FEW-SHOT BIOACTIVITY PREDICTION WITH META-LEARNING UNDER ASSAY HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate bioactivity prediction is a central challenge in early-stage drug discovery, as individual assays often contain too few measurements to train reliable models independently. Meta-learning offers a principled approach to this few-shot setting, but assay heterogeneity may limit its effectiveness. Here, we test this hypothesis and show that meta-learning performance degrades as meta-training tasks become more heterogeneous. To address this, we introduce `MetaHeta`, a meta-learning framework that accounts for assay heterogeneity by conditioning predictions on auxiliary data from related assays, with relatedness defined flexibly from available assay information. The architecture of `MetaHeta` combines linear attention over large auxiliary datasets with exact attention over scarce task-specific context, enabling efficient scaling to the former without compromising exact attention over the latter. We demonstrate the benefits of our approach on assays from ChEMBL and BindingDB, improving few-shot bioactivity prediction and downstream compound prioritization in retrospective Bayesian optimization.

1 INTRODUCTION

Accurately predicting compound activity in biological assays is a central challenge in early-stage drug discovery. Reliable *in silico* predictors can guide candidate selection and therefore reduce the time and resources required for experimental evaluation (Jiménez-Luna et al., 2021; Scalia et al., 2026). Machine learning models have been widely employed for this task, but their performance typically relies on the availability of large amounts of high-quality data (Tian et al., 2024; Mayr et al., 2018). However, since bioactivity measurements are expensive to acquire, individual assays often contain too little data to train accurate models in isolation, placing assay-specific bioactivity prediction in a few-shot learning regime (Vella & Ebejer, 2022; Altae-Tran et al., 2017). A natural way to alleviate this data scarcity is to share information across assays, but naively pooling their measurements risks conflating distinct assay-specific relationships.

A particularly promising approach to this problem is meta-learning (Schmidhuber, 1987; Hochreiter et al., 2001), which aims to learn shared structure across a collection of related *tasks* to facilitate adaptation to new ones, but without directly pooling their measurements. This collection of tasks is commonly referred to as the *meta-dataset*. In bioactivity prediction, each such task can naturally correspond to an individual assay, with a small number of observed compound–activity pairs forming a *context set* from which activities are predicted for new compounds (*target inputs*).

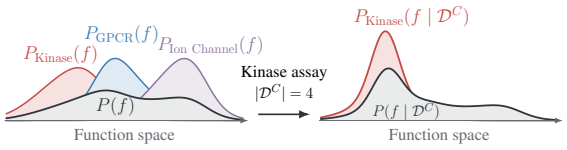


Figure 1: **Failure mode of meta-learning under assay heterogeneity.** A heterogeneous meta-dataset can induce a broad prior, such that limited task-specific data are insufficient to concentrate the posterior around functions relevant to a new assay.

A variety of meta-learning approaches have been developed to enable information sharing across tasks. Finn et al. (2017) introduced MAML, a gradient-based algorithm that learns a shared parameter initialization from which the model can adapt to new tasks in a few gradient steps. With a similar bi-level optimization objective, Chen et al. (2023) learn a shared feature representation on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

top of which task-specific Gaussian Processes (GPs) are fit. Neural Processes (NPs) (Garnelo et al., 2018b;a) instead amortize task adaptation by learning a mapping from a set of context points to predictive distributions at target locations, offering a scalable alternative to GPs. Transformer Neural Processes (TNPs) (Nguyen & Grover, 2022) parameterize this mapping using attention mechanisms, making them well suited to modeling sets while capturing rich interactions between context and target points. This architecture also underpins Prior-Data Fitted Networks (PFNs), which pre-train TNPs on massive synthetic datasets for off-the-shelf prediction on new tasks (Müller et al., 2022).

A key assumption underlying meta-learning is that tasks in the meta-dataset are sufficiently related to induce a shared prior that supports efficient adaptation to new tasks from limited data (Baxter, 2000). However, the heterogeneity of assays in large-scale bioactivity datasets, arising from differences in experimental conditions, endpoints, and measurement protocols (Landrum & Riniker, 2024), can render this assumption questionable, potentially limiting the effectiveness of meta-learning. We illustrate the intuition behind this potential failure mode in Figure 1. Indeed, prior work suggests that meta-learning methods struggle to adapt to task-specific bioactivity data (Stanley et al., 2021), with Schimunek et al. (2023) reporting that even a naive baseline that ignores the context set can outperform several established meta-learning approaches. Here, we hypothesize that assay heterogeneity is a key factor limiting the effectiveness of meta-learning for bioactivity prediction, and explicitly account for it in a dedicated modeling framework.

Our contributions are as follows:
(i) we systematically study the effect of meta-training task heterogeneity on few-shot bioactivity prediction, demonstrating in a controlled setting that increasing heterogeneity can degrade meta-learning performance across a range of methods and context sizes;
(ii) motivated by these findings, we introduce `MetaHeta`, a meta-learning framework in which predictions are additionally conditioned on data from related assays. This allows the model to approximate the posterior predictive under a more informative prior relevant to the target task, rather than under that induced by the entire heterogeneous meta-training distribution. Assay relatedness can be defined flexibly based on target homology or known experimental conditions;
(iii) to make conditioning on such auxiliary data computationally practical, we develop a TNP-based architecture that combines exact attention over scarce task-specific context with linear attention over potentially large auxiliary datasets, enabling efficient conditioning at scale.

We further demonstrate that findings from our controlled heterogeneity study translate into practical gains on assays from real-world bioactivity databases, improving few-shot bioactivity prediction and retrospective Bayesian optimization.

2 Related work

Early meta-learning approaches include metric-based methods (Vinyals et al., 2016; Snell et al., 2017), developed primarily for few-shot classification, and adaptation-based methods that learn shared structure across tasks to enable rapid task-specific adaptation from limited data (Finn et al., 2017; Nichol et al., 2018). These ideas have also been applied to few-shot molecular property prediction (Nguyen et al., 2020; Altae-Tran et al., 2017; Schimunek et al., 2023; Wang et al., 2021; 2024), with much of this work focusing on classification. More directly relevant to our probabilistic regression setting, deep kernel-based methods meta-learn shared feature representations while retaining task-specific Gaussian process inference (Chen et al., 2023; Patacchiola et al., 2020). A complementary line of work instead amortizes task adaptation by directly learning a mapping from task-specific context and target inputs to predictive distributions, with Neural Processes (NPs) forming a canonical family of such methods (Garnelo et al., 2018b;a; Kim et al., 2019; Rochussen & Fortuin, 2026). In recent years, NPs equipped with transformer architectures (Vaswani et al., 2017) (TNPs) have emerged as a prominent variant (Nguyen & Grover, 2022; Feng et al., 2023; Ashman, 2024), with Prior-Data Fitted Networks (PFNs) further extending this paradigm by pre-training TNPs on broad synthetic task distributions (Müller et al., 2022). This line of work has been particularly influential in tabular learning (Hollmann et al., 2025; Grinsztajn et al., 2026; Qu et al., 2026), with recent studies showing that tabular foundation models can exhibit capabilities beyond conventional tabular settings (Nakerst et al., 2026), including molecular property prediction (Hicham et al., 2026). We also discuss methods that explicitly account for task heterogeneity in Appendix D.

3 SETUP

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the space of d-dimensional molecular representations and $\mathcal{Y} \subseteq \mathbb{R}$ the space of assay outcomes. We model each assay τ as a task, corresponding to an unknown function $f_\tau : \mathcal{X} \to \mathcal{Y}$, where $f_\tau \sim P(f)$, and $P(f)$ denotes an underlying stochastic process on $\mathcal{X}$. For each τ, we observe a finite dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N \in \mathfrak{D}$ generated according to $y_n = f_\tau(x_n) + \epsilon_n$, with $\epsilon_n \sim \mathcal{N}(0, \sigma_\tau^2)$. Here, $\mathfrak{D} := \bigcup_{N\in\mathbb{N}}(\mathcal{X} \times \mathcal{Y})^N$ denotes the collection of all such datasets. Each dataset can be partitioned into a context set $\mathcal{D}^C$ and a target set $\mathcal{D}^T$ of sizes $N_c := |\mathcal{D}^C|$ and $N_t := |\mathcal{D}^T|$, respectively. We denote the corresponding inputs and outputs by $(X^C, Y^C) \in \mathbb{R}^{N_c\times d} \times \mathbb{R}^{N_c}$ and $(X^T, Y^T) \in \mathbb{R}^{N_t\times d} \times \mathbb{R}^{N_t}$. We assume access to a meta-dataset $\mathcal{D}_{\text{meta}} := \{\mathcal{D}_\tau\}_{\tau=1}^M$ consisting of M assay-specific datasets, each containing a potentially small number of observations. Our goal is to meta-learn a predictor that maps an assay-specific context set and target inputs to a predictive distribution over target outcomes. Formally, we seek to learn the posterior predictive map $\pi : \mathfrak{D} \times \mathcal{X}^{N_t} \to \mathcal{P}(\mathcal{Y}^{N_t})$, which we parameterize by $p_\theta(Y^T \mid X^T, \mathcal{D}^C)$.

4 METAHETA

4.1 BEYOND A SINGLE SHARED PRIOR

Failure mode of meta-learning under assay heterogeneity Consider a probabilistic meta-learner p_θ trained by maximizing the expected log-likelihood:

$$\mathcal{L}(\theta) = \mathbb{E}_{X^T,Y^T,\mathcal{D}^C}\left[\log p_\theta\left(Y^T \mid X^T, \mathcal{D}^C\right)\right] \tag{1}$$

$$= -\mathbb{E}_{X^T,\mathcal{D}^C}\left[\mathrm{KL}\big[p\left(Y^T \mid X^T, \mathcal{D}^C\right) \,\|\, p_\theta\left(Y^T \mid X^T, \mathcal{D}^C\right)\big]\right] + \text{const}, \tag{2}$$

where the expectations over meta-training tasks and their context-target splits are estimated using $\mathcal{D}_{\text{meta}}$, and $p(Y^T \mid X^T, \mathcal{D}^C)$ denotes the posterior predictive distribution (PPD) that $p_\theta(Y^T \mid X^T, \mathcal{D}^C)$ is trained to approximate (Nagler, 2023). The PPD can be written as:

$$p(Y^T \mid X^T, \mathcal{D}^C) = \int p(Y^T \mid X^T, f) P(df \mid \mathcal{D}^C), \tag{3}$$

where the prior $P(f)$ underlying the posterior $P(\cdot \mid \mathcal{D}^C)$ reflects the distribution of assay functions represented in $\mathcal{D}_{\text{meta}}$. Consequently, the composition of tasks in $\mathcal{D}_{\text{meta}}$ shapes the shared prior $P(f)$ and, in turn, influences the ability of the meta-learner to adapt to new tasks given limited task-specific context $\mathcal{D}^C$. Since bioactivity assays are often highly heterogeneous (Landrum & Riniker, 2024), the prior induced by $\mathcal{D}_{\text{meta}}$ may be too broad for the task-specific context to sufficiently concentrate the posterior on the relevant assay regime, limiting the effectiveness of meta-learning.

Accounting for heterogeneity through assay relatedness Motivated by this, we propose to model tasks as arising from a mixture of latent assay regimes $\xi_k \sim p(\xi)$, reflecting their shared characteristics such as the biological target or assay conditions. Each such component ξ_k induces its own prior $P_{\xi_k}(f)$, giving rise to a corresponding component-specific PPD $p_{\xi_k}(Y^T \mid X^T, \mathcal{D}^C)$. The prior induced by the heterogeneous $\mathcal{D}_{\text{meta}}$ can then be viewed as a mixture of these component priors, with the corresponding PPD being a mixture of the component-specific PPDs. Since learning a separate posterior predictive map for each of potentially many latent components is impractical, we follow Ashman et al. (2024) and instead use a single meta-learner $p_\theta(Y^T \mid X^T, \mathcal{D}^C, \mathcal{D}^{\text{aux}})$ to amortize prediction across different latent components. Here, $\mathcal{D}^{\text{aux}} = \{\mathcal{D}_{\tau'}\}$ denotes a set of auxiliary datasets generated from the same latent component ξ_k as the target task τ:

$$\xi_k \sim p(\xi), \quad f_\tau, \{f_{\tau'}\} \sim P_{\xi_k}(f), \quad \left(X^T, Y^T\right), \mathcal{D}^C \sim f_\tau, \quad \mathcal{D}_{\tau'} \sim f_{\tau'}. \tag{4}$$

For a single target input-output pair (x, y), Ashman et al. (2024) show that:

$$\begin{aligned} &\mathbb{E}_{\xi_k,\,\mathcal{D}^C,\,\mathcal{D}^{\text{aux}}}\Big[\mathrm{KL}\big(p_{\xi_k}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}})\big)\Big] \\ &\quad\le\; \mathbb{E}_{\xi_k,\,\mathcal{D}^C}\Big[\mathrm{KL}\big(p_{\xi_k}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C)\big)\Big]. \end{aligned} \tag{5}$$

Intuitively, conditioning on auxiliary data provides additional evidence about the underlying assay regime ξ_k and can yield a closer approximation to the corresponding PPD p_{ξ_k} than conditioning on task-specific context alone. Such additional evidence can be particularly valuable when $\mathcal{D}^C$ is small and may not be sufficiently informative on its own. This raises two practical challenges: identifying related assays to construct $\mathcal{D}^{\text{aux}}$, and efficiently conditioning a meta-learner on both $\mathcal{D}^C$ and $\mathcal{D}^{\text{aux}}$.

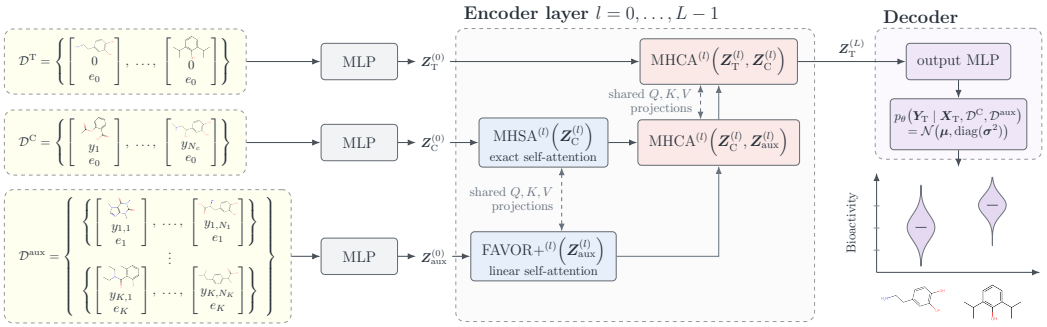


Figure 2: Architecture overview of `MetaHeta`. A shared input MLP embeds context, target, and concatenated auxiliary observations, augmented with task-of-origin indicators, with target activities masked to 0. Encoder layers use exact self-attention for the context and FAVOR+ for auxiliary data, followed by context-to-auxiliary and target-to-context cross-attention. When $\mathcal{D}^{\text{aux}} = \varnothing$, the auxiliary branch is skipped. Finally, an output MLP parameterizes a mean-field Gaussian predictive distribution. Formal definitions of the attention operators are given in Appendix F.

4.2 Defining assay relatedness

Depending on the modeling problem and available information, it might be useful to define assay relatedness at different levels. We consider two classes of settings.

Target-based relatedness Consider a setting where ξ_k is determined by target identity, such that assays sharing the same target are drawn from $P_{\xi_k}(f)$. Then, Eq. 5 directly suggests constructing $\mathcal{D}^{\text{aux}}$ by collecting tasks with the same biological target as τ. However, this construction can be overly restrictive with, for example, understudied protein targets, which are particularly relevant from a few-shot modeling perspective precisely because their assay data are often scarce (Sharma et al., 2023). In such settings, known biological relationships between targets can be leveraged to construct $\mathcal{D}^{\text{aux}}$ instead, as formalised in Proposition 1:

Proposition 1. *Consider an extension of the data-generating process in Eq. 4 in which the latent components are structured hierarchically as* $\xi_{\text{family}} \sim p(\xi)$ *and* $\xi_{\text{prot}}, \{\xi_{\text{prot}'}\} \sim p(\cdot \mid \xi_{\text{family}})$*, with* $f_\tau \sim P_{\xi_{\text{prot}}}(f)$ *and* $\{f_{\tau'}\} \sim P_{\xi_{\text{prot}'}}(f)$*. Under this generative structure, assume that auxiliary data from target homologs provide no additional information for predicting* y *once the target's protein-specific component is known. Then:*

$$\mathbb{E}_{\xi_{\text{family}},\, \xi_{\text{prot}},\, \mathcal{D}^C,\, \mathcal{D}^{\text{aux}}} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \;\|\; p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \Big) \right]$$
$$\leq \mathbb{E}_{\xi_{\text{family}},\, \xi_{\text{prot}},\, \mathcal{D}^C} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \;\|\; p(y \mid x, \mathcal{D}^C) \Big) \right].$$

Intuitively, assays of distinct but homologous proteins can still inform predictions for the target through their shared family-level component. The proof is in Appendix G.

Description-based relatedness Target identities or known biological relationships may not always be available to practitioners. In such cases, we instead define relatedness based on broadly available information about assay protocols or experimental conditions. Specifically, for each τ, we assume access to its natural-language description a_τ, which may be noisy or incomplete. Prior to training, we map each a_τ to a d-dimensional dense representation $h_\tau = \phi(a_\tau) \in \mathbb{R}^d$, where ϕ denotes an embedding model; in our experiments ϕ is instantiated as PubMedBERT (Gu et al., 2021). We then quantify relatedness between assays τ and τ' by cosine similarity between their description embeddings, $s_{\tau,\tau'} = \langle h_\tau, h_{\tau'} \rangle / (\|h_\tau\| \, \|h_{\tau'}\|)$. Since $s_{\tau,\tau'}$ provides only a heuristic for identifying useful auxiliary assays, we control retrieval selectivity using a threshold γ and include $\mathcal{D}_{\tau'}$ in $\mathcal{D}^{\text{aux}}$ if $s_{\tau,\tau'} \geq \gamma$. Meta-training with description-based relatedness is shown in Algorithm 1.

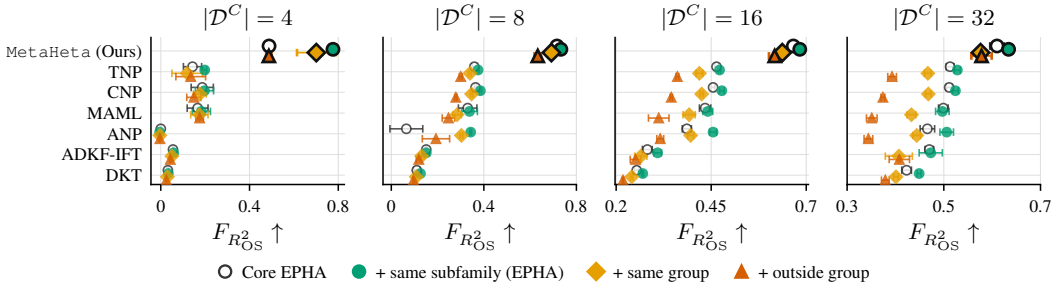


Figure 3: Effect of meta-training task heterogeneity on few-shot bioactivity prediction. Performance of all the considered methods generally deteriorates as $\mathcal{D}^{\text{core}}_{\text{meta}}$ is augmented with progressively less related tasks. `MetaHeta` remains the best-performing approach regardless of the training regime.

4.3 Model

The architecture is shaped by two competing design constraints arising from the variable availability of auxiliary data. First, because $\mathcal{D}^{\text{aux}}$ may frequently be empty, parameters dedicated exclusively to auxiliary data would often receive no learning signal, motivating parameter sharing between the processing of $\mathcal{D}^C$ and $\mathcal{D}^{\text{aux}}$. Second, $\mathcal{D}^{\text{aux}}$ may instead contain many observations across multiple related datasets, making quadratic attention prohibitively expensive, while $\mathcal{D}^C$ is typically small. Parameter sharing between $\mathcal{D}^C$ and $\mathcal{D}^{\text{aux}}$ is therefore necessary, but their different scales call for different computational treatment. Our architecture, shown in Figure 2, addresses this using two coupled TNP-like branches that share attention parameters while employing different self-attention mechanisms. For both branches, each observation $[x, y, e_j]$ is first embedded with an MLP, with target observations having their activities masked to 0. Here, e_j is a one-hot indicator of the task of origin, allowing observations from all auxiliary datasets to be concatenated into a single sequence without ambiguity. To preserve invariance to permutations of auxiliary datasets, the indicators are assigned using a canonical ordering of $\mathcal{D}^{\text{aux}}$, for example by task relatedness (Proposition 2 in Appendix G). Within each encoder layer, context representations attend to the auxiliary representations, before target representations attend to the augmented context. For the context branch, we employ standard softmax self-attention, whereas the auxiliary branch uses FAVOR+ (Choromanski et al., 2021), a linear-time approximation based on positive orthogonal random features. Since FAVOR+ provides an unbiased estimator of the softmax kernel, the same parameters can be used under both attention schemes while still targeting the same underlying operation. This allows us to scale efficiently to large $\mathcal{D}^{\text{aux}}$ without sacrificing exact attention over the task-specific context. The use of coupled TNP-like branches builds on ICICL-TNP (Ashman et al., 2024); we provide a detailed comparison of the two architectures in Appendix D.

5 Experiments

We structure the experiments as follows. In Section 5.1, we investigate the effect of meta-training task heterogeneity on meta-learning performance in a controlled environment. In Section 5.2, we show the few-shot bioactivity prediction capabilities of `MetaHeta` on real-world bioactivity assays from ChEMBL (Gaulton et al., 2012). In Section 5.3, we demonstrate its effectiveness for downstream compound prioritization through retrospective Bayesian optimization. Finally, in Section 5.4, we ablate the auxiliary-conditioning mechanism and examine its computational cost. All results are averaged over five random seeds and reported with standard errors.

5.1 Meta-learning under controlled assay heterogeneity

To study the effect of task heterogeneity in isolation, we construct a semi-synthetic benchmark based on kinase activity screening data from Drewry et al. (2017). Kinases provide a well-suited testbed: their common ATP-binding function is reflected in a highly conserved catalytic domain, while inhibitor response depends largely on variation at a small subset of active-site residues (Sheridan et al.,

Table 1: Few-shot prediction on EPHB test tasks after meta-training on $\mathcal{D}_{\text{meta}}^{\text{core}}$. `MetaHeta` matches or exceeds all baselines across both metrics and all context sizes.

	Context size = 4		Context size = 8		Context size = 16		Context size = 32	
Method	F_{NLL} ↑	$F_{R^2_{os}}$ ↑	F_{NLL} ↑	$F_{R^2_{os}}$ ↑	F_{NLL} ↑	$F_{R^2_{os}}$ ↑	F_{NLL} ↑	$F_{R^2_{os}}$ ↑
TNP	0.611 ± 0.019	-0.077 ± 0.029	0.827 ± 0.004	0.110 ± 0.025	0.900 ± 0.002	0.226 ± 0.020	0.931 ± 0.001	0.309 ± 0.015
ANP	0.601 ± 0.014	-0.001 ± 0.002	0.812 ± 0.004	0.025 ± 0.035	0.899 ± 0.001	0.228 ± 0.011	0.931 ± 0.001	0.321 ± 0.007
CNP	0.609 ± 0.014	0.124 ± 0.010	0.827 ± 0.004	0.156 ± 0.004	$\mathbf{0.901 \pm 0.002}$	0.252 ± 0.016	0.930 ± 0.001	0.313 ± 0.006
ADKF-IFT	0.262 ± 0.027	0.028 ± 0.003	0.760 ± 0.007	0.079 ± 0.013	0.882 ± 0.002	0.174 ± 0.013	0.925 ± 0.002	0.282 ± 0.019
DKT	0.121 ± 0.018	0.021 ± 0.002	0.646 ± 0.011	0.071 ± 0.002	0.806 ± 0.006	0.158 ± 0.012	0.874 ± 0.003	0.285 ± 0.010
MAML	–	-0.101 ± 0.119	–	0.108 ± 0.045	–	0.257 ± 0.013	–	0.314 ± 0.023
`MetaHeta`	$\mathbf{0.829 \pm 0.015}$	$\mathbf{0.501 \pm 0.018}$	$\mathbf{0.882 \pm 0.004}$	$\mathbf{0.346 \pm 0.053}$	$\mathbf{0.901 \pm 0.004}$	$\mathbf{0.327 \pm 0.084}$	$\mathbf{0.934 \pm 0.003}$	$\mathbf{0.333 \pm 0.041}$

2009; Born et al., 2022). Coupled with the well-defined phylogenetic organization of the kinome, these properties provide a natural, biologically meaningful hierarchy of task relatedness.

Benchmark construction We first fit a separate GP to each kinase's activity measurements and sample synthetic tasks from the resulting posteriors. We consider four meta-training regimes sharing a common meta-dataset $\mathcal{D}_{\text{meta}}^{\text{core}}$ of 5000 tasks sampled from GPs corresponding to EPHA kinases, with the unaugmented $\mathcal{D}_{\text{meta}}^{\text{core}}$ serving as a baseline. The three additional regimes extend this common core, each with another 5000 tasks sampled from GPs corresponding to: (i) EPHA kinases, (ii) non-EPHA tyrosine kinases (TK), and (iii) non-tyrosine kinases (non-TK), yielding progressively more heterogeneous task distributions, as reflected by relatedness measures reported in Table 3 in Appendix A.1. Crucially, the 5000 tasks added per regime differ only in the kinase-specific GP from which activities are sampled, so heterogeneity is induced solely through differences in the task-generating functions. Evaluation is likewise fixed across regimes and performed on held-out validation and test sets, each comprising 250 tasks sampled from GPs corresponding to EPHA kinases. A full description of the synthetic task-generation procedure is provided in Appendix A.1.

Experimental setup We evaluate all methods in the few-shot regime at context sizes $|\mathcal{D}^C| \in \{4, 8, 16, 32\}$. We compare our approach against a set of strong probabilistic meta-learning methods, including members of the NP family, CNP, ANP, and TNP (Garnelo et al., 2018a; Kim et al., 2019; Nguyen & Grover, 2022), as well as deep kernel-based methods, ADKF-IFT and DKT (Chen et al., 2023; Patacchiola et al., 2020). We additionally include MAML (Finn et al., 2017) as a widely used gradient-based baseline. Since the benchmark tasks are sampled from fitted GP posteriors, we can evaluate predictions relative to the corresponding true data-generating functions. We therefore report the fraction of oracle performance in terms of R^2_{os} and NLL, denoted by $F_{R^2_{os}}$ and F_{NLL}, respectively. For our method, $\mathcal{D}^{\text{aux}}$ contains at most five same-target datasets, ordered by decreasing average 5-nearest-neighbor Tanimoto similarity to the current task. Auxiliary datasets exceeding 512 observations are randomly subsampled to that size. Full details are provided in Appendix A.4.

Assay heterogeneity can hurt meta-learning performance The effect of increasing task heterogeneity is remarkably consistent across the considered methods and context sizes, as shown for $F_{R^2_{\text{os}}}$ in Figure 3. Across all 28 method–context combinations, augmenting $\mathcal{D}_{\text{meta}}^{\text{core}}$ with additional tasks from the same kinase subfamily improves $F_{R^2_{\text{os}}}$ in 25 cases, whereas augmentation with tasks from other tyrosine and non-tyrosine kinases decreases it in 24 and 26 cases, respectively. Moreover, the latter leads to a larger performance deterioration than the former in 23 comparisons. The effects on F_{NLL} are smaller in magnitude but follow the same overall pattern, as shown in Figure 6 of Appendix B. Same-subfamily augmentation improves performance in 20 of 24 method–context combinations; within-group and outside-group augmentations decrease it in 15 and 18 cases, respectively, and outside-group augmentation performs worse than within-group augmentation in 18 comparisons. For both metrics, differences between regimes become more pronounced with increasing context size, as generally weaker performance at small context sizes can partially mask the effect of task heterogeneity. Overall, these results demonstrate that increasing meta-training task heterogeneity can impair meta-learning performance, despite the increase in available training data.

Accounting for assay relatedness benefits few-shot meta-learning Across all heterogeneity regimes, `MetaHeta`, which explicitly accounts for task relatedness, achieves the best performance on both metrics at every context size. Notably, it retains this advantage even in a deliberately unfavorable cross-regime comparison: when trained under the most heterogeneous regime ($\mathcal{D}_{\text{meta}}^{\text{core}}$ aug-

Table 2: Performance comparison on FS-Mol. `MetaHeta` ranks best across seven of eight metric–context-size combinations, and ranks second on the remaining one.

	Context size = 4		Context size = 8		Context size = 16		Context size = 32	
Method	NLL ↓	R^2_{os} ↑	NLL ↓	R^2_{os} ↑	NLL ↓	R^2_{os} ↑	NLL ↓	R^2_{os} ↑
TNP	7.406 ± 2.519	-0.070 ± 0.070	2.381 ± 0.216	-0.048 ± 0.037	1.537 ± 0.026	0.153 ± 0.010	1.303 ± 0.028	0.199 ± 0.017
ICICL-TNP	2.291 ± 0.112	0.069 ± 0.038	1.786 ± 0.022	0.117 ± 0.020	1.485 ± 0.007	0.155 ± 0.020	1.362 ± 0.044	0.131 ± 0.026
ANP	8.020 ± 4.795	-0.048 ± 0.047	2.718 ± 0.265	-0.071 ± 0.045	1.697 ± 0.075	0.056 ± 0.044	1.418 ± 0.098	0.131 ± 0.057
CNP	5.945 ± 2.766	-0.176 ± 0.043	2.399 ± 0.229	-0.013 ± 0.010	1.715 ± 0.028	0.063 ± 0.003	1.472 ± 0.029	0.120 ± 0.008
ADKF-IFT	22.605 ± 9.400	0.060 ± 0.001	4.509 ± 0.374	0.110 ± 0.005	2.235 ± 0.029	0.173 ± 0.003	1.582 ± 0.052	$\mathbf{0.265 \pm 0.009}$
DKT	37.880 ± 18.236	0.022 ± 0.007	6.682 ± 0.564	0.068 ± 0.024	3.075 ± 0.066	0.148 ± 0.016	2.059 ± 0.070	0.224 ± 0.025
TabICLv2	-	-0.011 ± 0.006	-	-0.008 ± 0.003	-	-0.018 ± 0.003	-	0.010 ± 0.005
TabPFNv3	-	0.026 ± 0.003	-	0.037 ± 0.003	-	0.073 ± 0.003	-	0.111 ± 0.001
MAML	-	-0.214 ± 0.054	-	-0.022 ± 0.025	-	0.050 ± 0.013	-	0.065 ± 0.007
ST-MAML	-	-0.048 ± 0.005	-	0.043 ± 0.005	-	0.061 ± 0.003	-	0.062 ± 0.006
Frequent Hitters	-	0.086 ± 0.007	-	0.095 ± 0.014	-	0.123 ± 0.010	-	0.131 ± 0.008
Similarity Search	-	0.012 ± 0.001	-	0.029 ± 0.001	-	0.041 ± 0.000	-	0.050 ± 0.000
RF-$\mathcal{D}^{\text{aux}}$	-	-0.228 ± 0.009	-	-0.196 ± 0.006	-	-0.194 ± 0.009	-	-0.213 ± 0.023
`MetaHeta`	$\mathbf{2.216 \pm 0.050}$	$\mathbf{0.129 \pm 0.028}$	$\mathbf{1.765 \pm 0.037}$	$\mathbf{0.140 \pm 0.029}$	$\mathbf{1.466 \pm 0.020}$	$\mathbf{0.187 \pm 0.018}$	$\mathbf{1.277 \pm 0.010}$	0.228 ± 0.011
`MetaHeta` ($\gamma = 0.7$)	2.325 ± 0.091	0.127 ± 0.026	1.778 ± 0.039	0.129 ± 0.026	1.476 ± 0.024	0.179 ± 0.016	$\mathbf{1.277 \pm 0.012}$	0.223 ± 0.010
`MetaHeta` ($\gamma = 0.8$)	2.469 ± 0.124	0.119 ± 0.024	1.808 ± 0.036	0.112 ± 0.025	1.497 ± 0.029	0.170 ± 0.016	1.295 ± 0.025	0.218 ± 0.009
`MetaHeta` ($\gamma = 0.9$)	4.423 ± 1.938	0.098 ± 0.032	2.090 ± 0.163	0.074 ± 0.025	1.585 ± 0.051	0.151 ± 0.019	1.422 ± 0.076	0.204 ± 0.013
`MetaHeta`-R	2.438 ± 0.112	-0.256 ± 0.164	1.959 ± 0.030	-0.089 ± 0.076	1.606 ± 0.084	0.099 ± 0.045	1.363 ± 0.048	0.185 ± 0.013

mented with non-TK tasks), it still outperforms baselines trained under the most favorable regime ($\mathcal{D}^{\text{core}}_{\text{meta}}$ augmented with additional EPHA tasks). Even at the smallest context size $|\mathcal{D}^C| = 4$, `MetaHeta` already approaches oracle performance in terms of F_{NLL}. This suggests that conditioning on data from related assays provides a valuable signal beyond the task-specific context, yielding informative predictive distributions even with very few observations from the target assay.

Auxiliary information transfers across homologous targets To examine cross-target transfer of auxiliary information, we retain the preceding setup with meta-training on $\mathcal{D}^{\text{core}}_{\text{meta}}$, but perform validation and testing on tasks from the EPHB kinase subfamily. Since these targets are not represented in $\mathcal{D}^{\text{core}}_{\text{meta}}$, no same-target assays are available. However, because the EPHA and EPHB subfamilies both belong to the EPH kinase family, we can use Proposition 1 to construct $\mathcal{D}^{\text{aux}}$ from the homologous EPHA tasks present in $\mathcal{D}^{\text{core}}_{\text{meta}}$ (see Appendix A.4.1 for full details). The results in Table 1 demonstrate that `MetaHeta` matches or exceeds all baselines across both metrics and all context sizes, with particularly large gains in the lowest-context regime. This suggests that information from homologous targets can support prediction, particularly when task-specific data are limited.

5.2 Meta-learning evaluation on real-world bioactivity assays

To assess whether the findings from the controlled assay-heterogeneity setting extend beyond the semi-synthetic benchmark, we evaluate `MetaHeta` on FS-Mol (Stanley et al., 2021), a few-shot learning benchmark built from ChEMBL (Gaulton et al., 2012) bioactivity assays.

Experimental setup As in Section 5.1, we consider a highly data-scarce regime with $|\mathcal{D}^C| \in \{4, 8, 16, 32\}$. We construct $\mathcal{D}^{\text{aux}}$ based on assay-description similarity with threshold $\gamma = 0.5$, selecting at most 5 auxiliary datasets per task and subsampling each to at most 512 observations. We compare `MetaHeta` against previously introduced meta-learning baselines, together with ST-MAML (Wang et al., 2022), a MAML variant designed to account for task heterogeneity, and ICICL-TNP (Ashman et al., 2024), evaluated with the same $\mathcal{D}^{\text{aux}}$ as our method. We additionally include two state-of-the-art tabular foundation models, TabICLv2 (Qu et al., 2026) and TabPFNv3 (Grinsztajn et al., 2026), as well as three non-meta-learning sanity checks: Frequent Hitters and Similarity Search (Schimunek et al., 2023), and RF-$\mathcal{D}^{\text{aux}}$, a random forest fitted per target assay after combining its context with the observations in $\mathcal{D}^{\text{aux}}$. Since the underlying data-generating functions are unavailable, we report the standard R^2_{os} and NLL where applicable. Details are in Appendix A.4.

Results Across the four context sizes and both metrics, our method achieves the best performance in 7 out of 8 comparisons. The gains are the largest in the most data-scarce regime and diminish as more task-specific context becomes available. Consistent with the observations of Schimunek et al. (2023), Frequent Hitters, which ignores the task context and instead learns from observations naively pooled across all assays in $\mathcal{D}_{\text{meta}}$, outperforms several meta-learning approaches. This suggests that standard meta-learners may struggle to exploit highly heterogeneous assay collections, particularly when only limited context data are available. In contrast, our approach is the only method that out-

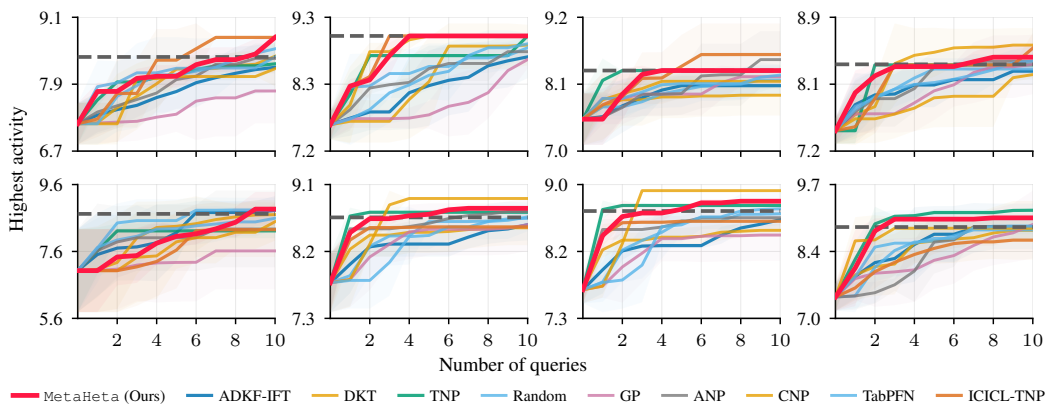


Figure 4: Retrospective Bayesian optimization trajectories on eight BindingDB assays. Dashed horizontal lines (- - -) mark the activity threshold for the top 5% of compounds in each assay; MetaHeta is the only method to identify such compounds in every assay.

performs Frequent Hitters across all context sizes. Both tabular foundation models perform poorly, particularly at the smallest context sizes, providing further evidence that broad generic priors can be insufficient when limited context is available to identify the relevant task regime. MetaHeta consistently outperforms ICICL-TNP, whose R^2_{os} declines as the context size increases from 16 to 32 while ours improves, potentially reflecting the benefit of retaining exact self-attention over the context rather than processing it through pseudo-tokens. Additionally, the poor performance of RF-$\mathcal{D}^{\text{aux}}$ suggests that access to related auxiliary data alone is insufficient for effective prediction.

5.3 Retrospective Bayesian optimization

To assess the practical utility of MetaHeta, we investigate whether its improved predictive performance in low-data settings translates into more effective compound prioritization when used as a surrogate model in retrospective Bayesian optimization on BindingDB assays (Liu et al., 2025).

Experimental setup We conduct the evaluation with initial conditions designed to resemble an early-stage compound-optimization scenario, in which activity measurements are available only for compounds belonging to a single, weakly active chemical series. Specifically, for each of the eight considered assays and each random seed, we initialize optimization with four randomly selected compounds sharing a Bemis–Murcko scaffold (Bemis & Murcko, 1996) with median activity below the assay-specific 25th percentile. We then perform 10 optimization rounds using the LogEI acquisition function (Ament et al., 2023). We compare MetaHeta against most of the probabilistic baselines introduced in Section 5.2, a task-specific GP with a Tanimoto kernel, and random selection. Further details on assay selection and candidate-pool construction are provided in Appendix A.2.

Results As shown in Figure 4, MetaHeta finds the most active compounds among the compared methods in three of the eight assays and ranks second in three others. It is also the only approach to retrieve compounds that rank among the 5% most active in each of the eight assays. Table 4 in Appendix B.2 further demonstrates the strong and consistent performance of MetaHeta across all the considered assays: our approach achieves an average rank of 1.875, leaving a substantial gap to the second-ranked method, TabPFN, with an average rank of 4.0. Notably, the task-specific GP ranks last in four of the eight assays, suggesting that fitting a surrogate independently from assay-specific observations alone can be unreliable in such a low-data regime.

5.4 Ablation

In this section, we examine: (i) how the trade-off between predictive performance and inference time of MetaHeta depends on the number of auxiliary datasets and the choice of architecture; (ii) its sensitivity to the similarity threshold γ; and (iii) whether assay-description similarity provides a useful proxy for task relatedness.

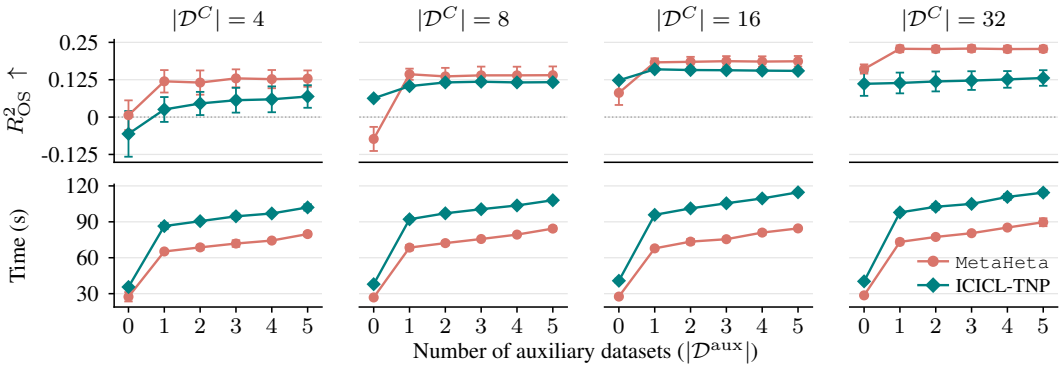


Figure 5: R^2_{os} and inference time of MetaHeta and ICICL-TNP with varying cardinality of $\mathcal{D}^{\text{aux}}$. Auxiliary conditioning improves predictive performance with modest incremental cost. Our architecture processes five auxiliary datasets faster than ICICL-TNP processes a single one.

Experimental setup We use the models trained on FS-Mol (Section 5.2). To answer (i), we compare predictive performance and inference time of MetaHeta and ICICL-TNP as we vary $|\mathcal{D}^{\text{aux}}|$ at inference time, keeping $\gamma = 0$ and using the same auxiliary datasets for both methods. For (ii), we evaluate stricter similarity thresholds by increasing γ while retaining a maximum of five auxiliary datasets. Lastly, to address (iii), we evaluate MetaHeta with the default $\gamma = 0.5$ and a maximum of five auxiliary datasets, but shuffle the assignment of retrieved auxiliary assays to target assays. We denote this variant as MetaHeta-R.

Results Figure 5 shows that conditioning on a single auxiliary dataset suffices to substantially improve R^2_{os}, with little further change as more datasets are added. Figure 7 in Appendix B.3 shows continued reductions in NLL, suggesting that additional datasets mainly refine uncertainty estimates rather than improve point predictions. Although conditioning on a single auxiliary dataset increases inference time by approximately $2.3 - 2.5$ times, each additional dataset adds only around three seconds, resulting in 20% overhead when expanding $\mathcal{D}^{\text{aux}}$ from 1 to 5. We attribute this modest incremental cost to FAVOR+ attention, which allows our architecture to condition on five auxiliary datasets faster than ICICL-TNP does on a single one. It also makes more effective use of the auxiliary signal, achieving higher mean R^2_{os} across all context sizes and nonzero auxiliary-dataset counts. Table 2 shows that increasing γ degrades performance, with the largest deterioration at $\gamma = 0.9$, where many target assays lack even a single auxiliary assay satisfying the similarity criterion. Notably, shuffling auxiliary-assay assignments in MetaHeta-R leads to comparatively modest increases in NLL but substantially degrades R^2_{os}, with the sharpest drops at small context sizes. This provides strong evidence that assay-description similarity offers a useful basis for constructing $\mathcal{D}^{\text{aux}}$ by capturing relatedness relevant to prediction, particularly when task-specific data are scarce.

6 Conclusion

In this work, we first investigate the effect of meta-training task heterogeneity on few-shot bioactivity prediction in a controlled setting, and show that increasing heterogeneity can systematically degrade meta-learning performance across methods and context sizes. To address this, we introduce MetaHeta, a meta-learning framework which explicitly accounts for assay heterogeneity by conditioning predictions on auxiliary data from related assays. Its architecture enables this auxiliary conditioning to scale efficiently to large datasets while preserving exact attention over scarce task-specific context. We further demonstrate the effectiveness of our method under three notions of assay relatedness: shared target identity, phylogenetic relationships between targets, and assay-description similarity. Finally, we demonstrate the effectiveness of MetaHeta on bioactivity assays from ChEMBL and BindingDB, improving both few-shot prediction and downstream compound prioritization through retrospective Bayesian optimization. We discuss the limitations of our approach and directions for future work in Appendix E. Taken together, our results identify assay heterogeneity as an important challenge for few-shot bioactivity meta-learning and show that conditioning on related assays provides an effective and scalable way to address it.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the writing and presentation of the manuscript, including the writing and presentation of proofs. We also used generative AI tools to provide feedback on experimental design, to assist with the implementation of baseline methods, and to generate and edit code used for data cleaning, preprocessing, and visualization. For the experiments in Section 5.3, we used generative AI tools to generate assay descriptions by summarizing the corresponding patent documents. Full details, including the prompt and the resulting assay descriptions, are provided in Appendix A.2. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving those claims, propose or refine hypotheses, interpret experimental results, assist with translation, or support qualitative and thematic data analysis. All AI-assisted outputs were reviewed by the authors. AI-generated or edited code was checked for correctness, proof-related assistance was verified, and generated assay descriptions were checked against the corresponding patent documents to ensure that they did not introduce unsupported information. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release the code used for our experiments at `https://anonymous.4open.science/r/metaheta`. The repository also contains the semi-synthetic data generated for the controlled-heterogeneity experiments in Section 5.1. Appendix A provides detailed experimental specifications, including the construction of the controlled-heterogeneity study (Appendix A.1), data preprocessing and the retrospective Bayesian optimization setup (Appendix A.2), and implementation and training details (Appendix A.4). In particular, Appendix A.4.1 provides implementation details for `MetaHeta`, while Appendix A.4.2 describes the implementation of the baseline methods.

REFERENCES

Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with One-Shot learning. *ACS Cent Sci*, 3(4):283–293, April 2017.

Sebastian Ament, Sam Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1vyAG6j9PE.

Matthew Ashman. *On Transformer Neural Processes*. PhD thesis, Apollo - University of Cambridge Repository, 2024. URL https://www.repository.cam.ac.uk/handle/1810/399768.

Matthew Ashman, Cristiana Diaconu, Adrian Weller, and Richard E. Turner. In-context in-context learning with transformer neural processes. In Javier Antorán and Christian A. Naesseth (eds.), *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, volume 253 of *Proceedings of Machine Learning Research*, pp. 1–29. PMLR, 21 Jul 2024. URL https://proceedings.mlr.press/v253/ashman24a.html.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, March 2000. ISSN 1076-9757. doi: 10.1613/jair.731. URL http://dx.doi.org/10.1613/jair.731.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 07 1996. ISSN 0022-2623. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

Jannis Born, Tien Huynh, Astrid Stroobants, Wendy D Cornell, and Matteo Manica. Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3D effects in a 1D model. *J. Chem. Inf. Model.*, 62(2):240–257, January 2022.

Lucian Chan, Marcel Verdonk, and Carl Poelking. Embracing assay heterogeneity with neural processes for markedly improved bioactivity predictions, 2023. URL https://arxiv.org/abs/2308.09086.

Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KXRSh0sdVTP.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

David H Drewry, Carrow I Wells, David M Andrews, Richard Angell, Hassan Al-Ali, Alison D Axtman, Stephen J Capuzzi, Jonathan M Elkins, Peter Ettmayer, Mathias Frederiksen, Opher Gileadi, Nathanael Gray, Alice Hooper, Stefan Knapp, Stefan Laufer, Ulrich Luecking, Michael Michaelides, Susanne Müller, Eugene Muratov, R Aldrin Denny, Kumar S Saikatendu, Daniel K Treiber, William J Zuercher, and Timothy M Willson. Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PLoS One*, 12(8):e0181585, August 2017.

Leo Feng, Hossein Hajimirsadeghi, Yoshua Bengio, and Mohamed Osama Ahmed. Latent bottlenecked attentive neural processes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yIxtevizEA.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/garnelo18a.html.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes, 2018b. URL https://arxiv.org/abs/1807.01622.

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 (Database issue):D1100–7, January 2012.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), October 2021. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Karim K. Ben Hicham, Jan G. Rittig, Martin Grohe, and Alexander Mitsos. Tabular foundation models for in-context prediction of molecular properties, 2026. URL https://arxiv.org/abs/2604.16123.

Sepp Hochreiter, A. Steven Younger, and Peter R. Conwell. Learning to learn using gradient descent. In Georg Dorffner, Horst Bischof, and Kurt Hornik (eds.), *Artificial Neural Networks — ICANN 2001*, pp. 87–94, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44668-2.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, Jan 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Adam Izdebski, Jan Olszewski, Pankhil Gawade, Krzysztof Koras, Serra Korkmaz, Valentin Rauscher, Jakub M. Tomczak, and Ewa Szczurek. Synergistic benefits of joint molecule generation and property prediction. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=jnzCOLyGOA. J2C Certification.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Ghassen Jerfel, Erin Grant, Tom Griffiths, and Katherine Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7a9a322cbe0d06a98667fdc5160dc6f8-Paper.pdf.

José Jiménez-Luna, Francesca Grisoni, Nils Weskamp, and Gisbert Schneider. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery*, 16(9):949–959, 2021. doi: 10.1080/17460441.2021.1909567. URL https://doi.org/10.1080/17460441.2021.1909567. PMID: 33779453.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkE6PjC9KX.

Greg Landrum, Paolo Tosco, Ricardo Rodriguez, Brian Kelley, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, Dan Nealschneider, Eisuke Kawashima, NadineSchneider, tadhurst cdd, Andrew Dalke, Niels Maeder, Matt Swain, Yakov Pechersky, Brian Cole, Kevin Boyd, Samo Turk, Aleksandr Savelev, Rachel Walker, Alain Vaucher, Maciej Wójcikowski, Hussein Faara, Ichiru Take, Vincent F. Scalfani, Steven Kearnes, Kazuya Ujihara, and Daniel Probst. rdkit/rdkit: 2026_03_6 (q1 2026) release, August 2026. URL https://doi.org/10.5281/zenodo.22140358.

Gregory A. Landrum and Sereina Riniker. Combining ic50 or ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5):1560–1567, Mar 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.4c00049. URL https://doi.org/10.1021/acs.jcim.4c00049.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp.

3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Tiqing Liu, Linda Hwang, Stephen K Burley, Carmen I Nitsche, Christopher Southan, W Patrick Walters, and Michael K Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1):D1633–D1644, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1075. URL https://doi.org/10.1093/nar/gkae1075.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science*, 9(24):5441–5451, 06 2018. ISSN 2041-6520. doi: 10.1039/c8sc00148k. URL https://doi.org/10.1039/c8sc00148k.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25660–25676. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nagler23a.html.

Goran Nakerst, John Brennan, Wouter Beugeling, and Masudul Haque. Tabular foundation models for non-tabular tasks, 2026. URL https://arxiv.org/abs/2608.22594.

Cuong Q. Nguyen, Constantine Kreatsoulas, and Kim M. Branson. Meta-learning gnn initializations for low-resource molecular property prediction, 2020. URL https://arxiv.org/abs/2003.05996.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16569–16594. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nguyen22b.html.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. URL https://arxiv.org/abs/1803.02999.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026.

Massimiliano Patacchiola, Jack Turner, Elliot J. Crowley, Michael O' Boyle, and Amos Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16108–16118. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf.

Pavel Polishchuk. Crem: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, 12(1):28, Apr 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00431-w. URL https://doi.org/10.1186/s13321-020-00431-w.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Tommy Rochussen and Vincent Fortuin. Amortising inference and meta-learning priors in neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KG6SSTz2GJ.

Gabriele Scalia, Steven T. Rutherford, Ziqing Lu, Kerry R. Buchholz, Nicholas Skelton, Kangway Chuang, Nathaniel Diamant, Jan-Christian Hütter, Jerome-Maxim Luescher, Anh Miu, Jeff Blaney, Leo Gendelev, Elizabeth Skippington, Greg Zynda, Nia Dickson, Michał Koziarski, Yoshua Bengio, Aviv Regev, Man-Wah Tan, and Tommaso Biancalani. Deep-learning-based virtual screening of antibacterial compounds. *Nature Biotechnology*, 44(7):1154–1167, Jul 2026. ISSN 1546-1696. doi: 10.1038/s41587-025-02814-6. URL https://doi.org/10.1038/s41587-025-02814-6.

Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XrMWUuEevr.

Jurgen Schmidhuber. Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May 1987. URL http://www.idsia.ch/~juergen/diploma.html.

Karlie R Sharma, Christine M Colvis, Griffin P Rodgers, and Douglas M Sheeley. Illuminating the druggable genome: Pathways to progress. *Drug Discov Today*, 29(3):103805, October 2023.

Robert P Sheridan, Kiyean Nam, Vladimir N Maiorov, Daniel R McMasters, and Wendy D Cornell. QSAR models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *J. Chem. Inf. Model.*, 49(8):1974–1985, August 2009.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Megan Stanley, John F Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-mol: A few-shot learning dataset of molecules. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=701FtuyLlAd.

Tingzhong Tian, Shuya Li, Ziting Zhang, Lin Chen, Ziheng Zou, Dan Zhao, and Jianyang Zeng. Benchmarking compound activity prediction for real-world drug discovery applications. *Communications Chemistry*, 7(1):127, Jun 2024. ISSN 2399-3669. doi: 10.1038/s42004-024-01204-4. URL https://doi.org/10.1038/s42004-024-01204-4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Daniel Vella and Jean-Paul Ebejer. Few-Shot learning for Low-Data drug discovery. *J Chem Inf Model*, 63(1):27–42, November 2022.

Anna Vettoruzzo, Mohamed-Rafik Bouguelia, and Thorsteinn Rögnvaldsson. Multimodal meta-learning through meta-learned task representations. *Neural Computing and Applications*, 36(15): 8519–8529, May 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-09540-1. URL https://doi.org/10.1007/s00521-024-09540-1.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3637–3645, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e4da3b7fbbce2345d7772b0674a318d5-Paper.pdf.

Liang Wang, Qiang Liu, Shaozhen Liu, Xin Sun, Shu Wu, and Liang Wang. Pin-tuning: Parameter-efficient in-context tuning for few-shot molecular property prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=859DtlwnAD.

Qi Wang and Herke van Hoof. Learning expressive meta-representations with mixture of expert neural processes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ju38DG3sbg6.

Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17441–17454. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/91bc333f6967019ac47b49ca0f2fa757-Paper.pdf.

Zhe Wang, Jake Grigsby, Arshdeep Sekhon, and Yanjun Qi. St-maml : A stochastic-task based method for task-heterogeneous meta-learning. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2066–2074. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/wang22c.html.

Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7045–7054. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yao19b.html.

Pan Zhou, Yingtian Zou, Xiao-Tong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. Task similarity aware meta learning: theory-inspired improvement on MAML. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 23–33. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/zhou21a.html.

A EXPERIMENTAL DETAILS

A.1 CONTROLLED-HETEROGENEITY BENCHMARK CONSTRUCTION

Data preparation We construct the semi-synthetic kinase benchmark from the PKIS2 profiling data of Drewry et al. (2017), comprising 645 compounds from 86 chemotypes screened against 392 wild-type human kinases. We first restrict the panel to the kinases considered in our benchmark and preprocess the molecular structures with RDKit (Landrum et al., 2026) by selecting the parent fragment, neutralizing charges, and canonicalizing tautomers. Activity values are transformed from the unit interval to logit space after clipping extreme values. Molecules are represented using 1024-bit Morgan fingerprints with radius 2.

Kinase-specific GP fitting For each kinase, we standardize the transformed activity values and fit an exact Gaussian process with a constant mean and a scaled Tanimoto kernel over Morgan fingerprints. This yields a separate posterior distribution over activity functions for each kinase. We assess each kinase-specific GP using leave-one-chemotype-out cross-validation. In each fold, one chemotype is held out, the GP is refitted on the remaining compounds, and predictive performance is compared against a baseline that predicts the mean activity of the training fold. We retain only kinases with positive balanced leave-one-chemotype-out.

Molecule generation To obtain a sufficiently large molecular domain for synthetic task generation, we expand the original kinase-panel chemistry using CReM (Polishchuk, 2020) mutation and growth operations. Generated molecules are filtered to satisfy basic drug-like property constraints, requiring molecular weight between 150 and 650, $\log P$ between -1 and 7, topological polar surface area below 180, at most 6 hydrogen-bond donors, 12 hydrogen-bond acceptors, and 15 rotatable bonds. We further retain only molecules whose maximum Tanimoto similarity to any molecule in the original panel lies between 0.40 and 0.60, yielding approximately 10^6 candidate molecules. These molecules are then partitioned into training, validation, and test molecular domains using an $80/10/10$ Bemis–Murcko scaffold split, with scaffolds kept disjoint across the three partitions.

Task construction Each synthetic assay is generated by first sampling its size from a PERT distribution with minimum 40, mode 100, maximum 1024, and concentration 20. We then sample the corresponding number of molecules from the relevant molecular split and select a kinase-specific GP. Activities are generated by drawing jointly from the GP posterior over the selected molecules and adding Gaussian observation noise, with the fitted observation-noise standard deviation scaled by 0.5. Using this procedure, we first generate a core meta-dataset of 5000 tasks from GPs corresponding to EPHA kinases. We then generate an additional 5000 tasks for each heterogeneity regime while keeping the sampled task sizes and molecular inputs fixed across regimes. Only the kinase-specific GP used to generate the activity values is varied, with source GPs corresponding to (i) EPHA kinases, (ii) non-EPHA tyrosine kinases, or (iii) non-tyrosine kinases. We additionally generate 250 validation and 250 test assays from EPHA kinase GPs. For the homologous-target transfer experiment, we generate 250 validation and 250 test assays from EPHB kinase GPs, while keeping training restricted to EPHA tasks.

Kinase relatedness across regimes To quantify the increasing relatedness gap across the three kinase regimes, we consider two measures: similarity of kinase activity profiles across the PKIS2 compound panel and sequence divergence at the 29 active-site residues identified by Sheridan et al. (2009), with the resulting values reported in Table 3.

Table 3: Quantitative characterization of kinase relatedness across heterogeneity regimes. ρ_{act}: average pairwise Pearson correlation between kinase activity profiles over the PKIS2 compound panel; d_{site}: average pairwise normalized Hamming distance across the 29 active-site residues identified by Sheridan et al. (2009).

Regime	ρ_{act}	d_{site}
EPHA	0.466	0.149
TK	0.316	0.367
EPHA + non-TK	0.181	0.512

A.2 Retrospective Bayesian optimization

Data preparation We curate assays from BindingDB (Liu et al., 2025) for retrospective Bayesian optimization experiments. Molecular structures are first processed with RDKit (Landrum et al., 2026) using the same procedure as described in Appendix A.1. We restrict the data to measurements against human, single-chain protein targets and remove target names associated with more than one UniProt (Consortium, 2025) identifier or more than one target-chain sequence, ensuring an unambiguous mapping between target name and protein identity. We further retain only BindingDB entries for which all measurements associated with a given DOI correspond to a single target. We consider affinity measurements reported as K_i, IC_{50}, or K_d, and retain only observations for which exactly one of these three endpoints is available, discarding observations with either no usable measurement or multiple reported endpoints. The remaining measurements are converted to a common p-activity scale as $Y = 9 - \log_{10}(x)$, where x is the reported affinity in nM. Finally, compounds with molecular weight above 900 Da are removed.

Assay selection We select assays that are sufficiently large and chemically diverse to support retrospective compound optimization while retaining a non-degenerate activity landscape. Specifically, we require each assay to contain between 500 and 2000 compounds and at least 30 Bemis–Murcko scaffolds (Bemis & Murcko, 1996) represented by at least five compounds each. We further exclude assays for which any single activity value accounts for more than 25% of the measurements. Finally, since the optimization experiments are commonly initialized from a single weakly active chemical series, we require each assay to contain at least five eligible starting scaffolds, defined as Bemis–Murcko scaffolds represented by at least five compounds whose median activity lies at or below the assay-specific 25th percentile. This ensures that each retained assay admits multiple chemically distinct low-activity starting series from which to simulate early-stage compound optimization. Since BindingDB contains records originating from both ChEMBL and the patent literature, we restrict assay selection to those derived from the latter to avoid potential overlap with the ChEMBL-based FS-Mol training data. We conduct the experiments on eight assays satisfying all the above criteria with the following identifiers: `10.7270/Q2348QXR`, `10.7270/Q2N018M8`, `10.7270/Q2S46W6V`, `10.7270/Q2FN152J`, `10.7270/Q2BZ68G7`, `10.7270/Q2B27XNR`, `10.7270/Q2TB1B01`, and `10.7270/Q2PV6PFT`.

Assay descriptions Assay descriptions were generated using GPT-5.6 Sol (OpenAI, 2026) by parsing the patent document corresponding to each of the eight BindingDB entries with the following prompt:

> Given the provided patent, write a single concise paragraph describing the experimental assay used to measure compound bioactivity. Include, when stated, the biological target and relevant construct or species, assay type and biological context (e.g., biochemical, cellular, binding, functional), key reagents or substrates, experimental conditions such as concentrations, incubation times, temperature, buffer or medium, the measurement technology or equipment, and the measured endpoint or readout. Describe only information explicitly supported by the source and do not infer missing experimental details. Do not report or summarize compound-specific activity values, potency measurements, active/inactive classifications, hit rates, structure–activity relationships, rankings, or conclusions about which com-

pounds performed well or poorly. The description should characterize how the assay was performed and what it measured, independently of the observed compound outcomes. Keep the output to one short paragraph.

The resulting descriptions used in the experiment were:

- `10.7270/Q2348QXR`: A biochemical ATPase assay measured ATP hydrolysis by recombinant human DHX9 in the presence of an annealed double-stranded RNA substrate. The 10-μL reactions used DHX9 delivered at 1.25 nM in 40 mM HEPES, pH 7.5, 0.01% Tween-20, 0.01% BSA, 1 mM DTT, 5 mM $MgCl_2$, and RNase inhibitor and were initiated with 30 nM double-stranded RNA and 10 μM ATP. After one hour at room temperature, ADP-Glo reagent was added for 40 minutes, followed by kinase-detection reagent for another 40 minutes. Luminescence was recorded on an EnVision plate reader, providing an ADP-production readout of DHX9 ATPase activity.
- `10.7270/Q2N018M8`: Endothelial and hepatic lipase activity was measured biochemically using the fluorescent FRET substrate A10070 incorporated into DMPG vesicles. Endothelial lipase was supplied in conditioned medium from HT-1080 cells engineered to overexpress endogenous EL, while hepatic lipase was partially purified from conditioned medium of HL-expressing COS cells. Enzyme was preincubated with compound for 30 minutes at 37°C in 50 mM HEPES, pH 8.0, 50 mM NaCl, and 1 mM $CaCl_2$, after which diluted substrate vesicles initiated the reaction. Fluorescence was recorded for 60 minutes on a Gemini plate reader at 490-nm excitation and 530-nm emission, and the slope between 300 and 900 seconds represented the reaction rate.
- `10.7270/Q2S46W6V`: A fluorescence-polarization competition assay measured binding of recombinant eIF4E to the fluorescent cap analogue EDA-m^7GDP-ATTO-550. Final reactions contained 75 nM eIF4E, 20 nM fluorescent ligand, and 1% DMSO in 20 mM Tris, pH 7.5, 100 mM NaCl, 1 mM DTT, and 0.02% Tween-20. Protein and ligand were preincubated for five minutes before compound addition, and 100-μL reactions in black half-area 96-well plates were equilibrated for 30 minutes at room temperature while protected from light. A PerkinElmer Victor 2 multi-label counter measured fluorescence polarization as the competitive-binding readout.
- `10.7270/Q2FN152J`: An isolated-enzyme HTRF kinase assay measured phosphorylation by the recombinant His-tagged cytoplasmic domain of TrkA/NTRK1, residues 441–796, using the Cisbio KinEASE-TK kit. The assay used 4 nM TrkA, 40 μM ATP, and kit buffer supplemented with 10 mM $MgCl_2$ and 2 mM DTT; enzyme and ATP were preactivated for 20 minutes at room temperature, exposed to compound for 20 minutes, and then incubated with 1 μM biotinylated tyrosine-kinase substrate for 35 minutes. Detection buffer containing streptavidin-XL665 and a cryptate-labelled phosphotyrosine antibody stopped the reaction, followed by 60 minutes at room temperature and EnVision measurement of the 665/620-nm emission ratio.
- `10.7270/Q2BZ68G7`: A biochemical radiometric methyltransferase assay measured incorporation of tritium from ^{3}H-SAM into GST-tagged MEKK2 by purified full-length human SMYD3. Compounds and 0.4 nM SMYD3 were preincubated for 30 minutes at room temperature in 25 mM Tris-Cl, pH 8.0, 1 mM TCEP, 0.005% bovine skin gelatin, and 0.005% Tween-20; 8 nM ^{3}H-SAM and 12 nM MEKK2 then initiated the reaction in 2% DMSO. Non-radiolabelled SAM stopped incorporation, and radiolabelled MEKK2 was precipitated onto scintillation-proximity beads in 0.5 M citric acid. A PerkinElmer TopCount reader measured incorporated radioactivity as disintegrations or counts per minute.
- `10.7270/Q2B27XNR`: A biochemical time-resolved fluorescence assay measured ERK2-catalysed phosphorylation of a truncated ATF2 substrate fused to GFP. Reactions contained 0.25 nM ERK2, 0.4 mM ATF2-GFP, 20 μM ATP, 50 mM Tris at pH 7.5, 10 mM $MgCl_2$, 1 mM EGTA, 0.01% Triton X-100, 1 mM DTT, and 2.5% DMSO and proceeded for 30 minutes at room temperature. Reactions were stopped with TR-FRET dilution buffer containing 25 mM EDTA and 2 nM terbium-labelled anti-phospho-ATF2 (Thr71), incubated for at least another 30 minutes, and read on a Pherastar at 340-nm excitation and 520/495-nm emission; the emission ratio was the phosphorylation readout.
- `10.7270/Q2TB1B01`: A biochemical time-resolved fluorescence assay measured ERK2 phosphorylation of truncated ATF2-GFP. Reactions contained 0.25 nM ERK2, 0.4 μM

ATF2-GFP, 20 μM ATP, 50 mM Tris at pH 7.5, 10 mM "MgC_2" as printed in the patent, 1 mM EGTA, 0.01% Triton X-100, 1 mM DTT, and 2.5% DMSO and were incubated for 30 minutes at room temperature. TR-FRET dilution buffer, 25 mM EDTA, and 2 nM terbium-labelled anti-phospho-ATF2 (Thr71) stopped the reaction; after at least 30 additional minutes, a Pherastar measured fluorescence at 520 and 495 nm following 340-nm excitation, and the emission ratio represented phosphorylation.

- `10.7270/Q2PV6PFT`: A low-ATP biochemical kinase assay used recombinant full-length N-terminally His-tagged human TBK1 expressed in insect cells and a biotinylated peptide substrate, biotin-Ahx-GDEDFSSFAEPG-amide. TBK1 and compound were preincubated for 15 minutes at 22°C in HEPES buffer containing $MgCl_2$, DTT, BSA, Nonidet-P40, and protease inhibitors; phosphorylation then proceeded for 30 minutes at 22°C with 10 μM ATP and 1 μM peptide. EDTA stopped the reaction, and streptavidin-XL665, an anti-phosphoserine antibody, and europium-labelled anti-mouse IgG were added for one hour. A Pherastar or Viewlux TR-FRET reader measured 620- and 665-nm emissions after 350-nm excitation, with their ratio representing phosphorylated substrate.

A.3 Metrics

We evaluate all methods using negative log-likelihood (NLL), where applicable, and the out-of-sample coefficient of determination R^2_{os}, following Chen et al. (2023). For a task with context set $\mathcal{D}^C = (X^C, Y^C)$ and target set $\mathcal{D}^T = (X^T, Y^T)$, we compute

$$R^2_{\text{os}} = 1 - \frac{\sum_{n=1}^{N_t} \left(y_n^T - \hat{y}_n^T\right)^2}{\sum_{n=1}^{N_t} \left(y_n^T - \bar{y}^C\right)^2}, \qquad \bar{y}^C = \frac{1}{N_c} \sum_{n=1}^{N_c} y_n^C,$$

where $\hat{y}_n^T$ denotes the predictive mean at target input x_n^T.

Since the synthetic assays in Section 5.1 are sampled from known kinase-specific GP posteriors, we have access to their corresponding data-generating functions. This allows us to compute an exact GP oracle by conditioning the generating GP on the same context set $\mathcal{D}^C$ available to each method. The oracle uses the same observation-noise model as the synthetic data-generation process. We report performance relative to this oracle. We further define:

$$F_{R^2_{\text{os}}} = \frac{R^2_{\text{os}}}{R^{2,\text{oracle}}_{\text{os}}}. \tag{6}$$

For NLL, we instead compare the corresponding average predictive densities:

$$F_{\text{NLL}} = \frac{\exp(-\text{NLL})}{\exp(-\text{NLL}_{\text{oracle}})} = \exp(\text{NLL}_{\text{oracle}} - \text{NLL})\,. \tag{7}$$

Since $\exp(-\text{NLL})$ is the geometric mean predictive density, F_{NLL} has a direct interpretation as the ratio between the geometric mean predictive density of a method and that of the oracle. In both cases, a value of 1 corresponds to oracle-level performance.

A.4 Implementation details

We evaluate all methods at context sizes $|\mathcal{D}^C| \in \{4, 8, 16, 32\}$. We largely follow Stanley et al. (2021), but deviate in how context size is handled during training and validation. Specifically, for each context size, we train and validate a separate model using that same fixed context size, and evaluate it only at the corresponding context size. Given the highly data-scarce regime considered here, this ensures that every method is optimized for the same task-specific data budget on which it is evaluated, and avoids conflating few-shot performance with generalization across context sizes. This also places amortized methods, which can naturally accommodate varying context sizes, on the same footing as prior FS-Mol evaluations in which methods are trained using a fixed support-set size (Stanley et al., 2021; Chen et al., 2023). For each evaluation assay, we sample 10 context–target splits and average evaluation metrics across splits before averaging across assays. For the controlled-heterogeneity experiments, all methods are given the same molecular representation used by the task-generating GPs, namely 1024-bit radius-2 Morgan fingerprints. This ensures that differences in performance relative to the GP oracle cannot be attributed to a mismatch in molecular

representation. For the experiments on FS-Mol, we do not enforce a single representation across methods: we align representations where appropriate, while retaining method-specific feature extractors when they are part of the original approach or using alternative representations when these perform better. This is because imposing a single shared representation may disadvantage particular baselines whose performance is strongly tied to the choice of molecular representation. For the retrospective Bayesian optimization experiments, we train models with the context size sampled uniformly from $\{4, \ldots, 32\}$. During validation, we evaluate each model on three independently sampled context–target splits at each context size $|\mathcal{D}^C| \in \{4, 8, 16, 32\}$, and average performance across all splits and context sizes.

A.4.1 METAHETA

Controlled-heterogeneity experiment Our model is trained with meta-batch size 32 for 10 epochs using AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4}, weight decay 0.01 and a cosine learning-rate schedule with 10% warm-up. We evaluate on the validation set after each epoch and retain the checkpoint with the lowest validation NLL. We use at most 5 auxiliary datasets per task and subsample each auxiliary dataset to 512 observations. The input embedder consists of 2 layers with hidden dimension 128, and the output decoder consists of a single layer with dimension 64. The main network contains 6 attention layers with embedding dimension 128, using 8 attention heads with head dimension 16, and a dropout rate of 0.25.

FS-Mol experiment For FS-Mol, we use a meta-batch size of 16 and train for 100 epochs with learning rate 3×10^{-4}. We evaluate on the validation set after each epoch and retain the checkpoint with the lowest validation NLL. The architecture is increased to an embedding and feed-forward dimension of 256, decoder dimension 128, 8 attention layers, and head dimension 32, while retaining 8 attention heads, a 2-layer input embedder, and a 1-layer decoder. We use at most 5 auxiliary datasets per task and subsample each to at most 512 observations, with cosine-similarity threshold $\gamma = 0.5$. For molecular representations, we use Hyformer embeddings (Izdebski et al., 2026).

Auxiliary dataset construction The construction of $\mathcal{D}^{\text{aux}}$ in `MetaHeta` is adapted to the notion of assay relatedness available in each experimental setting:

- Controlled-heterogeneity experiment. We construct $\mathcal{D}^{\text{aux}}$ according to Eq. 5, selecting at most five same-target datasets, ordered by decreasing average 5-nearest-neighbor Tanimoto similarity to the current task. The similarity is computed exclusively from molecular inputs; target outputs are never used in auxiliary-dataset selection. Auxiliary datasets containing more than 512 observations are randomly subsampled to that size.
- Homologous-target transfer experiment. For the EPHB evaluation, where same-target auxiliary assays are unavailable, we instead construct $\mathcal{D}^{\text{aux}}$ according to Proposition 1. For each such EPHB assay, we sample five auxiliary tasks from the pool of assays targeting any member of the EPHA subfamily and order them by decreasing activity variance. Each auxiliary dataset is again subsampled to at most 512 observations.
- FS-Mol & retrospective Bayesian optimization. We construct $\mathcal{D}^{\text{aux}}$ using assay-description similarity, as described in Algorithm 1, retaining at most five auxiliary datasets per task and subsampling each to at most 512 observations.

A.4.2 BASELINES

Implementations of the baseline methods were based primarily on the code released by Chen et al. (2023), Stanley et al. (2021), and implementations of several Neural Process variants available at `https://github.com/cambridge-mlg/tnp`, released under the MIT License.

- Transformer Neural Process (TNP) (Nguyen & Grover, 2022). We use the same architecture and training setup as for our model. In particular, setting the number of auxiliary datasets to zero reduces our model to the standard TNP-D (using the efficient implementation described by Feng et al. (2023)).
- Conditional Neural Process (CNP) (Garnelo et al., 2018a). For FS-Mol, we use mean aggregation and a 256-dimensional context representation, with a decoder hidden dimension

of 256 and with Hyformer embeddings as molecular representations. For the controlled-heterogeneity experiment we reduce both the context representation and decoder hidden dimensions to 128.

- Attentive Neural Process (ANP) (Kim et al., 2019). We match the TNP baseline in architectural parameters, including the embedding, feed-forward, decoder, and attention dimensions, as well as the number of attention heads and layers, while omitting the latent path.
- Adaptive Deep Kernel Fitting with Implicit Function Theorem (ADKF-IFT) (Chen et al., 2023). For FS-Mol, we retain the feature extractor used in the original implementation, combining ECFP fingerprints with a GNN-MT from Stanley et al. (2021). For the controlled-heterogeneity experiment, we replace the GNN with a two-layer MLP with hidden and output dimensions of 1024, operating on the same 1024-bit fingerprints used by the GP oracle.
- Deep Kernel Transfer (DKT) (Patacchiola et al., 2020). We use the implementation provided by Chen et al. (2023) and follow the same experimental setup and feature-extractor choices as for ADKF-IFT.
- Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017). We use the implementation from Stanley et al. (2021) based on GNN-ST. For the controlled-heterogeneity experiment, we replace the GNN with the same two-layer MLP used for ADKF-IFT and DKT. The regression head for the kinase experiments is an MLP with hidden dimensions 256 and 64.
- ST-MAML (Wang et al., 2022). We preserve the MAML setup, adding a stochastic task encoder with hidden dimension 128, a 32-dimensional latent representation, and a 32-dimensional task feature. We weight the KL regularization term by 0.1 and average predictions over five context-conditioned latent samples at evaluation.
- Frequent Hitters (Schimunek et al., 2023). Frequent Hitters is simply an MLP trained on data pooled across all training assays, without using the support set of the target assay. Following a hyperparameter sweep, we use four hidden layers of width 16, batch size 256, learning rate 10^{-3}, no dropout or weight decay, and train for up to 100 epochs with an early-stopping patience of 10 epochs. Molecular inputs are represented using ECFP fingerprints.
- Similarity Search. As an additional sanity-check baseline proposed by Schimunek et al. (2023), we use Morgan fingerprints with radius 2 and 1024 bits, and predict each target as the Tanimoto-similarity-weighted average of the context labels.
- TabPFNv3 (Grinsztajn et al., 2026). We use Morgan fingerprints with radius 2 and 1024 bits as molecular representations, which we found to perform better than Hyformer embeddings.
- TabICLv2 (Qu et al., 2026). We use the same molecular representation as for TabPFNv3.
- ICICL-TNP (Ashman et al., 2024). Across all settings, we always construct $\mathcal{D}^{\text{aux}}$ in the same way as for `MetaHeta`, such that both models condition on exactly the same auxiliary data. As in `MetaHeta`, we share the projection parameters between the corresponding MHSA and MHCA blocks of the context and auxiliary branches. We do not include dataset-of-origin one-hot indicators, as auxiliary datasets are represented separately through their pseudo-token sets. We use 64 pseudo-tokens for each auxiliary dataset and set the number of context pseudo-tokens equal to the context size, $|\mathcal{D}^C|$.
- RF-$\mathcal{D}^{\text{aux}}$. We construct $\mathcal{D}^{\text{aux}}$ in the same way as for `MetaHeta` and then concatenate its datasets with the task-specific context set. A random forest with 250 trees is then fit to that combined dataset.

A.5 Computational resources

A single training run of `MetaHeta` was performed on an NVIDIA H100 GPU and required approximately 24 hours. The runtime ablation in Section 5.4 was performed on an NVIDIA Tesla V100S 32 GB GPU with an Intel Xeon Gold 6226R CPU @ 2.90 GHz.

B Extended results

B.1 Controlled-heterogeneity experiment

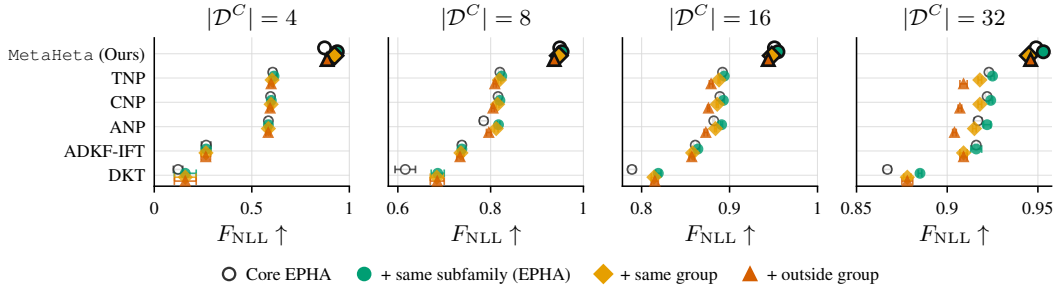


Figure 6: Effect of meta-training task heterogeneity on few-shot bioactivity prediction. Performance of all the considered methods generally deteriorates as $\mathcal{D}^{\text{core}}_{\text{meta}}$ is augmented with progressively less related tasks. `MetaHeta` remains the best-performing approach regardless of the training regime.

B.2 Retrospective Bayesian optimization

Table 4: Average rank across retrospective Bayesian optimization tasks. For each assay, methods are ranked according to their final best-observed activity averaged over five random seeds; ranks are then averaged across the eight assays.

Method	Average rank
TNP	4.375
ANP	5.375
CNP	5.375
ADKF-IFT	7.250
DKT	6.625
TabPFN	4.000
ICICL-TNP	5.000
Random	6.375
GP	7.750
`MetaHeta`	**1.875**

B.3 Ablation

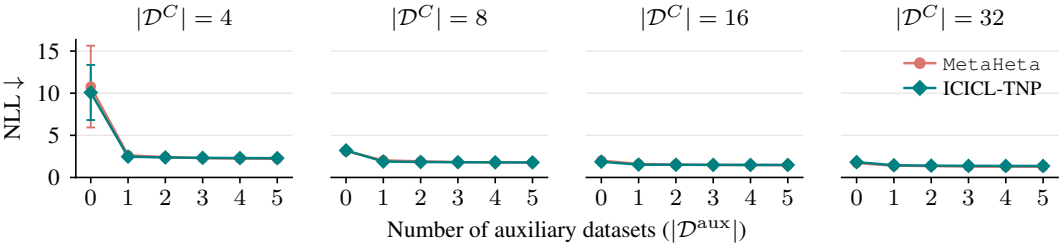


Figure 7: NLL of `MetaHeta` and ICICL-TNP with varying cardinality of $\mathcal{D}^{\text{aux}}$.

C ALGORITHMS

Algorithm 1 outlines the core meta-training procedure, with $\mathcal{D}^{\text{aux}}$ constructed based on assay-description similarity. For clarity, auxiliary-dataset limits, ordering, and subsampling are omitted; full experimental details are provided in Appendix A.

Algorithm 1: Description-based auxiliary dataset selection and meta-training

Input: Meta-dataset $\mathcal{D}_{\text{meta}} = \{\mathcal{D}_\tau\}_{\tau=1}^{M}$, assay descriptions $\{a_\tau\}_{\tau=1}^{M}$, meta-learner p_θ, embedding model ϕ, similarity threshold γ

Compute assay-description embeddings:
$h_\tau \leftarrow \phi(a_\tau) \qquad \forall \tau \in \{1, \dots, M\}$

Compute pairwise assay similarities:
$s_{\tau,\tau'} \leftarrow \frac{\langle h_\tau, h_{\tau'} \rangle}{\|h_\tau\| \, \|h_{\tau'}\|} \qquad \forall \tau, \tau' \in \{1, \dots, M\}$

repeat

  Sample a task τ from $\mathcal{D}_{\text{meta}}$:
  $\left((X^T, Y^T), \mathcal{D}^C\right) \sim \mathcal{D}_{\text{meta}}$

  Initialize auxiliary data:
  $\mathcal{D}^{\text{aux}} \leftarrow \emptyset$

  foreach $\tau' \in \{1, \dots, M\} \setminus \{\tau\}$ **do**

    if $s_{\tau,\tau'} \geq \gamma$ **then**

      $\mathcal{D}^{\text{aux}} \leftarrow \mathcal{D}^{\text{aux}} \cup \{\mathcal{D}_{\tau'}\}$

  Compute the negative log-likelihood:
  $\mathcal{L}_{\text{NLL}} \leftarrow -\log p_\theta \left(Y^T \mid X^T, \mathcal{D}^C, \mathcal{D}^{\text{aux}}\right)$

  Update θ:
  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{NLL}}$

until *convergence*

D EXTENDED RELATED WORK

Task heterogeneity in meta-learning Task heterogeneity in meta-learning has been addressed by prior work through two broad strategies. One line of work explicitly models multiple task regimes using mixtures of meta-learners or hierarchical task structures, with individual components specializing to subsets of related tasks (Wang & van Hoof, 2022; Yao et al., 2019; Jerfel et al., 2019; Zhou et al., 2021; Chan et al., 2023). These approaches typically rely on a prespecified number of task clusters or mixture components, thereby imposing some prior structure on the task distribution. This is difficult to justify for bioactivity prediction, where assay heterogeneity arises from multiple interacting factors and may not admit a small, disjoint set of task regimes. In large assay collections the relevant regimes may be numerous and overlapping, making a global mixture cardinality difficult to specify. Among these approaches, Jerfel et al. (2019) is an exception, using a Dirichlet process mixture that can introduce additional components as it encounters new task regimes, but still assumes a discrete mixture structure over tasks. A complementary line of work, more directly comparable to our setting, instead infers task-specific representations from the available context set and uses them to condition or modulate the meta-learner, thereby adapting to heterogeneous task distributions without explicitly enumerating the underlying task regimes (Wang et al., 2022; Vettoruzzo et al., 2024; Vuorio et al., 2019). Notably, Wang et al. (2022) explicitly address the difficulty of inferring task identity from scarce context by representing it as a stochastic latent variable, which captures uncertainty over the underlying task regime.

Comparison with ICICL-TNP (Ashman et al., 2024) The architecture of `MetaHeta` is related to ICICL-TNP, but differs primarily in how information from multiple auxiliary datasets is represented and exchanged. Both architectures use coupled TNP-style processing paths for the task-specific context and auxiliary data, while avoiding quadratic attention over the potentially large auxiliary datasets. ICICL-TNP achieves this using a Perceiver-style (Jaegle et al., 2021) pseudo-token architecture: (i) the task-specific context set and each auxiliary dataset are separately en-

coded into their own sets of pseudo-tokens; (ii) the auxiliary pseudo-tokens attend to the context pseudo-tokens; (iii) self-attention is applied separately within each set of pseudo-tokens; and (iv) the context pseudo-tokens attend back to the auxiliary pseudo-tokens before conditioning the target representations. Consequently, observations from different auxiliary datasets exchange information only indirectly. `MetaHeta` concatenates auxiliary observations into a single sequence and applies linear self-attention across them, enabling direct information exchange between observations from different auxiliary datasets. This is made possible by augmenting each observation with a one-hot indicator of its task of origin, which preserves task identity after concatenation, while the use of FAVOR+ (Choromanski et al., 2021) in the auxiliary branch keeps self-attention computationally efficient even for large auxiliary sets. This removes the pseudo-token bottleneck and, in our setting, enables substantially more efficient conditioning on large auxiliary collections. However, as a result of representing task identity explicitly through one-hot indicators, the input embedding dimension fixes the maximum number of auxiliary datasets that can be accommodated without modifying the architecture. ICICL-TNP does not have this restriction, since task identity is represented structurally through separate pseudo-token sets.

E Discussion

E.1 Limitations

Scope of the theoretical guarantees Both the result of Ashman et al. (2024) in Eq. 5 and Proposition 1, which motivate our selection of auxiliary datasets, are derived for exact Bayesian posterior predictive distributions under their respective generative assumptions. They therefore do not directly guarantee improvements for finite-capacity models trained on a finite meta-dataset. Likewise, description-based relatedness is only a heuristic for identifying useful auxiliary assays. However, we demonstrate that auxiliary conditioning remains effective across all three settings considered in this work: when relatedness is defined by shared target identity, by homology between distinct targets, and heuristically from assay descriptions. In the latter setting, randomly shuffling the retrieved auxiliary assays (Section 5.4) substantially degrades predictive performance, indicating that the observed gains depend on meaningful assay relatedness.

Fixed auxiliary-dataset capacity Because `MetaHeta` encodes dataset identity using one-hot indicators, the maximum number of auxiliary datasets is fixed by the input dimensionality and cannot be increased after training without modifying the model, as elaborated in Appendix D. In practice, however, we found this limitation to be mild: the model can simply be trained with a conservative upper bound and evaluated with fewer auxiliary datasets. Our ablation in Section 5.4 demonstrates that performance remains strong across varying inference-time auxiliary-set cardinalities.

E.2 Future work

Proposition 1 considers a two-level hierarchy in which protein-specific components share a common family-level component. In practice, biological relatedness can exhibit richer structure. A natural extension would therefore be to incorporate multiple levels of relatedness, allowing auxiliary information to be shared across increasingly broad groups of related targets. Beyond hierarchical structure, future work could also incorporate additional data modalities, such as protein sequence or structure and richer assay metadata, to define relatedness more comprehensively.

F Attention operators

MHSA and MHCA (Vaswani et al., 2017) Let $Z \in \mathbb{R}^{N \times d_z}$ and $Z' \in \mathbb{R}^{M \times d_z}$ denote query and conditioning representations, respectively, where d_z is the embedding dimension. Let H denote the number of attention heads, and let d_k and d_v denote the query/key and value dimensions within each head. Layer indices are omitted for clarity.

For each head $h = 1, \dots, H$, the query, key, and value matrices are defined as

$$Q_h = ZW_{Q,h}, \qquad K_h = Z'W_{K,h}, \qquad V_h = Z'W_{V,h}, \tag{8}$$

where $W_{Q,h}, W_{K,h} \in \mathbb{R}^{d_z \times d_k}$ and $W_{V,h} \in \mathbb{R}^{d_z \times d_v}$. Multi-head cross-attention is defined as

$$A_h(Z, Z') = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right) V_h, \tag{9}$$

$$\text{MHCA}(Z, Z') = \text{Concat}\big(A_1(Z, Z'), \ldots, A_H(Z, Z')\big) W_O, \tag{10}$$

where softmax is applied rowwise, concatenation is along the feature dimension, and $W_O \in \mathbb{R}^{Hd_v \times d_z}$. Multi-head self-attention is defined as

$$\text{MHSA}(Z) = \text{MHCA}(Z, Z). \tag{11}$$

FAVOR+ (Choromanski et al., 2021) The positive random feature map $\phi : \mathbb{R}^{d_k} \to \mathbb{R}^r_{>0}$, where r is the number of random features, is defined as

$$\phi(u) = \frac{\exp(-\|u\|^2/2)}{\sqrt{r}} \begin{bmatrix} \exp(\omega_1^\top u) \\ \vdots \\ \exp(\omega_r^\top u) \end{bmatrix}, \tag{12}$$

where the projection vectors have marginal distribution $\omega_i \sim \mathcal{N}(0, I_{d_k})$ and are sampled using orthogonal random features, in independent blocks when $r > d_k$. This yields

$$\mathbb{E}\big[\phi(u)^\top \phi(v)\big] = \exp(u^\top v). \tag{13}$$

Writing Φ for the rowwise application of ϕ, the feature matrices are defined as

$$Q'_h = \Phi(d_k^{-1/4} Q_h) \in \mathbb{R}^{N \times r}, \qquad K'_h = \Phi(d_k^{-1/4} K_h) \in \mathbb{R}^{M \times r}. \tag{14}$$

Bidirectional FAVOR+ approximates each attention head by

$$\widehat{A}_h(Z, Z') = D_h^{-1} Q'_h \big((K'_h)^\top V_h\big), \qquad D_h = \text{diag}\left(Q'_h \big((K'_h)^\top \mathbf{1}_M\big)\right), \tag{15}$$

where $\mathbf{1}_M$ is the length-M vector of ones. By FAVOR+(Z), we denote the corresponding multi-head self-attention operation:

$$\text{FAVOR+}(Z) = \text{Concat}\big(\widehat{A}_1(Z, Z), \ldots, \widehat{A}_H(Z, Z)\big) W_O. \tag{16}$$

G Proofs

G.1 Permutation invariance

Proposition 2 (Permutation invariance). *Let $\mathcal{D}^{\text{aux}} = \{\mathcal{D}_j\}_{j=1}^K$, and let $S : \mathcal{D}^K \to \mathcal{D}^K$ be a map satisfying*

$$S(\mathcal{D}_{\rho(1)}, \ldots, \mathcal{D}_{\rho(K)}) = S(\mathcal{D}_1, \ldots, \mathcal{D}_K) \qquad \textit{for every permutation } \rho. \tag{17}$$

With the dataset indicator e_j attached to each observation in the j-th dataset after sorting, predictions are invariant to permutations of datapoints within each $\mathcal{D}_j \in \mathcal{D}^{\text{aux}}$ and to permutations of datasets in $\mathcal{D}^{\text{aux}}$, at inference with fixed FAVOR+ features.

Proof. For fixed random features, $\Phi(PQ) = P\Phi(Q)$ for any permutation matrix P, and

$$(PK')^\top (PV) = K'^\top V, \qquad (PK')^\top \mathbf{1} = K'^\top \mathbf{1}. \tag{18}$$

Thus bidirectional FAVOR+ is permutation equivariant. Permutation-equivariant self-attention and cross-attention invariant to key/value ordering yield predictions invariant to observation order for fixed dataset indicators (Lee et al., 2019). It therefore remains to consider dataset order in $\mathcal{D}^{\text{aux}}$. By assumption, S produces the same canonical ordering under any dataset permutation, preserving each dataset's indicator and hence the predictive distribution. □

G.2 PROPOSITION 1

For convenience, we restate the proposition here.

Proposition 1. *Consider an extension of the data-generating process in Eq. 4 in which:*

$$\xi_{\text{family}} \sim p(\xi), \qquad \xi_{\text{prot}}, \{\xi_{\text{prot}'}\} \sim p(\cdot \mid \xi_{\text{family}}),$$

with $f_\tau \sim P_{\xi_{\text{prot}}}(f)$ and $f_{\tau'} \sim P_{\xi_{\text{prot}'}}(f)$. Assume that auxiliary data from target homologs provide no additional information for predicting y once the target's protein-specific component is known:

$$p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) = p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C).$$

Then:

$$\begin{aligned}
&\mathbb{E}_{\xi_{\text{family}}, \xi_{\text{prot}}, \mathcal{D}^C, \mathcal{D}^{\text{aux}}} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \Big) \right] \\
&\quad \le \mathbb{E}_{\xi_{\text{family}}, \xi_{\text{prot}}, \mathcal{D}^C} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C) \Big) \right].
\end{aligned}$$

Proof. The integrands on both sides do not depend explicitly on ξ_{family}, so this variable can be integrated out:

$$\begin{aligned}
&\mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C, \mathcal{D}^{\text{aux}}} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \Big) \right] \\
&\quad \le \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C} \left[\mathrm{KL}\Big(p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \,\|\, p(y \mid x, \mathcal{D}^C) \Big) \right].
\end{aligned} \tag{19}$$

The rest of the argument follows the proof of Ashman et al. (2024), with all expectations taken under the marginal distributions induced by the hierarchical generative process.

LHS. Expanding the KL divergence gives

$$\begin{aligned}
\text{LHS} = \; &- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C, \mathcal{D}^{\text{aux}}} H[p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C)] \\
&- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C, \mathcal{D}^{\text{aux}}} \int p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \log p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \, dy.
\end{aligned} \tag{20}$$

The first integrand does not depend on $\mathcal{D}^{\text{aux}}$, so this variable also can be integrated out. For the second term, conditional independence and marginalization yield

$$\begin{aligned}
&- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C, \mathcal{D}^{\text{aux}}} \int p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \log p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \, dy \\
&= -\mathbb{E}_{\mathcal{D}^C, \mathcal{D}^{\text{aux}}} \int \mathbb{E}_{\xi_{\text{prot}} \mid \mathcal{D}^C, \mathcal{D}^{\text{aux}}} \left[p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \right] \log p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \, dy \\
&= -\mathbb{E}_{\mathcal{D}^C, \mathcal{D}^{\text{aux}}} \int p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \log p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}}) \, dy \\
&= \mathbb{E}_{\mathcal{D}^C, \mathcal{D}^{\text{aux}}} H[p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}})].
\end{aligned} \tag{21}$$

Hence

$$\begin{aligned}
\text{LHS} = \; &- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C} H[p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C)] \\
&+ \mathbb{E}_{\mathcal{D}^C, \mathcal{D}^{\text{aux}}} H[p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\text{aux}})].
\end{aligned} \tag{22}$$

RHS. Similarly,

$$\begin{aligned}
\text{RHS} = \; &- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C} H[p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C)] \\
&- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C} \int p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \log p(y \mid x, \mathcal{D}^C) \, dy.
\end{aligned} \tag{23}$$

Marginalizing ξ_{prot} in the second term gives

$$\begin{aligned}
&- \mathbb{E}_{\xi_{\text{prot}}, \mathcal{D}^C} \int p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \log p(y \mid x, \mathcal{D}^C) \, dy \\
&= -\mathbb{E}_{\mathcal{D}^C} \int \mathbb{E}_{\xi_{\text{prot}} \mid \mathcal{D}^C} \left[p_{\xi_{\text{prot}}}(y \mid x, \mathcal{D}^C) \right] \log p(y \mid x, \mathcal{D}^C) \, dy \\
&= -\mathbb{E}_{\mathcal{D}^C} \int p(y \mid x, \mathcal{D}^C) \log p(y \mid x, \mathcal{D}^C) \, dy \\
&= \mathbb{E}_{\mathcal{D}^C} H[p(y \mid x, \mathcal{D}^C)].
\end{aligned} \tag{24}$$

Thus

$$\begin{aligned}\mathrm{RHS} = & -\mathbb{E}_{\xi_{\mathrm{prot}},\mathcal{D}^C} H[p_{\xi_{\mathrm{prot}}}(y \mid x, \mathcal{D}^C)] \\ & + \mathbb{E}_{\mathcal{D}^C} H[p(y \mid x, \mathcal{D}^C)].\end{aligned} \tag{25}$$

Finally, conditioning does not increase entropy in expectation:

$$\mathbb{E}_{\mathcal{D}^C,\mathcal{D}^{\mathrm{aux}}} H[p(y \mid x, \mathcal{D}^C, \mathcal{D}^{\mathrm{aux}})] \leq \mathbb{E}_{\mathcal{D}^C} H[p(y \mid x, \mathcal{D}^C)]. \tag{26}$$

The remaining terms are identical, so $\mathrm{LHS} \leq \mathrm{RHS}$. □