$/3.8$^{(0.0)}$	63.3$^{(0.0)}$/5.9$^{(0.2)}$	**92.3$^{(8.9)}$**/1.8$^{(0.0)}$	**49.8$^{(1.4)}$**/49.4$^{(6.1)}$	**95.7$^{(1.6)}$**/14.8$^{(0.6)}$	**2.8**/5.2	**14.5**/-45.8
ROMO	95.6$^{(0.0)}$/**55.7$^{(0.3)}$**	67.0$^{(0.2)}$/56.3$^{(0.1)}$	63.3$^{(0.0)}$/53.5$^{(0.1)}$	90.4$^{(0.0)}$/50.7$^{(0.0)}$	31.8$^{(3.1)}$/25.5$^{(20.3)}$	71.0$^{(0.6)}$/7.2$^{(3.9)}$	4.8/2.8	6.4/-20.5
GAMBO	94.0$^{(2.2)}$/15.1$^{(11.2)}$	60.0$^{(12.6)}$/10.3$^{(11.5)}$	**60.9$^{(8.7)}$**/12.1$^{(11.3)}$	**91.4$^{(6.3)}$**/19.6$^{(15.2)}$	37.8$^{(6.3)}$/0.3$^{(0.8)}$	**88.4$^{(13.8)}$**/2.6$^{(3.9)}$	5.3/5.8	8.6/-51.9
DynAMO	**95.2$^{(1.7)}$**/**54.8$^{(8.9)}$**	**86.2$^{(0.0)}$**/**72.3$^{(3.4)}$**	**65.2$^{(1.1)}$**/**84.8$^{(9.2)}$**	91.2$^{(0.0)}$/**89.9$^{(5.3)}$**	**45.5$^{(5.7)}$**/**158$^{(37.3)}$**	**84.9$^{(12.0)}$**/**126$^{(5.7)}$**	**2.8**/**1.2**	**14.5**/**35.7**

CMA-ES	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**87.6$^{(8.3)}$**/47.2$^{(11.2)}$	**86.2$^{(0.0)}$**/44.6$^{(15.9)}$	**66.1$^{(1.0)}$**/**93.5$^{(2.0)}$**	106$^{(5.9)}$/66.2$^{(9.4)}$	**49.0$^{(1.0)}$**/12.8$^{(0.6)}$	72.2$^{(0.1)}$/164$^{(10.6)}$	3.7/3.8	14.4/9.5
COMs$^-$	75.6$^{(10.2)}$/46.0$^{(17.6)}$	**85.7$^{(1.3)}$**/**56.2$^{(15.8)}$**	**64.8$^{(1.0)}$**/63.1$^{(23.0)}$**	**119$^{(3.3)}$**/58.8$^{(24.2)}$**	18.8$^{(7.9)}$/22.0$^{(7.9)}$	62.9$^{(2.1)}$/67.2$^{(8.0)}$	5.7/5.2	7.6/-9.7
COMs$^+$	68.0$^{(6.0)}$/24.8$^{(11.3)}$	**77.2$^{(9.7)}$**/35.4$^{(16.5)}$	63.6$^{(0.5)}$/36.7$^{(9.0)}$	**116$^{(5.6)}$**/45.8$^{(16.1)}$**	36.8$^{(3.5)}$/0.0$^{(0.0)}$	**62.2$^{(15.5)}$**/0.0$^{(0.0)}$**	7.3/7.8	7.1/-38.2
RoMA$^-$	**87.6$^{(8.3)}$**/46.7$^{(11.2)}$**	**86.2$^{(0.0)}$**/44.8$^{(15.8)}$**	**66.1$^{(1.0)}$**/**93.5$^{(2.1)}$**	106$^{(5.9)}$/66.2$^{(9.4)}$	**49.0$^{(1.0)}$**/12.8$^{(0.6)}$**	72.2$^{(0.1)}$/164$^{(10.6)}$**	3.7/3.5	14.4/9.5
RoMA$^+$	**85.9$^{(7.0)}$**/53.1$^{(15.0)}$**	79.8$^{(3.7)}$/31.9$^{(15.2)}$**	**64.6$^{(1.1)}$**/60.5$^{(14.9)}$**	**118$^{(6.6)}$**/63.7$^{(21.6)}$**	**44.6$^{(3.2)}$**/98.2$^{(18.9)}$**	72.2$^{(0.1)}$/112$^{(86.4)}$**	5.0/4.8	14.1/8.0
ROMO	**88.3$^{(6.0)}$**/57.5$^{(11.6)}$**	**86.2$^{(0.0)}$**/40.2$^{(13.1)}$**	**64.5$^{(0.9)}$**/66.5$^{(13.5)}$**	113$^{(6.0)}$/70.2$^{(11.5)}$**	**45.7$^{(1.3)}$**/97.7$^{(15.4)}$**	**77.3$^{(3.2)}$**/20.9$^{(40.9)}$**	4.2/4.3	15.7/-3.1
GAMBO	**90.4$^{(4.4)}$**/39.6$^{(15.5)}$**	**86.2$^{(0.0)}$**/53.4$^{(8.4)}$**	**66.2$^{(1.6)}$**/84.8$^{(4.8)}$**	**121$^{(0.0)}$**/61.3$^{(14.6)}$**	**45.2$^{(3.5)}$**/**173$^{(19.4)}$**	72.2$^{(0.1)}$/59.9$^{(19.6)}$**	**2.2**/4.3	16.7/16.8
DynAMO	**89.8$^{(3.6)}$**/**73.6$^{(0.6)}$**	**85.7$^{(5.8)}$**/**73.1$^{(3.1)}$**	63.9$^{(0.9)}$/72.0$^{(3.1)}$	**117$^{(6.7)}$**/**94.0$^{(0.5)}$**	**50.6$^{(4.8)}$**/97.8$^{(13.2)}$**	**78.5$^{(5.5)}$**/**292$^{(83.5)}$**	3.3/**1.8**	**17.5**/**55.2**

CoSyNE	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	61.7$^{(10.0)}$/**5.6$^{(5.0)}$**	57.3$^{(9.6)}$/**12.7$^{(9.8)}$**	**63.6$^{(0.4)}$**/**28.2$^{(11.3)}$**	94.8$^{(10.1)}$/**12.2$^{(7.3)}$**	**37.0$^{(4.1)}$**/0.0$^{(0.0)}$**	62.7$^{(1.3)}$/0.0$^{(0.0)}$**	5.3/4.5	-0.6/-52.1
COMs$^-$	70.1$^{(12.8)}$/**16.5$^{(8.5)}$**	**73.2$^{(8.9)}$**/8.7$^{(10.2)}$**	**63.5$^{(0.4)}$**/**19.3$^{(14.1)}$**	88.4$^{(11.7)}$/**17.4$^{(4.1)}$**	**28.6$^{(5.5)}$**/0.0$^{(0.0)}$**	63.8$^{(1.6)}$/0.0$^{(0.0)}$**	5.7/5.7	1.1/-51.6
COMs$^+$	66.9$^{(5.1)}$/**15.4$^{(4.9)}$**	56.2$^{(9.5)}$/**26.5$^{(6.8)}$**	**63.3$^{(0.0)}$**/18.5$^{(7.0)}$**	**117$^{(3.6)}$**/**17.8$^{(8.1)}$**	**28.0$^{(7.8)}$**/**34.3$^{(13.5)}$**	**70.6$^{(1.1)}$**/16.6$^{(9.0)}$**	5.7/3.3	3.5/-40.4
RoMA$^-$	61.7$^{(10.0)}$/**5.9$^{(5.2)}$**	57.3$^{(9.6)}$/**12.7$^{(9.9)}$**	**63.6$^{(0.4)}$**/**27.7$^{(10.9)}$**	94.8$^{(10.1)}$/**12.2$^{(7.3)}$**	**37.0$^{(4.1)}$**/0.0$^{(0.0)}$**	62.7$^{(1.3)}$/0.0$^{(0.0)}$**	5.3/4.5	-0.6/-52.2
RoMA$^+$	70.4$^{(7.2)}$/**17.7$^{(5.5)}$**	**77.8$^{(4.4)}$**/11.9$^{(4.0)}$**	**64.4$^{(2.5)}$**/19.8$^{(3.9)}$**	**117$^{(6.5)}$**/**10.2$^{(4.4)}$**	**38.0$^{(8.1)}$**/0.0$^{(0.0)}$**	50.7$^{(1.5)}$/0.0$^{(0.0)}$**	2.8/5.0	6.2/-52.0
ROMO	**79.7$^{(12.7)}$**/**15.7$^{(10.1)}$**	62.0$^{(9.8)}$/5.8$^{(4.6)}$**	**64.1$^{(0.6)}$**/**27.6$^{(12.4)}$**	90.8$^{(3.9)}$/**17.3$^{(9.8)}$**	**30.6$^{(4.5)}$**/0.0$^{(0.0)}$**	**72.1$^{(1.1)}$**/0.1$^{(0.2)}$**	4.2/5.2	3.1/-50.8
GAMBO	**79.8$^{(10.6)}$**/**5.2$^{(5.7)}$**	**68.0$^{(12.5)}$**/9.1$^{(9.0)}$**	**64.2$^{(0.9)}$**/**28.4$^{(15.7)}$**	**99.4$^{(15.0)}$**/**7.1$^{(8.0)}$**	**37.0$^{(4.1)}$**/0.0$^{(0.0)}$**	62.7$^{(1.3)}$/0.0$^{(0.0)}$**	3.7/6.3	5.0/-53.6
DynAMO	**91.3$^{(4.4)}$**/**18.1$^{(13.0)}$**	**77.2$^{(11.6)}$**/**20.3$^{(2.3)}$**	**63.9$^{(0.9)}$**/**35.0$^{(17.9)}$**	**114$^{(7.0)}$**/**22.8$^{(11.9)}$**	**40.6$^{(8.6)}$**/**74.4$^{(46.3)}$**	67.5$^{(1.4)}$/**77.0$^{(35.9)}$**	**2.3**/**1.2**	**12.3**/**-20.7**

BO-qEI	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**87.3$^{(5.8)}$**/73.7$^{(0.6)}$**	**86.2$^{(0.0)}$**/**73.8$^{(0.5)}$**	**65.4$^{(1.0)}$**/99.3$^{(0.1)}$**	116$^{(3.1)}$/**93.0$^{(0.5)}$**	**53.1$^{(3.3)}$**/190$^{(0.8)}$**	**84.4$^{(0.9)}$**/124$^{(7.4)}$**	5.8/5.2	18.7/47.1
COMs$^-$	**93.2$^{(2.7)}$**/73.7$^{(0.7)}$**	**86.2$^{(0.0)}$**/**74.3$^{(0.5)}$**	**66.4$^{(0.4)}$**/**99.3$^{(0.1)}$**	**121$^{(0.0)}$**/**93.2$^{(0.4)}$**	43.2$^{(5.1)}$/192$^{(10.2)}$**	**86.2$^{(0.9)}$**/147$^{(8.9)}$**	4.3/4.0	19.2/51.4
COMs$^+$	84.5$^{(5.5)}$/68.8$^{(0.8)}$**	**85.6$^{(0.8)}$**/71.8$^{(0.3)}$**	**65.1$^{(0.8)}$**/96.6$^{(0.9)}$**	**121$^{(0.0)}$**/90.8$^{(0.8)}$**	**47.3$^{(4.1)}$**/**206$^{(1.2)}$**	**84.9$^{(1.1)}$**/79.0$^{(2.8)}$**	6.0/5.7	17.9/40.3
RoMA$^-$	**95.2$^{(2.2)}$**/74.1$^{(0.4)}$**	**86.3$^{(0.1)}$**/**74.1$^{(0.3)}$**	**65.4$^{(1.1)}$**/**99.3$^{(0.1)}$**	**121$^{(0.0)}$**/**93.5$^{(0.6)}$**	**53.1$^{(3.3)}$**/190$^{(0.8)}$**	**85.8$^{(0.7)}$**/131$^{(15.9)}$**	2.7/3.3	**21.0**/48.4
RoMA$^+$	82.9$^{(5.2)}$/67.5$^{(2.0)}$**	84.1$^{(1.0)}$/64.2$^{(1.0)}$**	**66.6$^{(0.9)}$**/98.6$^{(0.2)}$**	**121$^{(0.1)}$**/78.0$^{(3.8)}$**	**50.9$^{(2.1)}$**/196$^{(0.5)}$**	**84.8$^{(1.3)}$**/115$^{(15.3)}$**	5.2/6.3	18.3/32.9
ROMO	**93.8$^{(1.6)}$**/73.8$^{(0.6)}$**	**86.3$^{(0.1)}$**/68.7$^{(1.6)}$**	63.9$^{(0.8)}$/94.8$^{(1.6)}$**	**118$^{(5.5)}$**/**92.5$^{(1.0)}$**	**48.5$^{(3.7)}$**/196$^{(2.7)}$**	**85.5$^{(1.7)}$**/55.2$^{(36.3)}$**	5.0/6.0	19.2/34.9
GAMBO	**94.1$^{(1.9)}$**/74.0$^{(0.6)}$**	**86.3$^{(0.2)}$**/**74.3$^{(0.4)}$**	**66.8$^{(0.7)}$**/**99.3$^{(0.1)}$**	**121$^{(0.0)}$**/**93.3$^{(0.4)}$**	**50.8$^{(3.3)}$**/193$^{(1.2)}$**	**86.7$^{(1.1)}$**/17.7$^{(3.5)}$**	**2.2**/4.0	20.8/30.0
DynAMO	**91.9$^{(4.4)}$**/**74.8$^{(0.2)}$**	**86.2$^{(0.0)}$**/**74.6$^{(0.3)}$**	**67.0$^{(1.3)}$**/**99.4$^{(0.1)}$**	**121$^{(0.0)}$**/**93.5$^{(0.4)}$**	**53.5$^{(5.0)}$**/198$^{(1.9)}$**	**85.5$^{(1.1)}$**/**277$^{(59.7)}$**	3.0/**1.3**	20.7/**74.2**

BO-qUCB	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	88.1$^{(5.3)}$/**73.9$^{(0.5)}$**	**86.2$^{(0.1)}$**/**74.3$^{(0.4)}$**	**66.4$^{(0.7)}$**/**99.4$^{(0.1)}$**	**121$^{(1.3)}$**/**93.6$^{(0.5)}$**	**51.3$^{(3.6)}$**/198$^{(10.3)}$**	**84.5$^{(0.8)}$**/94.1$^{(3.9)}$**	3.7/3.0	19.4/43.5
COMs$^-$	**88.5$^{(6.4)}$**/**73.4$^{(0.6)}$**	**86.2$^{(0.0)}$**/**74.2$^{(0.7)}$**	**66.0$^{(1.1)}$**/**99.2$^{(0.1)}$**	121$^{(0.0)}$**/**93.3$^{(0.4)}$**	**47.7$^{(3.5)}$**/198$^{(1.6)}$**	**85.4$^{(1.8)}$**/107$^{(5.5)}$**	4.5/4.5	19.0/45.5
COMs$^+$	**89.1$^{(7.1)}$**/69.0$^{(0.8)}$**	**85.9$^{(0.4)}$**/72.3$^{(0.5)}$**	**65.6$^{(1.1)}$**/97.1$^{(0.9)}$**	**122$^{(0.4)}$**/91.2$^{(0.6)}$**	**45.7$^{(3.7)}$**/**261$^{(50.0)}$**	**84.7$^{(1.6)}$**/89.2$^{(14.0)}$**	5.2/5.7	18.6/51.3
RoMA$^-$	86.9$^{(5.0)}$/**73.9$^{(0.5)}$**	**86.2$^{(0.1)}$**/**74.4$^{(0.4)}$**	**66.4$^{(0.7)}$**/**99.4$^{(0.0)}$**	120$^{(1.3)}$/**93.7$^{(0.5)}$**	**51.3$^{(3.6)}$**/198$^{(10.3)}$**	**84.5$^{(0.8)}$**/94.1$^{(3.9)}$**	3.8/3.0	19.2/43.6
RoMA$^+$	84.6$^{(5.9)}$/68.2$^{(2.2)}$**	84.3$^{(1.1)}$/63.3$^{(2.5)}$**	**66.9$^{(1.0)}$**/98.3$^{(0.3)}$**	121$^{(0.2)}$/78.3$^{(4.5)}$**	**52.1$^{(3.2)}$**/194$^{(0.8)}$**	**82.9$^{(1.2)}$**/109$^{(8.3)}$**	4.7/6.5	18.5/39.9
ROMO	**95.2$^{(2.5)}$**/**74.0$^{(0.5)}$**	**86.2$^{(0.0)}$**/67.2$^{(2.0)}$**	**64.7$^{(1.0)}$**/94.9$^{(1.3)}$**	118$^{(2.1)}$/**92.4$^{(1.0)}$**	**50.2$^{(4.7)}$**/197$^{(1.3)}$**	**85.5$^{(1.1)}$**/45.2$^{(5.6)}$**	4.7/6.2	19.9/33.2
GAMBO	**95.4$^{(1.6)}$**/**74.0$^{(0.5)}$**	**86.2$^{(0.0)}$**/**74.3$^{(0.3)}$**	**66.3$^{(1.1)}$**/**99.3$^{(0.1)}$**	**121$^{(1.3)}$**/**93.4$^{(0.4)}$**	**50.2$^{(2.8)}$**/190$^{(9.3)}$**	**83.6$^{(1.0)}$**/22.0$^{(2.1)}$**	4.7/5.0	20.2/30.3
DynAMO	**95.1$^{(1.9)}$**/**74.3$^{(0.5)}$**	**86.2$^{(0.0)}$**/**74.4$^{(0.6)}$**	**66.7$^{(1.5)}$**/**99.3$^{(0.1)}$**	121$^{(0.0)}$**/**93.5$^{(0.6)}$**	**48.1$^{(4.0)}$**/**211$^{(22.8)}$**	**86.9$^{(4.5)}$**/**175$^{(44.7)}$**	**3.5**/**1.8**	**20.5**/**59.4**

Table D.2: **Additional model-based optimization quality results.** Each cell is the **Median@128** oracle score (i.e., the median oracle score achieved by 128 sampled design candidates), reported as mean$^{(95\% \text{ confidence interval})}$ across 10 seeds (here, higher is better). **Bolded** entries indicate overlapping 95% confidence intervals with the best performing algorithm (according to the mean) per optimizer. **Bolded** (resp., Underlined) Rank and Optimality Gap metrics indicate the best (resp., second best) for a given backbone optimizer.

Grad.	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Dataset $\mathcal{D}$	33.7	42.8	50.9	87.6	6.7	77.8	—/—	—/—
Baseline	**58.1**$^{(6.1)}$	58.6$^{(13.1)}$	**59.3**$^{(8.6)}$	**85.3**$^{(7.7)}$	**36.0**$^{(6.7)}$	**65.1**$^{(14.4)}$	4.3	10.5
COMs$^-$	**53.0**$^{(8.5)}$	58.3$^{(13.2)}$	**59.1**$^{(8.6)}$	**84.0**$^{(7.2)}$	22.5$^{(3.2)}$	**71.0**$^{(10.7)}$	5.8	8.1
COMs$^+$	43.9$^{(0.0)}$	59.0$^{(0.5)}$	**63.3**$^{(0.0)}$	**93.2**$^{(7.7)}$	21.3$^{(5.6)}$	**89.9**$^{(1.0)}$	4.0	11.8
RoMA$^-$	**51.1**$^{(6.1)}$	58.6$^{(13.1)}$	**59.1**$^{(8.6)}$	**85.3**$^{(7.7)}$	**36.0**$^{(6.7)}$	**65.1**$^{(14.4)}$	5.0	9.3
RoMA$^+$	**48.2**$^{(4.3)}$	**77.4**$^{(0.0)}$	**63.3**$^{(0.0)}$	84.5$^{(0.0)}$	**38.2**$^{(0.8)}$	**88.5**$^{(0.1)}$	3.2	16.8
ROMO	**58.7**$^{(3.3)}$	37.7$^{(0.3)}$	27.4$^{(1.2)}$	61.8$^{(2.6)}$	27.0$^{(0.6)}$	46.0$^{(11.7)}$	6.5	-6.8
GAMBO	**63.8**$^{(13.7)}$	**75.3**$^{(9.9)}$	**60.1**$^{(3.3)}$	**91.6**$^{(11.2)}$	**46.0**$^{(6.7)}$	**90.1**$^{(14.4)}$	**1.8**	**21.2**
DynAMO	47.0$^{(2.8)}$	69.8$^{(6.0)}$	**61.9**$^{(2.2)}$	**85.9**$^{(0.4)}$	23.4$^{(8.5)}$	**68.7**$^{(12.1)}$	4.5	9.5

Adam	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**54.7**$^{(8.8)}$	60.4$^{(12.7)}$	**59.2**$^{(8.6)}$	**87.9**$^{(10.0)}$	**37.4**$^{(6.2)}$	56.8$^{(19.8)}$	**3.3**	9.5
COMs$^-$	**54.8**$^{(8.8)}$	59.5$^{(12.7)}$	**59.2**$^{(8.6)}$	**90.8**$^{(10.4)}$	22.5$^{(3.2)}$	57.1$^{(19.6)}$	4.0	7.4
COMs$^+$	**48.0**$^{(1.7)}$	59.1$^{(0.5)}$	**63.3**$^{(0.0)}$	**89.3**$^{(10.4)}$	23.3$^{(3.7)}$	56.4$^{(13.3)}$	4.8	6.6
RoMA$^-$	**54.7**$^{(8.8)}$	60.4$^{(12.7)}$	**59.2**$^{(8.6)}$	**87.9**$^{(10.0)}$	**37.4**$^{(6.2)}$	56.8$^{(19.8)}$	**3.3**	9.5
RoMA$^+$	**50.1**$^{(4.3)}$	**77.4**$^{(0.0)}$	**63.3**$^{(0.0)}$	**84.7**$^{(0.0)}$	**34.9**$^{(1.8)}$	63.7$^{(6.2)}$	**3.3**	**12.4**
ROMO	**54.0**$^{(0.0)}$	36.8$^{(0.1)}$	**63.3**$^{(0.0)}$	50.5$^{(0.3)}$	26.1$^{(0.5)}$	30.9$^{(0.0)}$	5.7	-6.3
GAMBO	**49.5**$^{(8.9)}$	55.7$^{(12.7)}$	**57.7**$^{(9.1)}$	**84.3**$^{(9.6)}$	**37.4**$^{(6.2)}$	**87.8**$^{(4.3)}$	5.0	12.1
DynAMO	**47.7**$^{(3.0)}$	69.0$^{(5.2)}$	**62.4**$^{(1.9)}$	**86.4**$^{(0.6)}$	23.0$^{(6.0)}$	65.6$^{(14.1)}$	4.7	9.1

CMA-ES	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**50.7**$^{(2.7)}$	**71.7**$^{(10.4)}$	**63.3**$^{(0.0)}$	83.9$^{(1.0)}$	**37.9**$^{(0.7)}$	**59.3**$^{(10.9)}$	**3.2**	**11.2**
COMs$^-$	45.0$^{(2.2)}$	**68.0**$^{(8.6)}$	**60.5**$^{(3.0)}$	**89.9**$^{(10.8)}$	18.8$^{(7.9)}$	**59.8**$^{(9.9)}$	5.3	7.1
COMs$^+$	44.3$^{(3.6)}$	**62.0**$^{(8.8)}$	**59.7**$^{(4.3)}$	**91.6**$^{(9.3)}$	29.0$^{(5.9)}$	**61.2**$^{(15.0)}$	5.0	8.0
RoMA$^-$	**50.7**$^{(2.7)}$	**71.7**$^{(10.4)}$	**63.3**$^{(0.0)}$	83.9$^{(1.0)}$	**37.9**$^{(0.7)}$	**59.3**$^{(10.9)}$	**3.2**	**11.2**
RoMA$^+$	**47.4**$^{(4.2)}$	**58.0**$^{(7.0)}$	**59.9**$^{(4.4)}$	**91.6**$^{(10.0)}$	31.6$^{(5.0)}$	**60.4**$^{(7.7)}$	4.5	8.2
ROMO	**48.9**$^{(3.1)}$	**74.0**$^{(9.2)}$	**60.0**$^{(3.4)}$	84.5$^{(1.7)}$	22.8$^{(1.6)}$	**61.6**$^{(15.3)}$	3.5	8.7
GAMBO	44.2$^{(0.8)}$	**72.7**$^{(3.8)}$	**62.7**$^{(1.1)}$	86.1$^{(0.5)}$	21.4$^{(2.0)}$	**54.9**$^{(9.6)}$	5.5	7.1
DynAMO	45.3$^{(2.4)}$	**65.8**$^{(8.9)}$	59.3$^{(3.8)}$	**99.0**$^{(12.1)}$	22.5$^{(5.1)}$	**60.6**$^{(15.0)}$	4.8	8.8

CoSyNE	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**55.3**$^{(8.0)}$	**53.6**$^{(10.2)}$	**60.8**$^{(3.2)}$	87.4$^{(16.6)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	4.3	8.9
COMs$^-$	**51.7**$^{(10.6)}$	**70.9**$^{(9.2)}$	**62.8**$^{(0.7)}$	83.1$^{(8.3)}$	**28.3**$^{(5.4)}$	**58.9**$^{(17.3)}$	5.3	9.3
COMs$^+$	**53.9**$^{(2.4)}$	41.1$^{(1.1)}$	**63.3**$^{(0.0)}$	**107**$^{(3.8)}$	**23.4**$^{(7.9)}$	**61.2**$^{(15.6)}$	4.2	8.4
RoMA$^-$	**55.3**$^{(8.0)}$	**53.6**$^{(10.2)}$	**60.8**$^{(3.2)}$	87.4$^{(16.6)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	4.3	8.9
RoMA$^+$	**60.2**$^{(7.1)}$	**67.7**$^{(8.7)}$	**60.2**$^{(4.5)}$	**103**$^{(11.2)}$	**37.8**$^{(8.1)}$	**48.9**$^{(13.9)}$	3.5	13.1
ROMO	**69.1**$^{(12.9)}$	**58.5**$^{(11.3)}$	**62.1**$^{(1.4)}$	88.4$^{(5.2)}$	**29.9**$^{(4.5)}$	**70.7**$^{(10.7)}$	**3.2**	**13.2**
GAMBO	**59.5**$^{(12.0)}$	**63.5**$^{(11.2)}$	**55.4**$^{(9.6)}$	84.2$^{(17.2)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	4.5	9.8
DynAMO	**53.8**$^{(11.0)}$	**63.4**$^{(11.5)}$	59.3$^{(3.8)}$	**99.0**$^{(12.1)}$	20.5$^{(5.8)}$	**60.6**$^{(15.0)}$	5.3	9.5

BO-qEI	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	48.5$^{(1.5)}$	59.9$^{(2.0)}$	**63.3**$^{(0.0)}$	86.7$^{(0.6)}$	**28.7**$^{(1.8)}$	72.4$^{(1.8)}$	4.3	10.0
COMs$^-$	**50.9**$^{(1.9)}$	59.6$^{(1.6)}$	**63.3**$^{(0.0)}$	86.6$^{(0.6)}$	19.4$^{(1.1)}$	**78.1**$^{(1.1)}$	4.7	9.7
COMs$^+$	43.6$^{(0.0)}$	**66.0**$^{(1.6)}$	**63.3**$^{(0.0)}$	**87.5**$^{(0.6)}$	20.6$^{(0.9)}$	66.3$^{(2.2)}$	4.5	8.0
RoMA$^-$	50.0$^{(1.6)}$	60.0$^{(2.1)}$	**63.3**$^{(0.0)}$	86.4$^{(0.5)}$	**28.7**$^{(1.8)}$	**78.5**$^{(1.2)}$	3.5	11.2
RoMA$^+$	**52.5**$^{(0.0)}$	61.0$^{(1.1)}$	**63.3**$^{(0.0)}$	**93.3**$^{(5.7)}$	26.4$^{(1.1)}$	74.3$^{(1.3)}$	**2.7**	**11.9**
ROMO	49.9$^{(2.2)}$	59.0$^{(1.4)}$	**63.3**$^{(0.0)}$	86.8$^{(0.6)}$	24.6$^{(0.8)}$	73.8$^{(2.0)}$	4.7	9.6
GAMBO	46.4$^{(1.8)}$	**63.4**$^{(3.3)}$	**63.3**$^{(0.0)}$	86.3$^{(0.5)}$	**28.9**$^{(1.1)}$	**79.1**$^{(0.7)}$	3.5	11.3
DynAMO	51.5$^{(0.9)}$	**65.6**$^{(3.1)}$	**63.3**$^{(0.0)}$	86.7$^{(0.6)}$	23.5$^{(2.4)}$	77.0$^{(0.7)}$	3.3	11.3

BO-qUCB	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	50.3$^{(1.8)}$	**62.1**$^{(3.4)}$	**63.3**$^{(0.0)}$	**86.6**$^{(0.6)}$	**31.7**$^{(1.2)}$	**74.4**$^{(0.6)}$	**2.7**	**11.5**
COMs$^-$	51.1$^{(1.0)}$	**61.0**$^{(2.9)}$	**63.3**$^{(0.0)}$	**86.3**$^{(0.7)}$	19.8$^{(1.3)}$	**74.2**$^{(1.3)}$	4.5	9.4
COMs$^+$	43.6$^{(0.0)}$	**65.6**$^{(1.4)}$	**63.3**$^{(0.0)}$	**87.5**$^{(0.9)}$	20.1$^{(1.0)}$	54.2$^{(11.9)}$	4.5	5.8
RoMA$^-$	50.0$^{(1.7)}$	**62.1**$^{(3.4)}$	**63.3**$^{(0.0)}$	**86.7**$^{(0.6)}$	**31.7**$^{(1.2)}$	**74.4**$^{(0.6)}$	**2.7**	11.4
RoMA$^+$	**52.5**$^{(0.0)}$	60.8$^{(0.9)}$	**63.3**$^{(0.0)}$	**91.4**$^{(5.6)}$	**29.5**$^{(1.4)}$	**65.8**$^{(9.3)}$	3.2	10.6
ROMO	49.8$^{(2.0)}$	58.2$^{(0.2)}$	**63.3**$^{(0.0)}$	**86.8**$^{(0.5)}$	24.3$^{(0.7)}$	**75.0**$^{(1.7)}$	3.8	9.6
GAMBO	47.9$^{(1.9)}$	59.8$^{(1.2)}$	**63.3**$^{(0.0)}$	**86.0**$^{(0.6)}$	**33.1**$^{(2.9)}$	**73.8**$^{(1.2)}$	4.8	10.7
DynAMO	48.8$^{(1.8)}$	**65.9**$^{(3.7)}$	63.3$^{(0.0)}$	**86.5**$^{(0.5)}$	22.7$^{(2.0)}$	50.4$^{(14.6)}$	4.7	6.3

D.2. Additional Design Diversity Results

We supplement the results shown in **Table 6.1** with the raw Pairwise Diversity scores reported for each of the 6 tasks in our evaluation suite in **Supp. Table D.1**.

In **Section 6.4**, we describe the **Pairwise Diversity** metric previously used in prior work (Kim et al., 2023; Jain et al., 2022; Maus et al., 2023) to measure the diversity of samples obtained from a given offline optimization method. We can think of Pairwise Diversity as measuring the *between-candidate* diversity of candidates proposed by a generative algorithm. However, this is far from the only relevant definition of diversity; other possible metrics might measure the following:

1. *Candidate-Dataset Diversity*: How ***novel*** is a proposed candidate compared to the real designs previously observed in the offline dataset?

2. *Aggregate Diversity*: How well does the batch of candidate designs collectively cover the possible search space?

To evaluate (1), we follow prior work by Kim et al. (2023) and Jain et al. (2022) and evaluate the **Minimum Novelty (MN)** for a batch of k final proposed candidates with respect to the offline dataset $\mathcal{D}$, defined as

$$\text{MN}(\{x_i^F\}_{i=1}^k; \mathcal{D}) := \mathbb{E}_{x_i^F}\left[\min_{x \in \mathcal{D}} d(x_i^F, x)\right] \tag{D.2}$$

where $\mathcal{D}$ is the task-specific dataset of offline sample designs and x_i^F is the ith candidate design proposed by an optimization experiment. Following (6.37), we define the distance function $d(\cdot, \cdot)$ as the normalized Levenshtein edit distance (Haldar and Mukhopadhyay, 2011) (resp., Euclidean distance) for discrete (resp., continuous) tasks.

For (2), we report the L_1 **Coverage (L1C)** of the candidate designs, defined as

$$\text{L1C}(\{x_i^F\}_{i=1}^B) := \frac{1}{\dim(x)} \sum_{k=1}^{\dim(x)} \max_{i \neq j} |x_{ik}^F - x_{jk}^F| \tag{D.3}$$

where $\dim(x)$ is the number of design dimensions and x_{ik}^F is the kth dimension of design x_i^F. Note

that the L1C metric is only defined for designs sampled from a continuous search space; to compute the L1C metric for discrete optimization tasks, we use task-specific foundation models to embed discrete designs into a continuous latent space. For DNA design tasks (i.e., **TFBind8** and **UTR**), we use the DNABERT-2 foundation model with 117M parameters (`zhihan1996/DNABERT-2-117M`) from Zhou et al. (2024b) to embed candidate DNA sequences into a continuous latent space. Similarly for molecule design tasks (i.e., **ChEMBL** and **Molecule**), we use the ChemBERT model (`jonghyunlee/ChemBERT_ChEMBL_pretrained`) from Zhang et al. (2022) to embed molecule candidates into a continuous latent space.

We report MN and L1C metric scores in **Supp. Table D.3**. We find that compared with other MBO objective-modifying methods, DynAMO achieves the best Rank and Optimality Gap for 3 of the 6 optimizers evaluated (Grad., Adam, and BO-qEI). For the remaining 3 optimizers evaluated, DynAMO is within the top 2 evaluated methods in terms of both average Rank and Optimality Gap for the L1C (L_1 coverage) metric. Altogether, our results support that DynAMO is competitive according to the MN and L1C diversity metrics in addition to the Pairwise Diversity metric reported in **Table 6.1**.

What is the *best* notion of diversity? In our work, we focus on the Pairwise Diversity metric in our main results (**Table 6.1**)—however, this does not mean that this metric is the best for all applications. Rather, our focus on the Pairwise Diversity metric is determined by our problem motivation. Compared with the minimum novelty and L_1 coverage diversity metrics, the definition of pairwise diversity best captures the notion of diversity that we are interested in—that is, capturing many possible 'modes of goodness' in optimizing the oracle reward function. We note that these modes of goodness may not necessarily be significantly 'novel' according to our task-specific distance metric, and so we treated the Minimum Novelty metric as only a secondary diversity objective for evaluating DynAMO. (Indeed, because DynAMO encourages a generative policy to match a distribution of designs constructed from the offline dataset, DynAMO may *not* increase the minimum novelty of designs compared to those proposed by the comparable baseline optimizer.) Similarly, we find that the L_1 Coverage metric is more sensitive to outlier designs when compared to the Pairwise

Diversity, and therefore also treat it as a secondary diversity evaluation metric for our experiments in **Supp. Table D.3**. Future work might explore other methods that focus on improving not only the Pairwise Diversity metric, but also other diversity metric(s).

D.3. Imposing Alternative f-Divergence Diversity Objectives via Mixed-Divergence Regularization

The MBO problem formulation proposed in (6.15) introduces a weighted KL-divergence regularization of the original MBO optimization objective. However, alternative distribution matching objectives have been used in prior work (Agarwal et al., 2024b; Gong et al., 2021; Ma et al., 2022), and one might hypothesize that we can similarly generalize (6.15) as

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J_f(\pi) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_f(q^\pi || p^\tau_{\mathcal{D}}) \\ \text{s.t.} \quad & \mathbb{E}_{x \sim p^\tau_{\mathcal{D}}(x)}[c^*(x)] - \mathbb{E}_{x \sim q^\pi(x)}[c^*(x)] \leq W_0 \end{aligned} \tag{D.4}$$

to any arbitrary f-divergence metric $D_f(\cdot||\cdot)$ that measures the difference between two probability distributions Q, P over a space Ω defined by $D_f(Q||P) := \int_\Omega dP \; f\left(\frac{dQ}{dP}\right)$ for a convex univariate *generator* function f. For example in our main text, we specialize to the KL-divergence where $f_{\text{KL}}(u) := u \log u$ traditionally used in the imitation learning literature.

However, we found that this naïve approach does ***not*** generalize well to alternative f-divergences: recall that a core contribution of our work was the ability to reformulate the optimization objective as a weighted sum over distribution entropy and divergence (i.e., **Lemma 5**) in order to admit an explicit, closed form solution for the dual function in **Lemma 6**. Such an approach is intractable using standard algebraic techniques. This is not ideal, as a number of prior works have proposed that alternative divergences—such as the χ^2-divergence defined by the generator $f_{\chi^2}(u) = (u-1)^2/2$—can better penalize out-of-distribution surrogate behavior and better quantify model uncertainty when compared to the KL-divergence (Tsybakov, 2008; Nishiyama and Sason, 2020; Ma et al., 2022; Wang et al., 2024a).

In this section, we show how to overcome this limitation and demonstrate how our theoretical

Table D.3: **Additional model-based optimization diversity results.** Each cell is a pair of values `mn`/`l1c`; where `mn` is the **Minimum Novelty** and `l1c` the L_1 **Coverage**. Metrics are reported as mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate overlapping 95% confidence intervals with the best performing algorithm (according to the mean) per optimizer. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Grad.	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Dataset $\mathcal{D}$	0.0/0.42	0.0/0.31	0.0/1.42	0.0/0.68	0.0/6.26	0.0/0.58	—/—	—/—
Baseline	**21.2**$^{(3.0)}$/0.16$^{(0.1)}$	**51.7**$^{(2.9)}$/0.20$^{(0.13)}$	**97.4**$^{(3.9)}$/0.21$^{(0.10)}$	**79.5**$^{(19.7)}$/0.42$^{(0.18)}$	95.0$^{(0.7)}$/0.00$^{(0.00)}$	102$^{(6.1)}$/0.00$^{(0.00)}$	3.3/6.3	74.5/-1.44
COMs$^-$	**22.1**$^{(2.7)}$/0.14$^{(0.11)}$	**51.4**$^{(3.2)}$/0.20$^{(0.12)}$	**97.4**$^{(3.9)}$/0.24$^{(0.11)}$	**89.0**$^{(7.7)}$/0.40$^{(0.10)}$	94.1$^{(0.7)}$/0.00$^{(0.00)}$	100$^{(4.8)}$/0.00$^{(0.00)}$	3.3/7.7	75.7/-1.45
COMs$^+$	10.9$^{(0.3)}$/**0.49**$^{(0.02)}$	31.7$^{(0.8)}$/0.31$^{(0.00)}$	52.4$^{(11.0)}$/**1.11**$^{(0.16)}$	13.7$^{(1.1)}$/0.61$^{(0.09)}$	**99.6**$^{(0.3)}$/0.37$^{(0.11)}$	100$^{(0.0)}$/0.80$^{(0.76)}$	6.2/2.7	51.4/-1.00
RoMA$^-$	**21.2**$^{(3.1)}$/0.16$^{(0.1)}$	**51.7**$^{(2.9)}$/0.21$^{(0.12)}$	**97.4**$^{(3.9)}$/0.26$^{(0.10)}$	**79.5**$^{(19.7)}$/0.40$^{(0.19)}$	95.0$^{(0.7)}$/0.00$^{(0.00)}$	102$^{(6.1)}$/0.00$^{(0.00)}$	2.8/5.8	74.5/-1.44
RoMA$^+$	18.1$^{(1.4)}$/0.27$^{(0.02)}$	40.1$^{(0.2)}$/0.44$^{(0.01)}$	18.7$^{(0.1)}$/0.41$^{(0.01)}$	**95.3**$^{(0.0)}$/0.41$^{(0.02)}$	7.1$^{(0.8)}$/1.28$^{(0.01)}$	0.2$^{(0.0)}$/0.45$^{(0.00)}$	5.8/3.5	29.9/-1.07
ROMO	16.1$^{(0.5)}$/0.33$^{(0.02)}$	32.9$^{(0.1)}$/0.30$^{(0.00)}$	5.0$^{(0.7)}$/**1.31**$^{(0.02)}$	23.1$^{(0.0)}$/0.61$^{(0.02)}$	78.5$^{(0.5)}$/0.34$^{(0.16)}$	**153**$^{(0.4)}$/**6.13**$^{(2.80)}$	6.0/2.7	51.5/-0.11
GAMBO	14.0$^{(2.0)}$/0.17$^{(0.10)}$	46.7$^{(2.7)}$/0.24$^{(0.13)}$	**96.8**$^{(3.9)}$/0.25$^{(0.16)}$	**76.8**$^{(19.7)}$/0.37$^{(0.11)}$	83.8$^{(6.8)}$/0.00$^{(0.00)}$	31.5$^{(3.6)}$/0.09$^{(0.14)}$	6.0/5.7	58.3/-1.42
DynAMO	**21.1**$^{(1.1)}$/0.36$^{(0.04)}$	**52.2**$^{(1.3)}$/**0.52**$^{(0.06)}$	**98.6**$^{(1.5)}$/**1.46**$^{(0.38)}$	85.8$^{(1.0)}$/**2.49**$^{(0.06)}$	95.0$^{(0.4)}$/**6.47**$^{(1.24)}$	107$^{(6.7)}$/**5.85**$^{(1.35)}$	**2.2**/**1.3**	**76.7**/**1.25**

Adam	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**23.7**$^{(2.8)}$/0.11$^{(0.06)}$	**51.1**$^{(3.5)}$/0.22$^{(0.09)}$	**95.5**$^{(5.3)}$/0.23$^{(0.15)}$	**79.3**$^{(21.2)}$/0.48$^{(0.31)}$	94.8$^{(0.7)}$/0.27$^{(0.55)}$	**103**$^{(6.3)}$/0.24$^{(0.49)}$	3.5/5.8	74.5/-1.35
COMs$^-$	**23.8**$^{(2.8)}$/0.13$^{(0.06)}$	**51.6**$^{(3.0)}$/0.20$^{(0.11)}$	**95.5**$^{(5.3)}$/0.24$^{(0.16)}$	**78.6**$^{(20.7)}$/0.49$^{(0.33)}$	94.1$^{(0.7)}$/0.00$^{(0.00)}$	**102**$^{(6.0)}$/0.03$^{(0.00)}$	3.3/6.8	74.2/-1.43
COMs$^+$	13.0$^{(0.4)}$/**0.44**$^{(0.02)}$	31.7$^{(0.8)}$/0.31$^{(0.00)}$	53.0$^{(12.8)}$/**1.10**$^{(0.20)}$	15.0$^{(1.6)}$/0.50$^{(0.12)}$	**99.7**$^{(0.2)}$/0.58$^{(0.28)}$	99.9$^{(0.1)}$/0.88$^{(0.95)}$	6.0/2.7	52.0/-0.98
RoMA$^-$	**23.6**$^{(2.8)}$/0.12$^{(0.06)}$	**51.1**$^{(3.5)}$/0.21$^{(0.09)}$	**95.5**$^{(5.3)}$/0.21$^{(0.16)}$	**79.3**$^{(21.2)}$/0.48$^{(0.32)}$	94.8$^{(0.7)}$/0.27$^{(0.55)}$	**103**$^{(6.3)}$/0.04$^{(0.03)}$	3.3/6.8	74.5/-1.39
RoMA$^+$	18.3$^{(0.5)}$/0.28$^{(0.00)}$	40.1$^{(0.2)}$/0.46$^{(0.01)}$	18.9$^{(0.2)}$/0.41$^{(0.02)}$	**95.3**$^{(0.0)}$/0.42$^{(0.01)}$	47.6$^{(2.4)}$/1.87$^{(0.06)}$	5.1$^{(0.2)}$/0.78$^{(0.01)}$	6.0/3.7	37.6/-0.91
ROMO	13.6$^{(0.2)}$/0.28$^{(0.01)}$	32.9$^{(0.1)}$/0.30$^{(0.00)}$	21.9$^{(0.0)}$/**1.05**$^{(0.02)}$	23.4$^{(0.0)}$/0.74$^{(0.0)}$	98.1$^{(0.0)}$/1.34$^{(0.03)}$	99.8$^{(0.1)}$/0.08$^{(0.00)}$	6.0/3.7	48.3/-0.98
GAMBO	**23.7**$^{(3.1)}$/0.14$^{(0.06)}$	**51.3**$^{(3.4)}$/0.22$^{(0.10)}$	**95.0**$^{(5.1)}$/0.35$^{(0.24)}$	**80.0**$^{(20.6)}$/0.50$^{(0.34)}$	84.8$^{(6.4)}$/0.26$^{(0.53)}$	27.3$^{(3.4)}$/0.09$^{(0.12)}$	4.3/5.2	60.4/-1.35
DynAMO	14.7$^{(1.9)}$/0.33$^{(0.05)}$	46.2$^{(0.5)}$/**0.55**$^{(0.03)}$	**98.7**$^{(1.2)}$/**1.44**$^{(0.39)}$	85.9$^{(1.8)}$/**2.40**$^{(0.16)}$	94.9$^{(0.4)}$/**7.06**$^{(0.73)}$	**108**$^{(7.2)}$/**6.91**$^{(0.71)}$	**3.0**/**1.2**	**74.7**/**1.50**

CMA-ES	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	16.5$^{(2.1)}$/**0.33**$^{(0.05)}$	47.8$^{(1.0)}$/0.48$^{(0.04)}$	96.5$^{(0.7)}$/**2.18**$^{(0.04)}$	**73.0**$^{(18.0)}$/1.82$^{(0.12)}$	**100**$^{(0.0)}$/**3.26**$^{(1.42)}$	100$^{(0.0)}$/**3.77**$^{(1.36)}$	3.8/4.2	72.3/0.36
COMs$^-$	13.6$^{(0.7)}$/**0.34**$^{(0.05)}$	46.9$^{(0.9)}$/**0.52**$^{(0.06)}$	**97.3**$^{(2.6)}$/1.81$^{(0.48)}$	**84.7**$^{(6.9)}$/2.10$^{(0.32)}$	**100**$^{(0.0)}$/0.31$^{(0.18)}$	100$^{(0.0)}$/0.43$^{(0.18)}$	4.7/4.7	73.7/-0.69
COMs$^+$	11.1$^{(2.1)}$/0.32$^{(0.04)}$	44.0$^{(2.8)}$/0.43$^{(0.06)}$	98.5$^{(1.1)}$/1.20$^{(0.27)}$	**77.4**$^{(19.7)}$/1.85$^{(0.16)}$	86.1$^{(3.5)}$/0.06$^{(0.01)}$	39.0$^{(7.9)}$/0.02$^{(0.00)}$	6.3/7.0	59.3/-0.96
RoMA$^-$	16.5$^{(2.2)}$/0.32$^{(0.04)}$	47.9$^{(0.9)}$/0.49$^{(0.03)}$	96.6$^{(0.7)}$/**2.16**$^{(0.05)}$	**73.0**$^{(18.0)}$/1.82$^{(0.12)}$	**100**$^{(0.0)}$/**4.15**$^{(1.93)}$	100$^{(0.0)}$/**3.77**$^{(1.36)}$	3.5/4.0	72.3/0.51
RoMA$^+$	13.4$^{(0.7)}$/**0.36**$^{(0.04)}$	47.7$^{(1.7)}$/0.42$^{(0.07)}$	**99.8**$^{(0.2)}$/1.35$^{(0.39)}$	**80.0**$^{(6.2)}$/1.80$^{(0.45)}$	**100**$^{(0.0)}$/0.30$^{(0.15)}$	100$^{(0.0)}$/**3.97**$^{(2.11)}$	**3.2**/5.7	73.5/-0.24
ROMO	16.2$^{(2.3)}$/**0.38**$^{(0.02)}$	47.0$^{(1.7)}$/0.49$^{(0.06)}$	97.7$^{(1.7)}$/**2.01**$^{(0.24)}$	**85.3**$^{(5.9)}$/2.13$^{(0.15)}$	**100**$^{(0.0)}$/**3.14**$^{(1.85)}$	51.9$^{(17.6)}$/0.50$^{(0.33)}$	3.5/4.0	66.4/-0.17
GAMBO	**24.3**$^{(0.9)}$/0.31$^{(0.04)}$	**53.3**$^{(1.4)}$/0.51$^{(0.01)}$	95.0$^{(1.5)}$/**2.17**$^{(0.06)}$	**72.5**$^{(23.6)}$/1.83$^{(0.15)}$	85.6$^{(3.0)}$/**3.37**$^{(0.49)}$	41.5$^{(2.0)}$/**3.12**$^{(0.40)}$	5.5/4.3	62.0/0.27
DynAMO	12.9$^{(0.8)}$/**0.40**$^{(0.03)}$	48.0$^{(1.6)}$/**0.56**$^{(0.01)}$	**96.7**$^{(3.5)}$/**1.82**$^{(0.72)}$	**81.8**$^{(13.4)}$/**2.54**$^{(0.05)}$	94.5$^{(0.7)}$/**4.75**$^{(2.16)}$	**112**$^{(7.8)}$/**3.29**$^{(1.56)}$	4.0/**2.2**	**74.3**/**0.62**

CoSyNE	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**24.5**$^{(3.5)}$/0.10$^{(0.07)}$	**49.7**$^{(3.1)}$/0.22$^{(0.10)}$	**98.5**$^{(1.6)}$/0.39$^{(0.20)}$	**86.6**$^{(12.7)}$/0.27$^{(0.13)}$	**93.2**$^{(1.0)}$/0.10$^{(0.00)}$	91.9$^{(2.0)}$/0.10$^{(0.00)}$	3.2/5.3	74.1/-1.41
COMs$^-$	15.4$^{(4.9)}$/**0.23**$^{(0.09)}$	44.9$^{(3.5)}$/0.20$^{(0.13)}$	**90.0**$^{(15.2)}$/0.38$^{(0.27)}$	**86.2**$^{(7.7)}$/0.39$^{(0.29)}$	**93.8**$^{(0.6)}$/0.07$^{(0.01)}$	**101**$^{(6.3)}$/0.06$^{(0.00)}$	5.0/5.8	71.9/-1.39
COMs$^+$	12.8$^{(0.6)}$/**0.24**$^{(0.04)}$	42.7$^{(1.4)}$/**0.32**$^{(0.03)}$	34.1$^{(14.4)}$/**1.31**$^{(0.22)}$	50.0$^{(1.9)}$/**1.39**$^{(0.24)}$	87.8$^{(3.1)}$/**1.29**$^{(0.37)}$	30.8$^{(5.7)}$/0.25$^{(0.09)}$	7.5/**1.5**	43.0/**-0.81**
RoMA$^-$	**24.6**$^{(3.5)}$/0.10$^{(0.07)}$	**49.7**$^{(3.0)}$/0.22$^{(0.11)}$	**98.6**$^{(1.5)}$/0.40$^{(0.20)}$	**86.6**$^{(12.7)}$/0.27$^{(0.12)}$	**93.2**$^{(1.0)}$/0.10$^{(0.00)}$	91.9$^{(2.0)}$/0.10$^{(0.00)}$	**2.5**/5.2	74.1/-1.41
RoMA$^+$	17.1$^{(3.3)}$/**0.22**$^{(0.05)}$	**49.3**$^{(3.0)}$/**0.42**$^{(0.09)}$	**99.5**$^{(0.9)}$/0.53$^{(0.13)}$	**89.6**$^{(3.4)}$/0.55$^{(0.22)}$	**93.2**$^{(1.6)}$/0.10$^{(0.00)}$	65.5$^{(25.2)}$/0.05$^{(0.03)}$	4.0/3.7	69.0/-1.30
ROMO	**22.9**$^{(6.3)}$/**0.18**$^{(0.09)}$	**48.9**$^{(1.5)}$/**0.25**$^{(0.10)}$	75.8$^{(6.1)}$/**1.20**$^{(0.21)}$	**92.2**$^{(3.3)}$/0.39$^{(0.12)}$	84.6$^{(4.4)}$/0.09$^{(0.00)}$	31.8$^{(5.6)}$/0.06$^{(0.10)}$	5.0/4.5	59.4/-1.25
GAMBO	**22.8**$^{(2.8)}$/0.12$^{(0.08)}$	**50.8**$^{(1.5)}$/0.22$^{(0.15)}$	**90.8**$^{(14.2)}$/0.53$^{(0.18)}$	**91.9**$^{(3.4)}$/0.14$^{(0.13)}$	86.0$^{(3.3)}$/0.10$^{(0.00)}$	29.6$^{(3.6)}$/0.02$^{(0.00)}$	4.5/6.3	62.0/-1.42
DynAMO	**17.8**$^{(5.5)}$/**0.21**$^{(0.11)}$	**48.4**$^{(2.3)}$/0.18$^{(0.09)}$	**96.7**$^{(3.5)}$/0.64$^{(0.42)}$	**80.2**$^{(12.9)}$/0.43$^{(0.34)}$	**94.5**$^{(0.7)}$/**1.85**$^{(0.22)}$	**112**$^{(7.8)}$/**0.94**$^{(0.17)}$	4.0/3.3	**75.0**/-0.90

BO-qEI	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**21.8**$^{(0.5)}$/0.41$^{(0.02)}$	**51.5**$^{(0.3)}$/0.55$^{(0.01)}$	97.6$^{(0.3)}$/2.37$^{(0.03)}$	85.4$^{(1.5)}$/2.11$^{(0.15)}$	94.6$^{(0.1)}$/7.84$^{(0.01)}$	106$^{(2.9)}$/6.61$^{(0.33)}$	3.3/5.0	76.2/1.70
COMs$^-$	**21.9**$^{(0.5)}$/**0.41**$^{(0.03)}$	**51.8**$^{(0.2)}$/**0.56**$^{(0.01)}$	97.3$^{(0.5)}$/**2.44**$^{(0.05)}$	85.2$^{(0.5)}$/**2.55**$^{(0.02)}$	92.5$^{(1.2)}$/7.61$^{(0.24)}$	105$^{(1.6)}$/7.19$^{(0.22)}$	4.0/3.0	75.6/1.85
COMs$^+$	12.7$^{(0.4)}$/**0.44**$^{(0.01)}$	43.2$^{(0.2)}$/**0.57**$^{(0.01)}$	83.3$^{(2.1)}$/**2.50**$^{(0.05)}$	79.4$^{(1.3)}$/2.41$^{(0.07)}$	85.5$^{(0.4)}$/7.50$^{(0.01)}$	35.5$^{(1.2)}$/1.66$^{(0.01)}$	7.5/3.8	56.6/0.90
RoMA$^-$	**21.6**$^{(0.3)}$/0.40$^{(0.03)}$	**51.7**$^{(0.3)}$/**0.55**$^{(0.02)}$	**97.7**$^{(0.5)}$/2.41$^{(0.04)}$	85.9$^{(1.1)}$/0.52$^{(0.02)}$	94.6$^{(0.1)}$/7.84$^{(0.01)}$	105$^{(3.5)}$/6.45$^{(0.66)}$	3.2/5.2	76.1/1.42
RoMA$^+$	13.6$^{(0.3)}$/0.31$^{(0.02)}$	45.0$^{(0.2)}$/**0.56**$^{(0.01)}$	**98.5**$^{(0.4)}$/1.86$^{(0.05)}$	**88.2**$^{(0.6)}$/1.99$^{(0.06)}$	94.3$^{(0.1)}$/**7.87**$^{(0.01)}$	**116**$^{(2.5)}$/7.09$^{(0.49)}$	3.7/5.0	76.0/1.67
ROMO	15.6$^{(0.4)}$/0.39$^{(0.02)}$	47.5$^{(0.2)}$/0.55$^{(0.01)}$	87.7$^{(1.3)}$/**2.50**$^{(0.03)}$	**87.9**$^{(1.0)}$/**2.48**$^{(0.05)}$	86.6$^{(2.2)}$/7.51$^{(0.09)}$	30.8$^{(25.4)}$/2.14$^{(1.34)}$	5.5/5.3	59.3/0.98
GAMBO	15.4$^{(0.3)}$/0.40$^{(0.03)}$	**51.8**$^{(0.2)}$/0.55$^{(0.01)}$	97.8$^{(0.3)}$/**2.38**$^{(0.10)}$	84.9$^{(0.9)}$/**2.53**$^{(0.05)}$	85.1$^{(0.4)}$/7.45$^{(0.01)}$	14.3$^{(1.5)}$/1.29$^{(0.08)}$	5.7/6.3	58.2/0.82
DynAMO	**21.0**$^{(0.5)}$/0.42$^{(0.01)}$	**51.9**$^{(0.2)}$/**0.56**$^{(0.01)}$	97.4$^{(0.4)}$/**2.47**$^{(0.03)}$	85.2$^{(0.9)}$/**2.54**$^{(0.03)}$	**94.8**$^{(0.1)}$/**7.87**$^{(0.01)}$	**126**$^{(14.6)}$/**7.92**$^{(0.04)}$	**3.0**/**2.2**	**79.4**/**2.02**

BO-qUCB	TFBind8	UTR	ChEMBL	Molecule	Superconductor	D'Kitty	Rank ↓	Opt. Gap ↑
Baseline	**21.6**$^{(0.3)}$/0.40$^{(0.02)}$	**51.7**$^{(0.2)}$/0.54$^{(0.01)}$	97.9$^{(0.4)}$/2.40$^{(0.05)}$	85.3$^{(1.1)}$/**2.52**$^{(0.07)}$	93.8$^{(0.6)}$/7.78$^{(0.04)}$	98.8$^{(1.1)}$/6.64$^{(0.09)}$	3.5/4.7	74.8/1.77
COMs$^-$	**21.7**$^{(0.3)}$/0.40$^{(0.02)}$	**51.7**$^{(0.2)}$/**0.56**$^{(0.01)}$	97.4$^{(0.4)}$/2.40$^{(0.06)}$	85.4$^{(1.4)}$/**2.49**$^{(0.04)}$	92.9$^{(1.1)}$/7.75$^{(0.09)}$	99.2$^{(1.6)}$/6.84$^{(0.11)}$	3.8/3.7	74.7/1.80
COMs$^+$	12.6$^{(0.3)}$/**0.44**$^{(0.02)}$	43.5$^{(0.3)}$/**0.57**$^{(0.01)}$	84.0$^{(2.3)}$/**2.52**$^{(0.05)}$	79.2$^{(1.4)}$/2.38$^{(0.08)}$	84.6$^{(1.1)}$/7.51$^{(0.02)}$	39.4$^{(6.0)}$/1.66$^{(0.01)}$	7.5/3.7	57.2/0.90
RoMA$^-$	**21.6**$^{(0.3)}$/0.39$^{(0.03)}$	**51.6**$^{(0.3)}$/0.55$^{(0.01)}$	97.8$^{(0.4)}$/2.37$^{(0.05)}$	85.5$^{(1.1)}$/**2.54**$^{(0.07)}$	93.8$^{(0.6)}$/7.78$^{(0.04)}$	98.8$^{(1.1)}$/6.64$^{(0.09)}$	3.5/4.8	74.9/1.77
RoMA$^+$	13.9$^{(0.3)}$/0.31$^{(0.02)}$	45.1$^{(0.5)}$/**0.56**$^{(0.01)}$	**98.8**$^{(0.3)}$/1.85$^{(0.02)}$	88.5$^{(0.5)}$/2.01$^{(0.06)}$	94.1$^{(0.1)}$/**7.86**$^{(0.01)}$	**112**$^{(2.5)}$/7.23$^{(0.16)}$	**3.3**/5.2	75.4/1.69
ROMO	16.0$^{(0.4)}$/0.40$^{(0.02)}$	47.7$^{(0.2)}$/0.55$^{(0.01)}$	88.1$^{(1.2)}$/**2.51**$^{(0.06)}$	**90.9**$^{(0.6)}$/**2.49**$^{(0.05)}$	85.4$^{(0.3)}$/7.48$^{(0.02)}$	20.9$^{(2.4)}$/1.60$^{(0.03)}$	5.7/5.3	58.2/0.89
GAMBO	**21.9**$^{(0.4)}$/0.40$^{(0.01)}$	**51.7**$^{(0.3)}$/**0.56**$^{(0.01)}$	97.5$^{(0.4)}$/2.39$^{(0.05)}$	85.2$^{(1.0)}$/**2.52**$^{(0.04)}$	81.9$^{(1.9)}$/7.37$^{(0.09)}$	25.9$^{(1.4)}$/1.34$^{(0.04)}$	4.7/6.0	60.7/0.82
DynAMO	**21.4**$^{(0.5)}$/0.40$^{(0.02)}$	**51.7**$^{(0.2)}$/0.55$^{(0.01)}$	97.1$^{(0.5)}$/**2.47**$^{(0.07)}$	85.3$^{(1.1)}$/**2.54**$^{(0.05)}$	**94.7**$^{(0.2)}$/**7.88**$^{(0.03)}$	**109**$^{(4.5)}$/**7.80**$^{(0.23)}$	3.7/**2.3**	**76.6**/**2.00**

and empirical results generalize to alternative f-divergence objectives for enforcing distribution matching in the sampling policy. Firstly, we look to recent work by Huang et al. (2024a) and others describing 'mixed f-regularization' defined by a mixed generator function $f_\gamma(u) := \gamma f(u) + u \log u$ for some weighting scalar $\gamma \in [1, +\infty)$, which admits a 'mixed f-divergence' given by

$$D_f(Q||P; \gamma) := \gamma D_f(Q||P) + D_{\text{KL}}(Q||P) \tag{D.5}$$

for probability distributions Q, P.[5] Given a mixed f-divergence, we can define a modified MBO objective as in (6.13):

$$\begin{aligned} J_f(\pi; \gamma) &:= \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_f(q^\pi||p^\tau_{\mathcal{D}}; \gamma) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_{\text{KL}}(q^\pi||p^\tau_{\mathcal{D}}) - \frac{\beta\gamma}{\tau} D_f(q^\pi||p^\tau_{\mathcal{D}}) \\ &= J(\pi) - \frac{\beta\gamma}{\tau} D_f(q^\pi||p^\tau_{\mathcal{D}}) \end{aligned} \tag{D.6}$$

where r_θ is again the forward surrogate model, $J(\pi)$ is as in (6.13), $p^\tau_{\mathcal{D}}(x)$ is the τ-weighted probability distribution as in **Definition 4**, and $q^\pi(x)$ is the sampled distribution over designs admitted by the realized sampling policy π. Given this expression for the modified MBO objective, it is easy to rewrite $J_f(\pi; \gamma)$ similar to **Lemma 5** in the main text:

Lemma 9 (Generalized Entropy-Divergence Formulation for Mixed f-Divergence)**.** *Define $J_f(\pi; \gamma)$ as in (D.6). An equivalent representation of $J_f(\pi; \gamma)$ is*

$$J_f(\pi) \simeq -\mathcal{H}(q^\pi(x)) - (1 + \beta) D_{KL}(q^\pi(x)||p^\tau_{\mathcal{D}}(x)) - \beta\gamma D_f(q^\pi(x)||p^\tau_{\mathcal{D}}(x)) \tag{D.7}$$

where $\mathcal{H}(\cdot)$ as the Shannon entropy and $D_f(\cdot||\cdot)$ as the f-divergence.

[5] It is trivial to verify both that $f_\gamma(u)$ is convex and that $0 \notin \text{dom}(f_\gamma)$ given a function $f(u)$ that also satisfies both of these conditions.

Proof. The proof is trivial using **Lemma 5**:

$$\begin{aligned} J_f(\pi) &:= \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_f(q^\pi || p_{\mathcal{D}}^\tau; \pi) = J(\pi) - \frac{\beta\gamma}{\tau} D_f(q^\pi || p_{\mathcal{D}}^\tau) \\ &\simeq \tau \cdot J(\pi) - \beta\gamma D_f(q^\pi || p_{\mathcal{D}}^\tau) \\ &\simeq -\mathcal{H}(q^\pi) - (1+\beta) D_{\mathrm{KL}}(q^\pi || p_{\mathcal{D}}^\tau) - \beta\gamma D_f(q^\pi || p_{\mathcal{D}}^\tau) \end{aligned} \tag{D.8}$$

up to a constant independent of the policy π. □

We now consider its derivative optimization problem constrained by source critic feedback analogous to (6.15):

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J_f(\pi; \gamma) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_f(q^\pi || p_{\mathcal{D}}^\tau; \gamma) \\ \text{s.t.} \quad & \mathbb{E}_{x \sim p_{\mathcal{D}}^\tau(x)} c^*(x) - \mathbb{E}_{x \sim q^\pi(x)} c^*(x) \leq W_0 \end{aligned} \tag{D.9}$$

where $c^*(x)$ is again an adversarial source critic and W_0 is some nonnegative constant. We can show that (D.9) admits an explicit dual function which can be used to tractably solve this optimization problem.

Lemma 10 (Explicit Dual Function of (D.9))**.** *Consider the primal problem*

$$\begin{aligned} \max_{\pi \in \Pi} \quad & J_f(\pi; \gamma) = \mathbb{E}_{q^\pi}[r_\theta(x)] - \frac{\beta}{\tau} D_f(q^\pi || p_{\mathcal{D}}^\tau; \gamma) \\ \textit{s.t.} \quad & \mathbb{E}_{x \sim p_{\mathcal{D}}^\tau(x)}[c^*(x)] - \mathbb{E}_{x \sim q^\pi(x)}[c^*(x)] \leq W_0 \end{aligned} \tag{D.10}$$

for some convex function f where $0 \notin dom(f)$. The Lagrangian dual function $g(\lambda)$ is bounded from below by the function $g_\ell(\lambda)$ given by

$$g_\ell(\lambda) := \beta \left[(1+\gamma)\lambda(\mathbb{E}_{p_{\mathcal{D}}^\tau}[c^*(x)] - W_0) - \mathbb{E}_{p_{\mathcal{D}}^\tau} e^{\lambda c^*(x) - 1} - \gamma \mathbb{E}_{p_{\mathcal{D}}^\tau} f^\star(\lambda c^*(x)) \right] \tag{D.11}$$

where $f^\star(\cdot)$ is the Fenchel conjugate of f.

Proof. Recall that the generator function of the mixed f-divergence penalty is given by $f_\gamma(u) = \gamma f(u) + u \log u$ for some weighting scalar $\gamma \in [1, +\infty)$. Define $f_{\mathrm{KL}}(u) := u \log u$. From (6.18), the

dual function $g(\lambda) : \mathbb{R}_+ \to \mathbb{R}$ of the primal problem is given by

$$\begin{aligned} g(\lambda) &:= \min_{\pi \in \Pi} \Big[(1+\beta)\mathbb{E}_{p_{\mathcal{D}}^\tau} f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + \beta\gamma \mathbb{E}_{p_{\mathcal{D}}^\tau} f\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) \\ &\qquad - \mathbb{E}_{q^\pi} \log\left(q^\pi\right) + \beta(1+\gamma)\lambda\left(\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \mathbb{E}_{q^\pi} c^*(x) - W_0\right) \Big] \\ &= \min_{\pi \in \Pi} \Big[(1+\beta)\mathbb{E}_{p_{\mathcal{D}}^\tau} f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + \beta\gamma \mathbb{E}_{p_{\mathcal{D}}^\tau} f\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) - \left(\mathbb{E}_{p_{\mathcal{D}}^\tau} f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + \mathbb{E}_{q^\pi} \log p_{\mathcal{D}}^\tau\right) \\ &\qquad + \beta(1+\gamma)\lambda\left(\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \mathbb{E}_{q^\pi} c^*(x) - W_0\right) \Big] \\ &= \min_{\pi \in \Pi} \Big[\beta\mathbb{E}_{p_{\mathcal{D}}^\tau} f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + \beta\gamma \mathbb{E}_{p_{\mathcal{D}}^\tau} f\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) - \mathbb{E}_{q^\pi} \log p_{\mathcal{D}}^\tau \\ &\qquad + \beta(1+\gamma)\lambda\left(\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \mathbb{E}_{q^\pi} c^*(x) - W_0\right) \Big] \end{aligned} \tag{D.12}$$

where we define $\beta(1+\gamma)\lambda \in \mathbb{R}_+$ as the Lagrangian multiplier associated with the constraint in (D.9) (recall that $\mathbb{R}_+$ is closed under multiplication). We rearrange terms to rewrite $g(\lambda)$ as

$$\begin{aligned} g(\lambda) = \min_{\pi \in \Pi} \Bigg[& \beta\mathbb{E}_{p_{\mathcal{D}}^\tau}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}}\right)\right] \\ & + \beta\gamma\mathbb{E}_{p_{\mathcal{D}}^\tau}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f\left(\frac{q^\pi}{p_{\mathcal{D}}^\tau}\right)\right] \\ & - \mathbb{E}_{q^\pi} \log p_{\mathcal{D}}^\tau + \beta(1+\gamma)\lambda\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \beta(1+\gamma)\lambda W_0 \Bigg] \end{aligned} \tag{D.13}$$

The sum of function minima is a lower bound on the minima of the sum:

$$\begin{aligned} g(\lambda) &\geq \beta\mathbb{E}_{p_{\mathcal{D}}^\tau} \min_{\pi \in \Pi}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}}\right)\right] \\ &\quad + \beta\gamma\mathbb{E}_{p_{\mathcal{D}}^\tau} \min_{\pi \in \Pi}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f\left(\frac{q^\pi}{p_{\mathcal{D}}}\right)\right] \\ &\quad - \max_{\pi \in \Pi} \mathbb{E}_{q^\pi} \log p_{\mathcal{D}}^\tau + \min_{\pi \in \Pi}\left[\beta(1+\gamma)\lambda\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \beta(1+\gamma)\lambda W_0\right] \\ &\sim \beta\mathbb{E}_{p_{\mathcal{D}}^\tau} \min_{\pi \in \Pi}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f_{\mathrm{KL}}\left(\frac{q^\pi}{p_{\mathcal{D}}}\right)\right] \\ &\quad + \beta\gamma\mathbb{E}_{p_{\mathcal{D}}^\tau} \min_{\pi \in \Pi}\left[-\left(\lambda c^*(x) \cdot \frac{q^\pi}{p_{\mathcal{D}}^\tau}\right) + f\left(\frac{q^\pi}{p_{\mathcal{D}}}\right)\right] \\ &\quad + \beta(1+\gamma)\lambda\mathbb{E}_{p_{\mathcal{D}}^\tau} c^*(x) - \beta(1+\gamma)\lambda W_0 \end{aligned} \tag{D.14}$$

ignoring the term $\max_{\pi \in \Pi}\left[\mathbb{E}_{q^\pi} \log p_{\mathcal{D}}^\tau\right]$ that is constant with respect to λ. We then perform the same

tactic of minimizing over the superset $\mathbb{R}_+ \supseteq \{z \mid \exists \pi \in \Pi \text{ s.t. } q^\pi(x)/p^\tau_{\mathcal{D}}(x) = z\}$ as in our proof for **Lemma 6**:

$$\begin{aligned} g(\lambda) &\geq \beta \mathbb{E}_{p^\tau_{\mathcal{D}}} \min_{z \in \mathbb{R}_+} \left[-\left(\lambda c^*(x) \cdot z\right) + f_{\mathrm{KL}}\left(z\right)\right] \\ &\quad + \beta\gamma \mathbb{E}_{p^\tau_{\mathcal{D}}} \min_{z \in \mathbb{R}_+} \left[-\left(\lambda c^*(x) \cdot z\right) + f(z)\right] + \beta(1+\gamma)\lambda(\mathbb{E}_{p^\tau_{\mathcal{D}}} c^*(x) - W_0) \\ &= \beta \left[-\mathbb{E}_{p^\tau_{\mathcal{D}}} f^\star_{\mathrm{KL}}(\lambda c^*(x)) - \gamma \mathbb{E}_{p^\tau_{\mathcal{D}}} f^\star(\lambda c^*(x)) + (1+\gamma)\lambda(\mathbb{E}_{p^\tau_{\mathcal{D}}} c^*(x) - W_0)\right] \end{aligned} \tag{D.15}$$

where $f^\star(\cdot)$ is the Fenchel conjugate of a convex function $f(\cdot)$. The Fenchel conjugate of $f_{\mathrm{KL}}(u) = u \log u$ is $f^\star_{\mathrm{KL}}(v) = e^{v-1}$ (Borwein and Lewis, 2006), so

$$g(\lambda) \geq \beta \left[-\mathbb{E}_{p^\tau_{\mathcal{D}}} e^{\lambda c^*(x) - 1} - \gamma \mathbb{E}_{p^\tau_{\mathcal{D}}} f^\star(\lambda c^*(x)) + (1+\gamma)\lambda(\mathbb{E}_{p^\tau_{\mathcal{D}}} c^*(x) - W_0)\right] \tag{D.16}$$

Define the right hand side of this inequality as the function $g_\ell(\lambda)$ and the result is immediate. □

Corollary 1 (Explicit Dual Function of (D.9) Using Mixed χ^2-Divergence)**.** *As an example, we can consider the mixed χ^2-divergence defined by $D_{\chi^2}(Q||P;\gamma) = \gamma D_{\chi^2}(Q||P) + D_{KL}(Q||P)$ (Huang et al., 2024a). The χ^2-divergence generator function is $f_{\chi^2}(u) = (u-1)^2/2$, and its Fenchel conjugate is $f^\star_{\chi^2}(v) = v + (v^2/2)$ from **Lemma 3**. Directly applying **Lemma 10**, our lower bound on the our dual function is*

$$\begin{aligned} g(\lambda) \geq g_\ell(\lambda) := \beta \Bigg[&- \mathbb{E}_{p^\tau_{\mathcal{D}}} e^{\lambda c^*(x) - 1} - \gamma \mathbb{E}_{p^\tau_{\mathcal{D}}} \left(\frac{1}{2}(\lambda c^*(x))^2 + \lambda c^*(x) \right) \\ &+ (1+\gamma)\lambda(\mathbb{E}_{p^\tau_{\mathcal{D}}} c^*(x) - W_0) \Bigg] \end{aligned} \tag{D.17}$$

To experimentally evaluate the utility of distribution matching using a mixed χ^2-KL-Divergence, we substitute the $D_{\mathrm{KL}}(\cdot||\cdot)$ divergence with the mixed χ^2-Divergence $D_{f_{\chi^2}}(\cdot||\cdot;\gamma)$ (setting $\gamma = 1.0$ for experimental evaluation) and its associated dual function bound from (D.17) into **Algorithm 3**. Practically, we find that this only requires updating the dual function in **Algorithm 3** per **Corol-**

lary 1 and the Lagrangian of (D.9) given by

$$\begin{aligned}\mathcal{L}(x;\lambda) = -\mathbb{E}_{q^\pi}[r_\theta(x)] + \frac{\beta}{\tau}\left[\gamma D_f(q^\pi||p^\tau_{\mathcal{D}}) + D_{\text{KL}}(q^\pi||p^\tau_{\mathcal{D}})\right] \\ + \beta(1+\gamma)\lambda\left[\mathbb{E}_{p^\tau_{\mathcal{D}}}c^*(x) - \mathbb{E}_{q^\pi}c^*(x) - W_0\right]\end{aligned} \tag{D.18}$$

Experimental Results. We compare DynAMO implemented with a mixed χ^2 divergence penalty (with $\gamma = 1.0$) against our original DynAMO implementation (i.e., $\gamma = 0$) in **Supp. Tables D.4-D.5**. Empirically, we find that using the mixed χ^2-divergence penalty offers limited utility compared with KL-divergence alone: the latter is non-inferior to the former according to both the Rank and Optimality Gap metrics for all 6 optimizers assessed according to the Best@128 oracle score. Furthermore, DynAMO outperforms DynAMO with mixed χ^2-divergence according to the Rank and Optimality Gap metrics for 5 out of the 6 optimizers assessed according to the Pairwise Diversity metric. Based on our qualitative analysis, we hypothesize that the over-conservatism often attributed to χ^2-divergence-based penalties in related literature (Ma et al., 2022; Huang et al., 2024a; Wang et al., 2024a) may adversely affect the generative policy's ability to sufficiently explore the design space when compared to using KL-divergence-base distribution matching alone. Further work is needed to tune the relative mixing parameter γ and/or explore how other alternative f-divergence metrics may be used with DynAMO.

D.4. Comparison with Offline Model-Free Optimization Methods

In our main experimental results reported in **Section 6.5**, we compare DynAMO against other *model-based optimization* (MBO) methods—that is, optimization methods that explicitly (1) learn a proxy forward surrogate model $r_\theta(x)$ for the oracle reward function from the offline dataset; and (2) optimize against $r_\theta(x)$ and rank final candidate designs according to a scoring metric involving r_θ. Alternatively, recent work have also proposed methods that instead do *not* learn a forward surrogate model $r_\theta(x)$; we refer to such methods as *model-free* algorithms.

Survey of Existing Model-Free and Additional Model-Based Methods. Krishnamoorthy et al. (2023b) introduce **BONET** (i.e., **B**lack-box **O**ptimization **Net**works), which learns an autoregressive model on synthetically constructed optimization trajectories that simulate runs of implicit

Table D.4: **Quality of design candidates using mixed χ^2-divergence DynAMO.** Using **Corollary 1** and (D.9), we show that it is possible to extend DynAMO to leverage a mixed χ^2-divergence that equally weights both χ^2-divergence and KL-divergence to penalize the original MBO objective. We evaluate this specialized implementation of DynAMO against baseline DynAMO and vanilla optimization methods, and report the Best@128 (resp., Median@128) oracle score achieved by the 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Best@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	43.9	59.4	60.5	88.9	40.0	88.4	—	—
Grad.	**90.0**$^{(4.3)}$	**80.9**$^{(12.1)}$	**60.2**$^{(8.9)}$	88.8$^{(4.0)}$	**36.0**$^{(6.8)}$	65.6$^{(14.5)}$	2.8	6.8
DynAMO-Grad.	**90.3**$^{(4.7)}$	**86.2**$^{(0.0)}$	**64.4**$^{(2.5)}$	91.2$^{(0.0)}$	**44.2**$^{(7.8)}$	**89.8**$^{(3.2)}$	**1.2**	**14.2**
Mixed χ^2 DynAMO-Grad.	59.3$^{(8.3)}$	**86.2**$^{(0.0)}$	**64.4**$^{(2.6)}$	**120**$^{(1.4)}$	**42.0**$^{(5.6)}$	83.6$^{(1.4)}$	2.0	12.5
Adam	62.9$^{(13.0)}$	69.7$^{(10.5)}$	**62.9**$^{(1.9)}$	92.3$^{(8.9)}$	**37.8**$^{(6.3)}$	**58.4**$^{(18.5)}$	2.7	0.5
DynAMO-Adam	**95.2**$^{(1.7)}$	**86.2**$^{(0.0)}$	**65.2**$^{(1.1)}$	91.2$^{(0.0)}$	**45.5**$^{(5.7)}$	**84.9**$^{(12.0)}$	**1.3**	**14.5**
Mixed χ^2 DynAMO-Adam	59.3$^{(8.3)}$	86.2$^{(0.0)}$	**64.4**$^{(2.6)}$	**120**$^{(1.4)}$	**42.0**$^{(5.6)}$	**83.6**$^{(1.4)}$	2.0	12.5
BO-qEI	**87.3**$^{(5.8)}$	**86.2**$^{(0.0)}$	**65.4**$^{(1.0)}$	116$^{(3.1)}$	**53.1**$^{(3.3)}$	**84.4**$^{(0.9)}$	2.8	18.7
DynAMO-BO-qEI	**91.9**$^{(4.4)}$	**86.2**$^{(0.0)}$	**67.0**$^{(1.3)}$	**121**$^{(0.0)}$	**53.5**$^{(5.0)}$	**85.5**$^{(1.1)}$	**1.5**	**20.7**
Mixed χ^2 DynAMO-BO-qEI	**92.2**$^{(4.1)}$	**86.3**$^{(0.1)}$	**66.8**$^{(1.2)}$	**123**$^{(3.0)}$	**51.7**$^{(4.4)}$	**85.1**$^{(1.5)}$	1.7	**20.7**
BO-qUCB	**88.1**$^{(5.3)}$	**86.2**$^{(0.1)}$	**66.4**$^{(0.7)}$	**121**$^{(1.3)}$	**51.3**$^{(3.6)}$	**84.5**$^{(0.8)}$	2.2	19.4
DynAMO-BO-qUCB	**95.1**$^{(1.9)}$	**86.2**$^{(0.0)}$	**66.7**$^{(1.5)}$	**121**$^{(0.0)}$	**48.1**$^{(4.0)}$	**86.9**$^{(4.5)}$	**1.7**	**20.5**
Mixed χ^2 DynAMO-BO-qUCB	85.7$^{(5.4)}$	**86.3**$^{(0.2)}$	**66.3**$^{(0.9)}$	**121**$^{(0.0)}$	**51.5**$^{(4.3)}$	**83.8**$^{(1.1)}$	2.0	19.0
CMA-ES	**87.6**$^{(8.3)}$	**86.2**$^{(0.0)}$	**66.1**$^{(1.0)}$	**106**$^{(5.9)}$	**49.0**$^{(1.0)}$	72.2$^{(0.1)}$	2.0	14.4
DynAMO-CMA-ES	**89.8**$^{(3.6)}$	**85.7**$^{(5.8)}$	63.9$^{(0.9)}$	**117**$^{(6.7)}$	**50.6**$^{(4.8)}$	**78.5**$^{(5.5)}$	**1.7**	**17.5**
Mixed χ^2 DynAMO-CMA-ES	**84.2**$^{(10.7)}$	**84.5**$^{(2.6)}$	**65.1**$^{(1.3)}$	**113**$^{(4.8)}$	**45.0**$^{(4.9)}$	**81.8**$^{(4.0)}$	2.3	15.4
CoSyNE	61.7$^{(10.0)}$	57.3$^{(9.6)}$	**63.6**$^{(0.4)}$	94.8$^{(10.1)}$	**37.0**$^{(4.1)}$	62.7$^{(1.3)}$	2.8	-0.6
DynAMO-CoSyNE	**91.3**$^{(4.4)}$	**77.2**$^{(11.6)}$	**63.9**$^{(0.9)}$	**114**$^{(7.0)}$	**40.6**$^{(8.6)}$	67.5$^{(1.4)}$	**1.5**	**12.3**
Mixed χ^2 DynAMO-CoSyNE	**94.3**$^{(2.3)}$	**78.3**$^{(8.3)}$	**63.1**$^{(2.2)}$	**100**$^{(10.9)}$	**37.4**$^{(9.7)}$	**82.3**$^{(4.1)}$	1.7	**12.4**

Median@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	33.7	42.8	50.9	87.6	6.7	77.8	—	—
Grad.	**58.1**$^{(6.1)}$	**58.6**$^{(13.1)}$	**59.3**$^{(8.6)}$	**85.3**$^{(7.7)}$	**36.0**$^{(6.7)}$	**65.1**$^{(14.4)}$	2.0	**10.5**
DynAMO-Grad.	47.0$^{(2.8)}$	**69.8**$^{(6.0)}$	**61.9**$^{(2.2)}$	**85.9**$^{(0.4)}$	**23.4**$^{(8.5)}$	**68.7**$^{(12.1)}$	**1.5**	9.5
Mixed χ^2 DynAMO-Grad.	45.2$^{(6.9)}$	**66.9**$^{(5.2)}$	**58.3**$^{(6.5)}$	**86.6**$^{(2.1)}$	20.6$^{(1.4)}$	**64.4**$^{(8.3)}$	2.5	7.1
Adam	**54.7**$^{(8.8)}$	**60.4**$^{(12.7)}$	**59.2**$^{(8.6)}$	**87.9**$^{(10.0)}$	**37.4**$^{(6.2)}$	**56.8**$^{(19.8)}$	1.8	**9.5**
DynAMO-Adam	**47.7**$^{(3.0)}$	**69.0**$^{(5.2)}$	**62.4**$^{(1.9)}$	**86.4**$^{(0.6)}$	23.0$^{(6.0)}$	**65.6**$^{(14.1)}$	**1.7**	9.1
Mixed χ^2 DynAMO-Adam	**45.2**$^{(6.9)}$	**66.9**$^{(5.2)}$	**58.3**$^{(6.5)}$	**86.6**$^{(2.1)}$	20.6$^{(1.4)}$	**64.4**$^{(8.3)}$	2.5	7.1
BO-qEI	48.5$^{(1.5)}$	59.9$^{(2.0)}$	**63.3**$^{(0.0)}$	**86.7**$^{(0.6)}$	**28.7**$^{(1.8)}$	72.4$^{(1.8)}$	1.8	10.0
DynAMO-BO-qEI	**51.5**$^{(0.9)}$	**65.6**$^{(3.1)}$	**63.3**$^{(0.0)}$	**86.7**$^{(0.6)}$	23.5$^{(2.4)}$	**77.0**$^{(0.7)}$	**1.7**	**11.3**
Mixed χ^2 DynAMO-BO-qEI	44.8$^{(1.4)}$	**66.1**$^{(3.0)}$	**63.3**$^{(0.0)}$	**86.4**$^{(0.6)}$	**27.7**$^{(3.5)}$	73.9$^{(2.8)}$	2.3	10.4
BO-qUCB	**50.3**$^{(1.8)}$	62.1$^{(3.4)}$	**63.3**$^{(0.0)}$	**86.6**$^{(0.6)}$	**31.7**$^{(1.2)}$	**74.4**$^{(0.6)}$	**1.5**	**11.5**
DynAMO-BO-qUCB	**48.8**$^{(1.8)}$	**65.9**$^{(3.7)}$	**63.3**$^{(0.0)}$	**86.5**$^{(0.5)}$	22.7$^{(2.0)}$	50.4$^{(14.6)}$	2.0	6.3
Mixed χ^2 DynAMO-BO-qUCB	44.0$^{(0.7)}$	**68.2**$^{(3.0)}$	**63.3**$^{(0.0)}$	**86.5**$^{(0.6)}$	20.9$^{(1.3)}$	**74.5**$^{(2.0)}$	2.0	9.6
CMA-ES	**50.7**$^{(2.7)}$	**71.7**$^{(10.4)}$	**63.3**$^{(0.0)}$	83.9$^{(1.0)}$	**37.9**$^{(0.7)}$	**59.3**$^{(10.9)}$	**1.5**	**11.2**
DynAMO-CMA-ES	**45.3**$^{(2.4)}$	**65.8**$^{(8.9)}$	59.3$^{(3.8)}$	**99.0**$^{(12.1)}$	22.5$^{(5.1)}$	**60.6**$^{(15.0)}$	2.2	8.8
Mixed χ^2 DynAMO-CMA-ES	**48.5**$^{(3.0)}$	**70.0**$^{(6.5)}$	**63.2**$^{(0.3)}$	**87.0**$^{(2.0)}$	19.4$^{(3.8)}$	**43.7**$^{(14.6)}$	2.3	5.4
CoSyNE	**55.3**$^{(8.0)}$	**53.6**$^{(10.2)}$	**60.8**$^{(3.2)}$	**87.4**$^{(16.6)}$	**36.6**$^{(4.4)}$	**59.3**$^{(14.5)}$	2.3	8.9
DynAMO-CoSyNE	**53.8**$^{(11.0)}$	**63.4**$^{(11.5)}$	**59.3**$^{(3.8)}$	**99.0**$^{(12.1)}$	20.5$^{(5.8)}$	**60.6**$^{(15.0)}$	2.3	9.5
Mixed χ^2 DynAMO-CoSyNE	**59.9**$^{(9.8)}$	**65.4**$^{(9.9)}$	**60.9**$^{(3.1)}$	**89.3**$^{(14.8)}$	23.4$^{(4.5)}$	**66.6**$^{(5.9)}$	**1.3**	**11.0**

Table D.5: **Diversity of design candidates using mixed χ^2-divergence DynAMO.** Using **Corollary 1** and (D.9), we show that it is possible to extend DynAMO to leverage a *mixed χ^2-divergence* that equally weights both χ^2-divergence and KL-divergence to penalize the original MBO objective. We evaluate this specialized implementation of DynAMO against baseline DynAMO and vanilla optimization methods, and report the pairwise diversity oracle score achieved by the 128 evaluated designs. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer. Minimum novelty and L_1 coverage scores are reported in **Supp. Table D.6**.

Pairwise Diversity@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	65.9	57.3	60.0	36.7	66.0	85.7	—	—
Grad.	$12.5^{(8.0)}$	$7.8^{(8.8)}$	$7.9^{(7.8)}$	$24.1^{(13.3)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	3.0	-53.2
DynAMO-Grad.	**66.9**$^{(6.9)}$	**68.2**$^{(10.8)}$	**77.2**$^{(21.5)}$	**93.0**$^{(1.2)}$	**129**$^{(55.3)}$	**104**$^{(56.1)}$	**1.3**	**27.8**
Mixed χ^2 DynAMO-Grad.	$16.8^{(12.6)}$	**72.6**$^{(1.1)}$	**47.0**$^{(31.1)}$	**91.0**$^{(1.6)}$	**182**$^{(45.9)}$	**74.2**$^{(3.0)}$	1.7	18.7
Adam	$12.0^{(12.3)}$	$11.0^{(12.1)}$	$4.8^{(3.8)}$	$16.8^{(12.4)}$	$6.4^{(14.5)}$	$6.2^{(14.0)}$	3.0	-52.4
DynAMO-Adam	**54.8**$^{(8.9)}$	**72.3**$^{(3.4)}$	$84.8^{(9.2)}$	**89.9**$^{(5.3)}$	**158**$^{(37.3)}$	**126**$^{(57.3)}$	1.7	**35.7**
Mixed χ^2 DynAMO-Adam	$16.8^{(12.6)}$	**72.6**$^{(1.1)}$	**99.2**$^{(0.7)}$	**91.0**$^{(1.6)}$	**182**$^{(45.9)}$	**74.2**$^{(3.0)}$	**1.3**	27.4
BO-qEI	$73.7^{(0.6)}$	$73.8^{(0.5)}$	**99.3**$^{(0.1)}$	$93.0^{(0.5)}$	$190^{(0.8)}$	$124^{(7.4)}$	2.5	47.1
DynAMO-BO-qEI	**74.8**$^{(0.2)}$	**74.6**$^{(0.3)}$	**99.4**$^{(0.1)}$	**93.5**$^{(0.4)}$	**198**$^{(1.9)}$	**277**$^{(59.7)}$	**1.3**	**74.2**
Mixed χ^2 DynAMO-BO-qEI	$73.6^{(0.7)}$	**74.3**$^{(0.5)}$	**99.4**$^{(0.0)}$	**93.9**$^{(0.4)}$	**167**$^{(41.5)}$	$29.1^{(7.7)}$	2.2	27.5
BO-qUCB	**73.9**$^{(0.5)}$	**74.3**$^{(0.4)}$	**99.4**$^{(0.1)}$	**93.6**$^{(0.5)}$	**198**$^{(10.3)}$	$94.1^{(3.9)}$	2.2	43.5
DynAMO-BO-qUCB	**74.3**$^{(0.5)}$	**74.4**$^{(0.6)}$	$99.3^{(0.1)}$	**93.5**$^{(0.6)}$	**211**$^{(22.8)}$	**175**$^{(44.7)}$	**1.7**	**59.4**
Mixed χ^2 DynAMO-BO-qUCB	**73.4**$^{(0.7)}$	**74.3**$^{(0.4)}$	**99.5**$^{(0.1)}$	**93.7**$^{(0.5)}$	**177**$^{(25.2)}$	$28.1^{(7.3)}$	2.2	29.0
CMA-ES	$47.2^{(11.2)}$	$44.6^{(15.9)}$	**93.5**$^{(2.0)}$	$66.2^{(9.4)}$	$12.8^{(0.6)}$	$164^{(10.6)}$	2.5	9.5
DynAMO-CMA-ES	**73.6**$^{(0.6)}$	**73.1**$^{(3.1)}$	$72.0^{(3.1)}$	**94.0**$^{(0.5)}$	**97.8**$^{(13.2)}$	**292**$^{(83.5)}$	**1.3**	**55.2**
Mixed χ^2 DynAMO-CMA-ES	**52.9**$^{(20.8)}$	**51.5**$^{(22.1)}$	**67.7**$^{(24.7)}$	$69.7^{(12.4)}$	**154**$^{(107)}$	$86.9^{(72.9)}$	2.2	18.6
CoSyNE	$5.6^{(5.0)}$	**12.7**$^{(9.8)}$	**28.2**$^{(11.3)}$	**12.2**$^{(7.3)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	3.0	-52.1
DynAMO-CoSyNE	**18.1**$^{(13.0)}$	**20.3**$^{(2.3)}$	**35.0**$^{(17.9)}$	**22.8**$^{(11.9)}$	**74.4**$^{(46.3)}$	**77.0**$^{(35.9)}$	1.7	-20.7
Mixed χ^2 DynAMO-CoSyNE	**22.4**$^{(8.1)}$	**34.1**$^{(20.1)}$	**34.1**$^{(20.1)}$	**23.0**$^{(17.4)}$	**104**$^{(77.9)}$	**49.4**$^{(25.6)}$	**1.3**	**-17.5**

Table D.6: **Diversity of design candidates using mixed χ^2-divergence DynAMO (cont.).** Using **Corollary 1** and (D.9), we show that it is possible to extend DynAMO to leverage a *mixed* χ^2-*divergence* that equally weights both χ^2-divergence and KL-divergence to penalize the original MBO objective. We evaluate this specialized implementation of DynAMO against baseline DynAMO and vanilla optimization methods, and report the minimum novelty and L_1 coverage oracle scores achieved by the 128 evaluated designs. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer. Pairwise diversity scores are reported in **Supp. Table D.5**.

Minimum Novelty@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	0.0	0.0	0.0	0.0	0.0	0.0	—	—
Grad.	**21.2**$^{(3.0)}$	**51.7**$^{(2.9)}$	**97.4**$^{(3.9)}$	**79.5**$^{(19.7)}$	**95.0**$^{(0.7)}$	**102**$^{(6.1)}$	2.3	74.5
DynAMO-Grad.	**21.1**$^{(1.1)}$	**52.2**$^{(1.3)}$	**98.6**$^{(1.5)}$	**85.8**$^{(1.0)}$	**95.0**$^{(0.4)}$	**107**$^{(6.7)}$	**1.3**	**76.7**
Mixed χ^2 DynAMO-Grad.	14.6$^{(2.8)}$	**51.9**$^{(0.4)}$	**99.2**$^{(0.7)}$	**85.2**$^{(2.1)}$	85.2$^{(1.6)}$	34.0$^{(1.1)}$	2.3	61.7
Adam	**23.7**$^{(2.8)}$	**51.1**$^{(3.5)}$	**95.5**$^{(5.3)}$	**79.3**$^{(21.2)}$	**94.8**$^{(0.7)}$	**103**$^{(6.3)}$	2.0	74.5
DynAMO-Adam	14.7$^{(1.9)}$	46.2$^{(0.5)}$	**98.7**$^{(1.2)}$	**85.9**$^{(1.8)}$	**94.9**$^{(0.4)}$	**108**$^{(7.2)}$	**1.7**	**74.7**
Mixed χ^2 DynAMO-Adam	**20.4**$^{(3.3)}$	**51.9**$^{(0.4)}$	**87.3**$^{(57.9)}$	**85.2**$^{(2.1)}$	85.2$^{(1.6)}$	34.0$^{(1.1)}$	2.3	60.7
BO-qEI	**21.8**$^{(0.5)}$	**51.5**$^{(0.3)}$	**97.6**$^{(0.3)}$	**85.4**$^{(1.5)}$	94.6$^{(0.1)}$	106$^{(2.9)}$	**1.8**	76.2
DynAMO-BO-qEI	**21.0**$^{(0.5)}$	**51.9**$^{(0.2)}$	**97.4**$^{(0.4)}$	**85.2**$^{(0.9)}$	**94.8**$^{(0.1)}$	**126**$^{(14.6)}$	2.0	**79.4**
Mixed χ^2 DynAMO-BO-qEI	14.6$^{(0.5)}$	**51.9**$^{(0.4)}$	**97.4**$^{(0.5)}$	**85.5**$^{(1.3)}$	80.8$^{(3.7)}$	16.6$^{(4.5)}$	2.2	57.8
BO-qUCB	**21.6**$^{(0.3)}$	51.7$^{(0.2)}$	**97.9**$^{(0.4)}$	**85.3**$^{(1.1)}$	93.8$^{(0.6)}$	98.8$^{(1.1)}$	**1.8**	74.8
DynAMO-BO-qUCB	**21.4**$^{(0.5)}$	51.7$^{(0.2)}$	**97.1**$^{(0.5)}$	**85.3**$^{(1.1)}$	**94.7**$^{(0.2)}$	**109**$^{(4.5)}$	2.0	**76.6**
Mixed χ^2 DynAMO-BO-qUCB	19.8$^{(0.3)}$	**52.0**$^{(0.1)}$	**97.4**$^{(0.4)}$	**85.6**$^{(1.3)}$	79.8$^{(3.5)}$	14.9$^{(3.3)}$	2.2	58.2
CMA-ES	16.5$^{(2.1)}$	47.8$^{(1.0)}$	96.5$^{(0.7)}$	**73.0**$^{(18.0)}$	**100**$^{(0.0)}$	100$^{(0.0)}$	2.3	72.3
DynAMO-CMA-ES	12.9$^{(0.8)}$	48.0$^{(1.6)}$	**96.7**$^{(3.5)}$	**81.8**$^{(13.4)}$	94.5$^{(0.7)}$	**112**$^{(7.8)}$	2.0	**74.3**
Mixed χ^2 DynAMO-CMA-ES	**23.0**$^{(1.6)}$	**51.8**$^{(0.4)}$	**98.0**$^{(0.9)}$	**83.8**$^{(9.9)}$	87.1$^{(2.1)}$	48.6$^{(16.8)}$	**1.7**	65.4
CoSyNE	**24.5**$^{(3.5)}$	**49.7**$^{(3.1)}$	**98.5**$^{(1.6)}$	**86.6**$^{(12.7)}$	**93.2**$^{(1.0)}$	91.9$^{(2.0)}$	**1.7**	74.1
DynAMO-CoSyNE	**17.8**$^{(5.5)}$	**48.4**$^{(2.3)}$	**96.7**$^{(3.5)}$	**80.2**$^{(12.9)}$	**94.5**$^{(0.7)}$	**112**$^{(7.8)}$	2.3	**75.0**
Mixed χ^2 DynAMO-CoSyNE	**23.5**$^{(3.7)}$	**52.5**$^{(2.0)}$	**99.9**$^{(0.1)}$	**80.6**$^{(21.5)}$	83.9$^{(2.6)}$	30.3$^{(4.2)}$	2.0	61.8

L_1 **Coverage@128**	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	0.42	0.31	1.42	0.68	6.26	0.58	—	—
Grad.	0.16$^{(0.10)}$	0.20$^{(0.13)}$	0.21$^{(0.10)}$	0.42$^{(0.18)}$	0.00$^{(0.00)}$	0.00$^{(0.00)}$	3.0	-1.44
DynAMO-Grad.	**0.36**$^{(0.04)}$	**0.52**$^{(0.06)}$	**1.46**$^{(0.38)}$	**2.49**$^{(0.06)}$	**6.47**$^{(1.24)}$	**5.85**$^{(1.35)}$	**1.3**	**1.25**
Mixed χ^2 DynAMO-Grad.	0.16$^{(0.09)}$	**0.54**$^{(0.02)}$	**0.87**$^{(0.58)}$	2.20$^{(0.10)}$	**6.67**$^{(1.68)}$	1.61$^{(0.03)}$	1.7	0.40
Adam	0.11$^{(0.06)}$	0.22$^{(0.09)}$	0.23$^{(0.15)}$	0.48$^{(0.31)}$	0.27$^{(0.55)}$	0.24$^{(0.49)}$	3.0	-1.35
DynAMO-Adam	**0.33**$^{(0.05)}$	**0.55**$^{(0.03)}$	**1.44**$^{(0.39)}$	**2.40**$^{(0.16)}$	**7.06**$^{(0.73)}$	**6.91**$^{(0.71)}$	**1.0**	**1.50**
Mixed χ^2 DynAMO-Adam	0.16$^{(0.09)}$	**0.54**$^{(0.02)}$	0.47$^{(0.31)}$	**2.20**$^{(0.10)}$	**6.67**$^{(1.68)}$	1.61$^{(0.03)}$	2.0	0.33
BO-qEI	**0.41**$^{(0.02)}$	**0.55**$^{(0.01)}$	2.37$^{(0.03)}$	2.11$^{(0.15)}$	7.84$^{(0.01)}$	6.61$^{(0.33)}$	2.3	1.70
DynAMO-BO-qEI	**0.42**$^{(0.01)}$	**0.56**$^{(0.01)}$	**2.47**$^{(0.03)}$	**2.54**$^{(0.03)}$	**7.87**$^{(0.01)}$	**7.92**$^{(0.04)}$	**1.2**	**2.02**
Mixed χ^2 DynAMO-BO-qEI	0.39$^{(0.02)}$	**0.55**$^{(0.01)}$	2.39$^{(0.04)}$	**2.56**$^{(0.02)}$	6.11$^{(0.64)}$	1.36$^{(0.12)}$	2.5	0.62
BO-qUCB	**0.40**$^{(0.02)}$	0.54$^{(0.01)}$	**2.40**$^{(0.05)}$	**2.52**$^{(0.07)}$	7.78$^{(0.04)}$	6.64$^{(0.09)}$	2.5	1.77
DynAMO-BO-qUCB	**0.40**$^{(0.02)}$	**0.55**$^{(0.01)}$	**2.47**$^{(0.07)}$	**2.54**$^{(0.05)}$	**7.88**$^{(0.03)}$	**7.80**$^{(0.23)}$	**1.2**	**2.00**
Mixed χ^2 DynAMO-BO-qUCB	**0.39**$^{(0.02)}$	**0.56**$^{(0.01)}$	**2.41**$^{(0.05)}$	**2.53**$^{(0.06)}$	5.96$^{(0.78)}$	1.38$^{(0.11)}$	2.3	0.59
CMA-ES	**0.33**$^{(0.05)}$	0.48$^{(0.04)}$	**2.18**$^{(0.04)}$	1.82$^{(0.12)}$	**3.26**$^{(1.42)}$	**3.77**$^{(1.36)}$	2.3	0.36
DynAMO-CMA-ES	**0.40**$^{(0.03)}$	**0.56**$^{(0.01)}$	**1.82**$^{(0.72)}$	**2.54**$^{(0.05)}$	**4.75**$^{(2.16)}$	**3.29**$^{(1.56)}$	**1.7**	0.62
Mixed χ^2 DynAMO-CMA-ES	**0.30**$^{(0.09)}$	**0.52**$^{(0.05)}$	**1.56**$^{(0.68)}$	1.97$^{(0.19)}$	**5.58**$^{(1.68)}$	**4.03**$^{(3.01)}$	2.0	**0.71**
CoSyNE	**0.10**$^{(0.07)}$	**0.22**$^{(0.10)}$	**0.39**$^{(0.20)}$	**0.27**$^{(0.13)}$	0.10$^{(0.00)}$	0.10$^{(0.00)}$	2.8	-1.41
DynAMO-CoSyNE	**0.21**$^{(0.11)}$	**0.18**$^{(0.09)}$	**0.64**$^{(0.42)}$	**0.43**$^{(0.34)}$	**1.85**$^{(0.22)}$	**0.94**$^{(0.17)}$	2.0	-0.90
Mixed χ^2 DynAMO-CoSyNE	**0.20**$^{(0.04)}$	**0.37**$^{(0.14)}$	**0.74**$^{(0.47)}$	**0.53**$^{(0.42)}$	**3.85**$^{(2.79)}$	**1.10**$^{(0.55)}$	**1.2**	**-0.48**

black-box optimization experiments. The model is trained to learn a rollout of monotonic transitions from low- to high- scoring design candidates using the offline dataset. Nguyen et al. (2023) propose **ExPT** (i.e., **Ex**periment **P**retrained **T**ransformers) as a task-agnostic method of pre-training a transformer foundation model to learn an inverse modeling of designs from input reward scores and associated contexts. **DDOM** (i.e., **D**enoising **D**iffusion **O**ptimization **M**odels) learns a diffusion model conditioned on the oracle values in the offline dataset (Krishnamoorthy et al., 2023a). Similarly, **GTG** (i.e., **G**uided **T**rajectory **G**eneration) trains a diffusion model to learn from synthetically constructed optimization trajectories conditioned on final scores. **MINs** (i.e., **M**odel **I**nversion **N**etworks) from Kumar and Levine (2019) learn and optimize against an inverse mapping from reward scores to candidate designs.[6] **Tri-Mentoring** and **ICT** (i.e., **I**mportance-aware **C**o-**T**eaching) co-learn an ensemble of multiple surrogate models (Chen et al., 2023a; Yuan et al., 2023). Separately, **PGS** (i.e., **P**olicy-**G**uided **S**earch) from Chemingui et al. (2024) learns a policy to optimize against a surrogate model (although only limit their method to first-order optimization algorithms), and **Match-Opt** from Hoang et al. (2024) proposes a black-box gradient matching algorithm to learn better forward surrogate models. Finally, **RGD** (i.e., **R**obust-**G**uided **D**iffusion) uses a forward surrogate model to guide the generative sampling process from a diffusion model (Chen et al., 2024). Other model-free optimization methods have been proposed specifically for the biological sequence design problems (Kim et al., 2023; Chen et al., 2023b; Jain et al., 2022); we exclude these from our analysis and instead focus on task-agnostic optimization algorithms. We also exclude Design Editing for offline Model-based Optimization (DEMO) (Yuan et al., 2024), Noise-intensified Telescoping density-Ratio Estimation (NTRE) (Yu et al., 2024), and Ranking Models (RaM) (Tan et al., 2025) from our analysis since there are no presently available open-source implementations.

Experimental Results. We compare representative implementations of DynAMO (i.e., DynAMO with Gradient Ascent (**DynAMO-Grad.**), Bayesian optimization with Upper Confidence Bound acquisition function (**DynAMO-BO-qUCB**), and Covariance Matrix Adaptation Evolution Strat-

[6] One might argue that MINs (Kumar and Levine, 2019) are also a form of model-based optimization, as the method involves learning a surrogate function $f_\theta^{-1} : \mathbb{R} \to \mathcal{X}$. However, the method proposes a design x given an input score value, and therefore does not make available an output proxy score by which to rank candidate designs. We therefore include MINs as a *model-free* optimization algorithm.

egy (**DynAMO-CMA-ES**)) against other model-based optimization methods using the respective backbone optimizer described by the original authors (i.e., **RoMA** from Yu et al. (2021) using Adam Ascent, **COMs** from Trabucco et al. (2021) using Gradient Ascent, **ROMO** from Chen et al. (2023c) using Gradient Ascent, **GAMBO** from Yao et al. (2024) using BO-qEI) against model-free optimization methods in **Supp. Tables D.7-D.9**. We find that DynAMO-augmented optimizers can be competitive in proposing high-quality designs—in particular, DynAMO-BO-qUCB achieves both the second best Rank and Optimality Gap across all six tasks according to the Best@128 oracle score metric. However, the improvement in *diversity* of designs using DynAMO is significant: DynAMO-BO-qUCB achieves the best Rank and Optimality gap according to both the Pairwise Diversity and L_1 Coverage metrics, and DynAMO-Grad. achieves the best Rank and Optimality gap according to the Minimum Novelty metric. Furthermore, DynAMO-BO-qUCB attains the best mean Pairwise Diversity score compared to the model-free optimization methods evaluated in 5 out of the 6 tasks assessed. Altogether, our results suggest that DynAMO is a promising technique to propose a diverse set of high-quality designs compared with existing state-of-the-art offline optimization methods.

D.5. τ-Weighted Distribution Visualization

In **Definition 4**, we define the τ-weighted probability distribution to serve as the reference distribution for a generative policy to learn from in (6.15). This reference probability distribution is important and should ideally capture the diversity of *high-quality* designs contained in the offline dataset. To investigate if this is indeed the case, we plot the empirical τ-weighted distributions for each of the six offline optimization tasks in our experimental evaluation suite using $\tau = 1.0$, which is the value of the temperature hyperparameter used in our experiments in **Table 6.1**. The resulting plots are shown in **Figure D.1**; in general, we can see that our τ-weighted reference distributions weight optimal and near-optimal designs more heavily (i.e., a distribution with negative skew), while still capturing a variety of different possible designs.

Table D.7: **Comparison of design quality against model-free optimization methods.** We evaluate DynAMO and other MBO methods against model-free optimization methods. We report the maximum (resp., median) oracle score achieved out of 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. All metrics are multiplied by 100 for easier legibility. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Best@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	43.9	59.4	60.5	88.9	40.0	88.4	—	—
Grad.	$90.0^{(4.3)}$	$80.9^{(12.1)}$	$60.2^{(8.9)}$	$88.8^{(4.0)}$	$36.0^{(6.8)}$	$65.6^{(14.5)}$	16.0	6.8
BO-qUCB	$88.1^{(5.3)}$	$86.2^{(0.1)}$	$66.4^{(0.7)}$	$\mathbf{121^{(1.3)}}$	$\mathbf{51.3^{(3.6)}}$	$84.5^{(0.8)}$	7.3	19.4
CMA-ES	$87.6^{(8.3)}$	$86.2^{(0.0)}$	$66.1^{(1.0)}$	$106^{(5.9)}$	$49.0^{(1.0)}$	$72.2^{(0.1)}$	10.2	14.4
BONET	$95.5^{(0.0)}$	$\mathbf{92.9^{(0.1)}}$	$63.3^{(0.0)}$	$97.3^{(0.0)}$	$39.0^{(0.7)}$	$93.7^{(0.2)}$	8.0	16.8
DDOM	$93.0^{(3.6)}$	$85.3^{(0.5)}$	$63.5^{(0.4)}$	$87.9^{(0.6)}$	$44.7^{(2.2)}$	$63.0^{(12.1)}$	13.0	9.4
ExPT	$89.3^{(5.7)}$	$84.2^{(2.4)}$	$63.3^{(0.0)}$	$93.0^{(0.8)}$	$\mathbf{48.5^{(11.0)}}$	$82.3^{(2.3)}$	12.0	13.3
MINs	$89.0^{(3.4)}$	$68.3^{(0.6)}$	$63.9^{(0.9)}$	$93.1^{(0.7)}$	$45.8^{(2.1)}$	$91.5^{(1.1)}$	11.2	11.8
GTG	$92.1^{(0.0)}$	$70.2^{(0.0)}$	$63.3^{(0.0)}$	$85.0^{(0.0)}$	$52.5^{(0.0)}$	$\mathbf{96.4^{(0.0)}}$	9.7	13.1
Tri-Mentoring	$82.4^{(0.0)}$	$66.6^{(0.0)}$	$\mathbf{68.4^{(0.0)}}$	$88.9^{(0.0)}$	$50.9^{(1.1)}$	$94.0^{(0.0)}$	10.2	11.7
ICT	$93.3^{(3.4)}$	$66.6^{(0.0)}$	$\mathbf{68.4^{(0.0)}}$	$88.9^{(0.0)}$	$48.9^{(1.4)}$	$\mathbf{95.5^{(1.1)}}$	9.0	13.4
PGS	$79.6^{(7.5)}$	$67.1^{(0.8)}$	$\mathbf{68.4^{(0.0)}}$	$88.9^{(0.0)}$	$\mathbf{54.8^{(0.8)}}$	$72.3^{(0.0)}$	11.3	8.3
Match-Opt	$90.9^{(3.4)}$	$68.4^{(0.8)}$	$63.3^{(0.1)}$	$87.8^{(0.6)}$	$35.2^{(2.3)}$	$72.2^{(0.1)}$	15.5	6.2
RGD	$87.9^{(4.2)}$	$68.7^{(0.6)}$	$63.4^{(0.2)}$	$90.2^{(0.3)}$	$43.0^{(2.7)}$	$88.5^{(1.1)}$	13.2	10.1
COMs	$93.1^{(3.4)}$	$67.0^{(0.9)}$	$64.6^{(1.0)}$	$97.1^{(1.6)}$	$41.2^{(4.8)}$	$91.8^{(0.9)}$	10.2	12.3
RoMA	$96.5^{(0.0)}$	$77.8^{(0.0)}$	$63.3^{(0.0)}$	$84.7^{(0.0)}$	$49.8^{(1.4)}$	$\mathbf{95.7^{(1.6)}}$	9.7	14.5
ROMO	$\mathbf{98.1^{(0.7)}}$	$66.8^{(1.0)}$	$63.0^{(0.8)}$	$91.8^{(0.9)}$	$38.7^{(2.5)}$	$87.8^{(0.9)}$	12.7	10.9
GAMBO	$94.1^{(1.9)}$	$86.3^{(0.2)}$	$66.8^{(0.7)}$	$\mathbf{121^{(0.0)}}$	$\mathbf{50.8^{(3.3)}}$	$86.7^{(1.1)}$	**4.8**	**20.8**
DynAMO-Grad.	$90.3^{(4.7)}$	$86.2^{(0.0)}$	$64.4^{(2.5)}$	$91.2^{(0.0)}$	$44.2^{(7.8)}$	$89.8^{(3.2)}$	9.5	14.2
DynAMO-BO-qUCB	$95.1^{(1.9)}$	$86.2^{(0.0)}$	$66.7^{(1.5)}$	$\mathbf{121^{(0.0)}}$	$48.1^{(4.0)}$	$86.9^{(4.5)}$	6.3	20.5
DynAMO-CMA-ES	$89.8^{(3.6)}$	$85.7^{(5.8)}$	$63.9^{(0.9)}$	$\mathbf{117^{(6.7)}}$	$\mathbf{50.6^{(4.8)}}$	$78.5^{(5.5)}$	9.3	17.5

Median@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	33.7	42.8	50.9	87.6	6.7	77.8	—	—
Grad.	$\mathbf{58.1^{(6.1)}}$	$58.6^{(13.1)}$	$\mathbf{59.3^{(8.6)}}$	$\mathbf{85.3^{(7.7)}}$	$\mathbf{36.0^{(6.7)}}$	$65.1^{(14.4)}$	10.7	10.5
BO-qUCB	$50.3^{(1.8)}$	$62.1^{(3.4)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{86.6^{(0.6)}}$	$31.7^{(1.2)}$	$74.4^{(0.6)}$	6.8	11.5
CMA-ES	$50.7^{(2.7)}$	$\mathbf{71.7^{(10.4)}}$	$\mathbf{63.3^{(0.0)}}$	$83.9^{(1.0)}$	$\mathbf{37.9^{(0.7)}}$	$59.3^{(10.9)}$	7.2	11.2
BONET	$53.1^{(0.0)}$	$46.5^{(0.6)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{91.2^{(0.1)}}$	$\mathbf{37.9^{(0.0)}}$	$\mathbf{92.1^{(0.0)}}$	**5.2**	14.1
DDOM	$\mathbf{55.9^{(0.7)}}$	$57.6^{(0.8)}$	$\mathbf{63.3^{(0.0)}}$	$83.4^{(0.1)}$	$21.9^{(1.7)}$	$57.6^{(13.2)}$	12.2	6.7
ExPT	$44.7^{(6.8)}$	$57.0^{(4.8)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{89.8^{(3.3)}}$	$\mathbf{34.1^{(12.3)}}$	$67.8^{(14.1)}$	9.3	9.5
MINs	$41.3^{(1.2)}$	$58.0^{(0.5)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{88.3^{(0.3)}}$	$32.3^{(1.6)}$	$68.9^{(15.9)}$	9.7	8.8
GTG	$43.4^{(0.3)}$	$64.2^{(0.0)}$	$\mathbf{63.3^{(0.0)}}$	$83.1^{(0.0)}$	$28.0^{(0.0)}$	$90.6^{(0.0)}$	9.3	12.2
Tri-Mentoring	$46.1^{(0.0)}$	$61.1^{(0.0)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{88.9^{(0.0)}}$	$34.2^{(1.2)}$	$88.4^{(0.0)}$	6.5	13.7
ICT	$\mathbf{59.5^{(3.0)}}$	$61.1^{(0.0)}$	$57.1^{(0.0)}$	$\mathbf{88.9^{(0.0)}}$	$\mathbf{37.0^{(1.2)}}$	$88.4^{(0.1)}$	7.0	**15.4**
PGS	$40.7^{(2.6)}$	$60.3^{(0.7)}$	$58.4^{(0.0)}$	$\mathbf{88.9^{(0.0)}}$	$28.2^{(0.5)}$	$70.9^{(0.7)}$	12.3	8.0
Match-Opt	$40.7^{(1.2)}$	$57.9^{(0.5)}$	$\mathbf{63.3^{(0.0)}}$	$83.2^{(0.1)}$	$14.5^{(0.9)}$	$60.5^{(9.0)}$	15.0	3.4
RGD	$41.1^{(1.3)}$	$57.4^{(0.9)}$	$\mathbf{63.3^{(0.0)}}$	$86.4^{(0.1)}$	$20.7^{(0.6)}$	$70.9^{(1.8)}$	12.3	6.7
COMs	$43.9^{(0.0)}$	$59.0^{(0.5)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{93.2^{(7.7)}}$	$21.3^{(5.6)}$	$89.9^{(1.0)}$	8.5	11.8
RoMA	$50.1^{(4.3)}$	$\mathbf{77.4^{(0.0)}}$	$\mathbf{63.3^{(0.0)}}$	$84.7^{(0.0)}$	$34.9^{(1.8)}$	$63.7^{(6.2)}$	7.3	12.4
ROMO	$\mathbf{58.7^{(3.3)}}$	$37.7^{(0.3)}$	$27.4^{(1.2)}$	$61.8^{(2.6)}$	$27.0^{(0.6)}$	$46.0^{(11.7)}$	15.8	-6.8
GAMBO	$46.4^{(1.8)}$	$63.4^{(3.3)}$	$\mathbf{63.3^{(0.0)}}$	$86.3^{(0.5)}$	$28.9^{(1.1)}$	$79.1^{(0.7)}$	7.8	11.3
DynAMO-Grad.	$47.0^{(2.8)}$	$69.8^{(6.0)}$	$61.9^{(2.2)}$	$85.9^{(0.4)}$	$23.4^{(8.5)}$	$68.7^{(12.1)}$	11.0	9.5
DynAMO-BO-qUCB	$48.8^{(1.8)}$	$65.9^{(3.7)}$	$\mathbf{63.3^{(0.0)}}$	$\mathbf{86.5^{(0.5)}}$	$22.7^{(2.0)}$	$50.4^{(14.6)}$	9.7	6.3
DynAMO-CMA-ES	$45.3^{(2.4)}$	$65.8^{(8.9)}$	$59.3^{(3.8)}$	$\mathbf{99.0^{(12.1)}}$	$22.5^{(5.1)}$	$60.6^{(15.0)}$	11.0	8.8

Table D.8: **Comparison of design diversity against model-free optimization methods.** We evaluate DynAMO and other model-based methods against model-free optimization methods. We report the pairwise diversity (resp., minimum novelty) oracle score achieved by the 128 evaluated designs in the top (resp., bottom) table. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. All metrics are multiplied by 100 for easier legibility. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given backbone optimizer.

Pairwise Diversity@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	65.9	57.3	60.0	36.7	66.0	85.7	—	—
Grad.	$12.5^{(8.0)}$	$7.8^{(8.8)}$	$7.9^{(7.8)}$	$24.1^{(13.3)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	18.7	-53.2
BO-qUCB	$\mathbf{73.9^{(0.5)}}$	$\mathbf{74.3^{(0.4)}}$	$99.4^{(0.1)}$	$93.6^{(0.5)}$	$\mathbf{198^{(10.3)}}$	$94.1^{(3.9)}$	<u>3.8</u>	43.5
CMA-ES	$47.2^{(11.2)}$	$44.6^{(15.9)}$	$93.5^{(2.0)}$	$66.2^{(9.4)}$	$12.8^{(0.6)}$	$164^{(10.6)}$	10.8	9.5
BONET	$46.7^{(2.5)}$	$24.6^{(0.4)}$	$14.9^{(1.6)}$	$5.5^{(0.2)}$	$0.2^{(0.0)}$	$0.1^{(0.0)}$	17.3	-46.6
DDOM	$51.3^{(1.0)}$	$47.1^{(0.3)}$	$21.9^{(3.6)}$	$97.2^{(0.0)}$	$1.9^{(0.1)}$	$50.7^{(13.1)}$	12.5	-16.9
ExPT	$15.3^{(5.6)}$	$16.5^{(1.6)}$	$21.3^{(1.8)}$	$5.3^{(0.7)}$	$8.1^{(2.3)}$	$0.2^{(0.0)}$	17.5	-50.8
MINs	$67.0^{(0.3)}$	$56.8^{(0.4)}$	$53.5^{(3.1)}$	$34.1^{(2.6)}$	$84.6^{(21.1)}$	$4.3^{(0.3)}$	11.8	-11.9
GTG	$60.9^{(0.0)}$	$44.6^{(0.0)}$	$2.8^{(0.0)}$	$0.9^{(0.0)}$	$114.8^{(0.1)}$	$2.7^{(0.0)}$	15.0	-24.2
Tri-Mentoring	$58.5^{(0.0)}$	$57.6^{(0.0)}$	$85.5^{(0.0)}$	$39.9^{(0.0)}$	$47.7^{(0.0)}$	$62.5^{(0.0)}$	10.5	-3.3
ICT	$44.8^{(6.0)}$	$57.5^{(0.0)}$	$89.9^{(1.8)}$	$70.3^{(8.6)}$	$78.9^{(3.7)}$	$164^{(0.8)}$	9.3	22.3
PGS	$65.8^{(1.6)}$	$57.4^{(0.3)}$	$63.2^{(0.0)}$	$39.9^{(0.0)}$	$36.7^{(0.6)}$	$162^{(0.7)}$	10.5	8.9
Match-Opt	$65.1^{(0.4)}$	$55.9^{(0.1)}$	$\mathbf{99.8^{(0.0)}}$	$97.2^{(0.0)}$	$10.9^{(0.4)}$	$202^{(0.5)}$	7.5	26.6
RGD	$67.1^{(0.2)}$	$58.4^{(0.2)}$	$\mathbf{99.8^{(0.0)}}$	$\mathbf{97.3^{(0.0)}}$	$88.4^{(3.8)}$	$76.2^{(0.7)}$	5.0	19.3
COMs	$66.6^{(1.0)}$	$57.4^{(0.2)}$	$81.6^{(4.9)}$	$3.8^{(0.9)}$	$99.5^{(25.8)}$	$21.1^{(23.5)}$	10.7	-6.9
RoMA	$21.3^{(0.3)}$	$3.8^{(0.0)}$	$5.9^{(0.2)}$	$1.8^{(0.0)}$	$49.4^{(6.1)}$	$14.8^{(0.6)}$	17.2	-45.8
ROMO	$62.1^{(0.8)}$	$57.1^{(0.1)}$	$53.9^{(0.6)}$	$48.7^{(0.1)}$	$51.7^{(31.7)}$	$22.1^{(5.5)}$	11.5	-12.7
GAMBO	$\mathbf{74.0^{(0.6)}}$	$\mathbf{74.3^{(0.4)}}$	$99.3^{(0.1)}$	$93.3^{(0.4)}$	$193^{(1.2)}$	$17.7^{(3.5)}$	5.5	30.0
DynAMO-Grad.	$66.9^{(6.9)}$	$\mathbf{68.2^{(10.8)}}$	$77.2^{(21.5)}$	$93.0^{(1.2)}$	$129^{(55.3)}$	$104^{(56.1)}$	6.8	27.8
DynAMO-BO-qUCB	$\mathbf{74.3^{(0.5)}}$	$\mathbf{74.4^{(0.6)}}$	$99.3^{(0.1)}$	$93.5^{(0.6)}$	$\mathbf{211^{(22.8)}}$	$\mathbf{175^{(44.7)}}$	**2.8**	**59.4**
DynAMO-CMA-ES	$73.6^{(0.6)}$	$\mathbf{73.1^{(3.1)}}$	$72.0^{(3.1)}$	$94.0^{(0.5)}$	$97.8^{(13.2)}$	$\mathbf{292^{(83.5)}}$	5.2	<u>55.2</u>

Minimum Novelty@128	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank ↓**	**Opt. Gap ↑**
Dataset $\mathcal{D}$	0.0	0.0	0.0	0.0	0.0	0.0	—	—
Grad.	$21.2^{(3.0)}$	$\mathbf{51.7^{(2.9)}}$	$\mathbf{97.4^{(3.9)}}$	$79.5^{(19.7)}$	$95^{(0.7)}$	$102.2^{(6.1)}$	5.8	74.5
BO-qUCB	$21.6^{(0.3)}$	$\mathbf{51.7^{(0.2)}}$	$97.9^{(0.4)}$	$85.3^{(1.1)}$	$93.8^{(0.6)}$	$98.8^{(1.1)}$	6.0	74.8
CMA-ES	$16.5^{(2.1)}$	$47.8^{(1.0)}$	$96.5^{(0.7)}$	$73.0^{(18.0)}$	$\mathbf{100^{(0.0)}}$	$100^{(0.0)}$	7.8	72.3
BONET	$\mathbf{94.6^{(1.3)}}$	$38.8^{(0.1)}$	$41.2^{(0.3)}$	$10.3^{(0.1)}$	$1.3^{(0.0)}$	$1.1^{(0.0)}$	14.7	31.2
DDOM	$11.1^{(0.4)}$	$38.6^{(0.1)}$	$96.5^{(0.9)}$	$94.2^{(0.1)}$	$98.0^{(0.1)}$	$100^{(0.0)}$	9.3	73.1
ExPT	$11.1^{(1.6)}$	$39.0^{(0.7)}$	$54.2^{(2.0)}$	$15.6^{(1.2)}$	$69.8^{(6.8)}$	$3.5^{(0.9)}$	15.2	32.2
MINs	$12.2^{(0.4)}$	$38.3^{(0.2)}$	$48.7^{(2.9)}$	$22.1^{(1.7)}$	$6.1^{(1.1)}$	$0.6^{(0.1)}$	16.5	21.3
GTG	$13.8^{(0.0)}$	$37.8^{(0.0)}$	$\mathbf{99.3^{(0.0)}}$	$\mathbf{99.4^{(0.0)}}$	$67.7^{(0.0)}$	$0.2^{(0.0)}$	10.5	53.0
Tri-Mentoring	$13.9^{(0.0)}$	$31.8^{(0.0)}$	$74.7^{(0.0)}$	$75.5^{(0.0)}$	$44.0^{(0.0)}$	$64.3^{(0.0)}$	14.3	50.7
ICT	$19.3^{(1.2)}$	$31.7^{(0.1)}$	$70.8^{(0.4)}$	$75.6^{(0.0)}$	$46.2^{(0.4)}$	$66.2^{(0.7)}$	13.2	51.6
PGS	$11.5^{(0.4)}$	$33.1^{(1.8)}$	$16.0^{(0.0)}$	$15.8^{(0.0)}$	$45.0^{(0.2)}$	$74.8^{(1.1)}$	16.3	32.7
Match-Opt	$12.1^{(0.5)}$	$40.0^{(0.1)}$	$98.5^{(0.1)}$	$94.9^{(0.1)}$	$91.9^{(0.1)}$	$85.8^{(2.6)}$	8.7	70.5
RGD	$12.3^{(0.5)}$	$39.0^{(0.2)}$	$98.6^{(0.1)}$	$94.2^{(0.1)}$	$90.4^{(0.3)}$	$90.5^{(2.3)}$	8.5	70.9
COMs	$10.9^{(0.3)}$	$31.7^{(0.8)}$	$52.4^{(11.0)}$	$13.7^{(1.1)}$	$99.6^{(0.3)}$	$100^{(0.0)}$	14.0	51.4
RoMA	$18.3^{(0.5)}$	$40.1^{(0.2)}$	$18.9^{(0.2)}$	$95.3^{(0.0)}$	$47.6^{(2.4)}$	$5.1^{(0.2)}$	11.0	37.6
ROMO	$16.1^{(0.5)}$	$32.9^{(0.1)}$	$5.0^{(0.7)}$	$23.1^{(0.0)}$	$78.5^{(0.5)}$	$\mathbf{153.3^{(0.4)}}$	12.3	51.5
GAMBO	$15.4^{(0.3)}$	$\mathbf{51.8^{(0.2)}}$	$97.8^{(0.3)}$	$84.9^{(0.9)}$	$85.1^{(0.4)}$	$14.3^{(1.5)}$	8.8	58.2
DynAMO-Grad.	$21.1^{(1.1)}$	$\mathbf{52.2^{(1.3)}}$	$\mathbf{98.6^{(1.5)}}$	$85.8^{(1.0)}$	$95.0^{(0.4)}$	$107.2^{(6.7)}$	**3.8**	**76.7**
DynAMO-BO-qUCB	$21.4^{(0.5)}$	$\mathbf{51.7^{(0.2)}}$	$97.1^{(0.5)}$	$85.3^{(1.1)}$	$94.7^{(0.2)}$	$109^{(4.5)}$	<u>5.3</u>	<u>76.6</u>
DynAMO-CMA-ES	$12.9^{(0.8)}$	$48.0^{(1.6)}$	$\mathbf{96.7^{(3.5)}}$	$81.8^{(13.4)}$	$94.5^{(0.7)}$	$112^{(7.8)}$	7.8	74.3

Table D.9: **Comparison of design diversity against model-free optimization methods (cont.).** We evaluate DynAMO and other model-based optimization methods against model-free optimization methods. We report the L_1 coverage score achieved by the 128 evaluated designs. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better. All metrics are multiplied by 100 for easier legibility. **Bolded** entries indicate average scores with an overlapping 95% confidence interval to the best performing method. **Bolded** (resp., Underlined) Rank and Optimality Gap (Opt. Gap) metrics indicate the best (resp., second best) for a given optimizer.

L_1 **Coverage@128**	**TFBind8**	**UTR**	**ChEMBL**	**Molecule**	**Superconductor**	**D'Kitty**	**Rank** $\downarrow$	**Opt. Gap** $\uparrow$
Dataset $\mathcal{D}$	0.42	0.31	1.42	0.68	6.26	0.58	—	—
Grad.	$0.16^{(0.10)}$	$0.20^{(0.13)}$	$0.21^{(0.10)}$	$0.42^{(0.18)}$	$0.00^{(0.00)}$	$0.00^{(0.00)}$	19.5	-1.44
BO-qUCB	$0.40^{(0.02)}$	$\mathbf{0.54^{(0.01)}}$	$\mathbf{2.40^{(0.05)}}$	$\mathbf{2.52^{(0.07)}}$	$7.79^{(0.04)}$	$6.64^{(0.09)}$	<u>4.3</u>	<u>1.77</u>
CMA-ES	$0.33^{(0.05)}$	$0.48^{(0.04)}$	$2.18^{(0.04)}$	$1.82^{(0.12)}$	$3.26^{(1.42)}$	$3.78^{(1.36)}$	8.5	0.36
BONET	$0.11^{(0.00)}$	$0.22^{(0.00)}$	$0.72^{(0.02)}$	$0.54^{(0.00)}$	$0.03^{(0.00)}$	$0.07^{(0.00)}$	17.8	-1.33
DDOM	$0.43^{(0.01)}$	$0.29^{(0.00)}$	$0.68^{(0.05)}$	$0.85^{(0.02)}$	$\mathbf{9.24^{(0.23)}}$	$0.66^{(0.11)}$	10.8	0.41
ExPT	$0.22^{(0.05)}$	$0.25^{(0.01)}$	$0.60^{(0.03)}$	$0.29^{(0.03)}$	$0.45^{(0.01)}$	$0.10^{(0.00)}$	18.0	-1.29
MINs	$0.43^{(0.02)}$	$0.30^{(0.01)}$	$1.32^{(0.03)}$	$0.70^{(0.03)}$	$0.86^{(0.13)}$	$0.43^{(0.01)}$	11.0	-0.94
GTG	$0.42^{(0.00)}$	$0.30^{(0.00)}$	$1.61^{(0.00)}$	$1.96^{(0.00)}$	$8.77^{(0.00)}$	$0.31^{(0.00)}$	8.7	0.62
Tri-Mentoring	$\mathbf{0.48^{(0.00)}}$	$0.30^{(0.00)}$	$1.08^{(0.00)}$	$0.53^{(0.00)}$	$1.94^{(0.00)}$	$3.79^{(0.00)}$	10.8	-0.26
ICT	$0.34^{(0.02)}$	$0.30^{(0.00)}$	$0.90^{(0.02)}$	$0.80^{(0.05)}$	$0.56^{(0.01)}$	$3.73^{(0.02)}$	12.8	-0.51
PGS	$\mathbf{0.45^{(0.04)}}$	$0.30^{(0.01)}$	$1.40^{(0.00)}$	$0.60^{(0.00)}$	$1.92^{(0.00)}$	$3.66^{(0.04)}$	10.0	-0.22
Match-Opt	$0.41^{(0.01)}$	$0.32^{(0.01)}$	$0.69^{(0.01)}$	$0.86^{(0.02)}$	$5.26^{(0.02)}$	$5.22^{(0.07)}$	8.7	0.52
RGD	$0.42^{(0.01)}$	$0.33^{(0.00)}$	$0.69^{(0.01)}$	$0.86^{(0.02)}$	$4.08^{(0.13)}$	$4.90^{(0.07)}$	9.0	0.27
COMs	$\mathbf{0.49^{(0.02)}}$	$0.31^{(0.00)}$	$1.11^{(0.16)}$	$0.61^{(0.09)}$	$0.37^{(0.11)}$	$0.81^{(0.76)}$	11.0	-1.00
RoMA	$0.28^{(0.00)}$	$0.46^{(0.01)}$	$0.41^{(0.02)}$	$0.42^{(0.01)}$	$1.87^{(0.06)}$	$0.79^{(0.01)}$	14.8	-0.91
ROMO	$0.33^{(0.02)}$	$0.30^{(0.00)}$	$1.31^{(0.02)}$	$0.62^{(0.02)}$	$0.34^{(0.16)}$	$\mathbf{6.13^{(2.81)}}$	11.8	-0.11
GAMBO	$0.40^{(0.03)}$	$\mathbf{0.55^{(0.01)}}$	$\mathbf{2.38^{(0.10)}}$	$\mathbf{2.53^{(0.05)}}$	$7.45^{(0.01)}$	$1.29^{(0.08)}$	6.3	0.82
DynAMO-Grad.	$0.36^{(0.04)}$	$\mathbf{0.53^{(0.06)}}$	$1.46^{(0.38)}$	$\mathbf{2.50^{(0.06)}}$	$6.47^{(1.24)}$	$5.85^{(1.35)}$	6.7	1.25
DynAMO-BO-qUCB	$0.40^{(0.03)}$	$\mathbf{0.55^{(0.01)}}$	$\mathbf{2.47^{(0.07)}}$	$\mathbf{2.54^{(0.04)}}$	$7.88^{(0.03)}$	$\mathbf{7.80^{(0.23)}}$	**2.8**	**2.00**
DynAMO-CMA-ES	$0.40^{(0.03)}$	$\mathbf{0.56^{(0.01)}}$	$\mathbf{1.82^{(0.72)}}$	$\mathbf{2.54^{(0.05)}}$	$4.75^{(2.16)}$	$3.29^{(1.56)}$	6.3	0.62

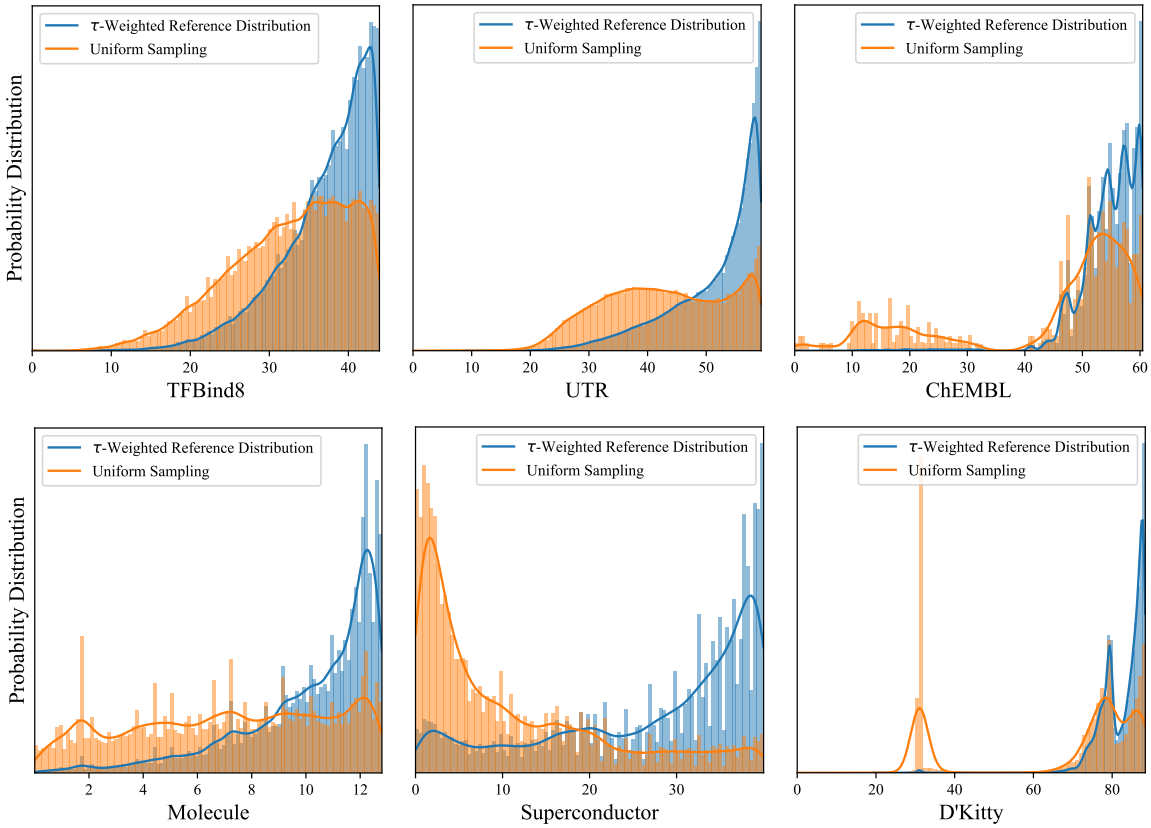

Figure D.1: **Sample τ-weighted probability distributions.** We plot ($\tau = 1.0$)-weighted distributions $p^{\tau}_{\mathcal{D}}(y)$ (blue) versus the original distribution of oracle scores y in the public offline dataset $\mathcal{D}$ (orange) for the 6 offline optimization tasks in our experimental evaluation suite: (1) **TFBind8** (top left); (2) **UTR** (top middle); (3) **ChEMBL** (top right); (4) **Molecule** (bottom left); (5) **Superconductor** (bottom middle); and (6) **D'Kitty** (bottom right). DynAMO penalizes a model-based optimization objective to encourage sampling policies to match the *diversity* of (high-scoring) designs in the τ-weighted distribution. The x-axis represents the normalized oracle scores.

D.6. Distribution Analysis of Quality and Diversity Results

In our experimental results in the main text and in the Appendices, we primarily focus on reporting summative statistics: for example, the Best@128 oracle score and the average Pairwise Diversity metric over the final batch of $k = 128$ samples. In this section, we isolate a single representative experimental run and plot the distribution of scores achieved by all $k = 128$ designs from a single experimental run to better interrogate the robustness and empirical properties of DynAMO.

In **Supp. Figure D.2**, we first plot the distributions of the oracle reward score $r(x_i^F)$ and minimum novelty score $\min_{x' \in \mathcal{D}} d(x_i^F, x')$ achieved by each of the $k = 128$ designs in the set $\{x_i^F\}_{i=1}^k$ proposed by the CMA-ES optimizer with and without DynAMO augmentation in a single experimental run. (Recall that $\mathcal{D}$ is the static, offline dataset of reference designs and $d(\cdot, \cdot)$ is the normalized Levenshtein distance metric for this task.) We see that in general, DynAMO not only enables the optimizer to discover more optimal designs with higher probability, but also yields a wider-tailed distribution of oracle scores compared to the baseline method. Separately, we see that DynAMO augmentation *decreases* both the median and mode Minimum Novelty score compared to the baseline method, in agreement with our discussion in **Section D.2**.

In the bottom row of **Supp. Figure D.2**, we visualize a heat map of pairwise diversity scores; that is, the color of pixel (i, j) is correlated with the distance $d(x_i^F, x_j^F)$ for any $1 \leq i, j \leq k$ pair of generated designs proposed by the optimization method. Even a cursory visual inspection reveals that DynAMO augmentation of the backbone CMA-ES optimizer significantly improves the pairwise diversity of candidate designs when compared to the baseline method.

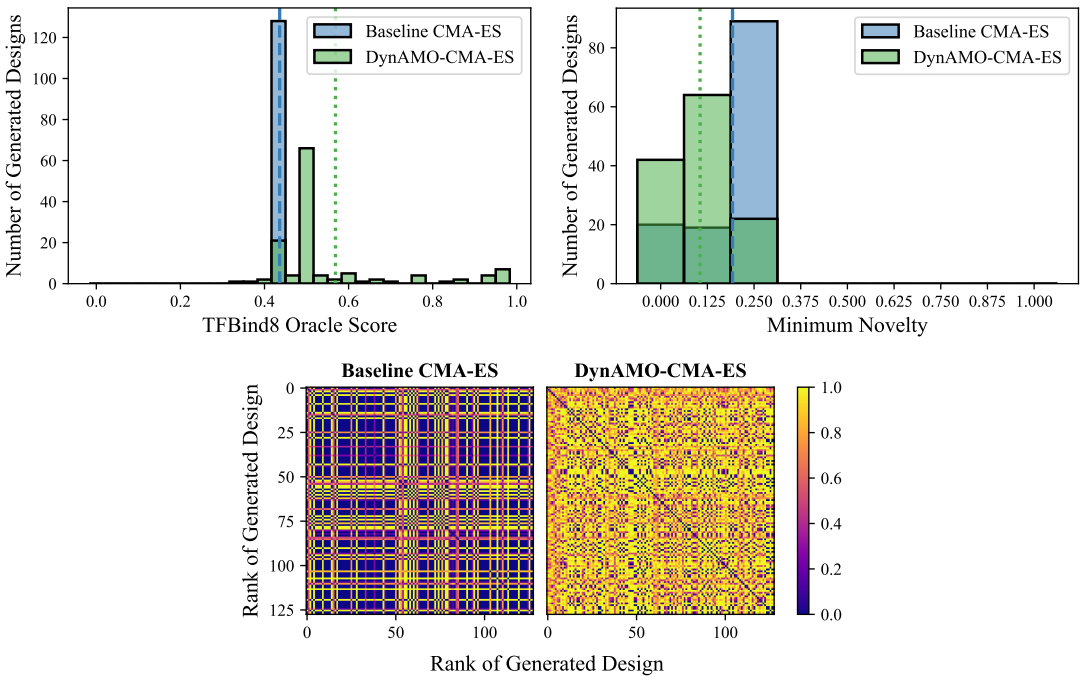

Figure D.2: **Distribution of generated design quality and diversity scores.** We plot the distributions of the (**top left**) oracle score; (**top right**) minimum novelty; and (**bottom**) pairwise diversity of the $k = 128$ proposed designs from a single representative experimental run using the CMA-ES backbone optimizers with and without DynAMO on the TFBind8 task. Dashed blue (resp., dotted green) lines in the top panels represent the mean score achieved by the Baseline CMA-ES (resp., DynAMO-CMA-ES) method from the experimental run.

D.7. Why Is Diversity Important?

Our principle motivation for obtaining a diverse sample of designs in offline MBO is to enable downstream **secondary exploration** of other objectives that we might care about in real-world applications. For example, given a batch of proposed candidates drugs that were optimized for maximal therapeutic efficacy in treating a disease, we might then try to quantify each candidate's manufacturing cost, difficulty of synthesis, profile of potential side effects, and other objectives. In this setting, obtaining highly similar designs from offline MBO may result in strong therapeutic efficacy, but also *equally* unacceptable values of other secondary objectives.

To validate this motivating claim that diversity is important to obtain a wide range of secondary objective values, we compare the range and variance of secondary objective values within a batch

of proposed candidate designs. We consider the following 3 offline MBO tasks:

1. **Vehicle Safety** is continuous, 5-dimensional optimization problem from Liao et al. (2008) to find an optimal set of car dimensions that minimize the total **Mass** of the vehicle. The problem initially stems from work on multi-objective optimization (Blank and Deb, 2020; Liao et al., 2008; Huo et al., 2022; Gonzalez de Oliveira et al., 2023), where the secondary goals are to (1) minimize the worst-case **Acceleration**-induced biomechanical damage of the car occupants in the event of a collision; and (2) minimize the worst-case toe board **Intrusion** of the vehicle in the event of an 'offset-frontal crash.' We treat the Mass as the offline MBO optimization objective and Acceleration and Intrusion as the downstream secondary objectives. We negate all objective values prior to max-min normalization as described in **Section 6.4.2** to frame this as a *maximization* problem in accordance with the setup in **Table 6.1**. An offline dataset of $n = 800$ designs was synthetically constructed, and we used the oracle function from Liao et al. (2008) to compute all 3 objective values.

2. **Welded Beam** is a continuous, 4-dimensional optimization problem (Ray and Liew, 2002) to find an optimal set of dimensions for a welded steel beam that minimizes the total manufacturing **Cost**. Similar to the Vehicle task, this problem was initially proposed in the multi-objective optimization literature (Blank and Deb, 2020; Liao et al., 2008; Kamil et al., 2021; Deb et al., 2006) where the secondary goal is to (1) minimize the end **Deflection** of the beam.[7] Again, we negate all objective values prior to max-min normalization as described in **Section 6.4.2** to frame this as a maximization problem. An offline dataset of $n = 800$ designs was synthetically constructed, and we used the oracle function from Ray and Liew (2002) to compute both objective values.

3. **UTR** is a discrete, 50-dimensional optimization problem from Angermueller et al. (2020a) and Sample et al. (2019) with the goal of finding an optimal 50-bp DNA sequence that maximizes the gene expression from a 5′ UTR DNA sequence. This is an offline MBO problem

[7] The original problem from Ray and Liew (2002) was proposed as a constrained optimization problem with 5 sets of constraints on the maximum considered shear stress, bending stress, buckling load, and other material testing parameters. To simplify our experimental setting, we consider the *unconstrained* version of the optimization problem here.

from the Design-Bench benchmarking suite (Trabucco et al., 2022) that we use to evaluate offline MBO algorithms in our main experimental results in **Table 6.1** and elsewhere. However, a secondary objective is to minimize the **GC Content** of the resulting DNA sequence, which is correlated with the difficulty of cloning and sequencing the DNA sequence using standard DNA amplification and analysis methods in the laboratory setting (Benita et al., 2003; Yakovchuk et al., 2006; Gardner et al., 2002). To evaluate this secondary objective, we use the same experimental setting as for the initial UTR experiments described in **Section 6.4.2** and evaluate the GC Content of the $k = 128$ proposed designs as the secondary objective according to Benita et al. (2003).

We used the standard deviation of secondary objective values achieved by a proposed set of designs to quantify the range of secondary objective values, and the pairwise diversity metric (PD@128) to quantify the diversity of designs. We evaluated both baseline and DynAMO-enhanced optimization methods on the three tasks above (**Supp. Table D.10**). Our results consistently demonstrate that a greater diversity score of the final proposed designs (i.e., higher PD@128 score) is correlated with a greater range of captured secondary objective values. As a result, a diverse set of designs (such as those proposed by DynAMO-enhanced optimization methods) can better enable downstream evaluation of the trade-offs between different objectives for a given design.

Table D.10: **Pairwise diversity as a predictor for downstream secondary exploration.** We report the pairwise diversity achieved by 128 proposed designs (**PD@128**); and also the variance of the distribution of oracle secondary objective values of those same 128 proposed designs. Note that the secondary objectives are *not* explicitly optimized against in the offline MBO setting. Metrics are reported mean$^{(95\% \text{ confidence interval})}$ across 10 random seeds, where higher is better (i.e., more diverse designs and better capture of the range of secondary objective values). All metrics are multiplied by 100 for easier legibility.

	Vehicle Safety			**Welded Beam**		**UTR**	
Method	PD@128	Acceleration	Intrusion	PD@128	Deflection	PD@128	GC Content
Grad.	$0.0^{(0.0)}$	$0.1^{(0.0)}$	$0.0^{(0.0)}$	$0.0^{(0.0)}$	$0.1^{(0.3)}$	$7.8^{(8.8)}$	$0.7^{(0.7)}$
DynAMO-Grad.	$2.5^{(0.2)}$	$12.6^{(0.5)}$	$10.7^{(0.9)}$	$19.6^{(33.3)}$	$31.7^{(9.3)}$	$68.2^{(10.8)}$	$3.4^{(0.7)}$
Adam	$0.0^{(0.0)}$	$0.1^{(0.1)}$	$0.1^{(0.1)}$	$0.0^{(0.0)}$	$1.5^{(1.6)}$	$11.0^{(12.1)}$	$4.0^{(5.9)}$
DynAMO-Adam	$2.2^{(0.1)}$	$12.6^{(0.6)}$	$10.3^{(1.0)}$	$11.1^{(3.5)}$	$49.5^{(21.3)}$	$72.3^{(3.4)}$	$14.0^{(3.1)}$
CMA-ES	$0.0^{(0.0)}$	$0.5^{(0.2)}$	$0.4^{(0.2)}$	$0.1^{(0.1)}$	$0.0^{(0.0)}$	$44.6^{(15.9)}$	$36.5^{(8.3)}$
DynAMO-CMA-ES	$8.6^{(6.0)}$	$28.5^{(10.8)}$	$30.8^{(20.0)}$	$43.9^{(6.0)}$	$19.8^{(18.8)}$	$73.1^{(3.1)}$	$42.0^{(2.3)}$
CoSyNE	$0.0^{(0.0)}$	$0.6^{(0.1)}$	$0.5^{(0.1)}$	$0.1^{(0.0)}$	$16.8^{(6.5)}$	$12.7^{(9.8)}$	$1.6^{(2.5)}$
DynAMO-CoSyNE	$1.9^{(2.7)}$	$4.0^{(4.2)}$	$2.2^{(2.6)}$	$55.1^{(65.3)}$	$65.8^{(13.2)}$	$20.3^{(2.3)}$	$1.0^{(1.3)}$
BO-qEI	$1.3^{(0.1)}$	$7.5^{(0.4)}$	$7.2^{(0.4)}$	$46.3^{(2.0)}$	$8.0^{(0.3)}$	$73.8^{(0.5)}$	$44.8^{(0.8)}$
DynAMO-BO-qEI	$2.4^{(0.1)}$	$13.3^{(0.2)}$	$10.5^{(0.2)}$	$78.1^{(18.0)}$	$26.6^{(2.1)}$	$74.6^{(0.3)}$	$45.7^{(0.6)}$
BO-qUCB	$1.2^{(0.1)}$	$5.7^{(0.4)}$	$6.1^{(0.4)}$	$46.5^{(7.5)}$	$7.8^{(0.2)}$	$74.3^{(0.2)}$	$45.3^{(0.4)}$
DynAMO-BO-qUCB	$2.8^{(0.1)}$	$12.3^{(0.2)}$	$10.9^{(0.1)}$	$63.9^{(4.2)}$	$29.0^{(1.9)}$	$74.4^{(0.6)}$	$45.2^{(0.5)}$

BIBLIOGRAPHY

Mohamad H Abedi, Michael S Yao, David R Mittelstein, Avinoam Bar-Zion, Margaret B Swift, Audrey Lee-Gosselin, Pierina Barturen-Larrea, Marjorie T Buss, and Mikhail G Shapiro. Ultrasound-controllable engineered bacteria for cancer immunotherapy. *Nat Commun*, 13, 2022. doi: 10.1038/s41467-022-29065-2.

Roy Adams, Katharine E Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, Eili Y Klein, Edward S Chen, Mustapha O Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C Linton, Anushree R Ahluwalia, Albert W Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med*, 28:1455–60, 2022. doi: 10.1038/s41591-022-01894-0.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proc NeurIPS*, pages 9525–36, 2018. doi: 10.5555/3327546.3327621.

Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *Proc ICLR*, 2022. URL https://openreview.net/forum?id=xNOVfCCvDpM.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv Preprint*, 2024a. doi: 10.48550/arXiv.2402.04614.

Nikhil Agarwal, Alex Moehring, Pranav Rapurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. *NBER Working Paper*, 2023. URL https://ssrn.com/abstract=4505053.

Siddhant Agarwal, Ishan Durugkar, Peter Stone, and Amy Zhang. f-Policy gradients: A general framework for goal conditioned RL using f-divergences. In *Proc NeurIPS*, pages 12100–23, 2024b. doi: 10.5555/3666122.3666652.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. In *Proc NeurIPS*, pages 9562–74, 2019. doi: 10.5555/3454287.3455145.

Pulkit Agrawal, Ashvin Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Proc NeurIPS*, page 5092–100, 2016. doi: 10.5555/3157382.3157666.

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *J Mach Learn Res*, 22(128):1–59, 2021. URL https://jmlr.csail.mit.edu/papers/volume22/20-867/20-867.pdf.

Michael Ahn, Henry Zhu, Kristian Hartikainen, Hugo Ponte, Abhishek Gupta, Levine Levine, and Vikash Kumar. ROBEL: Robotics benchmarks for learning with low-cost robots. In *Proc Conf Robot Learn*, volume 100, pages 1300–13, 2020.

Cohere For AI. c4ai-command-r-plus (revision 432fac1), 2024a.

Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024b. URL https://ai.meta.com/blog/meta-llama-3/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *Proc ICLR*, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Mohammed Albarakat and Ali Guzu. Prevalence of type 2 diabetes and their complications among home health care patients at Al-Kharj military industries corporation hospital. *J Family Med Prim Care*, 8(10):3303–12, 2019. doi: 10.4103/jfmpc.jfmpc_634_19.

Bibb Allen, Sheela Agarwal, Laura Coombs, Christoph Wald, and Keith Dreyer. 2020 ACR Data Science Institute Artificial Intelligence Survey. *J Am Coll Radiol*, 18(8):1153–9, 2021. doi: 10.1016/j.jacr.2021.04.002.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for Bayesian optimization. In *Proc NeurIPS*, pages 20577–612, 2023. doi: 10.5555/3666122.3667026.

Brandon Amos. On amortizing convex conjugates for optimal transport. In *Proc ICLR*, 2023. URL https://openreview.net/forum?id=TQ5WUwS_4ai.

Christof Angermueller, David Belanger, Andreea Gane, Zelda Mariet, David Dohan, Kevin Murphy, Lucy Colwell, and D Sculley. Population-based black-box optimization for biological sequence design. In *Proc ICML*, volume 119, pages 324–34, 2020a.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *Proc ICLR*, 2020b. URL https://openreview.net/forum?id=HklxbgBKvr.

Anthropic. Claude 3.5 Sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Mariamma Antony, Siva Teja Kakileti, Rachit Shah, Sabyasachi Sahoo, Chiranjib Bhattacharyya, and Geetha Manjunath. Challenges of AI driven diagnosis of chest X-rays transmitted through smart phones: A case study in COVID-19. *Sci Rep*, 13(18102), 2023. doi: 10.1038/s41598-023-44653-y.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc ICML*, volume 70, pages 214–23, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Shahriar Arman. On one-liners and doing no harm. *Acad Med*, 98(10):1184, 2023. doi: 10.1097/ACM.0000000000005324.

Raul Astudillo and Peter I Frazier. Bayesian optimization of composite functions. In *Proc Int Conf Mach Learn*, volume 97 of *ICML'17*, pages 354–63. PMLR, 2019.

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv Preprint*, 2021. doi: 10.48550/arXiv.2110.09485.

Cristiana Baloescu. Diagnostic imaging in emergency medicine: How much is too much? *Annals of Emergency Medicine*, 72(6):637–43, 2018. doi: 101016/j.annemergmed.2018.06.034.

Heejung Bang, Alison M Edwards, Andrew S Bomback, Christie M Ballantyne, David Brillon, Mark A Callahan, Steven M Teutsch, Alvin I Mushlin, and Lisa M Kern. Development and validation of a patient self-assessment score for diabetes risk. *Ann Intern Med*, 151(11):775–83, 2009. doi: 10.7326/0003-4819-151-11-200912010-00005.

Gioele Barabucci, Victor Shia, Eugene Chu, Benjamin Harack, Kyle Laskowski, and Nathan Fu. Combining multiple large language models improves diagnostic accuracy. *NEJM AI*, 1(11), 2024. doi: 10.1056/AIcs2400502.

I Bárány and Z Füredi. On the shape of the convex hull of random points. *Probab Theory Relat Fields*, 77:231–40, 1988. doi: 10.1007/BF00334039.

Paul R Barber, Rami Mustapha, Fabian Flores-Borja, Giovanna Alfano, Kenrick Ng, Gregory Weitsman, Luigi Dolcetti, Ali Abdulnabi Suwaidan, Felix Wong, Jose M Vicencio, Myria Galazi, James W Opzoomer, James N Arnold, Selvam Thavaraj, Shahram Kordasti, Jana Doyle, Jon Greenberg, Magnus T Dillon, Kevin J Harrington, Martin Forster, Anthony C C Coolen, and Tony Ng. Predicting progression-free survival after systemic therapy in advanced head and neck cancer: Bayesian regression and model development. *eLife*, 11:e73288, 2022. doi: 10.7554/eLife.73288.

Luis A Barrera, Anastasia Vedenko, Jesse V Kurland, Julia M Rogers, Stephen S Gisselbrecht, Elizabeth J Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, Trevor Siggers, Leila Shokri, Raluca Gordân, Nidhi Sahni, Chris Cotsapas, Tong Hao, Song Yi, Manolis Kellis, Mark J Daly, Marc Vidal, David E Hill, and Martha L Bulyk. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, 351(6280):1450–4, 2016. doi: 10.1126/science.aad2257.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J Mach Learn Res*, 3:463–82, 2003. doi: 10.5555/944919.944944.

Hamsa Bastani, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjichristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, and Sotirios Tsiodras. Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, 599:108–13,

2021. doi: 10.1038/s41586-021-04014-z.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc Nat Acad Sci*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.

Andre B Bautista, Anthony Burgos, Barbara J Nickel, John J Yoon, Amish A Tilara, and Judith K Amorosa. Do clinicians use the American College of Radiology Appropriateness Criteria in the management of their patients? *Am J Roentgenol*, 192(6):1581–5, 2009. doi: 10.2214/AJR.08.1622.

Yair Benita, Ronald S Oosting, Martin C Lok, Michael J Wise, and Ian Humphery-Smith. Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res*, 31(16):e99, 2003. doi: 10.1093/nar/gng101.

L Bergé. Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm. Technical Report 13, Department of Economics at the University of Luxembourg, 2018. URL https://EconPapers.repec.org/RePEc:luc:wpaper:18-13.

Jeroen Berrevoets, Sam Verboven, and Wouter Verbeke. Treatment effect optimisation in dynamic environments. *J Caus Infer*, 10(1):106–22, 2022. doi: 10.1515/jci-2020-0009.

Leonard Bickman, Henry J Domenico, Daniel W Byrne, Rebecca N Jerome, Terri L Edwards, Mary Stroud, Laurie Lebo, Kyle Mcguffin, Consuelo H Wilkins, and Paul A Harris. Effects of financial incentives on volunteering for clinical trials: A randomized vignette experiment. *Contemporary Clinical Trials*, 110:106584, 2021. doi: 10.1016/j.cct.2021.106584.

Julian Blank and Kalyanmoy Deb. pymoo: Multi-objective optimization in Python. *IEEE Access*, 8: 89497–509, 2020. doi: 10.1109/ACCESS.2020.2990567.

Jonathan Borwein and Adrian Lewis. *Convex analysis and nonlinear optimization: Theory and examples*. Springer, 2006. ISBN 978-0-387-29570-1. doi: 10.1007/978-0-387-31256-9.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proc NeurIPS*, pages 343–51, 2016. doi: 10.5555/3157096.3157135.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proc IEEE CVPR*, pages 95–104, 2017. doi: 10.1109/CVPR.2017.18.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. doi: 10.1017/CBO978051180444.

Adrian P Brady, Jaqueline A Bello, Lorenzo E Derchi, Michael Fuchsjäger, Stacy Goergen, Gabriel P Krestin, Emil J Y Lee, David C Levin, Josephine Pressacco, Vijay M Rao, John Slavotinek, Jacob J

Visser, Richard EA Walker, and James A Brink. Radiology in the era of value-based healthcare: A multi-society expert statement from the ACR, CAR, ESR, IS3R, RANZCR, and RSNA. *Radiology*, 298(3):486–91, 2020. doi: 10.1148/radiol.2020209027.

Franklin H Branin. Widely convergent method for finding multiple solutions of simultaneous nonlinear equations. *IBM Journal of Research and Development*, 16(5):504–22, 1972. doi: 10.1147/rd.165.0504.

Leah T Braun, Katharina F Borrmann, Christian Lottspeich, Daniel A Heinrich, Jan Kiesewetter, Martin R Fischer, and Ralf Schmidmaier. Guessing right – Whether and how medical students give incorrect reasons for their correct diagnoses. *GMS J Med Educ*, 36(6), 2019. doi: 10.3205/zma001293.

Brian W Bresnahan. Economic evaluation in radiology: Reviewing the literature and examples in oncology. *Acad Radiol*, 17(9):1090–5, 2010. doi: 10.1016/j.acra.2010.05.020.

David Brookes, Hahnbeom Park, and Jennifer Listgarten. Conditioning by adaptive sampling for robust design. In *Proc ICML*, volume 97, pages 773–82, 2019. URL https://proceedings.mlr.press/v97/brookes19a.html.

Nathan Brown, Marco Fiscato, Marwin H S Segler, and Alain C Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59:1096–108, 2019. doi: 10.1021/acs.jcim.8b00839.

Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Proc NeurIPS*, 2024. URL https://openreview.net/forum?id=Q5RYn6jagC.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proc ACM SIGKDD*, pages 1721–30, 2015. doi: 10.1145/2783258.2788613.

Allison Chae, Michael S Yao, Hersh Sagreiya, Ari D Goldberg, Neil Chatterjee, Matthew T MacLean, Jeffrey Duda, Ameena Elahi, Arijitt Borthakur, Marylyn D Ritchie, Daniel Rader, Charles E Kahn, Walter R Witschey, and James C Gee. Strategies for implementing machine learning algorithms in the clinical practice of radiology. *Radiology*, 310(1):e223170, 2024. doi: 10.1148/radiol.223170.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proc ICML*, volume 235 of *Proceedings of Machine Learning Research*, pages 6116–35. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Pierre Chambon, Tessa S Cook, and Curtis P Langlotz. Improved fine-tuning of in-domain transformer model for inferring COVID-19 presence in multi-institutional radiology reports. *J Digit Imaging*, 36(1):164–77, 2023. doi: 10.1007/s10278-022-00714-8.

Crystal T Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H Jacobson, Binglan Li, Celeste H Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A Fries, Jiapeng Xu, Joe Jamison, John S Tamaresis, Jonathan H Chen, Joshua Lazaro, Juan M Banda, Julie J Lee, Karen Ebert Matthys, Kirsten R Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T Truong, Scott L Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y Eulalio, Wenyuan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam Shah, and Roxana Daneshjou. Red teaming ChatGPT in medicine to yield real-world insights on model behavior. *npj Digit Med*, 8(149), 2025. doi: 10.1038/s41746-025-01542-0.

Yassine Chemingui, Aryan Deshwal, Trong Nghia Hoang, and Janardhan R Doppa. Offline model-based optimization via policy-guided gradient search. In *Proc AAAI*, pages 11230–9, 2024. doi: 10.1609/aaai.v38i10.29001.

Can Chen, Christopher Beckham, Zixuan Liu, Xue Liu Liu, and Christopher Pal. Parallel-mentoring for offline model-based optimization. In *Proc NeurIPS*, pages 76619–36, 2023a. doi: 10.5555/3666122.3669469.

Can Chen, Yingxue Zhang, Xue Liu, and Mark Coates. Bidirectional learning for offline model-based biological sequence design. In *Proc ICML*, volume 202, pages 5351–66. PMLR, 2023b. URL https://proceedings.mlr.press/v202/chen23ao.html.

Can Chen Chen, Christopher Beckham, Zixuan Liu, Xue Liu, and Christopher Pal. Robust guided diffusion for offline black-box optimization. *Trans Mach Learn Res*, 2024. URL https://openreview.net/forum?id=4JcqmEZ5zt.

Can (Sam) Chen, Yingxue Zhang, Jie Fu, Xue Liu, and Mark Coates. Bidirectional learning for offline infinite-width model-based optimization. In *Proc NeurIPS*, pages 29454–67, 2022. doi: 10.5555/3600270.3602406.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proc NAACL: Hum Lang Tech*, pages 3563–99, 2025a. doi: 10.18653/v1/2025.naacl-long.182.

Mingcheng Chen, Haoran Zhao, Yuxiang Zhao, Hulei Fan, Hongqiao Gao, Yong Yu, and Zheng Tian. ROMO: Retrieval-enhanced offline model-based optimization. In *Proc International Conf Dist Artif Intell*, pages 1–9, 2023c. doi: 10.1145/3627676.3627685.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proc ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*, pages 785–94, 2016. doi: 10.1145/2939672.2939785.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Aruthi Somani, Peter Hase, Misha Wagner, Fabian Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. Technical report, Anthropic AI, 2025b. URL https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70B: Scaling medical pretraining for large language models. *arXiv Preprint*, 2023d. doi: 10.48550/arXiv.2311.16079.

Ziqi Chen, Martin Renqiang Min, Srinivasan Parthasarathy, and Xia Ning. A deep generative model for molecule optimization via one fragment modification. *Nat Mach Intell*, 3:1040–9, 2021. doi: 10.1038/s42256-021-00410-2.

Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inform Fus*, 81:59–83, 2022. doi: 10.1016/j.inffus.2021.11.003.

Jan Clusmann, Fiona R Kolbinger, Hanna Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, Sophia J Wagner, and Jakob Nikolas Kather. The future landscape of large language models in medicine. *Commun Med*, 3(1):141, 2023. doi: 10.1038/s43856-023-00370-1.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B Tenenbaum, and Thomas L Griffiths. Building machines that learn and think with people. *Nat Human Behaviour*, 8:1851–63, 2024. doi: 10.1038/s41562-024-01991-9.

The International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–64, 2009. doi: 10.1056/NEJMoa0809329.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting.

In *Proc NeurIPS*, volume 23, pages 442–50, 2010. doi: 10.5555/2997189.2997239.

Jonathan Crabbé and Mihaela van der Schaar. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. In *Proc NeurIPS*, 2023. URL https://openreview.net/forum?id=5UwnKSgY6u.

Trisha Das, Zifeng Wang, and Jimeng Sun. TWIN: Personalized clinical trial digital twin generation. In *Proc ACM SIGKDD Conf*, pages 402–13, 2023. doi: 10.1145/3580305.3599534.

Alexander G de G Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proc Int Conf Artif Intell Stats*, volume 51, pages 231–9. PMLR, 2016. URL https://proceedings.mlr.press/v51/matthews16.html.

Kalyanmoy Deb, J Sunda, Udaya B N Rao, and Shamik Chaudhuri. Reference point based multi-objective optimization using evolutionary algorithms. *Int J Comp Intell Res*, 2(3):273–86, 2006.

Kevin Debeire, Andreas Gerhardus, Jakob Runge, and Veronika Eyring. Bootstrap aggregation and confidence measures to improve time series causal discovery. In *Proc Conf Causal Learning and Reasoning*, volume 236, pages 979–1007, 2024. URL https://proceedings.mlr.press/v236/debeire24a.html.

Henock M Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *Int J Environ Res Public Health*, 18(6):3317, 2021. doi: 10.3390/ijerph18063317.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell*, 3:610–9, 2021. doi: 10.1038/s42256-021-00338-7.

Ankur Deka, Changliu Liu, and Katia Sycara. ARC - Actor residual critic for adversarial imitation learning. In *Proc CoRL*, volume 205 of *Proc Mach Learn Res*, pages 1446–56, 2023. URL https://proceedings.mlr.press/v205/deka23a.html.

Fabrizio Dell'Acqua. Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters, 2022. URL https://www.almendron.com/tribuna/wp-content/uploads/2023/09/falling-asleep-at-the-whee.pdf.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality, 2023.

Laith A Derbas, Krishna K Patel, Preetham R Muskula, Jingyan Wang, Kensey Gosch, Robert Fitridge, John A Spertus, and Kim G Smolderen. Variability in utilization of diagnostic imaging tests in patients with symptomatic peripheral artery disease. *Int J Cardiol*, 330:200–6, 2021. doi: 10.1016/j.ijcard.2021.02.014.

Aryan Deshwal and Janardhan Rao Doppa. Combining latent space and structured kernels for Bayesian optimization over combinatorial spaces. In *Proc NeurIPS*, volume 34, pages 8185–200, 2021. doi: 10.5555/3540261.3540887.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Proc Neur Inf Proc Sys*, pages 10088–115, 2024. doi: 10.5555/3666122.3666563.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc NAACL-HLT*, pages 4171–86, 2019. doi: 10.48550/arXiv.1810.04805.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *J Mach Learn Res*, 17(1):2909–13, 2016. doi: 10.5555/2946645.3007036.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, and Bressem Keno K. Biomedical large languages models seem not to be superior to generalist models on unseen medical data, 2024.

Sebastian P D Dowhanik, Nicola Schieda, Michael N Patlas, Fateme Salehi, and Christian B van der Pol. Doing more with less: CT and MRI utilization in Canada 2003-2019. *Can Assoc Radiol J*, 73 (3):592–4, 2022. doi: 10.1177/08465371211052012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res*, 12(61):2121–59, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Trans Mach Learn Res*, 2024. URL https://openreview.net/forum?id=uHLDkQVtyC.

Deniz Dutz, Michael Greenstone, Ali Hortaçsu, Lacouture Santiago, Magne Mogstad, Azeem M Shaikh, Alexander Torgovitsky, and Winnie van Dijk. Representation and hesitancy in population health research: Evidence from a COVID-19 antibody study. *Nat Bureau Econ Res*, 2023. doi: 10.3386/w30880.

David Eriksson and Margin Jankowiak. High dimensional Bayesian optimization with sparse axis-aligned subspaces. In *Proc UAI*, volume 161, pages 493–503, 2021.

David Eriksson, Michael Pearce, Jacob R Gardner, Ryan Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Proc NeurIPS*, pages 5496–507, 2019. doi: 10.5555/3454287.3454780.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like

molecules based on molecular complexity and fragment contributions. *J Cheminform*, 1(8), 2009. doi: 10.1186/1758-2946-1-8.

Birhanu Eshete. Making machine learning trustworthy. *Science*, 373(6556):743–44, 2021. doi: 10.1126/science.abi5052.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings ACL*, pages 1181–93, 2023. doi: 10.18653/v1/2023.findings-eacl.88.

Harriet Evans and David Snead. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *NPJ Digit Med*, 7(89), 2024. doi: 10.1038/s41746-024-01093-w.

Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. In *Proc EMNLP*, pages 595–605, 2017. doi: 10.18653/v1/D17-1063.

Bassam Farran, Arshad Mohamed Channanath, Kazem Behbehani, and Thangavel Alphonse Thanaraj. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from kuwait–A cohort study. *BMJ Open*, 3(5), 2013. doi: 10.1136/bmjopen-2012-002457.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecule distributions. *Nat Commun*, 13(3293), 2022. doi: 10.1038/s41467-022-30839-x.

Luciano Floridi and Josh Cowls. *A unified framework of five principles for AI in society*, chapter 22. Wiley, 2022. doi: 10.1002/9781119815075.ch45.

Martina Zaguini Francisco, Stephan Altmayer, Lucas Carlesso, Matheus Zanon, Thales Eymael, Jose E Lima, Guilherme Watte, and Bruno Hochhegger. Appropriateness and imaging outcomes of ultrasound, CT, and MR in the emergency department: A retrospective analysis from an urban academic center. *Emergency Radiology*, 31:367–72, 2024. doi: 10.1007/s10140-024-02226-0.

Justin Fu and Sergey Levine. Offline model-based optimization via normalized maximum likelihood estimation. In *Proc ICLR*, 2021. URL https://openreview.net/forum?id=FmMKSO4e8JK.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health*, 2(9):e489–92, 2020. doi: 10.1016/S2589-7500(20)30186-2.

Sebastian Gabarin. Locutusque/gpt2-large-medical (revision dd3fb2e), 2023.

Michela Gabelloni, Matteo Di Nasso, Riccardo Morganti, Lorenzo Faggioni, Gianluca Masi, Alfredo Falcone, and Emanuele Neri. Application of the ESR iGuide clinical decision support system to the imaging pathway of patients with hepatocellular carcinoma and cholangiocarcinoma:

Preliminary findings. *Radiol Med*, 125(6):531–7, 2020. doi: 10.1007/s11547-020-01142-w.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J Mach Learn Res*, 17(1):2096–130, 2016. doi: 10.5555/2946645.2946704.

Ravi Ganti and Alexander G Gray. Building bridges: Viewing active learning from the multi-armed bandit lens. In *Proc UAI*, pages 232–241, 2013. doi: 10.5555/3023638.3023662.

Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Veidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419:498–511, 2002. doi: 10.1038/nature01097.

Andrew Garland, Kevin Weinfurt, and Jeremy Sugarman. Incentives and payments in pragmatic clinical trials: Scientific, ethical, and policy considerations. *Clin Trials*, 18(6):699–705, 2021. doi: 10.1177/17407745211048178.

Khashayar Gatmiry, Zhiyuan Li, Ching-Yao Chuang, Sashank Reddi, Tengyu Ma, and Stefanie Jegelka. The inductive bias of flatness regularization for deep matrix factorization. In *Proc NeurIPS*, pages 28040–52, 2023. doi: 10.5555/3666122.3667339.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 40:D1100–7, 2012. doi: 10.1093/nar/gkr777.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv Preprint*, 2025. doi: 10.48550/arXiv.2505.13400.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *eBioMedicine*, 89:104467, 2023. doi: 10.1016/j.ebiom.2023.104467.

Evgin Goceri. Medical image data augmentation: Techniques, comparisons and interpretations. *Artif Intell Rev*, pages 1–45, 2023. doi: 10.1007/s10462-023-10453-z.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Josephine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P J Olson, Adam Rodman, and Jonathan H Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw Open*, 7(10):e2440969, 2024. doi: 10.1001/jamanetworkopen.2024.40969.

Ethan Goh, Robert J Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A Freed, Joséphine A Cool, Zahir Kanjee, Kathleen P Lane, Andrew S Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew PJ Olson, Jason Hom, Jonathan H Chen, and Adam Rodman. GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nat Med*, 31:1233–8, 2025. doi: 10.1038/s41591-024-03456-y.

Faustino Gomez, Jürgen Schmidhuber, and Risto Miikkulainen. Accelerated neural evolution through cooperatively coevolved synapses. *J Mach Learn Res*, 9(31):937–65, 2008. URL http://jmlr.org/papers/v9/gomez08a.html.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4:268–76, 2018. doi: 10.1021/acscentsci.7b00572.

Chen Gong, Qiang He, Yunpeng Bai, Zhou Yang, Xiaoyu Chen, Xinwen Hou, Xianjie Zhang, Yu Liu, and Guoliang Fan. The f-Divergence reinforcement learning framework. *arXiv Preprint*, 2021. doi: 10.48550/arXiv.2109.11867.

Javier Gonzalez, Zhenwen Dai, Philipp Hennig, and Neil Lawrence. Batch Bayesian optimization via local penalization. In *Proc Int Conf AISTATS*, volume 51 of *Proc Mach Learn Res*, pages 648–57. PMLR, 2016. URL https://proceedings.mlr.press/v51/gonzalez16a.html.

Ricardo Gonzalez de Oliveira, Nicolas Navet, and Achim Henkel. Multi-objective optimization for safety-related available E/E architectures scoping highly automated driving vehicles. *ACM Transactions on Design Automation of Electronic Systems*, 28(41):1–37, 2023. doi: 10.1145/3582004.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc NeurIPS*, pages 2672–80, 2014. doi: 10.48550/arXiv.1406.2661.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc ICLR*, 2015. doi: 10.48550/arXiv.1412.6572.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Proc NeurIPS*, pages 18932–43, 2021. doi: 10.5555/3540261.3541708.

Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-

seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–93, 2013. doi: 10.1016/j.tics.2013.09.001.

Xiang Gu, Liwei Yang, Jian Sun, and Zongben Xu. Optimal transport-guided conditional score-based diffusion model. In *Proc NeurIPS*, pages 36540–52, 2023. doi: 10.5555/3666122.3667709.

Hao Guan, David Bates, and Li Zhou. Keeping medical AI healthy: A review of detection and correction methods for system degradation. *arXZiv Preprint*, 2025. doi: 10.48550/arXiv.2506.17442.

Jeffrey P Guenette, Elyse Lynch, Nooshin Abbasi, Kathryn Schulz, Shweta Kumar, Sebastien Haneuse, Neena Kapoor, Ronilda Lacson, and Ramin Khorasani. Recommendations for additional imaging on head and neck imaging examinations: Interradiologist variation and associated factors. *Am J Roentgenol*, 222(5):e2330511, 2024. doi: 10.2214/AJR.23.30511.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv Preprint*, 2017. doi: 10.48550/arXiv.1705.10843.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *Proc ICLR*, 2024. URL https://openreview.net/forum?id=ZG3RaNIsO8.

Ragini Gupta, Beitong Tian, Yaohui Wang, and Klara Nahrstedt. TWIN-ADAPT: Continuous learning for digital twin-enabled online anomaly classification in IoT-driven smart labs. *Future Internet*, 16(7):239, 2024. doi: 10.3390/fi16070239.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–53, 2025. doi: 10.1038/s41586-025-08744-2.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*, 30:2613–22, 2024. doi: 10.1038/s41591-024-03097-1.

Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv Preprint*, 2011. doi: 10.48550/arXiv.1101.1232.

Scott D Halpern, Marzana Chowdhury, Brian Bayes, Elizabeth Cooney, Brian L Hitsman, Robert A Schnoll, Su Fen Lubitz, Celine Reyes, Mitesh S Patel, S Ryan Greysen, Ashley Mercede, Catherine Reale, Frances K Barg, Kevin G Volpp, Jason Karlawish, and Alisa J Stephens-Shields. Effectiveness and ethics of incentives for research participation. *JAMA Intern Med*, 181(11):1479–88, 2021. doi: 10.1001/jamainternmed.2021.5450.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a super-

conductor. *Comp Mat Sci*, 154:346–54, 2018. doi: 10.1016/j.commatsci.2018.07.052.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. MedSafetyBench: Evaluating and improving the medical safety of large language models. In *Proc NeurIPS*, 2024. URL https://openreview.net/forum?id=cFyagd2Yh4.

Nikolaus Hansen. The CMA evolution strategy: A comparing review. *Towards a New Evolutionary Computation*, pages 75–102, 2006. doi: 10.1007/3-540-32494-1_4.

Nikolaus Hansen. The CMA evolution strategy: A tutorial. *arXiv Preprint*, 2016. doi: 10.48550/arXiv.1604.00772.

Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proc IEEE Int Conf on Evolutionary Comp*, pages 312–7, 1996. doi: 10.1109/ICEC.1996.542381.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statist Sci*, 1(3):297–310, 1986. doi: 10.1214/ss/1177013604.

David P. Helmbold and Philip M. Long. On the inductive bias of dropout. *J Mach Learn Res*, 16(1):3403–54, 2015. doi: 10.5555/2789272.2912107.

Alex Hernández-García, Nikita Saxena, Moksh Jain, Cheng-Hao Liu, and Yoshua Bengio. Multi-fidelity active learning with GFlowNets. *Trans Mach Learn Res*, 2024. URL https://openreview.net/forum?id=dLaazW9zuF.

Ben Hicks, Kirsty Kitto, Leonie Payne, and Simon Buckingham Shum. Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *Proc LAK Conf*, pages 250–9, 2022. doi: 10.1145/3506860.3506899.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Proc NeurIPS*, pages 4572–80, 2016. doi: 10.5555/3157382.3157608.

Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606:129–36, 2022. doi: 10.1038/s41586-022-04743-9.

Minh Hoang, Azza Fadhel, Aryan Deshwal, Jana Doppa, and Trong Nghia Hoang. Learning surrogates for offline black-box optimization via gradient matching. In *Proc ICML*, pages 18374–93, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J Am Stat Assoc*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Samuel Holt, Tennison Liu, and Mihaela van der Schaar. Automatically learning hybrid digital

twins of dynamical systems. In *Proc NeurIPS*, pages 72170–218, 2024. doi: 10.5555/3737916.3740220.

Arthur S Hong, David Levin, Laurence Parker, Vijay M Rao, Dennis Ross-Degnan, and J Frank Wharam. Trends in diagnostic imaging utilization among Medicare and commercially insured adults from 2003 through 2016. *Radiology*, 294(2):342–50, 2020. doi: 10.1148/radiol.2019191116.

Charles Hong, Sahil Bhatia, Alvin Cheung, and Yakun Sophia Shao. Autocomp: LLM-driven code optimization for tensor accelerators. *arXiv Preprint*, 2025. doi: 10.48550/arXiv.2505.18574.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, and Yuanzhi Li. LoRA: Low-rank adaptation of large language models. *Proc ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kai Hu, Klas Leino, Zifan Wang, and Matt Fredrikson. A recipe for improved certifiable robustness. In *Proc ICLR*, 2024. URL https://openreview.net/forum?id=qz3mcn99cu.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via Chi-squared preference optimization. *arXiv Preprint*, 2024a. doi: 10.48550/arXiv.2407.13399.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *Proc ICML*, 162:9118–47, 2022. doi: 10.48550/arXiv.2201.07207.

Xinmeng Huang, Shuo Li, Edgar Dobriban, Osbert Bastani, Hamed Hassani, and Dongsheng Ding. One-shot safety alignment for large language models via optimal dualization. In *Proc NeurIPS*, 2024b. doi: 10.48550/arXiv.2405.19544.

Danny R Hughes, Miao Jiang, and Richard Duszak Jr. A comparison of diagnostic imaging ordering patterns between advanced practice clinicians and primary care physicians following office-based evaluation and management visits. *JAMA Intern Med*, 175(1):101–7, 2015. doi: 10.1001/jamainternmed.2014.6349.

Zhiyi Huo, Weize Liu, and Qian Wang. Multi objective optimization method for collision safety of networked vehicles based on improved particle optimization. *J Control and Decision*, 10:134–42, 2022. doi: 10.1080/23307706.2022.2080771.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla Bayesian optimization performs great in high dimensions. In *Proc ICML*, volume 235, pages 20793–817. PMLR, 2024. URL https://proceedings.mlr.press/v235/hvarfner24a.html.

Javier Camacho Ibáñez and Mónica Villas Olmeda. Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI and Society*, 37:1663–

87, 2022. doi: 10.1007/s00146-021-01267-0.

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F P Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with GFlowNets. In *Proc ICML*, volume 162 of *Proc Mach Learn Res*, pages 9786–801. PMLR, 2022. URL https://proceedings.mlr.press/v162/jain22a.html.

Jacob C Jameson, Soroush Saghafian, Robert S Huckman, and Nicole Hodgson. Variation in batch ordering of imaging tests in the emergency department and the impact on care delivery. *Health Serv Res*, pages 1–7, 2024. doi: 10.1111/1475-6773.14406.

Seowoo Jang, Soyoung Yoo, and Namwoo Kang. Generative design by reinforcement learning: Enhancing the diversity of topology optimization designs. *Computer-Aided Design*, 146:103225, 2022. doi: 10.1016/j.cad.2022.103225.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? In *Proc Conf Emp Meth Nat Lang Proc*, pages 12143–70, 2024. doi: 10.18653/v1/2024.emnlp-main.677.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, 2024.

Zijian Jiang, Jianwen Zhou, and Haiping Huang. Relationship between manifold smoothness and adversarial vulnerability in deep learning with local errors. *Chinese Physics B*, 30(4), 2021. doi: 10.1088/1674-1056/abd68e.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proc Conf Emp Meth Nat Lang Proc*, pages 2567–77, 2019. doi: 10.18653/v1/D19-1259.

Qio Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nat*

Mach Intell, 1:389–99, 2019. doi: 10.1038/s42256-019-0088-2.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H Lehman, Leo A Celi, and Roger G Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*, 10(1), 2023. doi: 10.1038/s41597-022-01899-x.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. LLMs are prone to fallacies in causal inference. In *Proc EMNLP*, pages 10553–69, 2024. doi: 10.18653/v1/2024.emnlp-main.590.

Simone Maria Kagerbauer, Bernhard Ulm, Armin Horst Podtschaske, Dimislav Ivanov Andonov, Manfred Blobner, Bettina Jungwirth, and Martin Graessner. Susceptibility of AutoML mortality prediction algorithms to model drift caused by the COVID pandemic. *BMC Med Inform Decis Mak*, 24(1):34, 2024. doi: 10.1186/s12911-024-02428-z.

Ahmed T Kamil, Hadeel M Saleh, and Israa Hussain. A multi-swarm structure for particle swarm optimization: Solving the welded beam design problem. *J Phys: Conf Ser*, 1804:012012, 2021. doi: 10.1088/1742-6596/1804/1/012012.

Leonid Kantorovich and Gennady S Rubinstein. On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13:52–9, 1958.

Padma Kaul, Luan Manh Chu, Douglas C Dover, Roseanne O Yeung, Dean T Eurich, and Sonia Butalia. Disparities in adherence to diabetes screening guidelines among males and females in a universal care setting: A population-based study of 1,380,697 adults. *Lancet Regional Health*, 2022. doi: 10.1016/j.lana.2022.100320.

Ren Kawamura, Yukinori Harada, Shu Sugimoto, Yuichiro Nagase, Shinichi Katsukura, and Taro Shimizu. Incidence of diagnostic errors among unexpectedly hospitalized patients using an automated medical history–taking system with a differential diagnosis generator: Retrospective observational study. *JMIR Med Inform*, 10(1):e35225, 2022. doi: 10.2196/35225.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV*, pages 313–29, 2021. doi: 10.1007/978-3-030-66723-8_19.

Moien Abdul Basith Khan, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Juma Al Kaabi. Epidemiology of type 2 diabetes–Global burden of disease and forecasted trends. *J Epi Glob Health*, 10(1):107–11, 2020. doi: 10.2991/jegh.k.191028.001.

Minsu Kim, Federico Berto, Sungsoo Ahn, and Jinkyoo Park. Bootstrapped training of score-conditioned generator for offline design of biological sequences. In *Proc NeurIPS*, pages 67643–61, 2023. doi: 10.5555/3666122.3669080.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. MedExQA: Medical question answering benchmark with multiple explanations. *Proc Biomed Nat Lang Proc*, pages 167–81, 2024. doi: 10.18653/v1/2024.bionlp-1.14.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc Int Conf Learn Repr*, 2014. doi: 10.48550/arXiv.1412.6980.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proc ICLR*, 2013. doi: 10.48550/arXiv.1312.6114.

S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by simulated annealing. *Science*, 220 (4598):671–80, 1983. doi: 10.1126/science.220.4598.671.

Elin Kjelle, Eivind R Andersen, Arne M Krokeide, Lesley J J Soril, Leti van Bodegom-Vos, Fiona M Clement, and Bjørn M Hofmann. Characterizing and quantifying low-value diagnostic imaging internationally: A scoping review. *BMC Med Imaging*, 22(1):73, 2022. doi: 10.1186/s12880-022-00798-2.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proc ICML*, volume 119, pages 5338–48, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep*, 2020. doi: 10.1038/s41598-020-68771-z.

Robert Korom, Sarah Kiptinness, Najib Adan, Kassim Said, Catherine Ithuli, Oliver Rotich, Boniface Kimani, Irene King'ori, Stellah Kamau, Elizabeth Atemba, Muna Aden, Preston Bowman, Michael Sharman, Rebecca Soskin Hicks, Rebecca Distler, Johannes Heidecke, Rahul K Arora, and Karan Singhal. AI-based clinical decision support for primary care: A real-world study. *arXiv Preprint*, 2025. doi: 10.48550/arXiv.2507.16947.

Maria Korshunova, Niles Huang, Stephen Capuzzi, Dmytro S Radchenko, Olena Savych, Yuriy S Moroz, Carrow I Wells, Timothy M Willson, Alexander Tropsha, and Olexandr Isayev. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Commun Chem*, 5(129), 2022. doi: 10.1038/s42004-022-00733-0.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *Proc ICLR*, 2020. URL https://openreview.net/forum?id=Hyg-JC4FDr.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947.

Simone Kresevic, Mauro Giuffrè, Milos Ajcevic, Agostino Accardo, Lory S Crocè, and Dennis L

Shung. Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. *NPJ Digit Med*, 7(102), 2024. doi: 10.1038/s41746-024-01091-y.

Siddarth Krishnamoorthy, Satvik Mashkaria, and Aditya Grover. Diffusion models for black-box optimization. In *Proc ICML*, pages 17842–57, 2023a. doi: 10.5555/3618408.3619142.

Siddarth Krishnamoorthy, Satvik Mehul Mashkaria, and Aditya Grover. Generative pretraining for black-box optimization. In *Proc ICML*, volume 202, pages 24173–97. JMLR, 2023b.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Sci Data*, 10(170), 2023. doi: 10.1038/s41597-023-02068-4.

Aviral Kumar and Sergey Levine. Model inversion networks for model-based optimization. In *Proc NeurIPS*, pages 5126–37, 2019. doi: 10.5555/3495724.3496155.

Harold J Kushner. A new method of locating the maximum of an arbitrary multipeak curve in the presence of noise. *J Basic Eng*, 86(1):97–106, 1964. doi: 10.1115/1.3653121.

Robert M Kwee, Romy Toxopeus, and Thomas C Kwee. Imaging overuse in the emergency department: The view of radiologists and emergency physicians. *European J Radiol*, 176:111536, 2024. doi: 10.1016/j.ejrad.2024.111536.

Steven Labkoff, Bilikis Oladimeji, Joseph Kannry, Anthony Solomonides, Russell Leftwich, Eileen Koski, Amanda L Joseph, Monica Lopez-Gonzalez, Lee A Fleisher, Kimberly Nolen, Sayon Dutta, Deborah R Levy, Amy Price, Paul J Barr, Jonathan D Hron, Baihan Lin, Gyana Srivastava, Nuria Pastor, Unai Sanchez Luque, Tien Thi Thuy Bui, Reva Singh, Tayler Williams, Mark G Weiner, Tristan Naumann, Dean F Sittig, Gretchen Purcell Jackson, and Yuri Quintana. Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc*, 31(11):2730–9, 2024. doi: 10.1093/jamia/ocae209.

Reinhard C Laubenbacher, Anna Niarakis, Tomas Helikar, Gary An, Bruce Shapiro, Rahuman S Malik-Sheriff, T J Sego, Adam C Knapp, Paul Macklin, and James A Glazier. Building digital twins of the human immune system: Toward a roadmap. *npj Digit Med*, 5, 2022. doi: 10.1038/s41746-022-00610-z.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proc IJCAI*, pages 2801–7, 2019. doi: 10.5555/3367243.3367428.

B Laurent and P Massart. Adaptive estimation of a quadratic functional by model selection. *Ann Statist*, 28(5):1302–38, 2000. doi: 10.1214/aos/1015957395.

Joseph D Laurent, Jon M andJanizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling,

Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2407.10362.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved ttechniques for training LLMs as generalist embedding models. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Cindy S Lee, Paul G Nagy, Sallie J Weaver, and David E Newman-Toker. Cognitive and system factors contributing to diagnostic errors in radiology. *Am J Roentgenol*, 201(3), 2013. doi: 10.2214/AJR.12.10375.

Aaron Jiaxun Li, Satyapriya Krishna, and Himabindu Lakkaraju. More RLHF, more trust? On the impact of preference alignment on trustworthiness. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=FpiCLJrSW8.

Junqi Li, Steven G Ballmer, Eric P Gillis, Seiko Fujii, Michael J Schmidt, Andrea M E Palazzolo, Jonathan W Lehmann, Greg F Morehouse, and Martin D Burke. Synthesis of many different types of organic small molecules using one automated process. *Science*, 347(6227):1221–6, 2015. doi: 10.1126/science.aaa5414.

Shibo Li, Jeff M Phillips, Xin Yu, Robert M Kirby, and Shandian Zhe. Batch multi-fidelity active learning with budget constraints. In *Proc NeurIPS*, pages 995–1007, 2022a. doi: 10.5555/3600270.3600343.

Shibo Li, Zheng Wang, Robert Kirby, and Shandian Zhe. Deep multi-fidelity active learning of high-dimensional outputs. In *Proc Int Conf Artif Intell Stats*, volume 151, pages 1694–711, 2022b. URL https://proceedings.mlr.press/v151/li22b.html.

Shibo Li, Robert M Kirby, and Shandian Zhe. Batch multi-fidelity Bayesian optimization with deep auto-regressive networks. In *Proc NeurIPS*, pages 25463–75, 2024. doi: 10.5555/3540261.3542211.

Xingtao Liao, Qing Li, Xujing Yang, Weigang Zhang, and Wei Li. Multiobjective optimization for crash safety design of vehicles using stepwise regression model. *Structural and Multidisciplinary Optimization*, 35(6):561–9, 2008. doi: 10.1007/s00158-007-0163-x.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *Proc Assoc Comp Ling*, 60:3214–52, 2022. doi: 10.18653/v1/2022.acl-long.229.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–30, 2023. doi: 10.1126/science.ade2574.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. doi: 10.1145/3236386.3241340.

Patricia E Litkowski, Gerald W Smetana, Mark L Zeidel, and Melvin S Blanchard. Curbing the urge to image. *Am J Med*, 129(10):1131–5, 2016. doi: 10.1016/j.amjmed.2016.06.020.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–28, 1989. doi: 10.1007/BF01589116.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance Bayesian optimization. In *Proc ICLR*, 2024. URL https://openreview.net/forum?id=OOxotBmGol.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, Song Wu, Ping Zhang, Jian Yang, Chunli Song, and Guangyu Wang. A generalist medical language model for disease diagnosis assistance. *Nat Med*, 31:932–42, 2025a. doi: 10.1038/s41591-024-03416-6.

Yixiu Liu, Yang Nan, Weixian Xu, Xiangkun Hu, Lyumanshan Ye, Zhen Qin, and Pengfei Liu. AlphaGo moment for model architecture discovery. *arXiv Preprint*, 2025b. doi: 10.48550/arXiv.2507.18074.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proc NeurIPS*, pages 4768–77, 2017. doi: 10.5555/3295222.3295230.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proc Int Joint Conf Nat Lang Proc*, pages 305–29, 2023. doi: 10.18653/v1/2023.ijcnlp-main.20.

Shen-Huan Lyu, Lu Wang, and Zhi-Hua Zhou. Improving generalization of deep neural networks by leveraging margin distribution. *Neural Networks*, 151:48–60, 2022. doi: 10.1016/j.neunet.2022.03.019.

Yecheng Jason Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned reinforcement learning via f-advantage regression. In *Proc NeurIPS*, pages 310–23, 2022. doi: 10.5555/3600270.3600293.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *Proc ICLR*, 2023. doi: 10.48550/arXiv.2210.00030.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *Proc ICLR*, 2024. URL https://openreview.net/forum?id=IEduRUO55F.

Matthew T MacLean, Qasim Jehangir, Marijana Vujkovic, Yi-An Ko, Harold Litt, Arijitt Borthakur, Hersh Sagreiya, Mark Rosen, David A Mankoff, Mitchell D Schnall, Haochang Shou, Julio Chirinos, Scott M Damrauer, Drew A Torigian, Rotonya Carr, Daniel J Rader, and Walter R Witschey. Quantification of abdominal fat from computed tomography using deep learning and its association with electronic health records in an academic biobank. *J Am Med Inform Assoc*, 28(6): 1178–87, 2021. doi: 10.1093/jamia/ocaa342.

B P MacLeod, F G L Parlane, T D Morrissey, F Häse, L M Roch, K E Dettelbach, R Moreira, L P E Yunker, M B Rooney, J R Deeth, V Lai, G J Ng, H Situ, R H Zhang, M S Elliott, T H Haley, D J Dvorak, A Aspuru-Guzik, J E Hein, and C P Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Proc ACL*, pages 295–337, 2024. doi: 10.18653/v1/2024.findings-acl.19.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci Rep*, 14(14156), 2024. doi: 10.1038/s41598-024-64827-6.

Chaitanya Malaviya, Priyanka Agrawal, Kuzman Ganchev, Pranesh Srinivasan, Fantine Huot, Jonathan Berant, Mark Yatskar, Dipanjan Das, Mirella Lapata, and Chris Alberti. DOLOMITES: Domain-specific long-form methodical tasks. *Trans Assoc Comp Ling*, 13:1–29, 2025. doi: 10.1162/tacl_a_00727.

Natalie Maus, Haydn T Jones, Juston S Moore, Matt J Kusner, John Bradshaw, and Jacob R Gardner. Local latent space Bayesian optimization over structured inputs. In *Proc NeurIPS*, pages 34505–18, 2022. doi: 10.5555/3600270.3602770.

Natalie Maus, Kaiwen Wu, David Eriksson, and Jacob Gardner. Discovering many diverse solutions with Bayesian optimization. In *Proc Int Conf AI Stats*, volume 206, pages 1779–98. PMLR, 2023. URL https://proceedings.mlr.press/v206/maus23a.html.

Mathew W McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. Functional generalized additive models. *J Comput Graph Stat*, 23(1):249–69, 2014. doi: 10.1080/10618600.2012.729985.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum S Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Proc*

NeurIPS, 2024. doi: 10.48550/arXiv.2312.02119.

Jonas Mockus. The Bayesian approach to global optimization. In *System Modeling and Optimization*, pages 473–81. Springer, 1982.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. NV-Retriever: Improving text embedding models with effective hard-negative mining. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2407.15831.

Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines*, 31:239–56, 2021. doi: 10.1007/s11023-021-09563-w.

Depen Morwani and Harish G Ramaswamy. Inductive bias of gradient descent for weight normalized mmooth homogeneous neural nets. In *Proc Int Conf Algo Learn Theory*, volume 167, pages 827–80, 2022. URL https://proceedings.mlr.press/v167/morwani22a.html.

Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, and Seong Joon Oh. Trustworthy machine learning. *arXiv Preprint*, 2023. doi: 10.48550/arXiv.2310.08215.

Omer Mujahid, Ivan Contreras, Aleix Beneyto, and Josep Vehi. Generative deep learning for the development of a type 1 diabetes simulator. *Commun Med*, 4, 2024. doi: 10.1038/s43856-024-00476-0.

Megan M Mullis, Ian M Ramo, Brett J Baker, and Brandi K Reese. Diversity, ecology, and prevalence of antimicrobials in nature. *Front Microbiol*, 10, 2019. doi: 10.3389/fmicb.2019.02518.

Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022. doi: 10.1016/j.array.2022.100258.

Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv Preprint*, 2020. doi: 10.48550/arXiv.2001.01866.

Siddharth Narayanan, James D Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G Rodriques, and Andrew D White. Aviary: Training language agents on challenging scientific tasks. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2412.21154.

Lleayem Nazario-Johnson, Hossam A Zaki, and Glenn A Tung. Use of large language models to predict neuroimaging. *J Am Coll Radiol*, 20(10):1004–9, 2023. doi: 10.1016/j.jacr.2023.06.008.

Nathan Hoyen Ng, Neha Hulkund, Kyunghyun Cho, and Marzyeh Ghassemi. Predicting out-of-domain generalization with neighborhood invariance. *Trans Mach Learn Res*, 2023. URL https://openreview.net/forum?id=jYkWdJzTwn.

Giang Ngo, Rodney Beard, and Rohitash Chandra. Evolutionary bagging for ensemble learning. *Neurocomputing*, 510:1–14, 2022. doi: 10.1016/j.neucom.2022.08.055.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D Hsu, and Brian L Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Tung Nguyen, Sudhanshu Agrawal, and Aditya Grover. ExPT: Synthetic pretraining for few-shot experimental design. In *Proc NeurIPS*, pages 45856–69, 2023. doi: 10.5555/3666122.3668109.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and Chi-Squared-divergence, generalizations and applications. *Entropy*, 22(5), 2020. doi: 10.3390/e22050563.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv Preprint*, 2025. doi: 10.48550/arXiv.2506.22405.

The American Society of Addiction Medicine. The ASAM clinical practice guideline on alcohol withdrawal management. *J Addict Med*, 14(3):1–72, 2020. doi: 10.1097/ADM.0000000000000668.

American College of Radiology. ACR Appropriateness Criteria, 2024. URL https://acsearch.acr.org/list.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The Well: A large-scale collection of diverse physics simulations for machine learning. In *Proc NeurIPS*, pages 44989–5037, 2025. doi: 10.5555/3737916.3739346.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *NPJ Digit Med*, 6(195), 2023. doi: 10.1038/s41746-023-00939-z.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, and Daniel Shu Wei Ting. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*, 6(6):E428–32, 2024. doi: 10.1016/S2589-7500(24)00061-X.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and

Red Avila. GPT-4 technical report, 2024.

Samet Oymak, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural architecture search with train-validation split. In *Proc ICML*, volume 139, pages 8291–301, 2021. URL https://proceedings.mlr.press/v139/oymak21a.html.

Aini Palizhati, Steven B Torrisi, Muratahan Aykol, Santosh K Suram, Jens S Hummelshøj, and Joseph H Montoya. Agents for sequential learning using multiple-fidelity data. *Sci Rep*, 12 (4694), 2022. doi: 10.1038/s41598-022-08413-8.

Theodore P Papalexopoulos, Christian Tjandraatmadja, Ross Anderson, Juan Pablo Vielma, and David Belanger. Constrained discrete black-box optimization using mixed-integer programming. In *Proc ICML*, volume 162, pages 17295–322. PMLR, 2022. URL https://proceedings.mlr.press/v162/papalexopoulos22a.html.

J H Park. Moments of the generalized Rayleigh distribution. *Quarterly of Appl Math*, 19(1):45–9, 1961. doi: 10.1090/qam/119222.

Walter G Park, Nicholas J Shaheen, Jonathan Cohen, Irving M Pike, Douglas G Adler, John M Inadomi, Loren A Laine, John G Lieb, Maged K Rizk, Mandeep S Sawhney, and Sachin Wani. Quality indicators for EGD. *Am J Gastroenterol*, 110(1):60–71, 2015. doi: 10.1038/ajg.2014.384.

Swati Patel, Folasade May, Joseph C Anderson, Carol A Burke, Jason A Dominitz, Seth A Gross, Brian C Jacobson, Aasma Shaukat, and Douglas J Robertson. Updates on age to start and stop colorectal cancer screening: Recommendations from the U.S. Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol*, 117(1):57–69, 2022. doi: 10.14309/ajg.0000000000001548.

Ofir Pele and Michael Werman. Fast and robust Earth mover's distances. In *Proc ICCV*, pages 460–7, 2009. doi: 10.1109/ICCV.2009.5459199.

Argyrios Perivolaris, Chris Adams-McGavin, Yasmine Madan, Teruko Kishibe, Tony Antoniou, Muhammad Mamdani, and James J Jung. Quality of interaction between clinicians and artificial intelligence systems. A systematic review. *Fut Health J*, 11(3):100172, 2024. doi: 10.1016/j.fhj.2024.100172.

Antonio Pinto, Alfonso Reginelli, Fabio Pinto, Giuseppe Lo Re, Federico Midiri, Carlo Muzj, Luigia Romano, and Luca Brunese. Errors in imaging patients in the emergency setting. *Br J Radiol*, 89 (1061):20150914, 2016. doi: 10.1259/bjr.20150914.

Fernanda C G Polubriaginof, Ning Shang, George Hripcsak, Nicholas P Tatonetti, and David K Vawdrey. Low screening rates for diabetes mellitus among family members of affected relatives. *AMIA Annu Symp Proc*, pages 1471–7, 2019.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv Preprint*, 2019. doi: 10.48550/arxiv.1909.06312.

Justin Porter, Cynthia Boyd, M Reza Skandari, and Neda Laiteerapong. Revisiting the time needed to provide adult primary care. *J Gen Intern Med*, 2022. doi: 10.1007/s11606-022-07707-x.

Drew Prinster, Amama Mahmood, Suchi Saria, Jean Jeudy, Cheng Ting Lin, Paul H Yi, and Chieng-Ming Huang. Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology*, 313(2):e233261, 2024. doi: 10.1148/radiol.233261.

Ayis Pyrros, Stephen M Borstelmann, Ramana Mantravadi, Zachary Zaiman, Kaesha Thomas, Brandon Price, Eugene Greenstein, Nasir Siddiqui, Melinda Willis, Ihar Shulhan, John Hines-Shah, Jeanne M Horowitz, Paul Nikolaidis, Matthew P Lungren, Jorge Mario Rodríguez-Fernández, Judy Wawira Gichoya, Sanmi Koyejo, Adam E Flanders, Nishith Khandwala, Amit Gupta, John W Garrett, Joseph Paul Cohen, Brian T Layden, Perry J Pickhardt, and William Galanter. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nat Commun*, 14(4039), 2023. doi: 10.1038/s41467-023-39631-x.

Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Mary Wood, and Mihaela van der Schaar. SyncTwin: Treatment effect estimation with longitudinal outcomes. In *Proc NeurIPS*, pages 3178–90, 2021. doi: 10.5555/3540261.3540504.

Ruizhong Qiu, Zhe Xu, Wenxuan Bao, and Hanghang Tong. Ask, and it shall be given: On the Turing completeness of prompting. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=AS8SPTyBgw.

Laura Quinn, Konstantinos Tryposkiadis, Jon Deeks, Henrica CW De Vet, Sue Mallett, Lidwine B Mokkink, Yemisi Takwoingi, Sian Taylor-Phillips, and Alice Sitch. Interobserver variability studies in diagnostic imaging: A methodological systematic review. *Br J Radiol*, 96(1148):20220972, 2023. doi: 10.1259/bjr.20220972.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc ICML*, volume 139, pages 8748–63, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Proc CoRL*, 2022. doi: 10.48550/arXiv.2210.03109.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proc NeurIPS*, pages 53728–41, 2023. doi: 10.5555/3666122.3668460.

Ronilo Ragodos, Tong Wang, Lu Feng, and Yu Hu. From model explanation to data misinterpretation: Uncovering the pitfalls of post hoc explainers in business research. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2408.16987.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. In *Proc NeurIPS*, pages 9689–701, 2019. doi: 10.5555/3454287.3455156.

Ahlam Rashid. Iraqi diabetes dataset, 2020.

Alexander Rau, Stephan Rau, Daniela Z öller, Anna Fink, Hien Tran, Caroline Wilpert, Johanna Nattenmüller, Jakob Neubauer, Fabian Bamberg, Marco Reisert, and Maximilian F Russe. A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR Appropriateness Guidelines. *Radiology*, 308(1):e230970, 2023. doi: 10.1148/radiol.230970.

Tapabrata Ray and K M Liew. A swarm metaphor for multiobjective design optimization. *Engineering Optimization*, 34:141–53, 2002. doi: 10.1080/03052150210915.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc ACM SIGKDD*, pages 1135–44, 2016. doi: 10.1145/2939672.2939778.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–89, 2009. doi: 10.1561/1500000019.

Gelareh Sadigh, Paniz Charkhchi, A Mark Fendrick, Diana G Hassan, Annette Hatfield, and Ruth C Carlos. Financial burden of advanced imaging in radiology (FAIR Study). *J Am Coll Radiol*, 19 (2):254–8, 2022.

Sergio Salerno, Andrea Laghi, Marie-Claire Cantone, Paolo Sartori, Antonio Pinto, and Guy Frija. Overdiagnosis and overimaging: An ethical issue for radiological protection. *Radiol Med*, 124 (8):714–20, 2019. doi: 10.1007/s11547-019-01029-5.

Paul J Sample, Ban Wang, David W Reid, Vlad Presnyak, Iain J McFadyen, David R Morris, and Georg Seelig. Human 5'UTR design and variant effect prediction from a massively parallel translation study. *Nature Biotechnology*, 37:803–9, 2019. doi: 10.1038/s41587-019-0164-5.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med*, 7(20), 2024. doi: 10.1038/s41746-024-01010-1.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Proc NeurIPS*, pages 3839–48, 2018. doi: 10.5555/3327144.3327299.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity

and capacity in neural networks. *arXiv Preprint*, 2022. doi: 10.48550/arXiv.2210.01892.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *J Big Data*, 6, 2019. doi: 10.1186/s40537-019-0197-0.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172–80, 2023. doi: 10.1038/s41586-023-06291-2.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Med Inform*, 12:e55318, 2024. doi: 10.2196/55318.

Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proc Conf Human Factors Comp Sys*, pages 1–18, 2023. doi: 10.1145/3544548.3581075.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Proc NeurIPS*, volume 2, pages 2951–9, 2012. doi: 10.5555/2999325.2999464.

Ilya M Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *Mathematical Physics*, 7:86–112, 1967. doi: 10.1016/0041-5553(67)90144-9.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Proc Neur Inf Proc Sys*, pages 16857–67, 2020. doi: 10.5555/3495724.3497138.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Dev Sci*, 10:89–96, 2007. doi: 10.1111/j.1467-7687.2007.00569.x.

Matthew Spotnitz, Betina Idnay, Emily R Gordon, Rebecca Shyu, Gongbo Zhang, Cong Liu, James J Cimino, and Chunhua Weng. A survey of clinicians' views of the utility of large language models. *Appl Clin Inform*, 15(2):306–12, 2024. doi: 10.1055/a-2281-7092.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Proc NeurIPS*, 2024. URL https://openreview.net/forum?id=Jm2aK3sDJD.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV 2016 Workshops*, 2016.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=RC5FPYVQaH.

Shingo Suzuki, Keisuke Chosa, Cristina Barillà, Michael Yao, Orsetta Zuffardi, Kai Hirofumi, Tsuyoshi Shuto, Mary Ann Suico, Yuet W Kan, R Geoffrey Sargent, and Dieter C Gruener. Seamless gene correction in the human cystic fibrosis transmembrane conductance regulator locus by vector replacement and vector insertion events. *Front Genome Ed*, 4, 2022. doi: 10.3389/fgeed.2022.843885.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 2025. doi: 10.1038/s41586-025-09442-9.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–94, 2007. doi: 10.1214/009053607000000505.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624:86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Rong-Xi Tan, Ke Xue, Shen-Huan Lyu, Haopu Shang, Yao Wang, Yaoyuan Wang, Fu Sheng, and Chao Qian. Offline model-based optimization by learning to rank. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=sb1HgVDLjN.

Benjamin H Taragin, Lei Feng, and Carrie Ruzal-Shapiro. Online radiology appropriateness survey: Results and conclusions from an academic internal medicine residency. *Acad Radiol*, 10(7): 781–5, 2003. doi: 10.1016/s1076-6332(03)80123-x.

See Boon Tay, Guat Hwa Low, Gillian Jing En Wong, Han Jieh Tey, Fun Loon Leong, Constance Li, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Choon Hua Thng, Iain Bee Huat Tan, and Ryan Shea Ying Cong Tan. Use of natural language processing to infer sites of metastatic disease from radiology reports at scale. *JCO Clin Cancer Inform*, 8:e2300122, 2024. doi: 10.1200/CCI.23.00122.

The Mosaic Research Team. Introducing DBRX: A new state-of-the-art open LLM, 2024. URL https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for LLM agents. In *Proc ICLR*, 2025. URL https://openreview.net/forum?id=MeGDmZjUXy.

Dávid Terjék and Diego González-Sánchez. Optimal transport with f-divergence regularization and generalized Sinkhorn algorithm. In *Proc AISTATS*, volume 151 of *Proc Mach Learn Res*, pages

5135–65. PMLR, 2022. URL https://proceedings.mlr.press/v151/terjek22a.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *Proc Int Conf Intell Rob Sys*, pages 5026–33, 2012. doi: 10.1109/IROS.2012.6386109.

Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In *Proc ICML*, volume 139, pages 10358–68, 2021. URL https://proceedings.mlr.press/v139/trabucco21a.html.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-Bench: Benchmarks for data-driven offline model-based optimization. In *Proc ICML*, volume 162 of *ICML'22*, pages 21658–76. PMLR, 2022.

Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. In *Proc NeurIPS*, pages 11259–72, 2020. doi: 10.5555/3495724.3496669.

Gianluca Truda and Patrick Marais. Evaluating warfarin dosing models on multiple datasets with a novel software framework and evoluationary optimisation. *J Biomedical Informatics*, 113:103634, 2021. doi: 10.1016/j.jbi.2020.103634.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer New York, NY, 2008. doi: 10.1007/b13794.

Monica Tung, Ritu Sharma, Jeremiah S Hinson, Stephanie Nothelle, Jean Pannikottu, and Jodi B Segal. Factors associated with imaging overuse in the emergency department: A systematic review. *Am J Emerg Med*, 36(2):301–9, 2017. doi: 10.1016/j.ajem.2017.10.049.

Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nat Mach Intell*, 5:250–60, 2023. doi: 10.1038/s42256-023-00620-w.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proc NeurIPS*, pages 74952–65, 2023. doi: 10.5555/3666122.3669397.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc IEEE CVPR*, pages 2962–71, 2017. doi: 10.1109/CVPR.2017.316.

Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*, 19(281), 2019. doi: 10.1186/s12911-019-1004-8.

Vladimir I Valtchinov, Ivan K Ip, Ramin Khorasani, Jeremiah D Schuur, David Zurakowski, Jarone Lee, and Ali S Raja. Use of imaging in the emergency department: Do individual physicians

contribute to variation? *Am J Roentgenol*, 213(3):637–43, 2019. doi: 10.2214/AJR.18.21065.

Arjun K Venkatesh, Jeffrey A Kline, D Mark Courtney, Carlos A Camargo, Michael C Plewa, Kristen E Nordenholz, Christopher L Moore, Peter B Richman, Howard A Smithline, Daren M Deam, and Christopher Kabrhel. Evaluation of pulmonary embolism in the emergency department and consistency with a national quality measure: Quantifying the opportunity for improvement. *Arch Intern Med*, 172(13):1028–32, 2012. doi: 10.1001/archinternmed.2012.1804.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Wackesser, and Jonathan Bright. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*, 27:261–72, 2020. doi: 10.1038/s41592-019-0686-2.

Roberto Visentin, Chiara Dalla Man, Boris Kovatchev, and Claudio Cobelli. The University of Virginia/Padova type 1 diabetes simulator matches the glucose traces of a clinical trial. *Diabetes Technol Ther*, 16:428–34, 2014. doi: 10.1089/dia.2013.0377.

Chaoqi Wang, Yibo Jian, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *Proc ICLR*, 2024a. doi: 10.48550/arXiv.2309.16240.

Jindong Wang, Haoliang Li, Haohan Wang, Sinno Jialin Pan, and Xing Xie. Trustworthy machine learning: Robustness, generalization, and interpretability. In *Proc ACM SIGKDD*, pages 5827–28, 2023. doi: 10.1145/3580305.3599574.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. TWIN-GPT: Digital twins for clinical trials via large language model. *ACM Trans Multimedia Comput Commun Appl*, 2024b. doi: 10.1145/3674838.

Yuhui Wang, Hao He, Chao Wen, and Xiaoyang Tan. Truly proximal policy optimization. In *Proc Conf Unc Artif Intell*, volume 115, pages 113–22. PMLR, 2020. URL https://proceedings.mlr.press/v115/wang20b.html.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proc EMNLP*, pages 3876–87, 2022. doi: 10.18653/v1/2022.emnlp-main.256.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proc NeurIPS*, pages 24824–37, 2022. doi: 10.5555/3600270.3602070.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In *Findings NAACL*, pages 2148–64, 2024. doi: 10.18653/v1/2024.findings-naacl.138.

David Weininger. SMILES, A chemical language and information system. *J Chem Inf Comput Sci*, 28(1):31–6, 1988. doi: 10.1021/ci00057a005.

Colin White, Sam Nolen, and Yash Savani. Exploring the loss landscape in neural architecture search. In *Proc UAI*, volume 161, pages 654–64, 2021. URL https://proceedings.mlr.press/v161/white21a.html.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci*, 39:868–73, 1999.

Christopher Y K Williams, Brenda Y Miao, Aaron E Kornblith, and Atul J Butte. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat Commun*, 15(8236), 2024. doi: 10.1038/s41467-024-52415-1.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn*, 8(3–4):229–56, 1992. doi: 10.1007/BF00992696.

James T Wilson, Riccardo Moriconi, Frank Hutter, and Marc P Disenroth. The reparameterization trick for acquisition functions. *arXiv Preprint*, 2017. doi: 10.48550/arXiv.1712.00424.

James T Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Proc NeurIPS*, page 9906–9917, 2018. doi: 10.5555/3327546.3327655.

Max Wintermark, Nancy Fredericks, Judy Burleson, Jacqueline A Bello, Geraldine McGinty, Cynthia Smith, Steven E Weinberger, William T Thorwarth, and G Rebecca Haines. R-SCAN: Why we should care! *J Am Coll Radiol*, 13(10):1247–8, 2016. doi: 10.1016/j.jacr.2016.06.035.

Dongxia Wu, Ruijia Niu, Matteo Chinazzi, Yian Ma, and Rose Yu. Disentangled multi-fidelity deep Bayesian active learning. In *Proc ICML*, pages 37624–34, 2023. doi: 10.5555/3618408.3619975. URL https://proceedings.mlr.press/v202/wu23p.

Jiqing Wu and Viktor H Koelzer. Towards generative digital twins in biomedical research. *Comp Struct Biotechnol J*, 23:3481–8, 2024. doi: 10.1016/j.csbj.2024.09.030.

Yifan Wu, Wang Liu, Yue Yang, Michael S Yao, Wenli Yang, Xuehui Shi, Lihong Yang, Dongjun Li, Yueming Liu, Shiyi Yin, Chunyan Lei, Meixia Zhang, James C Gee, Xuan Yang, Wenbin Wei, and Shi Gu. A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. *Nat Comms*, 16(3504), 2025.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Huan He, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. MeLLaMA: Foundation large language models for medical applications. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2402.12749.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented

generation for medicine. *Findings Assoc Comp Ling*, pages 6233–51, 2024. URL https://aclanthology.org/2024.findings-acl.372.

Huan Xiong. Dpcd: Discrete principal coordinate descent for binary variable problems. *Proc AAAI Conf Artif Intell*, 36(9):10391–8, 2022. doi: 10.1609/aaai.v36i9.21281.

Guifeng Xu, Buyun Liu, Yangbo Sun, Yang Du, Linda G Snetselaar, Frank B Hu, and Wei Bao. Prevalence of diagnosed type 1 and type 2 diabetes among us adults in 2016 and 2017: Population based study. *BMJ*, 362, 2018. doi: 10.1136/bmj.k1497.

Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*, 34: 564–74, 2006. doi: 10.1093/nar/gkj454.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. *Radiol Artif Intell*, 4(4): e210258, 2022. doi: 10.1148/ryai.210258.

Tyler D Yan, Sabeena Jalal, and Alison Harris. Value-based radiology in Canada: Reducing low-value care and improving system efficiency. *Canadian Assoc Rad J*, 76(1):61–67, 2024. doi: 10.1177/08465371241277110.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *Proc ICLR*, 2024a. URL https://openreview.net/forum?id=Bb4VGOWELI.

Jenny Yang, Andrew A S Soltan, and David A Clifton. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *npj Digit Med*, 5(69), 2022. doi: 10.1038/s41746-022-00614-9.

Jenny Yang, Nguyen Thanh Dung, Pham Ngoc Thach, Nguyen Thanh Phong, Vu Dinh Phu, Khiem Dong Phu, Lam Minh Yen, Doan Bui Xuan Thy, Andrew AS Soltan, Louise Thwaites, and David A Clifton. Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nat Commun*, 15(8270), 2024b. doi: 10.1038/s41467-024-52618-6.

Ying Yang, Fang Yao, and Peng Zhao. Online smooth backfitting for generalized additive models. *J Am Stats Assoc*, 119:1215–28, 2023a. doi: 10.1080/01621459.2023.2182213.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proc IEEE CVPR*, pages 19187–97, 2023b. doi: 10.1109/CVPR52729.2023.01839.

Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S Yao, Chris Callison-Burch, James C Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image

analysis. In *Proc NeurIPS*, pages 90683–713, 2024c. doi: 10.5555/3737916.3740795.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. In *Proc NeurIPS*, volume 35, pages 10309–24, 2022. doi: 10.5555/3600270.3601019.

Michael S Yao and Michael S Hansen. A path towards clinical adaptation of accelerated MRI. *Proc Mach Learn Res*, 193:489–511, 2022.

Michael S Yao, Allison Chae, Matthew T MacLean, Anurag Verma, Jeffrey Duda, James C Gee, Drew A Torigian, Daniel Rader, Charles E Kahn, Walter R Witschey, and Hersh Sagreiya. SynthA1c: Towards clinically interpretable patient representations for diabetes risk stratification. In *Predictive Intelligence in Medicine*, pages 46–57, 2023. doi: 10.1007/978-3-031-46005-0_5.

Michael S Yao, Yimeng Zeng, Hamsa Bastani, Jacob R Gardner, James C Gee, and Osbert Bastani. Generative adversarial model-based optimization via source critic regularization. In *Proc NeurIPS*, 2024. doi: 10.48550/arXiv.2402.06532.

Michael S Yao, Allison Chae, Piya Saraiya, Charles E Kahn, Walter R Witschey, James C Gee, Hersh Sagreiya, and Osbert Bastani. Evaluating acute image ordering for real-world patient cases via language model alignment with radiological guidelines. *Commun Med*, 5, 2025a. doi: 10.1038/s43856-025-01061-9.

Michael S Yao, James C Gee, and Osbert Bastani. Diversity by design: Leveraging distribution matching for offline model-based optimization. *Proc ICML*, 2025b. doi: 10.48550/arXiv.2501.18768.

Michael S Yao, Lawrence Huang, Emily Leventhal, Clara Sun, Steve J Stephen, and Lathan Liou. Leveraging datathons to teach AI in undergraduate medical education: Case study. *JMIR Med Educ*, 11:e63602, 2025c. doi: 10.2196/63602.

Gary J Young, Stephen Flaherty, E David Zepeda, Koenraad J Mortele, and John L Griffith. Effects of physician experience, specialty training, and self-referral on inappropriate diagnostic imaging. *J Gen Intern Med*, 35(6):1661–7, 2020. doi: 10.1007/s11606-019-05621-3.

Peiyu Yu, Dinghuai Zhang, Hengzhi He, Xiaojian Ma, Ruiyao Miao, Yifan Lu, Yasi Zhang, Deqian Kong, Ruiqi Gao, Jianwen Xie, Guang Cheng, and Ying N Wu. Latent energy-based odyssey: Black-box optimization via expanded exploration in the energy-based latent space. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2405.16730.

Sihyun Yu, Sungsoo Ahn, Le Song, and Jinwoo Shin. RoMA: Robust model adaptation for offline model-based optimization. In *Proc NeurIPS*, pages 4619–31, 2021. doi: 10.5555/3540261.3540614.

Ye Yuan, Can Chen, Zixuan Liu, Willie Neiswanger, and Xue Liu. Importance-aware co-teaching for offline model-based optimization. In *Proc NeurIPS*, pages 55718–33, 2023. doi: 10.5555/3666122.

3668554.

Ye Yuan, Youyuan Zhang, Can Chen, Haolun Wu, Zixuan Li, Jianmo Li, James J Clark, and Xue Liu. Design editing for offline model-based optimization. *arXiv Preprint*, 2024. doi: 10.48550/arXiv.2405.13964.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *Proc ICLR*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Taeyoung Yun, Sujin Yun, Jaewoo Lee, and Jinkyoo Park. Guided trajectory generation with diffusion models for offline model-based optimization. In *Proc NeurIPS*, 2024. URL https://openreview.net/forum?id=ioKQzb8SMr.

Lorijn Zaadnoordijk, Tarek R Besold, and Rhodri Cusack. Lessons from infant learning for unsupervised machine learning. *Nat Mach Intell*, 4:510–20, 2022. doi: 10.1038/s42256-022-00488-2.

Hossam A Zaki, Andrew Aoun, Saminah Munshi, Hazem Abdel-Megid, Lleayem Nazario-Johnson, and Sun Ho Ahn. The application of large language models for radiologic decision making. *J Am Coll Radiol*, 21(7):1072–8, 2024. doi: 10.1016/j.jacr.2024.01.007.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 2025. doi: 10.1038/s41586-025-08628-5.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med*, 30:3129–41, 2024. doi: 10.1038/s41591-024-03185-2.

Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *Proc ICLR*, 2025a. URL https://openreview.net/forum?id=g6Qc3p7JH5.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-RLVR: Emerging medical reasoning from a 3B base model via reinforcement learning. *arXiv Preprint*, 2025b. doi: 10.48550/arXiv.2502.19655.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image-text pairs. *NEJM AI*, 2(1), 2025c. doi:

10.1056/AIoa2400640.

Xiao-Chen Zhang, Cheng-Kun Wu, Jia-Cai Yi, Xiang-Xiang Zeng, Can-Qun Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research*, 2022: 0004, 2022. doi: 10.34133/research.0004.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans Intell Syst Technol*, 15(2):1–38, 2024. doi: 10.1145/3639372.

Han Zhou, Xingchen Ma, and Matthew B Blaschko. A corrected expected improvement acquisition function under noisy observations. In *Proc Asian Conf Mach Learn*, volume 222, pages 1747–62. PMLR, 2024a. URL https://proceedings.mlr.press/v222/zhou24a.html.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(10752), 2019. doi: 10.1038/s41598-019-47148-x.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *Proc ICLR*, 2024b. URL https://openreview.net/forum?id=oMLQB4EZE1.

He Zhu, Jun Bai, Na Li, Xiaoxiao Li, Dianbo Liu, David L Buckeridge, and Yue Li. FedWeight: Mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *npj Digit Med*, 8, 2025. doi: 10.1038/s41746-025-01661-8.

Anna Zink, Hongzhou Luan, and Irene Y Chen. Access to care improves EHR reliability and clinical risk prediction model performance. In *Proc ML4H*, 2024. doi: 10.48550/arXiv.2412.07712.

Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *Proc ICLR*, 2017. URL https://openreview.net/forum?id=r1Ue8Hcxg.

Jay W Zussman, Jessica Y Ma, Jay G Bindman, Susannah Cornes, John A Davis, and Sam Brondfield. Identifying strategies for the use of gender and sex language in clinical one-liners. *LGBT Health*, 11(6):484–94, 2024. doi: 10.1089/lgbt.2023.0220.