Target-Free Compound Activity Prediction via Few-Shot Learning

Peter Eckmann [1]   Jake Anderson [2]   Michael K. Gilson [2 3]   Rose Yu [1]

Abstract

Predicting the activities of compounds against protein-based or phenotypic assays using only a few known compounds and their activities is a common task in target-free drug discovery. Existing few-shot learning approaches are limited to predicting binary labels (active/inactive). However, in real-world drug discovery, degrees of compound activity are highly relevant. We study *Few-Shot Compound Activity Prediction* (FS-CAP) and design a novel neural architecture to meta-learn continuous compound activities across large bioactivity datasets. Our model aggregates encodings generated from the known compounds and their activities to capture assay information. We also introduce a separate encoder for the unknown compound. We show that FS-CAP surpasses traditional similarity-based techniques as well as other state of the art few-shot learning methods on a variety of target-free drug discovery settings and datasets.

1. Introduction

A key task in machine learning for drug discovery is to predict the activity of compounds against a target-based or phenotypic assay, reducing the need for expensive lab-based experimental tests (Paul et al., 2021; Vamathevan et al., 2019). Most existing methods (Öztürk et al., 2018; Somnath et al., 2021; Ragoza et al., 2017; Stepniewska-Dziubinska et al., 2018; Jones et al., 2021) require information about the target protein, such as amino acid sequence or 3D structure. However, such information is not always available due to experimental difficulties or a lack of mechanistic disease understanding. Indeed, there is increasing interest in target-free drug discovery (Haasen et al., 2017; Swin-

ney & Lee, 2020) where only a few compounds with weak activity in an experimental assay are known (Loew et al., 1993; Acharya et al., 2011). These hit compounds, while not drug candidates themselves, offer a starting point for the discovery of more promising compounds. Traditional methods use chemical similarity, such as the Tanimoto similarity between structural compound fingerprints (Bajusz et al., 2015), to find new compounds most similar to the hit compounds. However, these compounds are often similarly undesirable as drug candidates, based on the principle that structurally similar compounds have similar properties (Johnson & Maggiora, 1990).

We cast the problem of target-free compound activity prediction as few-shot learning (Wang et al., 2020), a framework that enables a trained model to generalize to new domains (in this case, assays). Few-shot learning is usually investigated for multi-class classification problems. For drug discovery, these techniques have been applied for binary compound activity prediction (Vella & Ebejer, 2022; Altae-Tran et al., 2017). However, since experimental activity readouts are often continuous (Chandrasekaran et al., 2021), formulation as a binary classification problem requires ad-hoc activity thresholding and is overly simplistic. The few-shot regression problem studied here is more relevant for drug discovery applications (Joo et al., 2019; Lenhof et al., 2022; Lee et al., 2022), although it is significantly more challenging (Stanley et al., 2021).

In this paper, we propose *Few-Shot Compound Activity Prediction* (FS-CAP), a model-based few-shot learning approach for target-free compound activity regression. Our model bears some similarity to neural processes (NPs, Garnelo et al. (2018a)) but with several important differences that are relevant for compound activity prediction. Specifically, we use a deterministic neural encoder to represent context compounds and their activities via a new multiplication-based featurization. We also introduce a separate encoder for the unknown compound to represent its assay-independent binding characteristics. We concatenate these two encodings and feed them to a predictor network to produce a final prediction for the activity of the unknown compound, and train the entire model using mean squared error (MSE).

Despite the rich literature on few-shot classification, few-shot regression remains largely under-explored in drug dis-

[1]Department of Computer Science and Engineering, UC San Diego, La Jolla, California, United States [2]Department of Chemistry and Biochemistry, UC San Diego, La Jolla, California, United States [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, California, United States. Correspondence to: Michael Gilson <mgilson@health.ucsd.edu>, Rose Yu <roseyu@ucsd.edu>.

covery. To the best of our knowledge, only Lee et al. (2022) have explored few-shot regression in drug discovery, by applying an Attentive Neural Process (ANP, Kim et al. (2019)), a variant of neural processes, to the task. However, many design choices in ANPs are not tailored to compound activity prediction, including their probabilistic framing and lack of unknown compound encoding, which we will show leads to poor performance. Lee et al. (2022) also perform very limited comparison with other few-shot learning methods, and only measure model performance on a single dataset.

In summary, our contributions include

- designing a novel few-shot learning model, FS-CAP, that builds on existing neural process designs but with several important architectural and training changes that are specific to drug prediction considerations,

- introducing several new datasets and settings to the problem of few-shot compound activity regression that mimic the drug discovery challenges of hit and lead optimization, high-throughput screening, and anti-cancer drug activity prediction, and

- showing that FS-CAP outperforms both traditional chemical similarity techniques and modern deep learning-based few-shot learning techniques on this robust set of datasets.

2. Related Work

We discuss the related work in compound activity prediction and then summarize few-shot learning and its applications to target-free compound activity prediction.

Compound activity prediction. Much work focuses on the prediction of compound activities using knowledge of a protein target (e.g. Öztürk et al. (2018); Somnath et al. (2021); Ragoza et al. (2017); Stepniewska-Dziubinska et al. (2018); Jones et al. (2021)), but such information is not always available in practice (Haasen et al., 2017; Swinney & Lee, 2020). In the target-free, or "ligand-based" setting, our aim is to use existing compounds (the "context set") to predict the activity of unknown compounds (the "query set") against new assays. A common computational chemistry technique for this task is to measure chemical similarity between the context compound(s) and each compound in the query set. This is often performed with binary fingerprints (e.g. Rogers & Hahn (2010)), although such structure-based similarity can miss compounds with similar activity but different chemical scaffolds. Therefore, more complex chemical descriptors may also be used, such as polarity, molecular topology, and 3D shape (Khan et al., 2016; Li et al., 2012; Kohlbacher et al., 2021; Kearnes & Pande, 2016).

Machine learning techniques derive molecular representations in a data-driven fashion and thus promise to improve the quality of similarity measurements that use these representations. Much work focuses on the unsupervised learning of molecular representations that can later be used for downstream tasks such as the assessment of compound similarity (Jaeger et al., 2018; Huang et al., 2021; Li & Jiang, 2021; Morris et al., 2020). Due to their unsupervised nature, however, similarity measurements between the learned embeddings are not necessarily useful for activity prediction.

Few-Shot Learning. Few-shot learning is a framework that enables a trained model to generalize to new domains (Wang et al., 2020). Common techniques include metric-based, optimization-based, and model-based approaches.

Metric-based methods use a learned metric space that is trained specifically to reflect activity differences, as opposed to unsupervised similarity-based methods. Altae-Tran et al. (2017) propose an LSTM-based method to iteratively update context compound embeddings, which are used to compute a similarity metric. Schimunek et al. (2021) learn a Siamese network-like embedding for compounds in a metric space. The well-known prototypical network (Snell et al., 2017) and matching network (Vinyals et al., 2016) techniques have also been proposed for use on molecular graphs (Ding et al., 2020; Vella & Ebejer, 2022). However, these techniques only measure similarity between discrete classes (active/active), and cannot use continuous labels. This is problematic when the difference between weakly and highly active compounds is critical, therefore reducing the real-world applicability of such techniques (Stanley et al., 2021; Lee et al., 2022; Lenhof et al., 2022; Joo et al., 2019). Indeed, one of the main challenges of drug discovery is to optimize weakly active compounds into highly active ones (Hughes et al., 2011), yet binary methods like the ones above can make no such distinction.

Optimization-based techniques use gradients computed on the context set to adapt the weights of a "base" model, and then apply this adapted model to the query set. Techniques in this area include the LSTM meta-learner (Ravi & Larochelle, 2016), which uses a separate "learner" network to adapt the weights of the main network. Nguyen et al. (2020) proposed the use of model-agnostic meta-learning (MAML, Finn et al. (2017)) for few-shot binary compound activity prediction, which finds a set of model parameters that can most quickly be fine-tuned to new tasks.

Instead of updating network weights during test time, model-based approaches take both the query and context set as inputs to a single model. For example, MetaNets (Munkhdalai & Yu, 2017) use a memory module coupled with both a base and meta-learner to generate network weights adapted to a new task. Another method, Non-Gaussian Gaussian Pro-

cesses (NGGPs, Sendera et al. (2021)), expands on previous approaches (Tossou et al., 2019; Rothfuss et al., 2021) that use GPs for few-shot learning by parameterizing the Gaussian posterior with a normalizing flow. However, neither the optimization-based nor the model-based techniques have been applied to few-shot compound activity regression.

Neural processes (NPs, Garnelo et al. (2018b)), as well as their variants like attentive neural processes (ANPs, Kim et al. (2019)), combine GPs and neural networks for few-shot learning. To the best of our knowledge, the prediction of continuous compound activity values in the few-shot setting has been explored only once in the literature using the ANP-based MetaDTA (Lee et al., 2022). However, they include a limited number of experimental settings and baseline comparisons to other few-shot learning models. We propose a novel architecture with some similarity to neural processes but with several important modifications tailored to the prediction of compound activities, and perform a more rigorous comparison across multiple datasets.

3. Methodology

We cast the problem of compound activity prediction in new assays given known compounds as a few-shot regression task. To address this problem, we introduce FS-CAP, which is summarized in Figure 1.

Problem statement. We seek to predict the activity of a "query" compound in a new assay, given only a small set of "context" compounds and their activities in the same assay.

Mathematically, suppose our training dataset consists of K different assays. Each assay k consists of N different compounds that are measured against it, $M_k := \{m_1, \cdots, m_N\}$. The experimentally measured activity of a molecule m against an assay k is defined as $\pi_k(m) \in \mathbb{R}$. In training, we take a query molecule m_q that is an element of some M_k and aim to predict its activity $\pi_k(m_q)$. To aid in prediction, we randomly sample n context examples from the same assay, $C_k = \{(m_i, \pi_k(m_i))\}_{i=1}^n$, where each m_i is randomly sampled from M_k. n must be $\leq N$, and typically it is a small number, hence few-shot. Then, we train the model f to predict the activity of the query molecule given the context set, i.e. $f(m_q, C_k) = \hat{\pi}_k(m_q) \approx \pi_k(m_q)$.

In testing, our model has a similar task, which is to predict the activity of a query molecule given some context set. However, the query and context set come from an assay not seen in training, meaning we measure the ability of the model to adapt its predictions to an unseen assay.

Architecture. We employ two separate encoders, a query encoder f_q and a context encoder f_c. Consider a single assay k. We encode the query molecule $m_q \in M_k$ and

elements of the context set $(m_i, \pi_k(m_i)) \in C_k$ as follows:

$$f_q : m_q \mapsto x_q, \quad f_c : (m_i, \pi_k(m_i))) \mapsto r_i \quad (1)$$

where r_i is a representation of the i-th context example. The query encoder learns to encode the query molecule into a representation x_q, that is useful for predicting its activity. The context encoder learns to capture some information about assay k from each example in the context set. To aggregate each individual context encoding r_i into a single real-valued vector x_c that represents the context set as a whole, we take the average across each r_i:

$$x_c = \frac{1}{n} \sum_{i=1}^{n} r_i. \quad (2)$$

This maintains permutation invariance, as desired, since the order of the contexts should not affect their encoding. More complex aggregation techniques, such as self-attention, did not lead to improved performance (Table 6).

The predictor network g combines both encodings to generate an activity prediction for the query molecule:

$$g : x_c \oplus x_q \mapsto \hat{\pi}_k(m_q) \quad (3)$$

where $\oplus$ denotes vector concatenation.

We represent molecules using their 2048-bit Morgan fingerprints (Rogers & Hahn, 2010). f_c, f_q, and g are all multilayer perceptrons with ReLU activations. To pass both the context compound and its measured activity value to f_c, we multiply the measured activity scalar with the Morgan fingerprint. Specifically, f_c receives the following vector:

$$\text{Morgan}(m_i) \cdot \pi_k(m_i) \quad (4)$$

Since Morgan fingerprints are substructure-based, i.e. each element in the vector has a 1 bit if there is a certain substructure present and 0 otherwise, and substructures are known to contribute directly to binding characteristics, this featurization may make it easier for the model to learn which substructures contribute how much to activity. We later confirm this intuition by comparing our proposed multiplication approach with the more traditional concatenation of fingerprint and activity values (Table 6).

Differences to neural processes. Although our architecture builds on neural processes (NPs, Garnelo et al. (2018b)) and attentive neural processes (ANPs, Kim et al. (2019)) such as MetaDTA (Lee et al., 2022), it differs in several important aspects that are specific to the few-shot compound activity regression task. First, both NPs and ANPs are based upon a probabilistic framework, which would theoretically allow for the prediction of a distribution of possible activities for a given compound. However, such distributions are not very relevant in drug discovery, where one almost
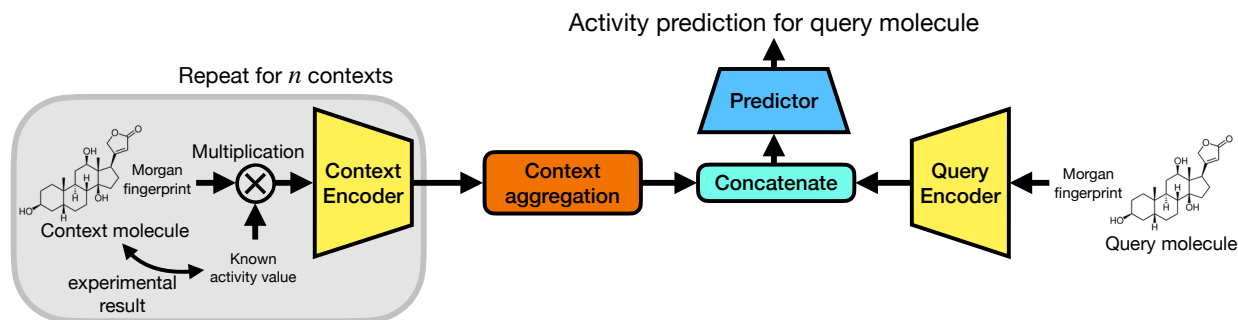
Figure 1. **Overview of the `FS-CAP` architecture.** The context encoder (left) receives the Morgan fingerprint of each context compound multiplied by its associated activity value. A final context encoding is produced by aggregating the individual encodings of each context compound. The query encoder (right), which has different weights, receives the Morgan fingerprint of the query compound. A predictor network receives the concatenated outputs of each encoder and produces a final scalar activity prediction of the query compound.

always works with point estimates of compound activity except perhaps in the special case of compound toxicity (Lazic & Williams, 2021). Avoiding a probabilistic framework stabilizes training, and allows us to simply minimize the mean squared error loss.

Second, NPs and ANPs do not perform query encoding, meaning the query features are fed directly along with the context embedding to the predictor network. However, in drug discovery, there are useful query features that may be extracted entirely independently of any assay, such as compound shape and electrostatics. Allowing the model to encode the query compound in a distinct query encoder, prior to receiving any assay information, is a novel step that appears to improve prediction performance over baselines that use no such encoding (Table 6).

Third, instead of concatenating the features of the context compound with its activity value, as in NPs or ANPs, we multiply the two before feeding into the context encoder, as described above. This novel featurization, which is only possible due to the unique binary nature of molecular fingerprints and the scalar nature of the activity value, appears to be more effective than concatenation (Table 6).

Training. We use a large assay dataset for training, but set aside some of these assays for testing. We train the model in an end-to-end fashion with Mean Squared Error (MSE), with the loss for each epoch defined as

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}(\pi_k(m_i) - \hat{\pi}_k(m_i))^2\right) \quad (5)$$

where each $m_i \in M_k$ is a query molecule.

4. Experiments

4.1. Tasks

We use four different datasets (Table 1) to test `FS-CAP` and baseline methods on three tasks related to drug discovery.

- **Hit and lead optimization:** In this scenario, one wishes to use knowledge of a few compounds with modest experimentally determined activities in a binding or phenotypic assay to predict the activities of additional candidate compounds. For this task, we train and test all methods on the BindingDB and PubChem BioAssay (PubChemBA) datasets, which contain continuous activity values across many different assays.

- **High-throughput screening:** In high-throughput screening (HTS), large numbers of compounds are assayed to provide a binary active/inactive label. In this scenario, one again has knowledge of a few compounds with modest activities in an assay of interest, but now the goal is to classify a large number of candidate compounds as active vs. inactive in the assay. Success in this task would provide the ability to use a small amount of data to guide the selection of a compound library for HTS that will have an enhanced fraction of novel actives than a library of randomly chosen compounds. To model this task, we train all methods on the PubChemBA dataset, but treat their outputs as unnormalized probabilities to compute metrics involving binary classification like ROC-AUC. We test all methods on the PubChem High-Throughput Screening (PubChemHTS) dataset, which contains binary active or inactive classifications for compounds in PubChem assays marked as "Screening." This dataset contains an entirely separate set of assays from the continuous ones of PubChemBA.

- **Anti-cancer drug activity prediction:** We explore whether a model trained on the PubChemBA dataset

Table 1. **Summary of datasets.** We report the number of assays in each dataset, the number of these assays excluded from training and used for testing, and the number of unique compounds present across all assays in the dataset. We also report the source and access date of the dataset, if applicable.

DATASET	TOTAL ASSAYS	TEST ASSAYS	UNIQUE COMPOUNDS	SOURCE	DATE
PUBCHEM BIOASSAYS (PUBCHEMBA)	98,593	1,000	1,108,355	WANG ET AL. (2012)	18 DEC. 2022
BINDINGDB	4,807	100	1,013,354	GILSON ET AL. (2016)	1 DEC. 2022
CANCER CELL LINE ENCYCLOPEDIA (CCLE)	275	275	24	BARRETINA ET AL. (2012)	N/A
PUBCHEM HIGH-THROUGHPUT SCREENING (PUBCHEMHTS)	100	100	34,716	WANG ET AL. (2012)	23 DEC. 2022

generalizes to the prediction of compound activity against cancer cell lines. For this task, we use the Cancer Cell Line Encyclopedia (CCLE), which contains IC50 measurements for 24 drugs against 275 patient-derived cancer cell lines. We further probe the biological understanding of the trained models with additional challenges on this dataset involving the generated context encodings.

We defer further dataset and preprocessing details to Appendix A. For all datasets, assay data were expressed as $\log_{10}$ of the activity in nanomolar (nM) units.

4.2. Baselines

As baselines for comparison, we include Tanimoto fingerprint similarity (a widely used traditional technique from computational chemistry) and several state-of-the-art approaches in few-shot learning. We applied both optimization-based (MAML, Finn et al. (2017)) and model-based (MetaNet, Munkhdalai & Yu (2017); ANP, Kim et al. (2019)) methods to the regression of compound activities. We omit similarity-based methods (e.g. Snell et al. (2017); Vinyals et al. (2016)) as they require binarizing the activity data of the context compounds, making for an unfair comparison. Details on the training and implementation of FS-CAP and baselines are reported in Appendix B.

- **Tanimoto similarity.** Traditional molecular structure-based similarity measure based on binary Morgan fingerprints (Rogers & Hahn, 2010; Bajusz et al., 2015). When given multiple context compounds, we use the highest similarity score between each of the contexts and the query.

- **MolBERT + attentive neural process (ANP).** Combines MolBERT, which is a start-of-the-art sequence-based molecular featurizer for property prediction tasks (Li & Jiang, 2021), with an attentive neural process model (Kim et al., 2019) for the few-shot prediction of activity values.

- **Non-Gaussian Gaussian process (NGGP)** (Sendera et al., 2021). Expands on basic Gaussian process techniques for few-shot learning by modeling the posterior distribution with an ODE-based normalizing flow.

- **MetaNet** (Munkhdalai & Yu, 2017). Uses two separate learners, the base learner and the meta-learner which utilizes a memory mechanism, to quickly adapt to new tasks in the few-shot setting via fast parameterization.

- **Model-agnostic meta-learning (MAML)** (Finn et al., 2017). Learns a model that can quickly adapt to a new task by training on a small set of context examples. For this paper, we use a simple multilayer perceptron that takes a Morgan fingerprint as input for the base model.

- **MetaDTA** (Lee et al., 2022). Applies attentive neural processes to the few-shot regression of continuous activity values. We use the MetaDTA(I) variant because its performance is superior to that of the other reported variants.

4.3. Hit and lead optimization

To explore the applicability of few-shot learning methods to the hit and lead optimization settings, we compare FS-CAP with baseline methods on the few-shot prediction of compound activity values against assays in PubChemBA and BindingDB. Compounds with high activity are often not known at the hit stage, so we only sampled context compounds (in both training and testing) that have activity values (i.e. effective concentrations) $> 10\ \mu M$, which is typical of hit compounds (Zhu et al., 2013). Note that a higher effective concentration means lower activity. Following training, we test each method against the assays in the held-out test set. Thus, each test set compound was treated as a query compound, with each query being used with a context set of 1-8 compounds randomly sampled from all compounds against the same assay as the query.

Table 2 reports the mean correlation of the predicted and ground truth activity values across all test-set assays for

Table 2. **Average per-assay correlation.** Mean Pearson's r between predicted and ground-truth compound activity values across all test-set assays in PubChemBA and BindingDB. To mimic a hit/lead optimization task, where compounds with high activity are not known, we only sampled context compounds with $> 10\ \mu M$ activity values. For each method, a separate model was trained on each dataset and for each different number of context compounds. We report the mean $\pm$ one standard deviation from three independent training runs with random seeds for the top three baselines. Due to computational constraints, we report results for all other baselines from one training run.

DATASET	PUBCHEMBA				BINDINGDB			
# CONTEXT COMPOUNDS	1	2	4	8	1	2	4	8
TANIMOTO SIMILARITY	0.01	0.13	0.21	0.27	-0.08	0.05	0.13	0.18
MOLBERT + ANP	-0.01	0.23	0.22	0.17	0.09	0.10	0.10	0.12
NGGP	0.17	0.20	0.25	0.30	0.12	0.17	0.17	0.18
METANET	0.02	0.05	0.06	0.02	0.06	-0.01	0.05	0.09
MAML	$0.39_{\pm0.00}$	$0.39_{\pm0.01}$	$0.39_{\pm0.00}$	$0.41_{\pm0.00}$	$0.34_{\pm0.02}$	$0.35_{\pm0.02}$	$0.36_{\pm0.00}$	$0.36_{\pm0.00}$
METADTA	$0.44_{\pm0.00}$	$0.45_{\pm0.00}$	$0.45_{\pm0.00}$	$0.45_{\pm0.01}$	$0.36_{\pm0.01}$	$0.36_{\pm0.00}$	$0.35_{\pm0.00}$	$0.34_{\pm0.01}$
FS-CAP	$\mathbf{0.48_{\pm0.01}}$	$\mathbf{0.48_{\pm0.01}}$	$\mathbf{0.49_{\pm0.00}}$	$\mathbf{0.49_{\pm0.01}}$	$\mathbf{0.38_{\pm0.01}}$	$\mathbf{0.38_{\pm0.00}}$	$\mathbf{0.38_{\pm0.02}}$	$\mathbf{0.39_{\pm0.01}}$

each method. Pearson's correlation coefficient measures the ability of each method to differentiate between compound activities against the same assay and is a standard metric in the literature (Jones et al., 2021; Wang et al., 2021). Other metrics, such as MSE, may appear favorable even if a method makes the same prediction for all compounds against a given assay.

As shown, FS-CAP consistently outperforms Tanimoto similarity, the de facto standard in medicinal chemistry, as well as deep learning-based few-shot learning baselines, across datasets and for different numbers of context compounds. This suggests that FS-CAP may be useful for hit and lead optimization, as it is the most successful in predicting the activities of unknown compounds using only weakly active context compounds. Similar results were obtained when we performed the same study without any limit on the context compound activities, except that the correlation coefficients were higher by about 0.10 (Appendix C). We used the version of the PubChemBA model trained without any context activity limits for Sections 4.4 and 4.5.

4.4. High-throughput screening

We evaluated the performance of FS-CAP and baseline methods on the few-shot prediction of compound activities in high-throughput screening (HTS) assays from PubChemHTS (Table 3). While the activity data for a given HTS assay are binary compound labels, more detailed confirmatory (dose-response) studies are often available for selected hit compounds, which can provide context compounds with continuous activity values. In this task, we obtained context compounds via separate dose-response assays not in PubChemHTS, but with the same targets as PubChemHTS assays (see Appendix A for details).

For this task, we train all models on PubChemBA. While the models predict a continuous activity value for each query compound, we treat their outputs as unnormalized proba-

Table 3. **Average ROC-AUC and enrichment statistics across all high-throughput screening assays.** ROC-AUC measures the ability of each method to classify compounds as active or inactive. Each percentage value indicates the $k\%$ enrichment. 8 context compounds were used. We report the mean ($\pm$ one standard deviation for ROC-AUC) from three independent training runs on the PubChemBA dataset with random seeds for the top three baselines.

METHOD	ROC-AUC	0.5%	1%	2%
TANIMOTO SIMILARITY	0.51	76%	98%	120%
MOLBERT + ANP	0.51	160%	130%	130%
NGGP	0.49	150%	110%	95%
METANET	0.49	100%	110%	120%
MAML	$0.51_{\pm0.01}$	**210%**	150%	150%
METADTA	$0.55_{\pm0.00}$	160%	150%	140%
FS-CAP	$\mathbf{0.57_{\pm0.00}}$	200%	**190%**	**180%**

bilities (that were inverted, because a low effective concentration corresponds to a high activity), so that classification metrics may be computed from the model output. In other words, we assumed that a high continuous compound activity prediction from the models corresponded to an "Active" classification, and vice-versa. Specifically, we measured performance through ROC-AUC using the ground-truth binary activity labels, a standard metric in the HTS literature (Triballeau et al., 2005). We also measured performance with $k\%$ enrichment, which is the percent increase of actives over the base rate in the top $k\%$ of scored compounds, also a standard metric in the HTS literature (Lopes et al., 2017).

We find that FS-CAP outperforms baselines both in ROC-AUC and in most enrichment measurements (Table 3). This suggests that FS-CAP is more capable of predicting compound activities in screening libraries than baseline methods, and maybe the most effective at raising the hit rate of a library selected from a much larger set of compounds to perform more targeted and cost-effective testing.

4.5. Anti-cancer drug activity prediction

In this task, we train all models on PubChemBA and test them on the prediction of anti-cancer drug activities against patient-derived cancer cell lines in the Cancer Cell Line Encyclopedia (CCLE, Barretina et al. (2012)). Context compounds were randomly sampled from all compounds with activity data against a given cell line, and were used to predict the activities against the same cell line of query compounds not in the context set. We report the mean correlation between predicted and experimentally determined IC50 values for drugs across all cell lines.

As shown in Table 4, FS-CAP is better than the baseline methods at predicting the phenotypic activities of anti-cancer drugs. Although the number of compounds tested in the CCLE is relatively small, the success of FS-CAP in predicting activity values in this dataset, despite being trained only on PubChemBA, suggests that it may learn fundamental relationships between compounds and assays that generalize across datasets.
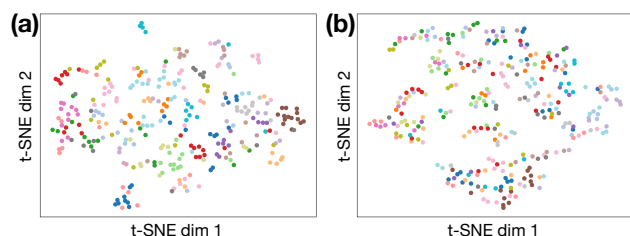

Figure 2. **t-SNE visualization of context encodings, colored by cell line, generated by (a) FS-CAP and (b) MetaDTA.** Each dot represents one context encoding using 8 randomly sampled context compounds and their associated activities against a given cell line. The color of the dot represents the cell line.

We further explored the properties of FS-CAP's trained context encoder using a new classification task. Here, the input was a set of context compounds and their activities against a given cell line, and the output was a prediction of which cell line these activities correspond to. For this task, we applied a simple logistic regression classifier on top of the context encoding generated by FS-CAP (i.e. x_c). For comparison, we apply a similar approach to the latent path prior of MetaDTA (z in Lee et al. (2022)), our most competitive baseline (Table 4).

We randomly selected 20 cell lines in the CCLE. For each of the 20 cell lines and for each number of context compounds, we conducted 15 trials, where each trial consisted of randomly sampling context compounds and their activities against the cell line. As not all compounds have measured activities against all cell lines in the CCLE, we only sampled contexts from the 15 compounds that have experimental activities measured against all 20 cell lines. This prevents

the logistic regression classifier from simply learning which compounds were tested against which cell lines. For training the classifier, we used a random 80/20 train/test split, where 80% of the context encodings and their associated cell lines were used to train the model and the remaining 20% were used to judge its accuracy.

As shown in Table 5, the classifier trained on top of FS-CAP significantly outperforms that of MetaDTA on the test set, suggesting that the context encodings generated by FS-CAP are more meaningful. In addition, Figure 2 shows the t-SNE (Van der Maaten & Hinton, 2008) projections of the context encodings generated by FS-CAP (left panel) and MetaDTA (right panel) using 8 context compounds. The encodings of FS-CAP appear to cluster by cell line (indicated by colors), while the corresponding projections of the MetaDTA encodings appear more scattered, helping to explain the high accuracy of the linear regression classifier trained on FS-CAP. Such clustering signifies that FS-CAP is able to produce similar encodings of context compounds when their associated activities are derived from the same assay, even if the identity of the context compounds themselves vary.

Particularly interesting is that such clustering is observed on the cell line dataset despite having been trained on the nonoverlapping PubChemBA dataset. This suggests that training on large assay datasets allows for the extraction of biologically relevant information on how functional drug responses relate to the unique aspects of various cancer cell lines, e.g. type of cancer or mutations present. Along with Table 5, these results help explain the observed superior performance of FS-CAP for compound activity prediction, as a meaningful encoding of assay information is a necessary first step towards predicting the activity of unknown compounds against that assay.

4.6. Model ablations

We report performance metrics of model ablations to the FS-CAP architecture in Table 6. For each ablation, we trained the model and then measured the mean correlation of the predicted and ground truth activity values across all test-set assays in PubChemBA and BindingDB. This experiment is similar to that presented in Section 4.3, except context compounds are selected at random and not constrained by their activity. 8 context compounds were used for all tests.

We test the significance of using a separate query encoder network ("Base model"), or feeding the query features directly to the predictor network ("No query encoding"), similar to a typical neural process model. The greater performance of the variation with the query encoder suggests that encoding the query independent of assay information is beneficial for prediction.

"Concatenated context" means that we feed the context en-

Table 4. **Average correlation per cell line.** Mean Pearson's r between ground truth and predicted drug activity values across all cell lines in the CCLE. Experiments were performed using 1, 2, 4, and 8 context compounds for each method tested. We report the mean $\pm$ one standard deviation from three independent training runs on the PubChemBA dataset with random seeds for the top three baselines.

# CONTEXT COMPOUNDS	1	2	4	8
TANIMOTO SIMILARITY	0.17	0.28	0.33	0.36
MOLBERT + ANP	0.04	0.11	-0.13	0.07
NGGP	0.12	0.18	0.25	0.32
METANET	-0.25	0.39	0.22	-0.04
MAML	$0.50_{\pm0.05}$	$0.45_{\pm0.04}$	$0.47_{\pm0.04}$	$0.17_{\pm0.03}$
METADTA	$0.52_{\pm0.03}$	$0.49_{\pm0.03}$	$\mathbf{0.51_{\pm0.02}}$	$0.39_{\pm0.03}$
FS-CAP	$\mathbf{0.58_{\pm0.02}}$	$\mathbf{0.56_{\pm0.03}}$	$\mathbf{0.51_{\pm0.03}}$	$\mathbf{0.46_{\pm0.03}}$

Table 5. **Accuracy of cell line identification using context encodings.** Accuracy scores of logistic regression models trained to classify the cell line based on context encodings generated by each method pretrained on PubChemBA. We included 20 randomly chosen cell lines, and performed 15 trials for each cell line and a number of context compounds, where a trial consisted of encoding randomly sampled context compounds and their associated activities. We trained a separate logistic regression classifier for each method and number of context compounds using 80% of the available encodings, and computed the reported accuracy scores on the remaining 20%. A random classifier would have 5% accuracy.

# CONTEXT COMPOUNDS	1	2	4	8
METADTA	5%	8%	10%	27 %
FS-CAP	**24%**	**39%**	**56%**	**81%**

Table 6. **Model ablations**. We measure the mean correlation between ground-truth and predicted activities across all test assays in PubChemBA and BindingDB using 8 context compounds. We report the mean $\pm$ one standard deviation from three independent training runs with random seeds.

ABLATION	PUBCHEMBA	BINDINGDB
BASE MODEL (FS-CAP)	$\mathbf{0.54_{\pm0.01}}$	$\mathbf{0.48_{\pm0.00}}$
NO QUERY ENCODING	$0.53_{\pm0.00}$	$0.46_{\pm0.01}$
CONCATENATED CONTEXT	$0.53_{\pm0.00}$	$0.45_{\pm0.00}$
NO CONTEXT	$0.48_{\pm0.00}$	$0.40_{\pm0.00}$
ATTENTIVE AGGREGATION	$0.51_{\pm0.01}$	$0.30_{\pm0.01}$

coder a binary compound fingerprint concatenated with its associated activity value, instead of multiplying the two. This is similar to a neural process model. This variation shows inferior performance, suggesting that combining the context compound fingerprint and activity value scalar via multiplication is a useful featurization for the activity prediction task. "No context" denotes that no context was fed to the model at all, and it made activity predictions based solely on the query compound. "Attentive aggregation" means that we applied 4-layer self-attention on the individual context encodings before taking the mean.

5. Discussion and Conclusions

The proposed few-shot learning model `FS-CAP` surpasses both a standard chemical similarity metric and prior few-shot learning baselines in multiple tasks of interest in early stage drug discovery. These tasks include prediction of compound activities based on a set of weak-binding context compounds, prediction of screening library compounds as active or inactive, and prediction of antitumor activity in cell-based assays, all performed with models trained on large activity datasets. Together, these results suggest that `FS-CAP` may be broadly useful for target-free, or ligand-based, drug discovery, which has become more common in recent years in comparison to target-based drug discovery that uses protein information (Haasen et al., 2017; Swinney & Lee, 2020).

`FS-CAP` may already be useful in its present form as a tool to leverage the limited compound activity data that is typically available in the earliest stages of drug discovery, focusing attention on candidate compounds that are much more likely than randomly chosen compounds to be active in an assay of interest. It thus offers a novel approach to speed drug discovery and reduce its costs. Exploring the use of `FS-CAP` for other compound properties might open further applications. For example, it may find applications in predicting pharmacokinetic parameters of candidate compounds, such as bioavailability and half-life; metabolic susceptibility; and toxicity.

Limitations of the present implementation of `FS-CAP` include its use of a relatively simple molecular representation (Morgan fingerprints), and a context aggregation technique with limited expressiveness. Additionally, the inherent limitations of training on experimental assay data, such as the limited tested dose range (Stanley et al., 2021) or systematic biases in which compounds are tested against which targets, may limit the applicability of few-shot methods like `FS-CAP` trained on these datasets to real-world drug discovery projects.

Future developments could include the exploration of more complex molecular representations (e.g. sequence or graph-based) and the application of more complex context aggregation methods beyond the mean. Finally, research into incorporating target information, when available, with few-shot methods may allow for increased prediction accuracy beyond using target information or context compounds alone.

References

Acharya, C., Coop, A., E Polli, J., and D MacKerell, A. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Current computer-aided drug design*, 7(1): 10–22, 2011.

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

Arnold, S. M. R., Mahajan, P., Datta, D., Bunner, I., and Zarkias, K. S. learn2learn: A library for Meta-Learning research. *arXiv*, August 2020.

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., and Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, 2021.

Ding, K., Wang, J., Li, J., Shu, K., Liu, C., and Liu, H. Graph prototypical networks for few-shot learning on attributed networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 295–304, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A. Conditional neural processes. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1): D1045–D1053, 2016.

Haasen, D., Schopfer, U., Antczak, C., Guy, C., Fuchs, F., and Selzer, P. How phenotypic screening influenced drug discovery: lessons from five years of practice. *Assay and drug development technologies*, 15(6):239–246, 2017.

Huang, K., Xiao, C., Glass, L. M., and Sun, J. Moltrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Jaeger, S., Fulle, S., and Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

Johnson, M. A. and Maggiora, G. M. *Concepts and applications of molecular similarity*. Wiley, 1990.

Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W. D., Kirshner, D., Wong, S. E., Lightstone, F. C., and Allen, J. E. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of chemical information and modeling*, 61 (4):1583–1592, 2021.

Joo, M., Park, A., Kim, K., Son, W.-J., Lee, H. S., Lim, G., Lee, J., Lee, D. H., An, J., Kim, J. H., et al. A deep learning model for cell growth inhibition ic50 prediction and its application for gastric cancer patients. *International journal of molecular sciences*, 20(24):6276, 2019.

Kearnes, S. and Pande, V. Rocs-derived features for virtual screening. *Journal of computer-aided molecular design*, 30(8):609–617, 2016.

Khan, A. U. et al. Descriptors and their selection methods in qsar analysis: paradigm for drug design. *Drug discovery today*, 21(8):1291–1302, 2016.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Kohlbacher, S. M., Langer, T., and Seidel, T. Qphar: quantitative pharmacophore activity relationship: method and validation. *Journal of cheminformatics*, 13(1):1–14, 2021.

Lazic, S. E. and Williams, D. P. Quantifying sources of uncertainty in drug discovery predictions with probabilistic models. *Artificial Intelligence in the Life Sciences*, 1: 100004, 2021.

Lee, E., Yoo, J., Lee, H., and Hong, S. Metadta: Meta-learning-based drug-target binding affinity prediction. In *ICLR Machine Learning for Drug Discovery Workshop*, 2022.

Lenhof, K., Eckhart, L., Gerstner, N., Kehl, T., and Lenhof, H.-P. Simultaneous regression and classification for drug sensitivity prediction using an advanced random forest method. *Scientific Reports*, 12(1):1–13, 2022.

Li, J. and Jiang, X. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021, 2021.

Li, X., Wang, Z., Liu, H., and Yu, H. Quantitative structure–activity relationship for prediction of the toxicity of phenols on photobacterium phosphoreum. *Bulletin of environmental contamination and toxicology*, 89(1):27–31, 2012.

Loew, G. H., Villar, H. O., and Alkorta, I. Strategies for indirect computer-aided drug design. *Pharmaceutical research*, 10(4):475–486, 1993.

Lopes, J. C. D., Dos Santos, F. M., Martins-José, A., Augustyns, K., and De Winter, H. The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *Journal of Cheminformatics*, 9(1):1–11, 2017.

Morris, P., St. Clair, R., Hahn, W. E., and Barenholtz, E. Predicting binding from screening assays with transformer network embeddings. *Journal of Chemical Information and Modeling*, 60(9):4191–4199, 2020.

Munkhdalai, T. and Yu, H. Meta networks. In *International Conference on Machine Learning*, pp. 2554–2563. PMLR, 2017.

Nguyen, C. Q., Kreatsoulas, C., and Branson, K. M. Meta-learning initializations for low-resource drug discovery. *ChemRxiv*, 2020.

Öztürk, H., Özgür, A., and Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R. K. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1):80, 2021.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. *International Conference on Learning Representations*, 2016.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pp. 9116–9126. PMLR, 2021.

Schimunek, J., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., and Klambauer, G. A generalized framework for embedding-based few-shot learning methods in drug discovery. *ELLIS Machine Learning for Molecules workshop*, 2021.

Sendera, M., Tabor, J., Nowak, A., Bedychaj, A., Patacchiola, M., Trzcinski, T., Spurek, P., and Zieba, M. Non-gaussian gaussian processes for few-shot regression. *Advances in Neural Information Processing Systems*, 34: 10285–10298, 2021.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Somnath, V. R., Bunne, C., and Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34:25244–25255, 2021.

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., Schneider, N., and Brockschmidt, M. 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.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

Swinney, D. C. and Lee, J. A. Recent advances in phenotypic drug discovery. *F1000Research*, 9, 2020.

Tossou, P., Dura, B., Laviolette, F., Marchand, M., and Lacoste, A. Adaptive deep kernel learning. *arXiv preprint arXiv:1905.12131*, 2019.

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., and Bertrand, H.-O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of medicinal chemistry*, 48(7):2534–2547, 2005.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vella, D. and Ebejer, J.-P. Few-shot learning for low-data drug discovery. *Journal of Chemical Information and Modeling*, 2022.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., et al. Pubchem's bioassay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Wang, Y., Wu, S., Duan, Y., and Huang, Y. Resatom system: protein and ligand affinity prediction model based on deep learning. *arXiv preprint arXiv:2105.05125*, 2021.

Zhu, T., Cao, S., Su, P.-C., Patel, R., Shah, D., Chokshi, H. B., Szukala, R., Johnson, M. E., and Hevener, K. E. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis: Miniperspective. *Journal of medicinal chemistry*, 56(17):6560–6572, 2013.

A. Dataset details

A.1. PubChemBA and BindingDB

We trained on PubChemBA (Wang et al., 2012) and BindingDB (Gilson et al., 2016) for their size, high quality, and broad coverage across many targets and assays. For both datasets, we excluded very small or very large molecules, defined as less than 10 atoms or more than 70. From BindingDB, we recorded activities in nanomolar units from either the K_D, K_i, IC50, or EC50 columns, if available. Similarly, we used the PubChem "activity value", which can be any dose-response activity value (either target-based or phenotypic), normalized to nanomolar units. We used such a broad range of different activity types because all values are similarly determined by an underlying binding mechanism, it increased the amount of data we can train on, and allowed the trained models to generalize to both target-based and phenotypic data types. If no continuous activity value was available for a given molecule, we discarded it. When activity was expressed as an upper or lower bound, we took the bound itself as the known activity. To reduce outlier activity values, we also clipped activity values with log10 nM values of < -2.5 or > 6.5, as values surpassing those limits were rare. Then, we excluded all assays that include less than 10 measured compounds. Assays were defined via protein sequence in BindingDB (although some protein targets may contain data aggregated from multiple experimental assays), and by bioassay (i.e. AssayID) in PubChemBA. We transformed all activity values using the base-10 logarithm, as activity often spans several orders of magnitude.

BindingDB data was taken directly from the file `BindingDB_All.tsv` (https://www.bindingdb.org/rwd/bind/chemsearch/marvin/SDFdownload.jsp?download_file=/bind/downloads/BindingDB_All_2D_2023m0.sdf.zip). PubChemBA data was downloaded via the FTP interface (https://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/Concise/JSON/). For each row in the downloaded files, the activity value was taken from the `PubChem Standard Value` column, and the SubstanceIDs were converted into corresponding SMILES strings via the files available at https://ftp.ncbi.nlm.nih.gov/pubchem/Substance/CURRENT-Full/SDF/.

A.2. Cancer Cell Line Encyclopedia

The Cancer Cell Line Encyclopedia (Barretina et al., 2012) consists of interaction data of 24 drugs against a wide array of 479 patient-derived cancer cell lines. For this paper, we used the dataset reported in Table S11 of Barretina et al. (2012), and extracted IC50 measurements for each drug measured against each cell line. We excluded compounds with less than 10 or more than 70 atoms, and cell lines with less than 10 drugs with measured activity. We also excluded all compound-activity pairs if there was no continuous activity value reported.

A.3. PubChemHTS

Starting with a list of Assay IDs (AIDs) obtained from the search function at https://pubchem.ncbi.nlm.nih.gov/, we downloaded the top 100 AIDs with the highest number of tested substances with "BioAssay Type" equal to "Screening" and a linked "Protein Target" section in the "BioAssay Record." For each linked protein in a given Screening assay, we obtained continuous activity values to be used as context compounds via the protein's "Chemicals and Bioactivities" section in PubChem. As we used these compounds and their activities for context compounds in our experiments, we excluded all proteins, and therefore assays, with less than 10 tested compounds with continuous activity values. After obtaining the context compounds, we downloaded the datatable for each assay, which contained Compound IDs (which were linked to SMILES strings, to be used as query compounds in our tests, via the PubChem API) and binary compound activity classifications ("Active" or "Inactive" in the datatable file, to be used for computing ROC-AUC and enrichment scores).

B. Implementation details

Unless specifically stated, all baselines were trained with the same molecular representation (2048-bit Morgan fingerprints with a radius of 3). For FS-CAP and all baseline methods, we tuned hyperparameters once for each model in the PubChemBA task discussed in Section 4.3 using 8 context compounds, except without limiting the context activity range, and used the same hyperparameters for all other datasets and subsequent tasks. For all methods, the reported model performance in each experiment is measured after 2^{27} query molecules had been seen in training, or until the average Pearson's r across all test assays stopped improving on PubChemBA with 8 context compounds. A grid search was performed for all sets of hyperparameters, which are listed below for each model (with the best hyperparameters bolded, according to the highest Pearson's r). All models were trained on a server with 8 NVIDIA GTX 3080 GPUs.

B.1. FS-CAP

`FS-CAP` was implemented in PyTorch. We used an Adam optimizer with a base learning rate of 10^{-5} and 128 steps for learning rate warmup, and then cosine annealed the learning rate to 0 over all training steps. We used dropout ($p = 0.1$) and batch normalization following each layer in the predictor network (except in the last 2 layers), while the encoder networks used neither.

Hyperparameters: `learning_rate`={1e-4, 5e-4, 1e-5, **5e-5**, 1e-6 5e-6}, `batch_size`={512, **1024**}, `encoding_dim`={256, **512**}, `n_layers`={4, 5, **6**}, `mlp_width`={**2048**}. We used the same number of layers, `n_layers`, and width of layers, `mlp_width`, in both the context encoder, query encoder, and predictor networks.

B.2. Tanimoto similarity

We used 2048-bit Morgan fingerprints with a radius of 3 for the calculation of Tanimoto similarity. When using multiple context compounds, we calculate the Tanimoto similarity between each context compound and all query compounds, but only use the highest similarity context compound for each query compound. This is because if a query compound is similar to one of, but not all, the known actives (the context set), it is still presumed to be active.

B.3. MolBERT + ANP

We used the pretrained MolBERT model available from `https://github.com/BenevolentAI/MolBERT` to encode SMILES strings into a 768-dimensional vector. We then used this featurizer (which was not made trainable) in an attentive neural process architecture to represent the context and query features, x_i and x_*, respectively (Kim et al., 2019). We re-implemented the attentive neural process architecture in PyTorch, following the original paper (Kim et al., 2019) and their published code (`https://github.com/deepmind/neural-processes/blob/master/attentive_neural_process.ipynb`) as closely as possible. We trained the model using an Adam optimizer with a base learning rate of 10^{-5} and 128 steps for learning rate warmup, and then cosine annealed the learning rate to 0 over all training steps.

Hyperparameters: `learning_rate`={1e-4, **1e-5**, 1e-6}, `batch_size`={**512**, 1024, 2048}, `num_attention_heads`={2, **4**, 8}, `encoding_dim`={256, **512**}, `decoder_layers`={4, **5**, 6}, `mlp_width`={**2048**}. We use the same `encoding_dim` for both the determinstic and latent paths.

B.4. NGGP

We used the official PyTorch implementation of NGGP available at `https://github.com/gmum/non-gaussian-gaussian-processes`. Using the existing code available for the `QMUL` dataset, we modified the datalaoders for our task by outputting 2048-bit Morgan fingerprints. We trained only two separate models, one for BindingDB and one for PubChemBA, because the size of the context set is only relevant at test time. We also expanded the `MLP2` model used in the code to more layers, so the number of parameters was about equivalent to other baselines.

Hyperparameters: `all_lr`={1e-2, **1e-3**, 1e-4}, `meta_batch_size`={**5**, 10}, `update_batch_size`={**5**, 10}, `noise`={gaussian, **none**}, `cnf_dims`={32, 64, **128**}, `mlp_layers`={4, 5, **6**}, `nonlinearity`={**tanh**, relu}, `batch_norm`={True, **False**}, `mlp_width`={**2048**}. We used the defaults provided in the code for the `QMUL` dataset for all other hyperparameters.

B.5. MetaNet

We adapted the Chainer code provided in the official MetaNet implementation (`https://bitbucket.org/tsendeemts/metanet/src/master/`) to PyTorch. Most of the architectural choices were kept the same as the original code, although we changed each Block network to include two 2048-wide linear layers with ReLU nonlinearities so that the entire model used about the same number of parameters as other baselines. We trained the model using an Adam optimizer with a base learning rate of 10^{-5} and 128 steps for learning rate warmup, and then cosine annealed the learning rate to 0 over all training steps.

Hyperparameters: `learning_rate`={1e-2, **1e-3**, 1e-4}, `num_blocks`={4, **5**, 6}, `mlp_width`={**2048**}, `hidden_dim`={512, 1024, **2048**}, `batch_size`={**8**, 16, 32}

B.6. MAML

We used the MAML implementation in the learn2learn library (Arnold et al., 2020). The base model was a simple multilayer perceptron that takes a 2048-bit Morgan fingerprint as input and produces a single scalar output, which is the activity value prediction. As in the original MAML paper (Finn et al., 2017), we used an SGD optimizer with a constant learning rate, as well as applied dropout with $p = 0.1$ after all layers of the network during training.

Hyperparameters: `learning_rate=`{**1e-4**, 1e-5, 1e-6}, `maml_learning_rate=`{**1e-1**, 1e-2, 1e-3}, `batch_size=`{512, **1024**, 2048}, `n_layers=`{6, 7, **8**}, `mlp_width=`{**2048**}

B.7. MetaDTA

Since there was no available implementation of MetaDTA, we re-implemented it in PyTorch. For information on the specifics of the MetaDTA architecture, see Section 3.2 of Lee et al. (2022). The context and query inputs, $\mathbf{x}_i$ and $\mathbf{x}_q$ as described in the paper, were represented with 2048-bit Morgan fingerprints, and the context target y_i used the same scalar activity representation as `FS-CAP`. As it does not specify in the original paper, similarly to `FS-CAP`, we used an Adam optimizer with a base learning rate of 10^{-5} and 128 steps for learning rate warmup, and then cosine annealed the learning rate to 0 over all training steps.

Hyerparameters: `learning_rate=`{1e-4, **1e-5**, 1e-6}, `batch_size=`{512, **1024**, 2048}, `encoding_dim=`{256, **512**}, `n_layers=`{4, 5, **6**}, `mlp_width=`{**2048**}, `attention_heads=`{1, 2, **4**, 8}. We used the same number of layers, `n_layers`, for the query and context set embedding networks, and the decoder network. We also used the same number of attention heads, `attention_heads`, for the multi-head cross and self-attention components of the model.

C. Additional results

Table 7. **Average per-assay correlation.** Mean Pearson's r between predicted and ground-truth compound activity values across all test-set assays in PubChemBA and BindingDB. Context compounds are drawn at random without respect to their activity value. Experiments were performed using 1, 2, 4, and 8 context compounds for each method tested.

DATASET # CONTEXT COMPOUNDS	PUBCHEMBA 1	 2	 4	 8	BINDINGDB 1	 2	 4	 8
TANIMOTO SIMILARITY	0.00	0.11	0.20	0.29	-0.04	0.06	0.15	0.18
MOLBERT + ANP	0.09	0.10	0.21	0.17	0.04	0.04	0.06	0.09
NGGP	0.05	0.11	0.24	0.37	0.04	0.06	0.10	0.15
METANET	0.01	0.01	-0.01	0.00	0.02	0.05	-0.01	0.01
MAML	0.40	0.38	0.37	0.38	0.37	0.35	0.35	0.35
METADTA	0.47	0.47	0.49	0.51	0.43	0.43	0.44	0.43
FS-CAP	**0.51**	**0.52**	**0.54**	**0.54**	**0.46**	**0.48**	**0.49**	**0.48**

Table 7 reports the same experimental setting as is reported in Section 4.3, except without any constraints placed on the activity of context compounds. Here, we simply drew context compounds at random, regardless of their activity value. The models trained on this task using the PubChemBA dataset were applied to the tasks presented in Sections 4.4 and 4.5, as the tasks presented in those sections similarly do not have activity constraints on the context compounds.