Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

GNN Pretraining

Master Thesis

Johannes Kurz

kurzj@student.ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Béni Egressy
Prof. Dr. Roger Wattenhofer

May 7, 2024

Acknowledgements

I want to thank my supervisor Béni for the helpful advice throughout the thesis.

Abstract

In this work we experiment with leveraging the knowledge that is incorporated in a collection of supervised molecular property prediction datasets by pretraining a shared GNN simultaneously on this collection of datasets in the hope of extracting knowledge generally relevant and applicable to different molecular property prediction downstream tasks. We further extended this by additionally adding self-supervised motif data to the set of pretraining datasets, we experiment with different ways of performing that extension as well as comparing it to pretraining purely on motifs. In order to do this we set up a framework to train on an arbitrary number of datasets simultaneously. We comprehensively evaluate the different pretraining configurations by looking at the performance on a variety of different downstream tasks after finetuning, analyzing the pretraining itself as well as testing the out-of-the-box usability of the GNN weights learned during pretraining. The results show that the multi-dataset-pretraining, especially with incorporation of motif data, leads to performance improvements on the majority of downstream tasks. We conclude that multi-dataset-pretraining is worth being investigated further and propose improvements as well as next steps.

Contents

Chapter 1

Introduction

By design Graph Neural Networks (GNNs) are well suited for structured data like social networks or knowledge graphs and established themselves as the method of choice for many applications due to their state of the art performance. As molecules can be naturally represented as graphs in which nodes are atoms and edges represent bonds between atoms GNNs are applied to a wide range of computational tasks on molecules. Probably the most relevant usage of GNNs in the molecular domain is drug discovery, where the most frequent type of task is the prediction of specific properties of a molecule, which is known as molecular property prediction. Such properties can range from basic ones like solubility in water to more complex ones like toxicity or even drug-target interaction and predicting them is useful to determine whether a molecule might be a potential fit for a drug. A reliable GNN could therefore make the discovery process more efficient by reducing cost as well as accelerating the process itself. Reliability in this field is crucial as no one wants a model that wrongly rules out a molecule that would have in the end led to a cure for a disease or on the other side make people waste resources on a molecule that should have been ruled out from the beginning. In conflict with the required reliability is the fact that task-specific labels in the molecular domain are scarce as they often require time-consuming and expensive wet-lab experiments. This makes the molecular domain challenging as despite this scarcity of labels one needs a reliable model i.e. good out-of-distribution generalization. One popular approach in the field of Neural Networks when facing label scarcity and the strong requirement for out-of-distribution generalization is pretraining, which has been proven itself in the recent years to be very effective in the domains of Computer Vision and Natural Language Processing (NLP). Thus there has been a lot of effort to develop pretraining approaches for GNNs in the hope of similar success. While many approaches achieve noticeable gain in performance on molecular property prediction tasks over the non-pretrained model, one is still far away from a general purpose pretraining framework that fulfills these challenging requirements. Thus in order to contribute to the research for a foundation model or general purpose pretraining framework we experiment with leveraging a collection of existing supervised datasets by simultaneously training on them to extract information generally useful for molecular downstream tasks.

Chapter 2

Preliminaries and Related Work

2.1 Molecules as Graphs

As mentioned in the introduction molecules can be represented as undirected graphs in which nodes are the atoms and edges represent the bonds between two atoms.

2.2 Motifs in Graphs

In the context of graphs motifs refer to subgraphs of statistical significance, these can be generic patterns like cycles or cliques of different sizes or more complex even domain specific patterns that involve node and edge features like functional groups in molecular graphs.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) have established themselves as extremely effective tools for tasks on structured data. Let $G = (V, E)$ be a graph with nodes V and corresponding node features x_v for $v \in V$ and edge features e_{uv} for $(u, v) \in E$. A GNN learns a representation h_v for each node $v \in G$ by repeatedly aggregating representations of its neighboring nodes and adjacent edges and then updating the node representation h_v based on the aggregated information and the previous node representation. A K-layer GNN performs K representation updates, following equation describes the update performed by the k-th layer with $k = 1, \ldots, K$ and $h_v^{(0)} = x_v$ for $v \in V$:

$$h_v^{(k)} = \text{UPDATE}\left(h_v^{(k-1)}, \text{AGGREGATE}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \text{N}(v)\right\}\right)\right) \tag{2.1}$$

Note that in equation 2.1 we use message passing from source to target, this is typically the default flow direction used for GNNs, however one is not restricted to do this and the other option is a target to source flow i.e. use edge feature e_{vu} in the above equation. As molecules are represented by undirected graphs changing the flow direction would not make a difference in our case.

If one wants to perform a node-level task (e.g. classifying each node) the node representations $h_v^{(K)}$ for $v \in V$ can be used directly, if one wants to perform a graph-level task (e.g. predicting a single label for the whole graph) one uses a pooling function to obtain a representation h_G of the entire graph, formally

$$h_G = POOL(\{h_v^{(K)} : v \in V\}) \quad , \tag{2.2}$$

which is then passed into a final linear layer. Pooling functions are permutation-invariant and often just simple operations like computing the mean or sum over all node representations in the graph, but can also be more complex functions.

2.4 Related Work

Given the success of pretraining in computer vision and natural language processing (NLP), there has been a lot of effort over the past few years to develop pretraining approaches for GNNs in the hope of similar results. In one of the earlier works that enable successful knowledge transfer to a variety of downstream datasets Hu et al. [1] combine self-supervised node-level and supervised graph-level pretraining. One year later GraphCL [2] proposed different graph augmentations and applying them in the form of constrastive learning between augmented graphs to achieve successful knowledge transfer. Many other approaches have been proposed including graph-autoregressive-models like GPT-GNN [3], and well-known concepts from NLP have been adapted to GNNs, MoleBert [4] adapts tokenization to molecular graphs to increase the vocabulary size for a harder and more informative pretraining task and GPPT [5] adapts prompting to graph data in order to narrow the training objective gap between pretraining task and downstream task. Of specific relevance for our work is GROVER [6] which introduces a new transformer based GNN architecture. The relevant part for our work is that for molecular property prediction tasks they pretrain this architecture with self-supervised graph-level motif prediction. They compute binary labels of 85 functional groups for an astonishing 10 million molecules and use them as a self-supervised graph-level pretraining task, in addition they perform self-supervised node- and edge-level pretraining via contextual property prediction and show that their combined pretraining strategy benefits the downstream performance of their architecture. In parts of our work we compute the same 85 binary labels for our datasets and use them as additional pretraining tasks.

Chapter 3

Datasets

3.1 Default Datasets

Throughout our work we build upon 11 datasets all introduced by MoleculeNet [7], we provide an overview over these datasets in Table 3.1. We often refer to these 11 datasets as default datasets, the reason for this is that we also experiment with variations of these datasets regarding their labels and thus want to be able to clearly differentiate between the default, i.e. untouched, and modified versions of the datasets. Given that all classification tasks are binary we use for classification datasets the Binary Cross Entropy (BCE) as loss function and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) as evaluation metric. For regression datasets we use the Root Mean Square Error (RMSE) as both the loss function and evaluation metric.

Task type	**Category**	**Dataset**	**# Tasks**	**# Compounds**	**Loss-Function**	**Eval-Metric**
Classification	Biophysics	MUV	17	93087	BCE	ROC-AUC
		HIV	1	41127	BCE	ROC-AUC
		BACE	1	1513	BCE	ROC-AUC
	Physiology	BBBP	1	2039	BCE	ROC-AUC
		TOX21	12	7831	BCE	ROC-AUC
		TOXCAST	617	8577	BCE	ROC-AUC
		SIDER	27	1427	BCE	ROC-AUC
		CLINTOX	2	1480	BCE	ROC-AUC
Regression	Physical Chemistry	ESOL	1	1128	RMSE	RMSE
		FREESOLV	1	642	RMSE	RMSE
		LIPOPHILICITY	1	4200	RMSE	RMSE

Table 3.1: Default Datasets used for Multi-Dataset-Training

To allow for a better understanding of the differences and similarities between datasets we provide a short description for each dataset (descriptions were partially carried over from [7]):

- **MUV** (Maximum Unbiased Validation): Provides 17 tasks that were designed for benchmarking virtual screening methods.

- **HIV**: Task is to predict compound's ability to inhibit HIV replication. Compounds are classified as active or inactive with regards to whether inhibition could be measured or not.
- **BACE**: Binary labels indicate whether or not a molecule is an inhibitor of the human beta-secretase-1 (BACE-1) enzyme.
- **BBBP** (Blood-Brain Barrier Penetration): Binary labels indicating whether it is likely or not that a compound is able to penetrate the blood-brain-barrier.
- **TOX21**: Chemical Compounds tested for various toxic effects, including nuclear receptors and stress response pathways.
- **TOXCAST** (Toxcitiy Forecaster): Various binary tasks to predict different toxic effects of compounds.
- **SIDER** (Side Effect Response): Database of marketed drugs and their recorded adverse drug reactions, grouped into 27 system organ classes.
- **CLINTOX**: Qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons. Provides binary label regarding toxicity and regarding FDA approval status.
- **ESOL** (Estimated Solubility): Water solubility (log solubility in mols per liter) for common small organic molecules.
- **FREESOLV**: Experimental and calculated hydration free energy of small molecules in water.
- **LIPOPHILICITY**: Experimental measures of the octanol-water distribution coefficient of compounds, which is a standard measure of a molecule's lipophilicity (note that we sometimes abbreviate the dataset name with LIPO).

3.2 Motif Data

As mentioned we also experiment with incorporating motif data in our pretraining, motif data in our work refers to labels indicating the presence of a functional group in a molecule. More specifically, for each graph in each dataset we compute 85 binary graph labels each indicating the presence of a different functional group in a molecule and the task is then to predict these labels. These labels have also been used for self-supervised molecular pretraining in [6]. The 85 functional groups are actually all functional groups that are provided by the `rdkit.Chem.Fragments` module of RDKit [8]. If one wants a detailed overview over the functional groups checked for we refer to the official documentation of the

`rdkit.Chem.Fragments` module and give a short overview over these functional group in the next paragraph.

The functional groups checked for include basic functional groups like alcohols (R-OH), carboxylic acids (R-COOH) and amines (R-NH_2, R_2NH, R_3N). They also include aromatic structure like benzene, phenols and anilines as well as more complex aromatic systems. In addition various functional groups containing nitrogen like pyridine as well as carbonyl containing functional groups like ketones and aldehydes. Besides that they also include a variety of other groups including less specific ring and chain like groups, groups containing phosphorus atoms and more. All in all many important and relevant functional groups.

3.2.1 Motif Datasets

When talking about a motif dataset we mean a dataset whose tasks only consist of these 85 binary labels, we mark them with $motif$ in the subscript i.e. MUV_{motif} corresponds to the MUV dataset relabeled with the 85 binary motif labels. As a result a motif dataset always corresponds to a classification dataset, so ESOL_{motif} is a classification dataset and not a regression dataset like ESOL. For all motif datasets we use BCE as the loss function and ROC-AUC as the evaluation metric.

3.2.2 Datasets with Additional Motif Labels

We also experiment with adding these motif labels as additional tasks to the default datasets by concatenating them to the default labels. We mark such dataset by appending + to the dataset name i.e. MUV+ refers to the MUV dataset with additional motif labels. MUV+ therefore now has $17 + 85$ tasks. Note that we also concatenate these labels to regression datasets where they correspond to 85 additional regression targets with value either 0.0 or 1.0, hence ESOL+ compared to ESOL has 85 additional regression targets all of which have either value 0.0 or value 1.0. For all datasets with additional motif labels we use the same loss function and evaluation metric as we use for the corresponding default dataset.

3.3 Features

For all datasets we use the same nine node features and the same three edge features, this is important for us because it allows sharing an encoder between pretraining datasets as well as between pretraining datasets and downstream dataset, which we will explain in Section 4.2. All features are categorical and we provide an overview of the node features in Table 3.2 and of the edge features in Table 3.3, note that the vocabulary size is the number of distinct values that are

possible for a specific feature, but not necessarily the number of distinct values that appeared in all used datasets.

The features can be computed for arbitrary molecular datasets, such that one is not restricted to specific datasets for pretraining which would otherwise harm the general applicability of the work. Given that one has a SMILES [9] (Simplified Molecular Input Line Entry System) string of the molecule, which is the case for most datasets, the features can be extracted using RDKIT [8]. In case no SMILES string is provided one can simply compute it from an RDKIT molecule, one should just make sure to correctly construct the RDKIT molecule from the given molecular graph.

Node/Atom Features		
feature	description	vocabulary size
atomic number	like in periodic table identifier of the atom type (e.g. C, H)	119
chirality tag	indicates the stereochemical configuration of a chiral atom	5
degree	number of bonds connected to an atom	12
formal charge	charge assigned to an atom in the molecule	11
number of H	number of hydrogen atoms directly bonded to the atom	9
number of radical electrons	number of unpaired electrons associated with the atom	6
hybridization type	indicates the type of hybridization (e.g. sp, sp2, sp3, sp3d)	6
is aromatic	whether the atom is part of an aromatic system	2
is in ring	whether the atom is part of a ring structure	2

Table 3.2: Node Features

Edge/Bond Features		
feature	description	vocabulary size
bond type	type of the bond (e.g. single, double, ...)	5
stereochemical configuration	spatial orientation of the chemical bond	6
is conjugated	flag indicating whether the bond is conjugated or not	2

Table 3.3: Edge Features

3.4 Label Distribution

As the last part of this chapter we provide an overview over the label distribution of the datasets, in order to make it comprehensive and at the same time well-aranged we compute for each classification dataset the fraction of the majority label per tasks as well as the mean over all tasks (see Table 3.4 as an example) and collect the means of all default classification datasets in Table 3.5, of the motif classification datasets in Table 3.6 and of the datasets with additional motif labels in Table 3.7. The fraction of the majority label for a specific task is computed as

$$\frac{\max(\text{num}(labels = 0), \text{num}(labels = 1))}{(\text{num}(labels = 0) + \text{num}(labels = 1))} \quad , \tag{3.1}$$

where num($labels = x$) returns the number of labels that are equal to x.

More detailed statistics about the label distributions can be found in Appendix A. For the default regression datasets we provide the mean and standard deviation of the targets in Table 3.8 (keep in mind that all default regression datasets only have one task) and for the regression datasets with additional motif targets we provide the mean and standard deviation over the target distribution of all tasks in Table 3.9.

CLINTOX, fraction of majority class per task (%)				
	train	val	test	total
task 1	93.2	96.6	93.9	93.6
task 2	91.9	95.9	93.2	92.4
mean	92.6 ± 0.7	96.3 ± 0.3	93.6 ± 0.3	93.0 ± 0.6

Table 3.4: Fraction of majority class (%) in the ClinTox dataset

Fraction of majority class averaged over all tasks (%)				
	train	val	test	total
MUV	99.8 ± 0.0	99.8 ± 0.1	99.8 ± 0.1	99.8 ± 0.0
HIV	96.3 ± 0.0	98.0 ± 0.0	96.8 ± 0.0	96.5 ± 0.0
BACE	60.3 ± 0.0	86.1 ± 0.0	53.3 ± 0.0	54.3 ± 0.0
BBBP	83.9 ± 0.0	59.3 ± 0.0	52.9 ± 0.0	76.5 ± 0.0
TOX21	92.8 ± 4.5	89.9 ± 6.3	89.9 ± 6.5	92.2 ± 4.7
TOXCAST	83.8 ± 14.2	82.6 ± 14.3	82.1 ± 14.2	83.4 ± 14.3
SIDER	74.3 ± 12.7	77.5 ± 12.3	76.9 ± 13.1	74.9 ± 12.7
CLINTOX	92.6 ± 0.7	96.3 ± 0.3	93.6 ± 0.3	93.0 ± 0.6

Table 3.5: Fraction of majority class averaged over all tasks for each default classification dataset.

Fraction of majority class averaged over all tasks (%)				
	train	val	test	total
$\text{MUV}_{\text{motif}}$	91.7 ± 12.1	92.0 ± 11.5	92.0 ± 11.4	91.7 ± 12.1
$\text{HIV}_{\text{motif}}$	90.7 ± 12.1	91.5 ± 11.3	91.8 ± 11.2	90.9 ± 12.0
$\text{BACE}_{\text{motif}}$	90.9 ± 14.1	91.8 ± 13.5	90.8 ± 13.3	90.8 ± 14.2
$\text{BBBP}_{\text{motif}}$	90.8 ± 12.2	89.0 ± 12.9	89.8 ± 12.6	90.4 ± 12.0
$\text{TOX21}_{\text{motif}}$	92.8 ± 10.5	90.8 ± 12.1	90.8 ± 11.9	92.3 ± 11.0
$\text{TOXCAST}_{\text{motif}}$	92.7 ± 10.6	90.8 ± 12.0	90.5 ± 12.4	92.1 ± 11.2
$\text{SIDER}_{\text{motif}}$	89.9 ± 12.8	89.3 ± 12.6	88.2 ± 13.2	89.6 ± 12.9
$\text{CLINTOX}_{\text{motif}}$	89.6 ± 13.2	89.6 ± 13.6	90.5 ± 12.7	89.6 ± 13.2
$\text{ESOL}_{\text{motif}}$	94.7 ± 8.6	92.7 ± 11.9	92.5 ± 12.0	94.2 ± 9.4
$\text{FREESOLV}_{\text{motif}}$	96.2 ± 6.7	95.5 ± 9.3	95.5 ± 8.1	96.0 ± 6.5
$\text{LIPO}_{\text{motif}}$	90.5 ± 13.0	91.2 ± 12.3	90.8 ± 12.6	90.6 ± 12.9

Table 3.6: Fraction of majority class averaged over all functional groups for each dataset relabeled with motif labels.

Fraction of majority class averaged over all tasks (%)				
	train	val	test	total
MUV+	93.0 ± 11.5	93.3 ± 10.9	93.3 ± 10.8	93.1 ± 11.4
HIV+	90.8 ± 12.0	91.6 ± 11.3	91.9 ± 11.2	90.9 ± 12.0
BACE+	90.5 ± 14.4	91.7 ± 13.5	90.4 ± 13.8	90.4 ± 14.7
BBBP+	90.7 ± 12.1	88.6 ± 13.2	89.4 ± 13.1	90.3 ± 12.0
TOX21+	92.8 ± 10.0	90.7 ± 11.5	90.7 ± 11.4	92.3 ± 10.4
TOXCAST+	84.9 ± 14.1	83.6 ± 14.3	83.1 ± 14.3	84.4 ± 14.2
SIDER+	86.1 ± 14.4	86.4 ± 13.5	85.5 ± 14.0	86.0 ± 14.3
CLINTOX+	89.7 ± 13.0	89.8 ± 13.5	90.6 ± 12.6	89.7 ± 13.1

Table 3.7: Fraction of majority class averaged over all tasks for each classification dataset with additional motif labels.

Mean and standard deviation of targets				
	train	val	test	total
ESOL	-2.87 ± 2.07	-3.77 ± 1.98	-3.80 ± 2.12	-3.05 ± 2.10
FREESOLV	-3.26 ± 3.28	-6.05 ± 6.08	-5.88 ± 3.64	-3.80 ± 3.84
LIPO	2.16 ± 1.21	2.20 ± 1.22	2.36 ± 1.10	2.19 ± 1.20

Table 3.8: Mean and standard deviation of regression target for each default regression datasets

Mean and standard deviation of targets				
	train	val	test	total
ESOL+	0.02 ± 0.32	0.03 ± 0.43	0.03 ± 0.44	0.02 ± 0.35
FREESOLV+	0.00 ± 0.36	-0.03 ± 0.66	-0.02 ± 0.64	0.00 ± 0.42
LIPO+	0.15 ± 0.30	0.15 ± 0.30	0.15 ± 0.31	0.15 ± 0.30

Table 3.9: Mean and standard deviation over distribution of regression tasks for each regression dataset with additional motif targets

Chapter 4

Multi-Dataset-Training

4.1 Idea

The idea is to train a single GNN on multiple datasets simultaneously in the hope of extracting knowledge generally relevant and applicable to molecular property prediction tasks. The thought behind is that if a single GNN has to perform as good as possible on multiple datasets simultaneously it has to learn to compute representations of molecules that are as useful as possible to as many datasets as possible at the same time. If the collection of datasets is comprehensive enough such that the distribution of possible molecular property prediction tasks as well as the distribution of molecules is sufficiently covered and the GNN is able to simultaneously perform good on all datasets then the GNN would have learned to compute representations that are generally useful for molecular property prediction tasks. And we think that such a GNN would then be a great foundation for finetuning. Of course for a given set of datasets there is the possibility that the GNN cannot cope with the variation of tasks, either because it does not have the capacity or because its simply not possible, and then its best option is to provide representations only useful for a subset of tasks or in the worst case to just provide random representations.

In order to evaluate whether our idea has potential we will introduce a framework in the next section to simultaneously train on an arbitrary number of datasets.

4.2 Framework

4.2.1 Train Batching

At the core of our idea is a shared GNN who simultaneously learns from/has to adapt to all datasets at the same time, thus during training every batch must contain data from every dataset. To enforce this we designed a custom batching procedure for the train-split, which we will explain now.

Let D_{train} denote the datasets on which we train on simultaneously, as the

framework works for an arbitrary number of datasets we distinguish between training on multiple datasets simultaneously i.e. $|D_{train}| > 1$ and training on a single dataset i.e. $|D_{train}| = 1$.

Multiple Datasets

Note that we perform data splitting prior to batching, so we only sample from the train split. When handling multiple datasets of potentially varying sizes at the same time we want to ensure that all datasets contribute equally to each gradient update in the training process. In order to do this we create multi-dataset-batches, such a multi-dataset-batch is basically just a list containing one batch from each dataset. Given that dataset-sizes vary strongly and therefore a different number of batches can be extracted from each dataset we use sampling to ensure that each dataset independent of its size is represented equally during training. Thus, at the beginning of each epoch we sample n samples uniformly with replacement from each training dataset $d \in D_{train}$, for every dataset these samples are then partitioned into batches and to each batch metadata about the corresponding dataset, including its name, the output dimension as well as the loss function and evaluation metric associated with it, is added. Finally, the multi-dataset-batches for this epoch are then obtained by zipping the list of batches of all datasets. We provide Pseudocode 1 as well as a graphical illustration of this batching procedure in Figure 4.1. The reason that we sample in every epoch and not just once before the training starts is that if the sample size n is much smaller than the size of the whole dataset we still want to make use of the data distribution of the whole dataset and not only of a small subset, the reason that we sample with replacement is that n might be larger than the size of a dataset in D_{train} or more generally formulated we sample with replacement so that n can be of arbitrary size. In the pseudocode the function **getDatasetInfo** extracts the metadata of the dataset.

Single Dataset

In the case that we train on a single dataset, we obviously do not sample and just use the whole training dataset as one would normally do. Note that the individual batches are still wrapped inside a multi-dataset-batch that just contains the batch of a single dataset and its associated metadata, this allows to use the same train and test code and makes the following section generally applicable.

Test and Evaluation Batching

For test and evaluation we obviously also use the whole available test and evaluation data, no sampling is performed here. We also do not wrap the batches

Algorithm 1 Pseudocode for Train Batching with more than one dataset

```
procedure GetTrainBatches(trainDatasets, sampleSize, batchSize)
    batchesPerDataset ← [ ]
    for (i, dataset) in enumerate(trainDatasets) do
        info ← getDatasetInfo(dataset)
        samples ← sample(dataset, sampleSize, replace=True)
        batches ← getBatches(samples, batchSize)
        # add dataset info to each batch
        batches ← [(batch, info) for batch in batches]
        batchesPerDataset[i] ← batches
    end for
    multiDatasetBatches ← zip(batchesPerDataset)
    return multiDatasetBatches
end procedure
```

inside a multi-dataset batches instead we use default data-loaders to which we add metadata of each dataset.

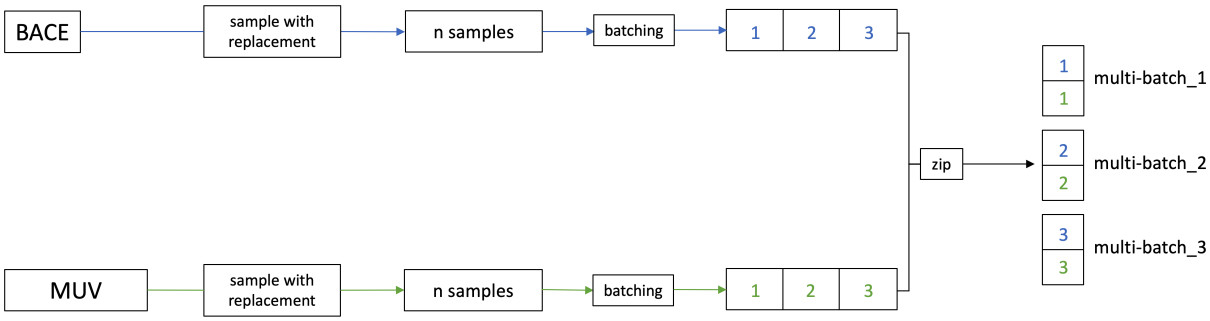

Figure 4.1: Batching for multi-dataset training, performed in every epoch

4.2.2 Model Structure

In this section we present the general structure of the model, illustrated by Figure 4.2 and motivate our design choices, the detailed implementations we use for each part throughout the experiments are provided in Section 4.3.

As mentioned before, the uniformity of features across the different datasets allows us to share the encoders of node and edge features across the datasets i.e. to use the same node feature and edge feature encoders for all datasets. Encoders are required as our features are categorical. The reason we decided to share the encoders is on one hand that if the shared GNN learns to compute

generally useful representations then so do the encoders, thus using the pretrained encoders during finetuning, which we do in all experiments, is coherent with the overall idea and might benefit the downstream performance. On the other hand if we would have separate encoders for each datasets the GNN might be able to learn that specific encodings belong to specific datasets and thus will not necessarily learn to compute representation that are useful for all datasets or in other words using separate encoders would increase the probability that the GNN learns on individual datasets rather than simultaneously on all. Nevertheless, experimenting with separate encoders might be an interesting avenue for future research. The node feature encoding is usually applied once before the first GNN layer, while dependent on the GNN architecture each layer might have its own edge decoder or just one applied before the first layer. Note that in our illustration (Figure 4.2) the encoder placed in front of the shared GNN has primarily the purpose to show that all datasets share the same encoder(s), we do not distinguish between node and edge encoders in this general illustration and it therefore is also not in perfect accordance with all possible GNN implementations. After the shared GNN each dataset is assigned its own dataset-specific decoder. Figure 4.2 already gives a glimpse about how the loss function is computed during training and in the next subsection we will explain the training procedure in more detail.

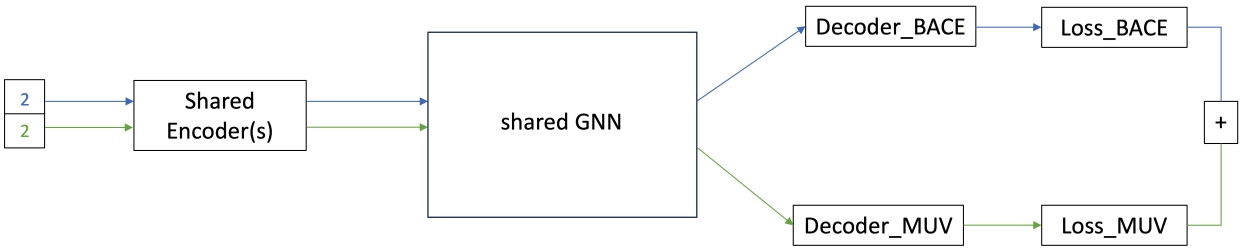

Figure 4.2: Multi-Dataset-Training-Setup with shared encoder

4.2.3 Training procedure

The training procedure is generally straight-forward, given a multi-dataset-batch we pass the batch of each dataset through the model and in the end uniformly add the individual losses together before computing the gradients and performing the optimization step, for a more detailed understanding we provide Pseudocode 2.

Formally the loss computation can be described as follows.

$$L_{multibatch} = \sum_{d \in D_{train}} L_d(\text{Decoder}_d(\text{GNN}(\text{multibatch}_d))) \tag{4.1}$$

In equation 4.1 L_d refers to the loss function used for dataset d, $Decoder_d$ to the encoder corresponding to dataset d, the GNN is shared between all datasets (encoders are part of the GNN), and $multibatch_d$ denotes the batch in the multi-batch corresponding to dataset d. We use Binary-Cross-Entropy (BCE) as the loss of choice for binary classification tasks and the Root-Mean-Squared-Error (RMSE) for regression tasks. Keep in mind that an overview over the loss and evaluation functions used for each dataset, as well as more information about the datasets, can be found in table 3.1.

Algorithm 2 Pseudocode for Training Procedure

```
1:  procedure TRAIN(model, loader, optimizer)
2:      for multiBatch in loader do
3:          optimizer.zerograd()
4:          loss ← 0
5:          # each batch in multiBatch has associated metdata in info
6:          for (batch, info) in multiBatch do
7:              # retrieve name of dataset of which the batch is
8:              datasetName ← info.name
9:              out ← model(batch, datasetName)
10:             # retrieve loss function associated with dataset
11:             criterion ← info.criterion
12:             loss ← loss + criterion(out, batch.y)
13:         end for
14:         loss.backward()
15:         optimizer.step()
16:     end for
17: end procedure
```

4.2.4 Evaluation Procedure

In contrast to loss functions for evaluation metrics a lower value does not always indicate better performance, this is exactly the case for the two evaluation metrics we use, namely the RMSE for regression tasks and ROC-AUC for classification tasks. While a lower value means better performance when using the RMSE, a larger score means better performance in the context of ROC-AUC. As a result when training on our collection of datasets simultaneously where for some datasets the RMSE is used and for some ROC-AUC we cannot simply use the average or sum over all datasets to judge the overall performance of our model when

training on multiple-datasets-simultaneously. Fortunately, the ROC-AUC score is by definition always between zero and one and therefore we use $1 - \text{ROC-AUC}$ instead to compute the average-evaluation-score during training as now for all datasets a lower value always means better performance. Note that throughout this work the performance on the individual classification datasets is still reported as the standard ROC-AUC value i.e. higher means better and that the main use of the average-evaluation-score is to pick the weights of the epoch with the lowest average-evaluation-score on the validation-split as pretrained weights to later finetune on. For the evaluation procedure, i.e. computation of validation or test scores, we perform deferred computations in the sense that we accumulate the output and targets of each batch and then compute the loss over all predictions and targets of the validation-split or test-split respectively. For this procedure we again provide Pseudocode 3. Let D be a collection of individual datasets d, formally we compute the average-evaluation-score as:

$$\text{Average-Evaluation-Score} = \frac{1}{|D|} \sum_{d \in D} Eval(d) \cdot \mathbb{1}_{\{\text{metric}_d = \text{RMSE}\}} + (1 - Eval(d)) \cdot \mathbb{1}_{\{\text{metric}_d = \text{ROCAUC}\}}, \quad (4.2)$$

where $Eval(d)$ refers to the deferred evaluation on validation-split or test-split of dataset d, metric_d to the evaluation metric used for dataset d and with $\mathbb{1}$ we specify an indicator variable.

4.3 Architectual Details

In this section we present the architectures we used for the different parts of the model i.e. for the encoders, the shared GNN and the decoders. We begin with the GNN.

4.3.1 Encoder

The encoders are part of the GNN and are therefore included in the next subsection.

4.3.2 GNN

For our K-layer GNN backbone we rely on the GIN implementation proposed by [1] which incorporates edge features in the node representation computation. However, we also slightly adapt it to first of all incorporate an arbitrary number of node and edge features. Secondly the original paper adds self loops with corresponding edge feature values to each node, because assigning the self-loop a stereochemical configuration as well as the is-conjugated flag cannot be done in a

Algorithm 3 Pseudocode for Evaluation Procedure

```
1: procedure Test(model, loaderList)
2:     averageScore ← 0
3:     scorePerDataset ← [ ]
4:     # loaderlist contains the dataloader of each dataset
5:     numDatasets ← len(loaderList)
6:     for (loader, info) in loaderList do
7:         datasetName ← info.name
8:         predictions ← [ ]
9:         targets ← [ ]
10:        for batch in loader do
11:            out ← model(batch, datasetName)
12:            predictions ← concatenate(predictions, out)
13:            targets ← concatenate(targets, batch.y)
14:        end for
15:        metric ← info.metric
16:        score ← metric(predictions, targets)
17:        scorePerDataset[datasetName] ← score
18:        if (metric == ROCAUC) then
19:            averageScore ← averageScore + (1 − score)
20:        else if (metric == RMSE) then
21:            averageScore ← averageScore + score
22:        else
23:            error("invalid metric")
24:        end if
25:    end for
26:    averageScore ← averageScore / numDatasets
27:    return averageScore, scorePerDataset
28: end procedure
```

well-reasoned manner we leave the addition of self-loops out. We will now take a closer look at how this GNN implementation computes the node representation update for the k-th layer (the following equations is adapted from [1]):

$$h_v^{(k)} = \text{ReLU}\left(\text{MLP}^{(k)}\left(\sum_{u \in N(v) \cup \{v\}} h_u^{(k-1)} + \sum_{e=(u,v):u \in N(v)} h_e^{(k-1)}\right)\right) \tag{4.3}$$

In order to embed the categorical node and edge features embedding layers (`torch.nn.Embedding`) are used. While the node features are only passed to the embedding layer during the first layer pass, the edge features are passed through an embedding layer during each layer pass. The edge feature encoders are not shared between layers i.e. each GNN layer has its seperate edge feature encoder. Let f_v be the l-dimensional feature vector of node v and f_e the p dimensional feature vector of edge e, formally this results in:

$$h_v^{(0)} = \sum_{i=1}^{l} \text{NodeEmbeddeingLayer}_i(f_{v,i}), \quad h_v^{(0)} \in \mathbb{R}^z$$

$$h_e^{(k)} = \sum_{i=1}^{p} \text{EdgeEmbeddingLayer}_i^{(k)}(f_{e,i}), \quad h_e^{(k)} \in \mathbb{R}^z, \quad \text{for} \quad k = 0, 1 \ldots, K-1$$

Where $\text{NodeEmbeddeingLayer}_i$ refers to the embedding layer of the i-th node feature, and $\text{EdgeEmbeddingLayer}_i^{(k)}$ to the embedding layer of the k-th layer in the GNN that embeds the i-th edge feature. The embedding dimension z is the same for all node and edge embedding layers and is a hyperparameter that we specify later. The MLP has two layers and upscales the input first to $2 \times z$ before applying ReLU and then downscales it again to z. Dropout is applied throughout all GNN layers. After passing the data through the GNN we perform mean pooling to obtain a graph presentation h_G, formally

$$h_G = \frac{1}{|V|} \sum_{v \in G} h_v^{(K)} \quad , \tag{4.4}$$

where $|V|$ denotes the number of nodes in the graph.

4.3.3 Decoder

The decoder of each dataset consists of a linear layer, so the final predictions for each dataset d in a set of Datasets D is obtained via

$$\text{predictions}_d = \text{Linear}_d(h_G), \quad \text{predictions}_d \in \mathbb{R}^{\text{out}_d} \ , \tag{4.5}$$

where out_d corresponds to the output dimension required by dataset d, which is part of the metadata stored with the batches.

Chapter 5

Experiments

5.1 Overview

In our experiments we evaluate the potential of training a shared GNN on multiple datasets simultaneously as a pretraining task. In addition to the simultaneous supervised pretraining on the default datasets we explore two ways of adding self-supervised motif pretraining and also perform pretraining purely on motif data as comparison. Furthermore we try out different sample sizes. For every pretraining configuration we analyze the downstream performance compared to no pretraining, how well tasks are learned during pretraining and and how the performance during pretraining relates to the downstream performance, as well as how useful the GNN-weights learned during pretraining are out-of-the-box i.e. when freezing the GNN during finetuning and only training the decoders on the downstream dataset for 5 epochs. Keep in mind that the encoders are part of the GNN and are thereby also frozen during finetuning. We present the experimental setup in the following sections

5.2 Data Splitting

Randomly splitting datasets into train and evaluation set does generally not reflect the strong requirement for out-of-distribution generalization of real-world applications in the molecular domain. As a consequence random-splitting often leads to overly optimistic scores. To achieve a more realistic assessment of the models performance the so called scaffold-splitting [10] is often used when evaluation GNNs in the molecular domain. Scaffold-splitting provides a more realistic evaluation environment by sorting the molecules according to its scaffold, which is the core structure of a molecule. Sorting the molecules this way places similar molecules close-by and different ones further apart. The sorted list is then split continuously (i.e. no reshuffling etc.) into train, validation and test set. Assuming there is enough variation in the scaffolds this allows for a more realistic evaluation. Note that there is no randomness in this process, so data splitting

this way is deterministic. We provide an illustration of scaffold-splitting in Figure 5.1.
For all experiments we use scaffold-splitting to split our datasets into train, validation and test set with a 80% : 10% : 10% split.

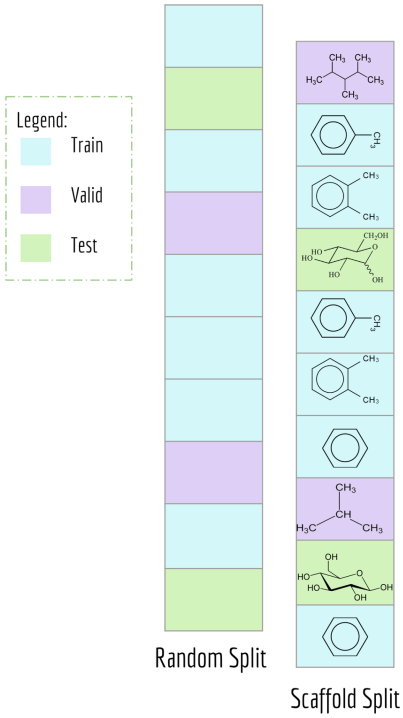

Figure 5.1: Illustration of scaffold splitting, adapted from [7]

5.3 Leave-one-out Experiments

For the rest of the chapter let D_{default} be the set of default datasets we presented in Table 3.1 and let D_{pretrain} denote a set of datasets we simultaneously pretrain on when evaluating downstream performance on a specific dataset $d \in D_{\text{default}}$. Note that for better readability we write d_+ instead of $d+$ to indicate the version with additional motif labels of a dataset $d \in D_{\text{default}}$. As described in Chapter 3 d_{motif} denotes the version of dataset $d \in D_{\text{default}}$ relabeled with the motif labels.

We perform the experiments in a leave-one-out manner in the sense that for every pretraining configuration we evaluate the downstream performance on every dataset $d \in D_{\text{default}}$ while during pretraining for the evaluation of the downstream performance on d we strictly leave out d, d_{motif} and d_+ from D_{pretrain} as well as have for every dataset $p \in D_{\text{default}} \setminus \{d\}$ at least one representative i.e. p itself, p_{motif} or p_+ in D_{pretrain}.

More formally when evaluating downstream performance of a pretrained model on an arbitrary downstream dataset $d \in D_{\text{default}}$ it always holds that

$$(D_{\text{pretrain}} \cap \{d,\ d_{\text{motif}},\ d_+\} = \emptyset) \quad \wedge$$
$$(\forall p \in (D_{\text{default}} \setminus \{d\})\ (p \in D_{\text{pretrain}} \vee p_{\text{motif}} \in D_{\text{pretrain}} \vee p_+ \in D_{\text{pretrain}})) \quad .$$

The reason we never add d_{motif} to the pretraining datasets when evaluating downstream performance on d is that we wanted to simulate the scenario that already a model pretrained simultaneously on a very large corpus of datasets exists and one then just finetunes it on the desired downstream dataset, thus its unlikely that a relabeled version of the downstream dataset is always already part of the pretraining datasets. But note that if one performs the pretraining process by them-self adding d_{motif} is feasible as the motif labels are of self-supervised nature. The reason that we do not have d_+ in the pretraining datasets when evaluating downstream performance on d is that $d+$ contains the labels of the downstream dataset.

In the next section we introduce the different pretraining methods, which will further clarify the previous paragraphs.

5.4 Pretraining Methods

In our experiments we analyze different multi-dataset-pretraining methods for all of which we perform leave-one-out experiments as described in the previous section. As mentioned in the previous chapter, we pick the weights from the epoch with the best-average-validation score (see 4.2) as our pretrained GNN weights that we use to finetune on the downstream dataset.
For this section let $d_{\text{leftout}} \in D_{\text{default}}$ denote the dataset we evaluate downstream performance on.

5.4.1 Default

In the default setup we pretrain on all default datasets except the left-out one i.e.

$$D_{\text{pretrain}} = (D_{\text{default}} \setminus \{d_{\text{leftout}}\})$$

In the tables of the results section we refer to this method as `pretrained (default)`.

5.4.2 Additional Motif Labels

We pretrain exclusively on the datasets with additional motif labels i.e.

$$D_{\text{pretrain}} = \{d_{+} : d \in (D_{\text{default}} \setminus \{d_{\text{leftout}}\})\}$$

In the tables of the results section we refer to this method as `pretrained (+ motif labels)`.

5.4.3 Additional Motif Datasets

Here we add relabeled versions of all datasets except for d_{leftout} to the pretraining dataset i.e.

$$D_{\text{pretrain}} = (D_{\text{default}} \setminus \{d_{\text{leftout}}\}) \cup \{d_{\text{motif}} \ : d \in (D_{\text{default}} \setminus \{d_{\text{leftout}}\})\}$$

In the tables of the results section we refer to this method as `pretrained (+ motif datasets)`.

5.4.4 Only Motif Datasets

As the title of the section suggest for this pretraining method we only use datasets relabeled with the functional group labels during pretraining i.e.

$$D_{\text{pretrain}} = \{d_{\text{motif}} \ : d \in (D_{\text{default}} \setminus \{d_{\text{leftout}}\})\}$$

In the tables of the results section we refer to this method as `pretrained (only motif datasets)`.

5.5 Terminology

We will use the terms pretraining configuration and pretraining method frequently throughout the rest of the thesis. A pretraining configuration consists of a pretraining method and the number of samples used during pretraining. So there is a clear distinction between the term method and configuration in the context of pretraining within this work. To indicate the sample size used for a pretraining configuration in tables we add a subscript to the pretraining method e.g. $\texttt{pretrained}_{\texttt{0.5k}}$ `(default)` would specify the configuration of pretraining on the default datasets with sample size 500.

5.6 Pretraining analysis

In addition to comparing downstream performance of different pretraining configurations and training from scratch, we also want to take a closer look at the pretraining, therefore we analyze the different pretraining runs. Note that we do not perform additional experiments for this, we just analyze the performance during pretraining of the different leave-one-out runs. We introduce this more thoroughly in the results section.

5.7 No GNN Finetune

We were also interested in evaluating the out-of-the-box usability of the pretrained weights and thus also perform experiments in which we which we freeze the GNN including encoders completely during finetuning and only train the decoders for 5 epochs. We again provide more details in the results section.

5.8 Hyperparameters

Throughout all experiments we use 5 GNN layers and an embedding dimension of 300. We always pretrain for 100 epochs with a batch size of 32, training on the downstream dataset is also performed with batch size of 32 and generally with 100 epochs, only for the "No GNN Finetune"-experiments we finetune for just 5 epochs on the downstream dataset. We train all models with the Adam optimizer using a learning rate of 0.001 and a `ReduceLROnPlateau` learning rate scheduler. During pretraining we use dropout of 0.2 and during finetuning we increase the dropout ratio to 0.5. For all pretraining methods we try out a sample size of 500 to create multi-dataset-batches during pretraining as well as sample size of 3000. As the "Additional Motif Datasets" pretraining method contains twice the number of datasets as all other methods in its pretraining corpus we decided to additionally perform experiments with sample size of 250 and 1500 for this pretraining method.

5.9 Result Reporting

Throughout all experiments when reporting test performance of a model we use the weights of the epoch with the best evaluation performance on the validation set. Every experiment is repeated with three random seeds, we report mean and standard deviation over the three runs.

Chapter 6

Results

6.1 Notation

An arrow next to a metric (ROC-AUC, RMSE, or AVG GAIN) indicates whether higher or lower values are better, i.e. an upwards pointing arrow (↑) indicates the higher the value of the metric the better the performance of the model and a downwards pointing arrow (↓) indicates the lower the value the better the performance of the model. For ROC-AUC the higher the value the better, for RMSE the lower the value the better.

For many tables we provide the average gain of a method compared to the no-pretraining baseline (the baseline is highlighted with a gray background), independent of the evaluation metric (ROC-AUC or RMSE) we ensure that a positive gain always indicates performance improvement and a negative one always performance decrease i.e. the gain is defined as

$$\text{gain} = \begin{cases} \text{pretrained} - \text{baseline}, & \text{if metric} = \text{ROC-AUC} \\ \text{baseline} - \text{pretrained}, & \text{if metric} = \text{RMSE} \end{cases}$$

6.2 Downstream Performance

The first and most important results we are looking at is whether our pretraining improved the downstream performance on the different datasets. We summarized all downstream results in Table 6.1, grouping them according to the pretraining method and the groups themselves are sorted by increasing sample size used during pretraining. The best performing pretraining configuration per group is marked as bold in black (i.e. black bold indicates which sample size performed the best for a given pretraining method) and the best performing pretraining configuration overall (i.e. across groups) is marked as bold in red. A green shaded cell indicates a positive knowledge transfer i.e. a performance improvement compared to the non-pretrained model. We also refer to the non-pretrained model as baseline.

6.2.1 Classification Datasets

First we focus on the classification datasets, looking at the average gain we can see that the overall best performing configuration is pretraining just on motif datasets with a sample size of 3000, showing an average gain of 2.38 and no negative transfer on all classification datasets. The second best performing pretraining configuration is pretraining with additional motif labels and 3000 samples with an average gain of 2.36 and the third best configuration is pretraining with additional motif labels and 3000 samples resulting in an average gain of 2.26. These three configurations have in common that they utilize motif prediction tasks (either purely or in combination with the default tasks) and use the highest number of samples, they all outperform the no-pretrainig baseline on 7 out of the 8 classification datasets. Pretraining purely on motif datasets with a sample size of 500 is able to outperform the baseline on all classification datasets, but with 2.15 it results in a lower average gain than the previous three configurations. In comparison pretraining purely on the default datasets, i.e. without incorporating any motif tasks, only results in an average gain of at most 0.92 and is able to outperform the baseline on only 5 of the 8 classification datasets, interestingly for this method using 500 samples performs better than using 3000 samples.

Now we are taking a closer look at the influence of increasing the sample size on the downstream performance we do this group by group and then make an overall conclusion. As we mentioned before, for the default pretraining method using the lower number of samples performs better than using the higher number of samples. For all other methods 3000 samples works the best, the performance improvement between lower and higher sample size is the most significant for the pretraining with additional/concatenated motif labels, whereas for pretraining only on motif datasets the improvement when increasing the sample size is the least significant. Interestingly for pretraining with additional motif datasets pretraining with 1500 samples performs the worst while pretraining with 3000 samples works the best. Overall there seems to be no clear relationship between sample size and downstream performance on the classification datasets, however using the highest number of samples, i.e. 3000, works best for most pretraining methods.

6.2.2 Regression Datasets

Looking at the performance on the regression datasets, pretraining with additional motif datasets and 3000 samples displays the best average gain. However, not a single pretrained model is able to outperform or match the baseline on the FreeSolv dataset. If one would not take the FreeSolv dataset into account pretraining with additional motif labels and 3000 samples would be the best performing pretraining configuration. For all pretraining methods using 3000 samples works the best, but similar to the classification datasets using 1500 sam-

ples when pretraining with additional motif datasets performs the worst for this method.

6.2.3 Overall

Overall pretraining methods that incorporate motif prediction tasks, either by combining them with supervised tasks or purely training on them, show the best downstream performance. Furthermore the highest sample size leads to the best performance in most cases, but no clear relationship between sample size and downstream performance can be derived. In the hope of better understanding our previous observations we are going to analyze the pretraining process next.

	ROC-AUC(%) ↑								
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	AVG GAIN ↑
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	
# tasks	17	1	1	1	12	617	27	2	
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	
pretrained$_{0.5k}$ (default)	$\mathbf{71.2 \pm 2.4}$	73.9 ± 1.1	$\mathbf{80.5 \pm 2.9}$	$\mathbf{68.0 \pm 1.0}$	73.6 ± 0.7	60.3 ± 0.8	60.4 ± 2.0	$\mathbf{83.0 \pm 1.7}$	$\mathbf{0.92 \pm 2.29}$
pretrained$_{3k}$ (default)	69.4 ± 1.5	$\mathbf{75.3 \pm 1.2}$	77.4 ± 4.0	66.5 ± 1.4	73.6 ± 1.0	$\mathbf{61.9 \pm 0.8}$	$\mathbf{60.9 \pm 1.9}$	80.3 ± 2.5	0.22 ± 2.64
pretrained$_{0.5k}$ (+ motif labels)	66.3 ± 2.3	73.7 ± 2.3	80.3 ± 2.2	65.7 ± 0.6	74.4 ± 0.4	63.4 ± 0.6	$\mathbf{62.9 \pm 0.5}$	82.2 ± 0.4	0.68 ± 3.26
pretrained$_{3k}$ (+ motif labels)	$\mathbf{72.4 \pm 5.5}$	$\mathbf{74.2 \pm 0.8}$	$\mathbf{80.7 \pm 2.6}$	$\mathbf{68.3 \pm 0.8}$	$\mathbf{76.3 \pm 0.5}$	$\mathbf{64.1 \pm 0.4}$	61.2 ± 1.3	$\mathbf{84.4 \pm 1.4}$	$\mathbf{2.26 \pm 2.05}$
pretrained$_{0.25k}$ (+ motif datasets)	68.3 ± 4.0	74.4 ± 0.4	79.8 ± 2.5	65.2 ± 2.5	72.6 ± 1.1	60.5 ± 1.1	59.9 ± 0.3	$\mathbf{86.5 \pm 2.5}$	0.46 ± 1.90
pretrained$_{0.5k}$ (+ motif datasets)	$\mathbf{73.5 \pm 0.8}$	$\mathbf{75.4 \pm 0.8}$	79.4 ± 2.4	67.9 ± 1.9	74.0 ± 0.3	61.2 ± 0.5	$\mathbf{61.5 \pm 1.8}$	84.0 ± 1.4	1.68 ± 2.12
pretrained$_{1.5k}$ (+ motif datasets)	67.6 ± 2.6	74.7 ± 1.7	76.6 ± 2.7	65.1 ± 1.0	$\mathbf{74.8 \pm 0.8}$	61.3 ± 0.4	60.6 ± 0.7	83.5 ± 2.2	0.09 ± 2.31
pretrained$_{3k}$ (+ motif datasets)	72.5 ± 4.0	74.3 ± 1.2	$\mathbf{82.1 \pm 2.7}$	$\mathbf{69.2 \pm 1.0}$	74.5 ± 0.5	$\mathbf{63.6 \pm 0.2}$	61.3 ± 1.5	84.9 ± 3.7	$\mathbf{2.36 \pm 2.18}$
pretrained$_{0.5k}$ (only motif datasets)	$\mathbf{73.7 \pm 3.4}$	75.5 ± 1.0	79.6 ± 1.1	66.2 ± 2.3	74.8 ± 0.9	62.7 ± 0.7	$\mathbf{63.2 \pm 1.5}$	85.0 ± 0.3	2.15 ± 2.03
pretrained$_{3k}$ (only motif datasets)	70.7 ± 3.6	$\mathbf{76.2 \pm 1.8}$	$\mathbf{79.7 \pm 0.9}$	$\mathbf{68.7 \pm 2.4}$	$\mathbf{76.7 \pm 1.1}$	$\mathbf{63.4 \pm 0.2}$	60.8 ± 2.1	$\mathbf{86.3 \pm 2.2}$	$\mathbf{2.38 \pm 1.86}$

	RMSE ↓			
Dataset	ESOL	FREESOLV	LIPO	AVG GAIN ↑
# compounds	1128	642	4200	
# tasks	1	1	1	
no pretrain	1.407 ± 0.186	$\mathbf{2.634 \pm 0.072}$	0.773 ± 0.013	
pretrained$_{0.5k}$ (default)	$\mathbf{1.249 \pm 0.073}$	3.328 ± 0.166	0.764 ± 0.020	-0.176 ± 0.455
pretrained$_{3k}$ (default)	1.306 ± 0.086	$\mathbf{2.654 \pm 0.396}$	$\mathbf{0.750 \pm 0.009}$	$\mathbf{0.035 \pm 0.061}$
pretrained$_{0.5k}$ (+ motif labels)	1.339 ± 0.074	$\mathbf{2.850 \pm 0.550}$	0.747 ± 0.021	-0.041 ± 0.153
pretrained$_{3k}$ (+ motif labels)	$\mathbf{1.216 \pm 0.061}$	2.995 ± 0.406	$\mathbf{0.724 \pm 0.019}$	$\mathbf{-0.040 \pm 0.287}$
pretrained$_{0.25k}$ (+ motif datasets)	1.305 ± 0.201	2.734 ± 0.249	0.759 ± 0.039	0.005 ± 0.101
pretrained$_{0.5k}$ (+ motif datasets)	1.365 ± 0.093	2.670 ± 0.247	0.756 ± 0.019	0.008 ± 0.040
pretrained$_{1.5k}$ (+ motif datasets)	1.295 ± 0.036	2.929 ± 0.589	0.742 ± 0.001	-0.051 ± 0.215
pretrained$_{3k}$ (+ motif datasets)	$\mathbf{1.238 \pm 0.055}$	$\mathbf{2.644 \pm 0.197}$	$\mathbf{0.734 \pm 0.014}$	$\mathbf{0.066 \pm 0.093}$
pretrained$_{0.5k}$ (only motif datasets)	$\mathbf{1.283 \pm 0.039}$	3.018 ± 0.065	0.749 ± 0.023	-0.079 ± 0.269
pretrained$_{3k}$ (only motif datasets)	1.303 ± 0.063	$\mathbf{2.761 \pm 0.087}$	$\mathbf{0.741 \pm 0.023}$	$\mathbf{0.003 \pm 0.118}$

Table 6.1: Downstream test performance on all datasets, all configurations.

6.3 Pretraining Analysis

In this section we evaluate how well the different datasets were learned during simultaneous multi-dataset-pretraining and whether we can see a clear relation between pretraining performance and downstream performance. In order to assess for a specific pretraining configuration how well datasets are learned during

pretraining we computed the average of the pretraining test scores per dataset over all leave-one-out runs with this configuration. As an example for this computation you can find in Table 6.5 the pretraining test scores for each leave-one-out run when pretraining on the default datasets with sample size of 500. Each non-shaded data row corresponds to one leave-one-out run indicated by the dataset that was left out during that run. In the second to last row we provide for each dataset the mean over all leave-one-out runs, this mean row is highlighted by the blue shading and is followed by the no-pretraining baseline (i.e. how well the datasets are learned when training on each datasets individually from scratch). Note that for this table the diagonal is empty as the left out dataset is never, neither in its default form nor in any modified form, part of the pretraining. The detailed pretraining tables for every configuration including test and validation scores can be found in Appendix C.
Table 6.2 provides for each configuration the average of the test scores on the default datasets during pretraining over all leave-one-out runs, that is the mean row we mentioned before (to emphasize this we kept the blue shading), the first blue row of this table thus corresponds to the mean row in Table 6.5. Green shading again indicates a better score than the baseline. Because some pretraining configurations train on motif datasets we provide the corresponding scores for the motif datasets in Table 6.3, note that for the configurations that pretrain with additional motif datasets the corresponding row in the default dataset table (6.2) and the corresponding row in the motif dataset table (6.3) belong to the same pretraining run i.e. these scores were achieved simultaneously as the model trained simultaneously on the default and motif datasets. Furthermore, we present in Table 6.4 the scores corresponding to the configurations that train with additional/concatenated motif labels (indicated by the + at the end of the dataset names), because these labels are concatenated to the default labels the scores for these datasets can neither be directly compared to the default datasets nor to the motif datasets and thus we present them in this additional table.

6.3.1 Analysis

When looking at the default datasets we can see that interestingly for some datasets during pretraining the performance is already better than from scratch, so they profit directly from simultaneous training on multiple datasets. There are multiple possible reasons for that, we think that training on multiple datasets might have a regularizing effect or that these datasets benefit directly from the knowledge implied by the gradients of some or all of the other datasets. This improvement of performance during pretraining can be especially seen for the configuration with additional motif datasets and sample size of 3000, which is able to outperform the baseline on 7 of the 8 classification datasets and is also able to deliver decent performance on the eighth one. We speculate, based on the fact that pretraining purely on motif datasets has shown to be an effective pre-

training task, that during simultaneous training these additional motif datasets support the model by directly providing gradients to learn to detect motifs which are useful for the actual task. This is supported by the fact that during pretraining this configurations achieves in addition to good performance on the default datasets also good performance on the motif datasets, which we can see when looking at the corresponding row in Table 6.3 where an average ROC-AUC of 90.1% is achieved, compared to an average ROC-AUC of 97.5% when training purely on motif datasets. Nevertheless, we just speculate that this is the reasons so further investigation is necessary to better understand the performance gain during pretraining. Simultaneously training only on the default datasets does benefit some datasets, but there is a large decrease in performance for MUV, HIV and CLINTOX, so it definitely seems that adding the motif datasets to the simultaneous training is key for the performance. It is also interesting that when adding the additional motif datasets to the simultaneous training with 3000 samples the GNN seems to be easily able to cope with training and performing on all datasets simultaneously, so the GNN weights learned while training with this configuration seem to be useful for all datasets and their corresponding decoders at the same time.

6.3.2 Influence of Sample Size

We can clearly see that the sample size is essential for the performance of simultaneous training on multiple datasets, because for every method increasing the number of samples always improved the performance during pretraining. This makes sense as with increasing sample size the model sees more data of each dataset. We think that it would be interesting to increase the sample size further, because increasing the number of samples from 500 to 3000 for the default datasets increased the average gain during pretraining from -7.30 to -3.42, and an average gain of zero with low standard deviation, i.e. similar performance as training from scratch on the individual datasets, seems desirable.

6.3.3 Relation between Pretraining and Downstream Performance

We cannot observe a clear relation between overall pretraining performance and downstream performance. We hoped that within a method, as its hard to compare across methods, a higher average gain or average score during pretraining would imply better downstream performance, but for pretraining on the default datasets using 500 samples showed an overall better performance on the downstream datasets although pretraining performance is better with 3000 samples. While for pretraining with additional motif datasets the configuration with the best pretraining performance (3000 samples) shows the best downstream performance, but the model pretrained with 500 samples shows better downstream performance than the model with 1500 samples. So there is no clear positive

or negative relationship between pretraining performance and downstream performance, at least not for the pretraining configurations that contain default datasets (i.e. pretraining only on default datasets and pretraining with additional motif datasets). For the pretraining configurations that explicitly contain motif tasks/labels in each dataset (i.e. pretraining with concatenated/additional motif labels as well as pure motif pretraining) better pretraining always implies better downstream performance, however one has to be careful when drawing a conclusion here as we did not evaluate these configurations with 250 and 1500 samples.

	ROC-AUC(%) ↑								
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	AVG GAIN ↑
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	
# tasks	17	1	1	1	12	617	27	2	
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	
pretrained$_{0.5k}$ (default)	48.1 ± 2.4	71.2 ± 1.6	75.3 ± 4.6	60.4 ± 1.6	68.7 ± 1.3	60.6 ± 0.8	54.9 ± 1.1	65.9 ± 3.6	-7.30 ± 8.31
pretrained$_{3k}$ (default)	53.7 ± 5.9	70.6 ± 2.8	78.7 ± 1.8	64.5 ± 3.1	71.1 ± 1.4	61.5 ± 1.0	59.6 ± 2.4	76.4 ± 7.7	-3.42 ± 6.46
pretrained$_{0.25k}$ (+ motif datasets)	45.5 ± 2.1	69.7 ± 1.6	76.4 ± 3.0	59.9 ± 1.3	67.6 ± 1.3	59.8 ± 0.8	55.9 ± 1.0	65.2 ± 1.3	-7.94 ± 9.08
pretrained$_{0.5k}$ (+ motif datasets)	48.5 ± 3.8	70.8 ± 1.6	80.2 ± 2.1	61.7 ± 1.7	68.5 ± 1.3	61.0 ± 0.7	57.4 ± 1.7	67.6 ± 5.5	-5.98 ± 8.71
pretrained$_{1.5k}$ (+ motif datasets)	58.3 ± 5.9	71.7 ± 1.0	80.9 ± 1.3	65.9 ± 1.6	71.0 ± 1.2	62.3 ± 0.7	62.2 ± 1.0	84.1 ± 2.9	-0.89 ± 5.43
pretrained$_{3k}$ (+ motif datasets)	70.9 ± 2.2	72.3 ± 1.2	80.2 ± 0.7	67.5 ± 1.7	74.0 ± 0.9	64.1 ± 0.5	63.3 ± 0.8	85.3 ± 3.4	1.76 ± 2.83

	RMSE ↓			
Dataset	ESOL	FREESOLV	LIPO	AVG GAIN ↑
# compounds	1128	642	4200	
# tasks	1	1	1	
no pretrain	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013	
pretrained$_{0.5k}$ (default)	1.394 ± 0.097	3.983 ± 0.218	1.004 ± 0.029	-0.522 ± 0.726
pretrained$_{3k}$ (default)	1.241 ± 0.096	4.344 ± 0.234	0.893 ± 0.076	-0.555 ± 1.011
pretrained$_{0.25k}$ (+ motif datasets)	1.396 ± 0.082	3.998 ± 0.305	1.029 ± 0.067	-0.536 ± 0.729
pretrained$_{0.5k}$ (+ motif datasets)	1.255 ± 0.026	4.250 ± 0.235	0.960 ± 0.056	-0.550 ± 0.938
pretrained$_{1.5k}$ (+ motif datasets)	1.201 ± 0.028	4.499 ± 0.076	0.842 ± 0.036	-0.576 ± 1.125
pretrained$_{3k}$ (+ motif datasets)	1.101 ± 0.025	4.646 ± 0.156	0.800 ± 0.030	-0.578 ± 1.253

Table 6.2: Test performance on default datasets during pretraining.

6.4 No GNN Finetuning on Downstream Tasks

In this section we analyze how useful the GNN-weights learned during pretraining are out-of-the-box i.e. when directly using them for predictions on the downstream dataset. Therefore during finetuning on the downstream task we train only the decoder for 5 epochs, the rest of the model is frozen throughout these 5 epochs of finetuning. Keep in mind that the decoder has not been pretrained i.e. is randomly initialized as the downstream dataset is not, neither in its default form nor in any modified form, part of the pretraining datasets. We also compare with a randomly initialized GNN for which we again only train the decoder for 5 epochs, thus this model has to work completely with random GNN-weights. For the baseline we use, as usual, fully training the model on each dataset individually from scratch for 100 epochs, so for the baseline no part of the model is frozen and no pretrained weights are used. Green shading of cells again indicates better

	ROC-AUC(%) ↑						
Dataset	MUV_{motif}	HIV_{motif}	$BACE_{motif}$	$BBBP_{motif}$	$TOX21_{motif}$	$TOXCAST_{motif}$	$SIDER_{motif}$
# compounds	93087	41127	1513	2039	7831	8577	1427
# tasks	85	85	85	85	85	85	85
$pretrained_{0.25k}$ (+ motif datasets)	69.4 ± 2.3	71.1 ± 2.3	71.9 ± 1.6	73.1 ± 1.9	69.6 ± 2.2	68.1 ± 2.3	73.8 ± 2.8
$pretrained_{0.5k}$ (+ motif datasets)	74.9 ± 2.9	75.6 ± 2.6	75.2 ± 2.4	78.0 ± 2.3	74.8 ± 3.1	73.1 ± 3.2	79.3 ± 2.5
$pretrained_{1.5k}$ (+ motif datasets)	85.3 ± 3.1	84.2 ± 2.5	81.8 ± 3.7	86.1 ± 2.6	85.3 ± 2.8	84.1 ± 2.9	86.2 ± 2.4
$pretrained_{3k}$ (+ motif datasets)	92.8 ± 1.7	90.0 ± 1.6	91.0 ± 1.9	91.4 ± 1.7	92.1 ± 1.7	91.1 ± 1.9	91.1 ± 1.9
$pretrained_{0.5k}$ (only motif datasets)	98.0 ± 0.4	96.4 ± 0.2	98.1 ± 0.1	97.9 ± 0.3	98.7 ± 0.1	98.1 ± 0.1	98.5 ± 0.2
$pretrained_{3k}$ (only motif datasets)	99.3 ± 0.2	99.0 ± 0.1	99.5 ± 0.1	99.1 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.5 ± 0.1

	ROC-AUC(%) ↑				
Dataset	$CLINTOX_{motif}$	$ESOL_{motif}$	$FREESOLV_{motif}$	$LIPO_{motif}$	AVG ↑
# compounds	1480	1128	642	4200	
# tasks	85	85	85	85	
$pretrained_{0.25k}$ (+ motif datasets)	68.7 ± 3.0	70.8 ± 1.9	69.0 ± 2.2	71.4 ± 2.6	70.6 ± 1.8
$pretrained_{0.5k}$ (+ motif datasets)	73.3 ± 3.0	75.3 ± 2.6	72.4 ± 1.2	75.6 ± 3.6	75.2 ± 2.0
$pretrained_{1.5k}$ (+ motif datasets)	82.8 ± 2.7	83.6 ± 1.5	79.4 ± 1.2	85.6 ± 3.0	84.0 ± 2.1
$pretrained_{3k}$ (+ motif datasets)	89.3 ± 1.6	88.0 ± 1.4	82.2 ± 1.1	92.6 ± 1.5	90.1 ± 3.0
$pretrained_{0.5k}$ (only motif datasets)	96.5 ± 0.4	93.3 ± 0.3	83.9 ± 0.7	97.9 ± 0.2	96.1 ± 4.3
$pretrained_{3k}$ (only motif datasets)	98.5 ± 0.2	93.7 ± 0.2	84.2 ± 0.4	99.8 ± 0.2	97.5 ± 4.7

Table 6.3: Test performance on motif datasets during pretraining (the average is over both tables).

	ROC-AUC(%) ↑								
Dataset	MUV+	HIV+	BACE+	BBBP+	TOX21+	TOXCAST+	SIDER+	CLINTOX+	AVG ↑
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	
# tasks	(17+85)	(1+85)	(1+85)	(1+85)	(12+85)	(617+85)	(27+85)	(2+85)	
$pretrained_{0.5k}$ (+ motif labels)	90.8 ± 0.7	93.5 ± 0.6	96.0 ± 0.6	95.8 ± 0.5	93.0 ± 0.5	65.3 ± 0.4	86.4 ± 0.6	92.7 ± 0.8	89.2 ± 10.1
$pretrained_{3k}$ (+ motif labels)	95.2 ± 0.3	98.2 ± 0.2	99.0 ± 0.2	98.6 ± 0.2	96.3 ± 0.1	68.0 ± 0.2	88.4 ± 0.2	98.3 ± 0.4	92.7 ± 10.6

	RMSE ↓			
Dataset	ESOL+	FREESOLV+	LIPO+	AVG ↓
# compounds	1128	642	4200	
# tasks	(1+85)	(1+85)	(1+85)	
$pretrained_{0.5k}$ (+ motif labels)	0.136 ± 0.002	0.165 ± 0.003	0.152 ± 0.004	0.151 ± 0.015
$pretrained_{3k}$ (+ motif labels)	0.110 ± 0.002	0.149 ± 0.003	0.119 ± 0.004	0.126 ± 0.020

Table 6.4: Test performance on datasets with additional motif labels during pretraining.

performance than the baseline and red shading of cells indicates that the performance is worse than the model with random GNN weights. We summarized the results in Table 6.6.

First of all, the pretrained models perform overall significantly better than the model with random GNN-weights. This shows us that was has been learned during pretraining is generally of use for the different downstream tasks. Nevertheless one has to note that for the BACE dataset the random GNN performs better than some pretraining configurations, interestingly this is only the case for configurations that include default datasets. For some classification datasets finetuning only the decoder for 5 epochs is already able to surpass the performance of the baseline. It is noteworthy that pretraining configurations that use motif data have a significantly higher rate of surpassing the baseline after 5 epochs of

	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
left out dataset											
muv		72.2 ± 0.7	78.4 ± 1.7	60.8 ± 1.4	69.7 ± 0.3	61.2 ± 0.5	56.4 ± 0.5	64.8 ± 6.3	1.266 ± 0.032	4.085 ± 0.108	0.984 ± 0.069
hiv	48.6 ± 1.1		64.9 ± 8.1	58.3 ± 2.3	67.1 ± 1.0	60.2 ± 0.6	53.4 ± 1.4	62.9 ± 2.0	1.519 ± 0.092	3.672 ± 0.369	1.057 ± 0.021
bace	46.2 ± 4.4	68.9 ± 2.1		60.1 ± 1.5	69.3 ± 0.7	60.4 ± 0.7	54.1 ± 0.3	62.2 ± 1.8	1.284 ± 0.068	3.983 ± 0.414	0.976 ± 0.038
bbbp	46.4 ± 1.3	72.7 ± 1.0	77.8 ± 2.0		69.9 ± 0.2	61.0 ± 1.0	54.9 ± 1.9	64.7 ± 3.4	1.454 ± 0.137	3.737 ± 0.265	1.006 ± 0.011
tox21	45.0 ± 1.9	72.7 ± 0.9	76.6 ± 4.0	61.7 ± 0.9		61.9 ± 0.0	55.8 ± 1.7	63.0 ± 0.9	1.372 ± 0.160	4.110 ± 0.317	0.972 ± 0.064
toxcast	47.8 ± 3.1	70.7 ± 1.1	79.0 ± 1.9	59.4 ± 0.9	69.0 ± 0.4		54.0 ± 0.8	67.9 ± 1.8	1.396 ± 0.041	4.128 ± 0.213	1.002 ± 0.023
sider	48.6 ± 2.0	72.2 ± 1.3	75.9 ± 8.4	59.7 ± 2.9	68.7 ± 0.2	61.1 ± 1.0		67.9 ± 3.2	1.286 ± 0.061	3.893 ± 0.399	1.027 ± 0.065
clintox	48.3 ± 4.6	71.3 ± 1.7	69.7 ± 18.0	59.0 ± 3.2	67.9 ± 2.6	59.2 ± 2.3	54.1 ± 0.6		1.463 ± 0.133	4.003 ± 0.478	0.987 ± 0.089
esol	53.5 ± 4.8	68.3 ± 2.7	77.2 ± 2.6	61.5 ± 2.5	66.8 ± 0.0	60.0 ± 0.6	56.1 ± 0.5	66.0 ± 3.1		3.808 ± 0.189	1.043 ± 0.054
freesolv	50.3 ± 6.6	70.5 ± 0.4	79.5 ± 3.1	63.7 ± 1.4	70.9 ± 0.3	60.2 ± 0.6	56.3 ± 0.9	74.5 ± 3.4	1.533 ± 0.199		0.987 ± 0.049
lipo	46.3 ± 3.1	72.1 ± 0.9	74.0 ± 4.1	59.4 ± 1.5	67.6 ± 0.3	60.4 ± 0.6	53.9 ± 1.2	65.2 ± 2.0	1.367 ± 0.061	4.413 ± 0.189	
mean	48.1 ± 2.4	71.2 ± 1.6	75.3 ± 4.6	60.4 ± 1.6	68.7 ± 1.3	60.6 ± 0.8	54.9 ± 1.1	65.9 ± 3.6	1.394 ± 0.097	3.983 ± 0.218	1.004 ± 0.029
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Table 6.5: Test performance during default pretraining with 500 samples, left most column indicates dataset left out during pretraining on which the model is subsequently finetuned on.

training only the decoder and also achieve a significantly better average gain with the same sample size. This is in accordance with our previous observations that incorporating motif prediction tasks into pretraining is effective. Pretraining with additional/concatenated motif labels and a sample size of 3000 has the highest rate of surpassing baseline performance with 5 out of 8 classification datasets and also displays the highest average gain of all pretraining methods with -2.08. The second best average gain is achieved by pretraining with additional motif datasets and 3000 samples, but with -4.17 it is significantly lower than the best one and this configuration also only achieves an improvement over the baseline on 3 out of the 8 classification datasets. Furthermore, we can see that pretraining with more samples improves the average gain for all methods on the classification datasets, this also means that for a pretraining method better pretraining performance implies better downstream performance on the classification datasets after 5 epochs of decoder training. For the regression datasets no pretrained model was able to surpass the baseline after 5 epochs of decoder training, this time pretraining with additional motif datasets is the best performing pretraining method across all sample sizes.

Overall for these experiments combining the default datasets with motif tasks (i.e. additional motif labels or additional motif datasets) for the pretraining task seems to provide the most useful out-of-the-box GNN weights.

	ROC-AUC(%) ↑								
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	AVG GAIN ↑
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	
# tasks	17	1	1	1	12	617	27	2	
no pretrain (full training)	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	
random GNN-weights	59.5 ± 2.1	51.3 ± 4.3	68.2 ± 4.7	53.7 ± 2.2	64.0 ± 2.3	54.1 ± 0.3	52.1 ± 0.4	41.0 ± 4.1	-14.95 ± 12.78
pretrained$_{0.5k}$ (default)	63.7 ± 1.7	68.6 ± 1.2	67.0 ± 2.5	54.8 ± 0.7	68.1 ± 0.8	61.3 ± 0.3	54.0 ± 0.9	63.7 ± 5.6	-7.79 ± 6.31
pretrained$_{3k}$ (default)	68.0 ± 4.9	69.9 ± 2.1	67.0 ± 2.3	57.2 ± 3.3	68.8 ± 0.6	61.8 ± 0.2	57.1 ± 0.6	62.9 ± 2.2	-6.35 ± 7.27
pretrained$_{0.5k}$ (+ motif labels)	74.5 ± 0.5	68.3 ± 3.4	72.2 ± 0.3	65.2 ± 1.1	71.6 ± 0.7	61.0 ± 0.2	59.3 ± 0.1	52.5 ± 1.9	-4.86 ± 11.60
pretrained$_{3k}$ (+ motif labels)	72.7 ± 0.6	71.5 ± 1.0	73.1 ± 1.9	67.1 ± 1.6	75.0 ± 0.1	63.0 ± 0.8	59.3 ± 0.8	65.2 ± 3.1	-2.08 ± 7.66
pretrained$_{0.25k}$ (+ motif datasets)	63.1 ± 1.5	67.8 ± 1.4	61.6 ± 1.7	56.1 ± 1.0	67.8 ± 0.1	60.4 ± 0.2	55.3 ± 0.5	59.0 ± 0.6	-9.05 ± 8.29
pretrained$_{0.5k}$ (+ motif datasets)	67.6 ± 3.0	69.5 ± 1.9	67.7 ± 0.8	55.6 ± 0.7	67.7 ± 1.1	60.3 ± 0.4	56.8 ± 0.5	58.8 ± 2.6	-7.44 ± 8.15
pretrained$_{1.5k}$ (+ motif datasets)	70.8 ± 0.7	67.0 ± 2.1	66.7 ± 2.4	57.6 ± 1.6	70.6 ± 0.3	61.1 ± 0.5	59.8 ± 0.5	61.4 ± 0.9	-6.06 ± 8.41
pretrained$_{3k}$ (+ motif datasets)	73.7 ± 1.0	67.9 ± 2.3	71.9 ± 1.8	60.2 ± 1.0	72.1 ± 1.4	63.2 ± 0.8	59.6 ± 1.0	61.5 ± 3.6	-4.17 ± 8.57
pretrained$_{0.5k}$ (only motif datasets)	73.0 ± 2.1	66.6 ± 1.2	74.0 ± 0.4	65.4 ± 1.3	70.0 ± 0.7	60.3 ± 0.1	58.4 ± 0.6	60.1 ± 1.0	-4.46 ± 8.85
pretrained$_{3k}$ (only motif datasets)	68.7 ± 1.4	64.3 ± 3.9	73.4 ± 1.4	67.0 ± 0.7	68.7 ± 0.6	59.8 ± 0.2	57.3 ± 0.6	68.8 ± 3.7	-4.44 ± 6.23

	RMSE ↓			
Dataset	ESOL	FREESOLV	LIPO	AVG GAIN ↑
# compounds	1128	642	4200	
# tasks	1	1	1	
no pretrain (full training)	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013	
random GNN-weights	3.801 ± 0.174	5.679 ± 0.109	1.795 ± 0.121	-2.154 ± 1.033
pretrained$_{0.5k}$ (default)	1.642 ± 0.018	4.681 ± 0.154	1.067 ± 0.020	-0.859 ± 1.030
pretrained$_{3k}$ (default)	1.636 ± 0.057	4.330 ± 0.116	1.071 ± 0.011	-0.741 ± 0.828
pretrained$_{0.5k}$ (+ motif labels)	1.802 ± 0.052	4.746 ± 0.132	0.984 ± 0.004	-0.906 ± 1.048
pretrained$_{3k}$ (+ motif labels)	1.468 ± 0.040	4.519 ± 0.122	0.952 ± 0.012	-0.708 ± 1.021
pretrained$_{0.25k}$ (+ motif datasets)	1.683 ± 0.035	3.958 ± 0.368	1.048 ± 0.024	-0.625 ± 0.605
pretrained$_{0.5k}$ (+ motif datasets)	1.496 ± 0.015	4.005 ± 0.028	1.047 ± 0.015	-0.578 ± 0.693
pretrained$_{1.5k}$ (+ motif datasets)	1.466 ± 0.092	4.081 ± 0.061	1.014 ± 0.017	-0.582 ± 0.754
pretrained$_{3k}$ (+ motif datasets)	1.415 ± 0.035	4.263 ± 0.093	0.943 ± 0.014	-0.602 ± 0.893
pretrained$_{0.5k}$ (only motif datasets)	1.899 ± 0.059	4.578 ± 0.313	1.011 ± 0.021	-0.891 ± 0.920
pretrained$_{3k}$ (only motif datasets)	1.866 ± 0.055	4.499 ± 0.088	0.988 ± 0.003	-0.846 ± 0.891

Table 6.6: Test performance when finetuning only the decoder for 5 epochs compared to full training from scratch each individual dataset

6.5 Key Observations

In this section we present the key observations, we group them according to their corresponding section in the results chapter. To avoid potential confusion we want to clarify that the three key observations listed in the subsection "Downstream Performance" refer to the performance after finetuning the complete pretrained model on the downstream dataset for 100 epochs and thereby do not include the "No GNN Finetune" (i.e. only finetuning the decoder for 5 epochs) experiments in their scope.

Downstream Performance

- **Observation (1):** All pretraining configurations achieve a positive average gain over the no-pretraining baseline and always outperform it on the majority of datasets i.e. on at least 6 out of the 11 datasets.

- **Observation (2):** The three pretraining configurations that perform best on the classification datasets, all of which display a comparable average gain, use motif prediction tasks (either purely or incorporated as concatenated tasks or additional datasets) and the highest number of samples i.e. 3000 samples.
- **Observation (3):** There is no clear relationship between pretraining performance and downstream performance across all configurations.

Simultaneous Pretraining

- **Observation (1):** Sample size is essential for simultaneously training on multiple datasets, increasing the sample size clearly and always improves the test performance during simultaneous pretraining.
- **Observation (2):** Adding motif datasets to the default pretraining datasets (i.e. pretraining with additional motif datasets) clearly improves the performance on the default datasets during simultaneous pretraining compared to training simultaneously on just the default datasets. During simultaneous pretraining with additional motif datasets and sample size of 3000 the performance on most datasets already outperforms pretraining on each dataset individually from scratch.

No GNN Finetune

- **Observation (1):** On some classification datasets finetuning only the decoder for 5 epochs on the downstream datasets after pretraining is already able to surpass the performance of the baseline. This is especially the case for pretraining configurations that use motif data, pretraining with additional motif labels and 3000 samples allows to outperform the baseline on 5 out of 8 classification datasets after just 5 epochs of decoder training.
- **Observation (2):** For a fixed pretraining method better pretraining performance implies better performance on the downstream classification datasets after 5 epochs of only training the decoder.

Chapter 7

Discussion

In this work we evaluated whether pretraining a GNN simultaneously on multiple datasets is a potential way of effectively improving the downstream performance in the molecular domain. We explained the intuition of our idea as the need of the GNN to learn to compute molecular representations during pretraining that are as useful as possible for as many pretraining datasets as possible and thereby in the optimal case generally useful for molecular downstream tasks. In order to investigate the idea we introduced a framework that enables simultaneous training of a single GNN on an arbitrary number of datasets. We chose a corpus of 11 supervised datasets on which we performed leave-one-out experiments and also experimented with different ways of incorporating self-supervised motif tasks to the pretraining datasets. We one one side looked at the direct effect of this pretraining on the downstream performance. On the other side we looked at the performance during pretraining as well as the out-of-the-box performance of the pretrained GNN-weights, because we want to understand what is learned during pretraining and whether it corresponds to our intuition.

The results of our experiments indicate that pretraining on multiple datasets at the same time, especially with incorporation of additional motif data, is beneficial for the downstream performance and that performing further experiments in this direction makes sense. Based on the fact that, especially when incorporating additional motif data, the GNN is able during pretraining to achieve good test performance on all pretraining datasets simultaneously and in addition to that the learned weights are useful out-of-the box for the majority of downstream datasets we speculate that the GNN indeed learns during our proposed pretraining to compute molecular representations that are generally useful for at least our set of molecular downstream tasks and that this is what benefits the downstream performance when finetuning. This supports our intuition and suggestion to further investigate the idea. However, from our results it is not clear whether the collection of supervised tasks in the corpus of datasets, the self-supervised motif prediction or the combination of both is the main driver of the positive results, although we believe that the motif data currently plays a crucial role. To further evaluate the potential of the idea, to allow for more ultimate conclusions and to answer open questions the most crucial next step is to to add more datasets to the

pretraining corpus in order to increase the variation of tasks and molecules during pretraining. We also suggest to increase the sample size used during pretraining, which might be even more important with a larger corpus of datasets. Given the increase in computation implied by our previous two suggestions we want to point out that the pass of a multi-dataset batch through the model can be parallelized as the dataset-specific batches in the multi-dataset-batch are independent of each other, currently, as one can see in Pseudocode 2 and 3, we sequentially predict on each dataset-specific batch. In order to further asses the general applicability of the approach more GNN backbones should be evaluated in future works to find out if the pretraining is agnostic to the GNN architecture or whether some architectures profit more than others from the pretraining. Furthermore we want to point out that the way we add the motif labels to the regression tasks, i.e. by concatenating default targets and motif labels, can be considered rather naive as binary labels are predicted as a regression task, so one can potentially come up with a better way of combining them or try out only adding them to the classification datasets and leaving the regression datasets untouched. Also the calculation of the average-evaluation-score for model selection is suboptimal because the values of $1 - \text{ROCAUC}$ can be much larger than the RMSE scores, meaning that the RMSE scores are less meaningful when choosing the optimal epoch. Lastly, also related to the previous point, we suggest to explore different dataset weighting schemes during pretraining for example nonuniform or even dynamic ones.

Bibliography

[1] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," 2020.

[2] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," 2021.

[3] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, and Y. Sun, "Gpt-gnn: Generative pre-training of graph neural networks," 2020.

[4] J. Xia, C. Zhao, B. Hu, Z. Gao, C. Tan, Y. Liu, S. Li, and S. Z. Li, "Mole-BERT: Rethinking pre-training graph neural networks for molecules," in *The Eleventh International Conference on Learning Representations*, 2023.

[5] M. Sun, K. Zhou, X. He, Y. Wang, and X. Wang, "Gppt: Graph pre-training and prompt tuning to generalize graph neural networks," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1717–1727.

[6] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, "Self-supervised graph transformer on large-scale molecular data," 2020.

[7] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: A benchmark for molecular machine learning," 2018.

[8] "RDKit: Open-source cheminformatics," https://www.rdkit.org.

[9] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[10] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of Medicinal Chemistry*, vol. 39, no. 15, pp. 2887–2893, 1996, pMID: 8709122.

Appendix A

Datasets

A.1 Dataset Label Distribution

We provide counts for the binary labels of each task in the default classification datasets, we omit ToxCast as it has 617 tasks. Not that the datasets not necessarily provide every label for every molecule i.e. sometimes a molecule does not have a specific label. We also omit the motif datasets as well as the datasets with additional motif labels as they all have upwards of 85 tasks.

A.1.1 Classification Datasets

MUV

MUV	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	11894	22	1431	5	1489	0	14814	27
Task 2	11445	21	1557	2	1703	6	14705	29
Task 3	11638	20	1465	3	1595	7	14698	30
Task 4	11677	23	1391	4	1525	3	14593	30
Task 5	11648	18	1569	4	1656	7	14873	29
Task 6	11568	25	1507	2	1497	2	14572	29
Task 7	11774	25	1379	3	1461	2	14614	30
Task 8	11418	21	1496	3	1469	4	14383	28
Task 9	12159	24	1294	3	1354	2	14807	29
Task 10	12011	21	1317	3	1326	4	14654	28
Task 11	11815	21	1387	2	1460	6	14662	29
Task 12	12002	23	1312	2	1301	4	14615	29
Task 13	11679	24	1443	6	1515	0	14637	30
Task 14	11684	26	1547	2	1450	2	14681	30
Task 15	11487	22	1600	4	1535	3	14622	29
Task 16	11957	23	1345	4	1443	2	14745	29
Task 17	11830	21	1386	0	1506	3	14722	24

HIV

HIV	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	31669	1232	4032	81	3983	130	39684	1443

BACE

BACE	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	730	480	21	130	71	81	822	691

BBBP

BBBP	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	262	1369	121	83	96	108	479	1560

TOX21

TOX21	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	5586	248	693	29	677	32	6956	309
Task 2	5323	190	602	25	596	22	6521	237
Task 3	4718	590	531	89	532	89	5781	768
Task 4	4585	210	469	46	467	44	5521	300
Task 5	4444	648	483	68	473	77	5400	793
Task 6	5347	297	632	30	626	23	6605	350
Task 7	5174	134	546	30	544	22	6264	186
Task 8	4177	717	360	107	353	118	4890	942
Task 9	5542	195	634	33	632	36	6808	264
Task 10	5054	282	520	44	521	46	6095	372
Task 11	4060	711	413	111	419	96	4892	918
Task 12	5223	278	574	75	554	70	6351	423

SIDER

SIDER	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	553	588	64	79	67	76	684	743
Task 2	354	787	34	109	43	100	431	996
Task 3	1125	16	140	3	140	3	1405	22
Task 4	440	701	48	95	63	80	551	876
Task 5	234	907	27	116	15	128	276	1151
Task 6	356	785	43	100	31	112	430	997
Task 7	108	1033	12	131	9	134	129	1298
Task 8	940	201	113	30	123	20	1176	251
Task 9	330	811	35	108	38	105	403	1024
Task 10	580	561	61	82	59	84	700	727
Task 11	839	302	107	36	105	38	1051	376
Task 12	109	1032	10	133	16	127	135	1292
Task 13	874	267	119	24	111	32	1104	323
Task 14	966	175	120	23	128	15	1214	213
Task 15	261	880	27	116	31	112	319	1108
Task 16	431	710	54	89	57	86	542	885
Task 17	99	1042	5	138	5	138	109	1318
Task 18	928	213	121	22	125	18	1174	253
Task 19	347	794	35	108	39	104	421	1006
Task 20	301	840	31	112	35	108	367	1060
Task 21	336	805	33	110	42	101	411	1016
Task 22	412	729	48	95	56	87	516	911
Task 23	1039	102	129	14	134	9	1302	125
Task 24	602	539	77	66	89	54	768	659
Task 25	358	783	37	106	44	99	439	988
Task 26	111	1030	2	141	10	133	123	1304
Task 27	394	747	44	99	43	100	481	946

CLINTOX

CLINTOX	train		val		test		total	
label	0	1	0	1	0	1	0	1
Task 1	80	1104	5	143	9	139	94	1386
Task 2	1088	96	142	6	138	10	1368	112

Appendix B

Downstream Performance

Here we provide additional validation scores for the downstream performances of the epoch chosen for test evaluation.

B.1 Low Sample Count

Downstream Performance, low sample count, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007
pretrained$_{0.5k}$ (default)	76.6 ± 0.2	80.8 ± 0.9	72.3 ± 2.1	93.8 ± 0.2	76.4 ± 0.6	64.0 ± 0.8	60.8 ± 1.3	98.8 ± 0.5	1.135 ± 0.032	1.321 ± 0.116	0.787 ± 0.019
pretrained$_{0.5k}$ (+ motif labels)	75.0 ± 0.7	82.1 ± 1.5	$\mathbf{75.9 \pm 1.8}$	92.8 ± 0.3	77.8 ± 0.0	64.0 ± 0.3	$\mathbf{64.9 \pm 0.4}$	98.3 ± 1.1	1.166 ± 0.074	1.840 ± 0.140	0.781 ± 0.013
pretrained$_{0.25k}$ (+ motif datasets)	75.9 ± 0.8	80.4 ± 0.8	74.5 ± 0.5	92.8 ± 1.0	76.6 ± 0.5	63.4 ± 0.5	60.8 ± 1.3	98.8 ± 1.1	1.162 ± 0.093	$\mathbf{1.309 \pm 0.191}$	0.796 ± 0.018
pretrained$_{0.5k}$ (+ motif datasets)	$\mathbf{77.0 \pm 2.5}$	80.2 ± 2.0	72.8 ± 0.5	93.4 ± 0.9	77.2 ± 0.5	$\mathbf{64.4 \pm 0.2}$	62.9 ± 0.6	$\mathbf{99.0 \pm 0.6}$	1.190 ± 0.050	1.416 ± 0.080	0.784 ± 0.010
pretrained$_{0.5k}$ (only motif datasets)	75.1 ± 1.0	$\mathbf{83.2 \pm 0.9}$	72.5 ± 0.6	$\mathbf{94.1 \pm 0.1}$	$\mathbf{78.3 \pm 0.4}$	64.3 ± 0.6	64.7 ± 0.6	98.7 ± 0.6	$\mathbf{1.116 \pm 0.016}$	2.216 ± 0.320	$\mathbf{0.769 \pm 0.015}$

Downstream Performance, low sample count, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	$\mathbf{2.634 \pm 0.072}$	0.773 ± 0.013
pretrained$_{0.5k}$ (default)	71.2 ± 2.4	73.9 ± 1.1	$\mathbf{80.5 \pm 2.9}$	$\mathbf{68.0 \pm 1.0}$	73.6 ± 0.7	60.3 ± 0.8	60.4 ± 2.0	83.0 ± 1.7	$\mathbf{1.249 \pm 0.073}$	3.328 ± 0.166	0.764 ± 0.020
pretrained$_{0.5k}$ (+ motif labels)	66.3 ± 2.3	73.7 ± 2.3	80.3 ± 2.2	65.7 ± 0.6	74.4 ± 0.4	$\mathbf{63.4 \pm 0.6}$	62.9 ± 0.5	82.2 ± 0.4	1.339 ± 0.074	2.850 ± 0.550	$\mathbf{0.747 \pm 0.021}$
pretrained$_{0.25k}$ (+ motif datasets)	68.3 ± 4.0	74.4 ± 0.4	79.8 ± 2.5	65.2 ± 2.5	72.6 ± 1.1	60.5 ± 1.1	59.9 ± 0.3	$\mathbf{86.5 \pm 2.5}$	1.305 ± 0.201	2.734 ± 0.249	0.759 ± 0.039
pretrained$_{0.5k}$ (+ motif datasets)	73.5 ± 0.8	75.4 ± 0.8	79.4 ± 2.4	67.9 ± 1.9	74.0 ± 0.3	61.2 ± 0.5	61.5 ± 1.8	84.0 ± 1.4	1.365 ± 0.093	2.670 ± 0.247	0.756 ± 0.019
pretrained$_{0.5k}$ (only motif datasets)	$\mathbf{73.7 \pm 3.4}$	$\mathbf{75.5 \pm 1.0}$	79.6 ± 1.1	66.2 ± 2.3	$\mathbf{74.8 \pm 0.9}$	62.7 ± 0.7	$\mathbf{63.2 \pm 1.5}$	85.0 ± 0.3	1.283 ± 0.039	3.018 ± 0.065	0.749 ± 0.023

Table B.1: Validation and Test downstream performance of configurations with low sample count.

B.2 Increased Sample Count

Downstream performance, increased sample count, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	$\mathbf{1.723 \pm 0.141}$	0.804 ± 0.007
pretrained_{3k} (default)	$\mathbf{76.2 \pm 1.9}$	81.9 ± 0.1	72.3 ± 0.4	$\mathbf{94.0 \pm 0.9}$	76.6 ± 0.5	64.6 ± 0.5	64.0 ± 1.7	98.4 ± 0.8	1.091 ± 0.052	1.987 ± 0.211	0.788 ± 0.020
pretrained_{3k} (+ motif labels)	76.1 ± 0.3	$\mathbf{82.7 \pm 0.3}$	74.4 ± 1.6	93.8 ± 0.4	$\mathbf{79.7 \pm 0.5}$	65.7 ± 0.6	$\mathbf{66.8 \pm 0.8}$	98.7 ± 0.4	1.095 ± 0.017	2.647 ± 0.150	$\mathbf{0.750 \pm 0.009}$
$\text{pretrained}_{1.5k}$ (+ motif datasets)	75.9 ± 1.5	82.2 ± 0.8	71.0 ± 1.0	93.5 ± 0.8	78.1 ± 0.5	64.6 ± 0.3	63.2 ± 2.0	$\mathbf{98.9 \pm 0.1}$	1.088 ± 0.026	1.879 ± 0.065	0.790 ± 0.009
pretrained_{3k} (+ motif datasets)	75.9 ± 0.8	80.7 ± 2.2	76.4 ± 2.5	93.8 ± 0.3	78.6 ± 0.4	65.3 ± 0.9	65.7 ± 1.1	98.4 ± 0.9	1.072 ± 0.036	2.414 ± 0.240	0.767 ± 0.008
pretrained_{3k} (only motif datasets)	75.6 ± 1.6	81.4 ± 1.3	$\mathbf{77.8 \pm 0.4}$	93.8 ± 0.6	79.3 ± 0.4	$\mathbf{65.9 \pm 0.4}$	66.4 ± 0.6	98.5 ± 0.6	$\mathbf{1.083 \pm 0.026}$	3.205 ± 0.256	0.770 ± 0.004

Downstream performance, increased sample count, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	$\mathbf{2.634 \pm 0.072}$	0.773 ± 0.013
pretrained_{3k} (default)	69.4 ± 1.5	75.3 ± 1.2	77.4 ± 4.0	66.5 ± 1.4	73.6 ± 1.0	61.9 ± 0.8	60.9 ± 1.9	80.3 ± 2.5	1.306 ± 0.086	2.654 ± 0.396	0.750 ± 0.009
pretrained_{3k} (+ motif labels)	$\mathbf{72.4 \pm 5.5}$	74.2 ± 0.8	80.7 ± 2.6	68.3 ± 0.8	76.3 ± 0.5	$\mathbf{64.1 \pm 0.4}$	61.2 ± 1.3	84.4 ± 1.4	$\mathbf{1.216 \pm 0.061}$	2.995 ± 0.406	$\mathbf{0.724 \pm 0.019}$
$\text{pretrained}_{1.5k}$ (+ motif datasets)	67.6 ± 2.6	74.7 ± 1.7	76.6 ± 2.7	65.1 ± 1.0	74.8 ± 0.8	61.3 ± 0.4	60.6 ± 0.7	83.5 ± 2.2	1.295 ± 0.036	2.929 ± 0.589	0.742 ± 0.001
pretrained_{3k} (+ motif datasets)	72.5 ± 4.0	74.3 ± 1.2	$\mathbf{82.1 \pm 2.7}$	69.2 ± 1.0	74.5 ± 0.5	63.6 ± 0.2	$\mathbf{61.3 \pm 1.5}$	84.9 ± 3.7	1.238 ± 0.055	2.644 ± 0.197	0.734 ± 0.014
pretrained_{3k} (only motif datasets)	70.7 ± 3.6	$\mathbf{76.2 \pm 1.8}$	79.7 ± 0.9	$\mathbf{68.7 \pm 2.4}$	$\mathbf{76.7 \pm 1.1}$	63.4 ± 0.2	60.8 ± 2.1	$\mathbf{86.3 \pm 2.2}$	1.303 ± 0.063	2.761 ± 0.087	0.741 ± 0.023

Table B.2: Validation and Test downstream performance of configurations with high sample count.

Appendix C

Pretraining

In this chapter of the Appendix we provide detailed pretraining tables, including validation and test performance, for all pretraining configurations. The chapter is divided into the different pretraining methods, which are sub-divided into the different sample sizes used for the corresponding method.

C.1 Pretraining Overview

Multi-Dataset-Pretraining performance, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007
$\text{pretrained}_{0.5k}$ (default)	52.2 ± 4.3	69.1 ± 1.6	63.8 ± 3.5	90.3 ± 1.1	71.3 ± 0.6	60.8 ± 1.2	56.0 ± 0.5	83.0 ± 4.9	1.278 ± 0.086	4.328 ± 0.331	1.102 ± 0.041
pretrained_{3k} (default)	59.2 ± 7.3	71.7 ± 1.9	69.5 ± 3.4	91.3 ± 0.5	72.5 ± 1.6	62.6 ± 1.3	57.3 ± 2.1	87.7 ± 3.5	1.152 ± 0.094	4.841 ± 0.285	0.965 ± 0.091
$\text{pretrained}_{0.25k}$ (+ motif datasets)	49.2 ± 1.9	69.4 ± 1.0	64.0 ± 3.8	91.0 ± 0.9	69.0 ± 1.4	60.4 ± 1.0	56.8 ± 0.4	82.8 ± 3.6	1.301 ± 0.076	4.276 ± 0.361	1.136 ± 0.070
$\text{pretrained}_{0.5k}$ (+ motif datasets)	51.9 ± 3.4	71.0 ± 1.7	67.8 ± 2.0	91.2 ± 1.0	70.4 ± 1.2	61.8 ± 0.8	56.6 ± 0.8	88.0 ± 2.8	1.172 ± 0.040	4.789 ± 0.160	1.050 ± 0.054
$\text{pretrained}_{1.5k}$ (+ motif datasets)	56.1 ± 6.5	73.3 ± 0.9	72.8 ± 2.3	92.3 ± 0.6	72.2 ± 1.7	64.0 ± 0.7	56.9 ± 1.6	94.1 ± 1.7	1.074 ± 0.043	5.464 ± 0.174	0.904 ± 0.047
pretrained_{3k} (+ motif datasets)	69.4 ± 2.5	75.9 ± 1.0	73.6 ± 1.5	92.6 ± 0.5	76.2 ± 1.0	64.9 ± 0.8	61.4 ± 1.8	92.2 ± 2.2	0.996 ± 0.026	5.447 ± 0.136	0.858 ± 0.031

Multi-Dataset-Pretraining performance, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013
$\text{pretrained}_{0.5k}$ (default)	48.1 ± 2.4	71.2 ± 1.6	75.3 ± 4.6	60.4 ± 1.6	68.7 ± 1.3	60.6 ± 0.8	54.9 ± 1.1	65.9 ± 3.6	1.394 ± 0.097	3.983 ± 0.218	1.004 ± 0.029
pretrained_{3k} (default)	53.7 ± 5.9	70.6 ± 2.8	78.7 ± 1.8	64.5 ± 3.1	71.1 ± 1.4	61.5 ± 1.0	59.6 ± 2.4	76.4 ± 7.7	1.241 ± 0.096	4.344 ± 0.234	0.893 ± 0.076
$\text{pretrained}_{0.25k}$ (+ motif datasets)	45.5 ± 2.1	69.7 ± 1.6	76.4 ± 3.0	59.9 ± 1.3	67.6 ± 1.3	59.8 ± 0.8	55.9 ± 1.0	65.2 ± 1.3	1.396 ± 0.082	3.998 ± 0.305	1.029 ± 0.067
$\text{pretrained}_{0.5k}$ (+ motif datasets)	48.5 ± 3.8	70.8 ± 1.6	80.2 ± 2.1	61.7 ± 1.7	68.5 ± 1.3	61.0 ± 0.7	57.4 ± 1.7	67.6 ± 5.5	1.255 ± 0.026	4.250 ± 0.235	0.960 ± 0.056
$\text{pretrained}_{1.5k}$ (+ motif datasets)	58.3 ± 5.9	71.7 ± 1.0	80.9 ± 1.3	65.9 ± 1.6	71.0 ± 1.2	62.3 ± 0.7	62.2 ± 1.0	84.1 ± 2.9	1.201 ± 0.028	4.499 ± 0.076	0.842 ± 0.036
pretrained_{3k} (+ motif datasets)	70.9 ± 2.2	72.3 ± 1.2	80.2 ± 0.7	67.5 ± 1.7	74.0 ± 0.9	64.1 ± 0.5	63.3 ± 0.8	85.3 ± 3.4	1.101 ± 0.025	4.646 ± 0.156	0.800 ± 0.030

Table C.1: Validation and Test performance on default datasets during pretraining.

C.2 Default

C.2.1 500 samples

Pretraining Default, 500 samples, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
left out dataset											
muv		70.1 ± 3.0	63.1 ± 4.4	90.7 ± 0.8	72.3 ± 1.0	61.4 ± 1.2	56.0 ± 0.6	74.8 ± 4.4	1.179 ± 0.038	4.547 ± 0.124	1.076 ± 0.081
hiv	51.3 ± 2.0		58.1 ± 12.2	88.2 ± 2.6	70.4 ± 1.0	60.1 ± 1.2	55.2 ± 0.8	77.7 ± 8.6	1.372 ± 0.076	3.638 ± 0.586	1.177 ± 0.004
bace	51.5 ± 9.4	65.4 ± 3.6		89.3 ± 1.6	72.1 ± 0.9	60.4 ± 0.4	55.3 ± 2.2	77.2 ± 3.3	1.170 ± 0.067	4.243 ± 0.542	1.077 ± 0.047
bbbp	55.1 ± 1.4	69.9 ± 1.4	65.3 ± 2.7		71.7 ± 1.0	61.4 ± 0.9	56.4 ± 0.8	84.1 ± 2.1	1.325 ± 0.139	4.594 ± 0.244	1.094 ± 0.009
tox21	47.1 ± 4.5	70.6 ± 1.8	67.9 ± 1.5	90.8 ± 0.4		62.4 ± 0.6	55.7 ± 1.0	84.3 ± 2.5	1.274 ± 0.095	4.585 ± 0.190	1.058 ± 0.094
toxcast	51.4 ± 2.0	68.2 ± 1.6	62.3 ± 2.1	90.4 ± 0.9	71.6 ± 0.8		56.2 ± 1.1	87.3 ± 1.2	1.258 ± 0.075	4.426 ± 0.405	1.100 ± 0.051
sider	52.7 ± 4.5	70.0 ± 1.6	62.9 ± 4.9	91.4 ± 0.7	71.4 ± 0.6	61.2 ± 0.6		84.5 ± 1.5	1.197 ± 0.084	4.475 ± 0.345	1.144 ± 0.081
clintox	48.7 ± 8.7	69.6 ± 2.0	64.7 ± 4.2	90.3 ± 1.2	70.9 ± 2.3	59.3 ± 2.1	57.0 ± 0.8		1.337 ± 0.102	4.376 ± 0.554	1.082 ± 0.108
esol	53.0 ± 11.1	68.5 ± 3.7	66.7 ± 1.7	89.2 ± 1.2	70.7 ± 0.4	58.9 ± 1.4	55.9 ± 0.8	83.6 ± 1.3		3.844 ± 0.142	1.151 ± 0.067
freesolv	62.4 ± 7.3	70.6 ± 1.0	68.0 ± 0.5	92.0 ± 0.2	71.4 ± 1.3	62.6 ± 1.7	56.1 ± 1.0	90.4 ± 0.9	1.426 ± 0.176		1.064 ± 0.037
lipo	48.6 ± 7.3	68.3 ± 0.5	58.6 ± 3.7	90.7 ± 0.8	70.8 ± 0.2	60.2 ± 0.6	55.7 ± 0.4	85.8 ± 2.8	1.239 ± 0.052	4.555 ± 0.116	
mean	52.2 ± 4.3	69.1 ± 1.6	63.8 ± 3.5	90.3 ± 1.1	71.3 ± 0.6	60.8 ± 1.2	56.0 ± 0.5	83.0 ± 4.9	1.278 ± 0.086	4.328 ± 0.331	1.102 ± 0.041
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Default, 500 samples, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
left out dataset											
muv		72.2 ± 0.7	78.4 ± 1.7	60.8 ± 1.4	69.7 ± 0.3	61.2 ± 0.5	56.4 ± 0.5	64.8 ± 6.3	1.266 ± 0.032	4.085 ± 0.108	0.984 ± 0.069
hiv	48.6 ± 1.1		64.9 ± 8.1	58.3 ± 2.3	67.1 ± 1.0	60.2 ± 0.6	53.4 ± 1.4	62.9 ± 2.0	1.519 ± 0.092	3.672 ± 0.369	1.057 ± 0.021
bace	46.2 ± 4.4	68.9 ± 2.1		60.1 ± 1.5	69.3 ± 0.7	60.4 ± 0.7	54.1 ± 0.3	62.2 ± 1.8	1.284 ± 0.068	3.983 ± 0.414	0.976 ± 0.038
bbbp	46.4 ± 1.3	72.7 ± 1.0	77.8 ± 2.0		69.9 ± 0.2	61.0 ± 1.0	54.9 ± 1.9	64.7 ± 3.4	1.454 ± 0.137	3.737 ± 0.265	1.006 ± 0.011
tox21	45.0 ± 1.9	72.7 ± 0.9	76.6 ± 4.0	61.7 ± 0.9		61.9 ± 0.0	55.8 ± 1.7	63.0 ± 0.9	1.372 ± 0.160	4.110 ± 0.317	0.972 ± 0.064
toxcast	47.8 ± 3.1	70.7 ± 1.1	79.0 ± 1.9	59.4 ± 0.9	69.0 ± 0.4		54.0 ± 0.8	67.9 ± 1.8	1.396 ± 0.041	4.128 ± 0.213	1.002 ± 0.023
sider	48.6 ± 2.0	72.2 ± 1.3	75.9 ± 8.4	59.7 ± 2.9	68.7 ± 0.2	61.1 ± 1.0		67.9 ± 3.2	1.286 ± 0.061	3.893 ± 0.399	1.027 ± 0.065
clintox	48.3 ± 4.6	71.3 ± 1.7	69.7 ± 18.0	59.0 ± 3.2	67.9 ± 2.6	59.2 ± 2.3	54.1 ± 0.6		1.463 ± 0.133	4.003 ± 0.478	0.987 ± 0.089
esol	53.5 ± 4.8	68.3 ± 2.7	77.2 ± 2.6	61.5 ± 2.5	66.8 ± 0.0	60.0 ± 0.6	56.1 ± 0.5	66.0 ± 3.1		3.808 ± 0.189	1.043 ± 0.054
freesolv	50.3 ± 6.6	70.5 ± 0.4	79.5 ± 3.1	63.7 ± 1.4	70.9 ± 0.3	60.2 ± 0.6	56.3 ± 0.9	74.5 ± 3.4	1.533 ± 0.199		0.987 ± 0.049
lipo	46.3 ± 3.1	72.1 ± 0.9	74.0 ± 4.1	59.4 ± 1.5	67.6 ± 0.3	60.4 ± 0.6	53.9 ± 1.2	65.2 ± 2.0	1.367 ± 0.061	4.413 ± 0.189	
mean	48.1 ± 2.4	71.2 ± 1.6	75.3 ± 4.6	60.4 ± 1.6	68.7 ± 1.3	60.6 ± 0.8	54.9 ± 1.1	65.9 ± 3.6	1.394 ± 0.097	3.983 ± 0.218	1.004 ± 0.029
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Table C.2: Validation and Test performance during default pretraining with 500 samples.

C.2.2 3000 samples

Pretraining Default, 3000 samples, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
left out dataset											
muv		71.0 ± 3.2	71.6 ± 2.2	91.1 ± 0.9	72.8 ± 0.9	63.2 ± 1.3	57.3 ± 2.4	89.6 ± 2.0	1.080 ± 0.048	4.793 ± 0.612	1.001 ± 0.137
hiv	60.2 ± 4.1		70.2 ± 1.6	91.8 ± 0.9	73.5 ± 1.1	63.7 ± 0.7	58.4 ± 1.9	92.9 ± 2.6	1.092 ± 0.073	5.147 ± 0.170	0.903 ± 0.071
bace	58.6 ± 5.0	71.1 ± 2.1		91.7 ± 0.3	72.2 ± 1.9	63.8 ± 0.5	56.8 ± 0.4	90.9 ± 3.5	1.099 ± 0.018	5.085 ± 0.139	0.913 ± 0.053
bbbp	64.5 ± 4.7	70.7 ± 4.2	72.4 ± 6.3		72.5 ± 0.7	63.4 ± 1.1	56.7 ± 0.9	84.6 ± 1.6	1.090 ± 0.068	5.023 ± 0.349	0.957 ± 0.099
tox21	46.1 ± 11.1	67.5 ± 4.8	62.0 ± 4.3	91.2 ± 1.7		59.9 ± 2.8	56.4 ± 0.9	83.4 ± 1.6	1.312 ± 0.057	4.369 ± 0.507	1.167 ± 0.077
toxcast	54.3 ± 4.1	74.0 ± 1.6	73.9 ± 0.4	91.2 ± 0.6	72.9 ± 0.4		57.4 ± 0.2	89.6 ± 4.6	1.120 ± 0.036	5.241 ± 0.387	0.903 ± 0.082
sider	56.5 ± 1.8	71.1 ± 1.4	70.0 ± 0.6	91.0 ± 0.9	71.0 ± 1.7	62.7 ± 0.4		83.0 ± 5.3	1.318 ± 0.164	4.525 ± 0.295	0.940 ± 0.043
clintox	60.2 ± 4.3	72.7 ± 1.4	70.7 ± 1.0	91.0 ± 1.0	72.7 ± 0.9	62.6 ± 0.2	55.5 ± 0.9		1.198 ± 0.155	4.698 ± 0.820	0.966 ± 0.171
esol	65.7 ± 5.2	73.5 ± 0.8	68.8 ± 2.7	90.7 ± 0.7	72.4 ± 1.1	61.2 ± 0.3	56.1 ± 0.3	90.7 ± 4.4		4.646 ± 0.203	1.050 ± 0.143
freesolv	72.4 ± 1.7	73.3 ± 1.8	70.3 ± 2.7	92.5 ± 0.3	75.8 ± 0.6	64.1 ± 0.9	62.8 ± 1.5	86.8 ± 2.4	1.063 ± 0.021		0.848 ± 0.030
lipo	53.4 ± 8.2	71.9 ± 2.5	65.6 ± 5.0	91.2 ± 0.7	69.6 ± 1.4	61.5 ± 0.1	55.3 ± 1.6	85.4 ± 3.8	1.144 ± 0.021	4.884 ± 0.306	
mean	59.2 ± 7.3	71.7 ± 1.9	69.5 ± 3.4	91.3 ± 0.5	72.5 ± 1.6	62.6 ± 1.3	57.3 ± 2.1	87.7 ± 3.5	1.152 ± 0.094	4.841 ± 0.285	0.965 ± 0.091
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Default, 3000 samples, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
left out dataset											
muv		70.7 ± 0.9	81.1 ± 1.9	63.6 ± 2.1	72.0 ± 1.7	62.3 ± 0.4	60.6 ± 1.4	78.3 ± 7.6	1.173 ± 0.057	4.386 ± 0.101	0.907 ± 0.118
hiv	57.8 ± 2.7		79.1 ± 3.2	64.7 ± 0.8	71.8 ± 1.3	61.9 ± 0.8	60.7 ± 3.1	78.4 ± 8.5	1.204 ± 0.076	4.275 ± 0.113	0.828 ± 0.061
bace	52.4 ± 2.8	68.8 ± 2.7		66.9 ± 1.4	71.5 ± 1.1	61.9 ± 0.4	60.3 ± 1.7	80.7 ± 1.2	1.182 ± 0.040	4.755 ± 0.147	0.872 ± 0.041
bbbp	56.9 ± 2.2	69.1 ± 2.3	79.6 ± 0.5		70.9 ± 1.1	61.7 ± 1.6	60.0 ± 3.4	74.0 ± 6.1	1.166 ± 0.066	4.412 ± 0.145	0.888 ± 0.081
tox21	46.2 ± 5.8	64.1 ± 10.2	74.9 ± 3.4	59.1 ± 2.0		59.6 ± 2.3	55.5 ± 1.6	65.4 ± 2.7	1.409 ± 0.064	4.184 ± 0.281	1.052 ± 0.039
toxcast	48.3 ± 3.4	70.7 ± 1.8	78.1 ± 1.0	65.4 ± 0.8	71.6 ± 1.4		60.6 ± 0.5	83.9 ± 1.0	1.197 ± 0.081	4.625 ± 0.339	0.843 ± 0.049
sider	53.9 ± 5.7	70.8 ± 1.4	80.5 ± 1.3	60.5 ± 2.0	70.0 ± 2.3	61.4 ± 0.6		67.3 ± 14.0	1.421 ± 0.185	3.915 ± 0.471	0.883 ± 0.038
clintox	56.2 ± 5.4	73.4 ± 0.9	79.9 ± 2.9	64.9 ± 2.6	71.2 ± 0.5	61.4 ± 0.2	57.3 ± 1.7		1.261 ± 0.192	4.398 ± 0.529	0.900 ± 0.137
esol	55.7 ± 4.8	71.9 ± 1.5	78.4 ± 1.6	65.1 ± 3.0	70.1 ± 2.3	60.8 ± 0.2	58.6 ± 0.8	78.2 ± 9.6		4.214 ± 0.120	0.972 ± 0.135
freesolv	64.4 ± 3.5	73.8 ± 1.9	78.5 ± 1.9	70.5 ± 1.4	73.6 ± 0.4	63.1 ± 0.4	64.4 ± 1.0	89.4 ± 2.5	1.171 ± 0.056		0.781 ± 0.043
lipo	44.8 ± 4.0	72.8 ± 1.1	76.8 ± 2.3	64.6 ± 3.3	68.4 ± 1.8	60.8 ± 0.7	58.1 ± 3.2	68.3 ± 5.4	1.230 ± 0.072	4.272 ± 0.123	
mean	53.7 ± 5.9	70.6 ± 2.8	78.7 ± 1.8	64.5 ± 3.1	71.1 ± 1.4	61.5 ± 1.0	59.6 ± 2.4	76.4 ± 7.7	1.241 ± 0.096	4.344 ± 0.234	0.893 ± 0.076
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Table C.3: Validation and Test performance during default pretraining with 3000 samples.

C.3 Additional Motif Labels

C.3.1 500 samples

Pretraining Additional Motif Labels, 500 samples, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV+	HIV+	BACE+	BBBP+	TOX21+	TOXCAST+	SIDER+	CLINTOX+	ESOL+	FREESOLV+	LIPO+
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	(17+85)	(1+85)	(1+85)	(1+85)	(12+85)	(617+85)	(27+85)	(2+85)	(1+85)	(1+85)	(1+85)
left out dataset											
muv		94.0 ± 0.4	96.5 ± 0.1	94.0 ± 0.2	92.7 ± 0.1	67.2 ± 1.0	86.5 ± 0.4	96.8 ± 0.2	0.126 ± 0.002	0.154 ± 0.004	0.158 ± 0.003
hiv	91.9 ± 0.9		96.5 ± 0.3	94.9 ± 0.0	93.0 ± 0.2	67.3 ± 0.4	87.0 ± 0.3	96.8 ± 0.3	0.125 ± 0.002	0.156 ± 0.004	0.156 ± 0.002
bace	90.4 ± 0.3	93.8 ± 0.1		94.1 ± 0.2	92.9 ± 0.3	67.3 ± 0.1	86.2 ± 0.4	96.5 ± 0.5	0.124 ± 0.002	0.152 ± 0.003	0.159 ± 0.002
bbbp	91.4 ± 0.3	93.7 ± 0.1	96.5 ± 0.4		92.6 ± 0.2	67.1 ± 0.4	86.0 ± 0.4	96.3 ± 0.5	0.127 ± 0.002	0.155 ± 0.006	0.156 ± 0.002
tox21	90.6 ± 0.3	93.8 ± 0.5	96.4 ± 0.3	94.5 ± 0.2		66.9 ± 0.6	86.5 ± 0.2	96.9 ± 0.1	0.128 ± 0.002	0.153 ± 0.001	0.154 ± 0.004
toxcast	92.2 ± 0.4	95.5 ± 0.1	98.2 ± 0.2	97.0 ± 0.1	94.2 ± 0.1		88.2 ± 0.3	98.0 ± 0.1	0.122 ± 0.001	0.157 ± 0.003	0.146 ± 0.004
sider	91.1 ± 0.9	93.9 ± 0.2	96.9 ± 0.3	94.2 ± 0.6	92.5 ± 0.3	66.6 ± 0.2		96.7 ± 0.4	0.128 ± 0.003	0.154 ± 0.003	0.161 ± 0.003
clintox	90.1 ± 0.1	93.8 ± 0.2	96.5 ± 0.1	93.6 ± 0.4	92.5 ± 0.5	67.1 ± 0.5	85.8 ± 0.6		0.125 ± 0.001	0.156 ± 0.001	0.155 ± 0.006
esol	90.7 ± 0.4	94.0 ± 0.2	97.0 ± 0.5	94.9 ± 0.1	93.2 ± 0.1	66.7 ± 0.6	86.7 ± 0.3	97.3 ± 0.3		0.156 ± 0.003	0.152 ± 0.002
freesolv	91.0 ± 0.7	94.4 ± 0.2	97.2 ± 0.1	95.2 ± 0.2	93.3 ± 0.5	66.9 ± 0.4	87.5 ± 0.2	97.1 ± 0.4	0.128 ± 0.004		0.150 ± 0.003
lipo	91.0 ± 0.9	93.8 ± 0.2	96.5 ± 0.4	94.0 ± 0.3	92.5 ± 0.1	66.4 ± 0.7	86.3 ± 0.1	96.8 ± 0.3	0.130 ± 0.004	0.156 ± 0.003	
mean	91.0 ± 0.7	94.1 ± 0.5	96.8 ± 0.6	94.6 ± 1.0	92.9 ± 0.5	67.0 ± 0.3	86.7 ± 0.7	96.9 ± 0.5	0.126 ± 0.002	0.155 ± 0.002	0.155 ± 0.004

Pretraining Additional Motif Labels, 500 samples, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV+	HIV+	BACE+	BBBP+	TOX21+	TOXCAST+	SIDER+	CLINTOX+	ESOL+	FREESOLV+	LIPO+
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	(17+85)	(1+85)	(1+85)	(1+85)	(12+85)	(617+85)	(27+85)	(2+85)	(1+85)	(1+85)	(1+85)
left out dataset											
muv		92.9 ± 0.3	95.5 ± 0.2	95.9 ± 0.5	92.7 ± 0.2	65.9 ± 0.8	86.0 ± 0.4	91.9 ± 0.7	0.135 ± 0.001	0.162 ± 0.001	0.154 ± 0.003
hiv	90.9 ± 1.0		95.8 ± 0.2	95.7 ± 0.2	93.0 ± 0.2	65.1 ± 0.5	86.5 ± 0.6	92.7 ± 0.4	0.136 ± 0.001	0.164 ± 0.004	0.154 ± 0.001
bace	90.4 ± 0.6	93.1 ± 0.4		95.7 ± 0.1	93.1 ± 0.2	65.1 ± 0.5	85.9 ± 0.5	92.1 ± 0.2	0.133 ± 0.002	0.164 ± 0.003	0.156 ± 0.001
bbbp	90.4 ± 0.4	92.9 ± 0.6	95.6 ± 0.5		92.6 ± 0.0	64.9 ± 0.6	85.3 ± 0.2	92.6 ± 0.1	0.137 ± 0.002	0.165 ± 0.003	0.153 ± 0.003
tox21	90.6 ± 0.4	93.1 ± 0.5	96.0 ± 0.1	95.4 ± 0.3		65.5 ± 0.7	86.5 ± 0.4	91.9 ± 0.1	0.137 ± 0.002	0.167 ± 0.003	0.152 ± 0.004
toxcast	91.8 ± 0.5	95.0 ± 0.3	97.5 ± 0.2	97.1 ± 0.0	94.2 ± 0.2		87.4 ± 0.2	94.7 ± 0.1	0.131 ± 0.001	0.169 ± 0.005	0.145 ± 0.004
sider	90.9 ± 0.7	93.6 ± 0.3	96.0 ± 0.3	95.4 ± 0.6	92.7 ± 0.1	65.8 ± 0.1		92.5 ± 0.5	0.136 ± 0.001	0.165 ± 0.003	0.159 ± 0.004
clintox	89.6 ± 0.5	93.4 ± 0.1	96.0 ± 0.2	95.5 ± 0.3	92.6 ± 0.2	65.4 ± 0.5	86.1 ± 0.7		0.134 ± 0.000	0.165 ± 0.001	0.153 ± 0.007
esol	92.0 ± 0.8	93.7 ± 0.3	96.3 ± 0.1	95.8 ± 0.6	93.4 ± 0.1	65.2 ± 0.9	86.3 ± 0.2	92.4 ± 0.3		0.170 ± 0.001	0.150 ± 0.002
freesolv	90.8 ± 0.1	93.7 ± 0.4	96.2 ± 0.4	96.1 ± 0.6	93.5 ± 0.2	65.3 ± 0.4	87.3 ± 0.2	93.4 ± 0.8	0.138 ± 0.002		0.148 ± 0.002
lipo	90.6 ± 0.8	93.3 ± 0.0	95.3 ± 0.4	95.6 ± 0.4	92.5 ± 0.3	64.7 ± 0.8	86.3 ± 0.1	92.7 ± 0.2	0.139 ± 0.002	0.163 ± 0.001	
mean	90.8 ± 0.7	93.5 ± 0.6	96.0 ± 0.6	95.8 ± 0.5	93.0 ± 0.5	65.3 ± 0.4	86.4 ± 0.6	92.7 ± 0.8	0.136 ± 0.002	0.165 ± 0.003	0.152 ± 0.004

Table C.4: Validation and Test Performance during pretraining with additional motif labels and 500 samples.

C.3.2 3000 samples

Pretraining Additional Motif Labels, 3000 samples, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV+	HIV+	BACE+	BBBP+	TOX21+	TOXCAST+	SIDER+	CLINTOX+	ESOL+	FREESOLV+	LIPO+
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	(17+85)	(1+85)	(1+85)	(1+85)	(12+85)	(617+85)	(27+85)	(2+85)	(1+85)	(1+85)	(1+85)
left out dataset											
muv		98.6 ± 0.1	99.2 ± 0.0	98.6 ± 0.1	96.5 ± 0.0	70.9 ± 0.2	89.3 ± 0.4	99.3 ± 0.1	0.101 ± 0.002	0.145 ± 0.003	0.131 ± 0.001
hiv	94.9 ± 0.1		99.0 ± 0.0	98.5 ± 0.1	96.4 ± 0.0	71.2 ± 0.2	89.4 ± 0.1	99.1 ± 0.2	0.101 ± 0.000	0.142 ± 0.004	0.129 ± 0.001
bace	94.3 ± 0.4	98.7 ± 0.1		99.0 ± 0.1	96.6 ± 0.1	70.8 ± 0.4	89.6 ± 0.3	99.1 ± 0.2	0.099 ± 0.002	0.145 ± 0.005	0.128 ± 0.003
bbbp	94.6 ± 0.5	98.5 ± 0.1	99.2 ± 0.1		96.6 ± 0.1	70.4 ± 0.3	89.3 ± 0.1	99.1 ± 0.2	0.098 ± 0.001	0.146 ± 0.001	0.127 ± 0.003
tox21	94.6 ± 0.3	98.7 ± 0.0	99.2 ± 0.1	99.0 ± 0.5		69.8 ± 0.3	89.6 ± 0.1	99.1 ± 0.2	0.102 ± 0.000	0.148 ± 0.003	0.124 ± 0.002
toxcast	94.6 ± 0.3	99.0 ± 0.1	99.3 ± 0.1	99.4 ± 0.2	96.8 ± 0.0		89.8 ± 0.1	99.4 ± 0.1	0.109 ± 0.005	0.144 ± 0.001	0.116 ± 0.004
sider	95.2 ± 0.5	98.8 ± 0.0	99.3 ± 0.0	99.5 ± 0.1	96.8 ± 0.1	70.7 ± 0.2		99.1 ± 0.1	0.105 ± 0.002	0.145 ± 0.001	0.121 ± 0.002
clintox	94.8 ± 0.2	98.6 ± 0.1	99.2 ± 0.1	98.9 ± 0.3	96.7 ± 0.2	71.0 ± 0.2	89.7 ± 0.1		0.099 ± 0.001	0.142 ± 0.003	0.127 ± 0.002
esol	94.5 ± 0.5	98.6 ± 0.1	99.1 ± 0.0	98.7 ± 0.1	96.6 ± 0.1	69.8 ± 0.1	89.6 ± 0.4	99.2 ± 0.1		0.148 ± 0.002	0.127 ± 0.001
freesolv	94.6 ± 0.5	98.7 ± 0.1	99.1 ± 0.0	99.0 ± 0.1	96.6 ± 0.0	70.2 ± 0.3	89.8 ± 0.2	99.2 ± 0.1	0.102 ± 0.001		0.125 ± 0.001
lipo	94.8 ± 0.4	98.7 ± 0.2	99.0 ± 0.1	99.0 ± 0.4	96.7 ± 0.0	70.5 ± 0.7	89.6 ± 0.3	99.0 ± 0.1	0.106 ± 0.000	0.143 ± 0.004	
mean	94.7 ± 0.2	98.7 ± 0.1	99.2 ± 0.1	99.0 ± 0.3	96.6 ± 0.1	70.5 ± 0.5	89.6 ± 0.2	99.2 ± 0.1	0.102 ± 0.003	0.145 ± 0.002	0.126 ± 0.004

Pretraining Additional Motif Labels, 3000 samples, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV+	HIV+	BACE+	BBBP+	TOX21+	TOXCAST+	SIDER+	CLINTOX+	ESOL+	FREESOLV+	LIPO+
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	(17+85)	(1+85)	(1+85)	(1+85)	(12+85)	(617+85)	(27+85)	(2+85)	(1+85)	(1+85)	(1+85)
left out dataset											
muv		98.1 ± 0.1	99.1 ± 0.1	98.5 ± 0.2	96.3 ± 0.1	68.1 ± 0.1	88.7 ± 0.2	97.8 ± 0.3	0.108 ± 0.002	0.147 ± 0.002	0.125 ± 0.001
hiv	95.0 ± 0.4		98.8 ± 0.1	98.7 ± 0.1	96.1 ± 0.1	67.9 ± 0.3	88.8 ± 0.2	97.9 ± 0.1	0.109 ± 0.001	0.147 ± 0.001	0.122 ± 0.001
bace	95.4 ± 0.2	98.1 ± 0.1		98.4 ± 0.1	96.2 ± 0.1	68.4 ± 0.3	88.4 ± 0.1	98.3 ± 0.2	0.107 ± 0.002	0.147 ± 0.003	0.121 ± 0.002
bbbp	95.3 ± 0.2	98.2 ± 0.1	99.1 ± 0.1		96.3 ± 0.1	67.9 ± 0.3	88.2 ± 0.1	98.2 ± 0.1	0.109 ± 0.001	0.148 ± 0.001	0.120 ± 0.003
tox21	95.2 ± 0.3	97.9 ± 0.2	99.0 ± 0.1	98.3 ± 0.2		67.7 ± 0.2	88.2 ± 0.4	98.4 ± 0.1	0.110 ± 0.000	0.151 ± 0.001	0.118 ± 0.002
toxcast	95.3 ± 0.2	98.6 ± 0.1	99.4 ± 0.1	99.1 ± 0.3	96.4 ± 0.1		88.3 ± 0.1	99.0 ± 0.1	0.111 ± 0.001	0.150 ± 0.002	0.109 ± 0.003
sider	94.6 ± 0.2	98.2 ± 0.0	99.3 ± 0.0	98.6 ± 0.2	96.4 ± 0.0	68.0 ± 0.3		98.8 ± 0.3	0.111 ± 0.000	0.156 ± 0.001	0.116 ± 0.002
clintox	95.6 ± 0.1	98.4 ± 0.1	98.8 ± 0.2	98.7 ± 0.2	96.3 ± 0.0	68.0 ± 0.5	88.7 ± 0.2		0.108 ± 0.001	0.149 ± 0.001	0.120 ± 0.002
esol	95.2 ± 0.1	98.3 ± 0.0	99.0 ± 0.1	98.7 ± 0.1	96.2 ± 0.2	68.0 ± 0.3	88.2 ± 0.3	98.1 ± 0.4		0.153 ± 0.001	0.119 ± 0.001
freesolv	95.4 ± 0.2	98.0 ± 0.2	99.1 ± 0.0	98.4 ± 0.2	96.4 ± 0.2	67.7 ± 0.3	88.4 ± 0.2	98.1 ± 0.2	0.111 ± 0.002		0.118 ± 0.002
lipo	95.2 ± 0.4	98.1 ± 0.2	98.9 ± 0.0	98.4 ± 0.2	96.3 ± 0.0	68.4 ± 0.3	88.4 ± 0.2	98.1 ± 0.3	0.114 ± 0.001	0.146 ± 0.003	
mean	95.2 ± 0.3	98.2 ± 0.2	99.0 ± 0.2	98.6 ± 0.2	96.3 ± 0.1	68.0 ± 0.2	88.4 ± 0.2	98.3 ± 0.4	0.110 ± 0.002	0.149 ± 0.003	0.119 ± 0.004

Table C.5: Validation and Test Performance during pretraining with additional motif labels and 3000 samples.

C.4 Additional Motif Datasets

In order to fit the whole table into one page we chose a to not give additional dataset information like metric, number of compounds and tasks in this table. The downwards pointing arrow in the left most column of the second row indicates that the left most column specifies the left out dataset so it has exceptionally no meaning of order.

C.4.1 250 samples

Pretraining Additional Motif Datasets, 250 samples, Validation											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		69.4 ± 1.6	60.0 ± 3.3	89.9 ± 0.7	66.3 ± 3.6	58.8 ± 2.1	56.6 ± 0.5	82.7 ± 1.7	1.236 ± 0.058	3.620 ± 0.455	1.276 ± 0.040
hiv	46.5 ± 7.2		62.0 ± 6.5	90.6 ± 0.7	68.9 ± 0.5	60.0 ± 1.1	57.2 ± 1.6	79.7 ± 0.9	1.412 ± 0.271	3.781 ± 0.585	1.136 ± 0.064
bace	49.9 ± 3.9	67.9 ± 2.1		90.1 ± 0.9	69.5 ± 0.5	59.8 ± 0.5	56.4 ± 0.4	81.7 ± 1.7	1.339 ± 0.040	4.510 ± 0.166	1.117 ± 0.057
bbbp	49.8 ± 2.9	68.8 ± 2.3	66.2 ± 1.0		70.7 ± 0.2	61.2 ± 0.3	57.3 ± 0.5	84.7 ± 2.0	1.246 ± 0.105	4.512 ± 0.029	1.055 ± 0.044
tox21	48.5 ± 0.9	67.5 ± 4.2	62.9 ± 1.2	90.8 ± 1.6		61.5 ± 0.4	56.3 ± 0.3	78.0 ± 1.3	1.441 ± 0.087	4.321 ± 0.123	1.176 ± 0.045
toxcast	47.6 ± 0.7	70.5 ± 1.8	65.6 ± 0.6	91.1 ± 0.5	69.4 ± 1.3		57.4 ± 0.1	83.6 ± 3.1	1.245 ± 0.089	4.786 ± 0.167	1.138 ± 0.006
sider	51.8 ± 0.8	70.1 ± 1.9	60.1 ± 6.4	90.9 ± 1.2	69.4 ± 0.7	61.2 ± 0.5		79.6 ± 5.0	1.282 ± 0.005	4.090 ± 0.537	1.149 ± 0.053
clintox	50.7 ± 1.5	70.3 ± 1.2	61.7 ± 2.2	90.5 ± 0.7	68.7 ± 0.8	59.7 ± 0.5	56.8 ± 0.2		1.272 ± 0.034	4.574 ± 0.169	1.195 ± 0.059
esol	49.0 ± 2.7	69.6 ± 1.0	65.8 ± 0.9	91.8 ± 0.2	68.2 ± 1.6	59.4 ± 1.5	56.6 ± 2.1	86.7 ± 1.0		4.242 ± 0.293	1.088 ± 0.023
freesolv	51.3 ± 0.4	69.7 ± 1.6	72.6 ± 1.0	93.2 ± 0.6	70.9 ± 0.7	61.8 ± 0.4	57.1 ± 1.0	89.9 ± 2.4	1.314 ± 0.091		1.033 ± 0.028
lipo	46.5 ± 0.5	70.2 ± 1.0	62.6 ± 1.8	91.1 ± 2.0	67.7 ± 0.9	60.3 ± 1.4	56.5 ± 0.9	81.6 ± 2.4	1.222 ± 0.099	4.322 ± 0.208	
mean	49.2 ± 1.9	69.4 ± 1.0	64.0 ± 3.8	91.0 ± 0.9	69.0 ± 1.4	60.4 ± 1.0	56.8 ± 0.4	82.8 ± 3.6	1.301 ± 0.076	4.276 ± 0.361	1.136 ± 0.070
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Additional Motif Datasets, 250 samples, Test											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		67.0 ± 3.7	72.4 ± 5.1	58.8 ± 0.7	65.4 ± 2.5	58.5 ± 2.4	55.3 ± 0.7	62.6 ± 2.0	1.415 ± 0.084	3.455 ± 0.380	1.177 ± 0.033
hiv	44.4 ± 4.2		74.1 ± 5.4	59.2 ± 2.1	67.3 ± 1.1	58.9 ± 1.1	54.9 ± 0.7	64.3 ± 3.6	1.516 ± 0.288	3.770 ± 0.333	1.028 ± 0.042
bace	45.5 ± 1.3	69.5 ± 1.3		59.8 ± 0.5	67.7 ± 1.2	60.4 ± 0.3	55.9 ± 0.5	64.7 ± 1.1	1.436 ± 0.050	4.382 ± 0.012	1.023 ± 0.047
bbbp	44.0 ± 1.4	66.9 ± 3.7	80.3 ± 0.7		68.9 ± 0.9	60.6 ± 0.5	55.3 ± 1.5	65.6 ± 1.8	1.332 ± 0.063	4.111 ± 0.306	0.961 ± 0.048
tox21	46.2 ± 6.5	68.7 ± 2.3	74.5 ± 0.8	59.2 ± 1.4		60.1 ± 0.2	55.6 ± 1.7	65.4 ± 0.6	1.544 ± 0.087	3.986 ± 0.092	1.061 ± 0.036
toxcast	41.5 ± 1.1	71.1 ± 1.6	78.1 ± 1.4	60.2 ± 0.4	67.9 ± 0.8		56.1 ± 0.7	65.6 ± 1.2	1.303 ± 0.080	4.437 ± 0.206	1.029 ± 0.005
sider	45.6 ± 2.2	71.5 ± 0.3	74.0 ± 7.7	58.4 ± 1.1	68.7 ± 0.5	60.5 ± 0.2		64.3 ± 2.3	1.387 ± 0.023	3.650 ± 0.604	1.022 ± 0.038
clintox	48.1 ± 1.1	70.7 ± 0.9	78.1 ± 2.8	59.6 ± 0.6	67.7 ± 0.4	60.2 ± 0.4	55.5 ± 0.2		1.335 ± 0.016	4.073 ± 0.025	1.076 ± 0.062
esol	45.9 ± 5.3	70.4 ± 1.7	77.4 ± 2.4	60.7 ± 0.9	66.3 ± 1.5	59.3 ± 0.7	56.1 ± 1.7	65.4 ± 1.5		4.104 ± 0.043	0.961 ± 0.010
freesolv	48.8 ± 2.2	70.7 ± 0.7	81.0 ± 1.3	63.2 ± 0.8	69.7 ± 0.3	60.3 ± 0.2	58.6 ± 0.4	67.5 ± 5.4	1.349 ± 0.079		0.951 ± 0.029
lipo	45.1 ± 2.5	70.4 ± 1.0	74.0 ± 4.4	59.7 ± 1.4	66.2 ± 0.5	58.8 ± 2.8	55.8 ± 1.0	66.6 ± 1.5	1.338 ± 0.138	4.016 ± 0.308	
mean	45.5 ± 2.1	69.7 ± 1.6	76.4 ± 3.0	59.9 ± 1.3	67.6 ± 1.3	59.8 ± 0.8	55.9 ± 1.0	65.2 ± 1.3	1.396 ± 0.082	3.998 ± 0.305	1.029 ± 0.067
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Pretraining Additional Motif Datasets, 250 samples, Validation											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		65.4 ± 1.7	74.8 ± 2.2	66.4 ± 1.7	65.8 ± 1.7	67.7 ± 1.8	67.6 ± 3.3	71.5 ± 2.6	73.7 ± 4.6	71.4 ± 1.5	65.2 ± 2.2
hiv	66.3 ± 3.7		77.0 ± 2.6	67.4 ± 3.2	66.7 ± 3.1	68.2 ± 3.8	69.8 ± 5.3	73.7 ± 2.7	75.3 ± 3.0	72.2 ± 3.1	65.1 ± 2.3
bace	69.1 ± 1.0	68.8 ± 0.7		69.7 ± 0.7	67.9 ± 0.8	69.9 ± 0.9	73.7 ± 0.3	74.1 ± 0.8	75.9 ± 0.8	76.1 ± 1.1	66.3 ± 0.5
bbbp	72.1 ± 0.7	71.8 ± 0.3	79.2 ± 0.9		71.3 ± 1.0	72.9 ± 0.8	74.5 ± 0.5	76.6 ± 1.2	80.2 ± 0.6	75.9 ± 2.6	67.9 ± 0.4
tox21	69.0 ± 1.5	69.7 ± 2.3	78.3 ± 1.0	70.2 ± 1.2		70.7 ± 2.1	71.7 ± 1.8	75.0 ± 0.6	76.0 ± 1.0	73.7 ± 1.6	66.7 ± 2.8
toxcast	71.0 ± 0.4	70.0 ± 0.3	77.8 ± 0.3	70.3 ± 1.3	69.4 ± 0.6		73.3 ± 0.4	75.9 ± 0.4	78.1 ± 0.8	77.0 ± 1.2	68.8 ± 1.8
sider	69.2 ± 3.7	68.5 ± 2.9	77.1 ± 3.1	69.6 ± 2.4	68.2 ± 3.0	69.5 ± 3.0		73.4 ± 4.1	75.8 ± 3.0	75.1 ± 2.3	66.5 ± 3.4
clintox	70.8 ± 0.5	70.0 ± 0.4	77.6 ± 0.2	69.7 ± 0.2	69.8 ± 0.6	71.7 ± 0.7	74.2 ± 0.8		79.6 ± 0.9	77.2 ± 1.1	67.0 ± 1.5
esol	70.1 ± 1.8	69.6 ± 1.2	78.1 ± 0.5	71.1 ± 2.4	68.6 ± 2.2	71.1 ± 1.2	72.4 ± 2.4	75.6 ± 0.4		75.0 ± 0.8	68.4 ± 0.7
freesolv	77.2 ± 0.6	76.0 ± 0.2	81.2 ± 0.6	76.3 ± 0.4	74.0 ± 1.0	76.4 ± 0.7	79.2 ± 0.4	79.1 ± 0.2	81.0 ± 1.1		75.2 ± 0.9
lipo	67.0 ± 2.7	68.7 ± 0.9	77.1 ± 1.7	68.6 ± 1.5	69.3 ± 1.2	70.7 ± 1.2	71.3 ± 3.1	74.2 ± 1.3	76.8 ± 2.1	73.5 ± 1.9	
mean	70.2 ± 3.0	69.8 ± 2.7	77.8 ± 1.7	69.9 ± 2.6	69.1 ± 2.3	70.9 ± 2.5	72.8 ± 3.1	74.9 ± 2.1	77.2 ± 2.4	74.7 ± 2.0	67.7 ± 2.9

Pretraining Additional Motif Datasets, 250 samples, Test											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		67.4 ± 2.0	69.9 ± 0.9	69.5 ± 1.6	66.8 ± 1.5	65.6 ± 1.6	69.5 ± 3.4	63.4 ± 2.4	69.1 ± 4.5	67.0 ± 2.0	68.1 ± 2.9
hiv	66.3 ± 3.2		70.3 ± 0.3	71.8 ± 2.4	67.7 ± 2.5	66.1 ± 2.9	71.2 ± 4.3	66.7 ± 4.0	68.6 ± 4.0	68.7 ± 2.5	69.6 ± 2.5
bace	68.8 ± 0.6	70.1 ± 0.6		72.3 ± 0.9	68.4 ± 1.0	67.4 ± 0.5	72.6 ± 0.2	68.3 ± 1.1	69.1 ± 0.3	68.9 ± 1.1	69.9 ± 0.4
bbbp	71.1 ± 0.5	73.1 ± 0.6	71.3 ± 0.1		71.3 ± 0.5	69.4 ± 0.3	76.3 ± 1.1	70.4 ± 0.3	72.2 ± 1.2	72.4 ± 0.5	73.0 ± 0.2
tox21	68.5 ± 2.0	71.2 ± 2.5	71.0 ± 1.8	72.4 ± 1.5		67.1 ± 1.8	73.5 ± 0.6	68.5 ± 1.7	70.7 ± 1.5	69.9 ± 1.8	71.1 ± 1.6
toxcast	69.9 ± 0.8	71.1 ± 0.2	73.2 ± 1.0	74.9 ± 0.8	70.1 ± 0.1		73.9 ± 0.4	69.8 ± 0.5	71.9 ± 1.1	70.6 ± 1.9	72.4 ± 0.9
sider	68.7 ± 3.9	70.1 ± 2.7	71.6 ± 1.9	72.7 ± 2.3	67.9 ± 3.1	66.3 ± 3.0		68.0 ± 2.3	69.7 ± 3.3	69.9 ± 4.5	69.7 ± 2.7
clintox	70.2 ± 0.7	71.3 ± 0.2	71.8 ± 1.3	74.1 ± 0.8	70.2 ± 0.3	68.9 ± 0.6	74.5 ± 1.0		72.0 ± 0.5	70.2 ± 0.6	71.4 ± 0.5
esol	69.0 ± 2.0	70.8 ± 0.7	73.1 ± 0.8	74.2 ± 0.8	69.5 ± 1.9	68.1 ± 2.2	73.9 ± 2.5	69.0 ± 1.0		64.4 ± 1.0	70.8 ± 1.3
freesolv	74.6 ± 0.9	76.2 ± 0.7	75.4 ± 1.3	76.3 ± 0.4	74.7 ± 1.2	73.6 ± 0.9	79.9 ± 0.2	75.1 ± 0.4	74.6 ± 1.0		77.7 ± 0.8
lipo	67.0 ± 1.9	69.9 ± 1.6	71.5 ± 1.6	72.8 ± 1.7	69.4 ± 1.2	68.1 ± 1.5	72.7 ± 2.6	67.9 ± 1.2	69.7 ± 1.3	68.4 ± 3.8	
mean	69.4 ± 2.3	71.1 ± 2.3	71.9 ± 1.6	73.1 ± 1.9	69.6 ± 2.2	68.1 ± 2.3	73.8 ± 2.8	68.7 ± 3.0	70.8 ± 1.9	69.0 ± 2.2	71.4 ± 2.6

Table C.6: Validation and Test Performance during pretraining with additional motif datasets and 250 samples.

C.4.2 500 samples

Pretraining Additional Motif Datasets, 500 samples, Validation											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		71.9 ± 0.2	67.0 ± 1.1	92.2 ± 0.1	71.1 ± 0.3	62.0 ± 1.1	58.1 ± 0.7	84.3 ± 8.9	1.155 ± 0.037	4.600 ± 0.430	1.060 ± 0.074
hiv	54.2 ± 6.1		68.2 ± 1.0	90.3 ± 0.4	71.1 ± 0.2	61.9 ± 0.2	57.6 ± 0.4	90.0 ± 2.3	1.205 ± 0.052	4.738 ± 0.209	1.077 ± 0.019
bace	52.0 ± 1.2	67.6 ± 1.7		89.1 ± 0.8	69.7 ± 0.5	61.7 ± 0.8	55.8 ± 0.6	88.4 ± 2.2	1.123 ± 0.031	4.783 ± 0.094	1.177 ± 0.059
bbbp	47.1 ± 3.9	69.5 ± 1.2	65.8 ± 7.5		69.9 ± 1.2	60.6 ± 0.8	55.9 ± 0.2	86.1 ± 6.8	1.213 ± 0.147	4.691 ± 0.848	1.039 ± 0.092
tox21	46.4 ± 2.4	70.5 ± 1.1	67.3 ± 1.8	91.1 ± 0.3		61.7 ± 0.4	56.7 ± 0.8	90.5 ± 1.2	1.195 ± 0.068	5.036 ± 0.105	1.009 ± 0.024
toxcast	53.1 ± 0.9	70.6 ± 0.4	67.4 ± 0.9	91.6 ± 0.8	70.2 ± 0.7		56.4 ± 0.0	87.7 ± 2.0	1.118 ± 0.009	4.924 ± 0.144	0.988 ± 0.034
sider	55.2 ± 2.7	70.8 ± 0.9	66.9 ± 1.2	91.0 ± 0.6	70.3 ± 0.1	61.7 ± 0.3		83.9 ± 5.0	1.158 ± 0.034	4.579 ± 0.219	1.036 ± 0.031
clintox	55.2 ± 3.4	72.9 ± 1.1	67.5 ± 1.5	90.5 ± 0.5	71.1 ± 0.1	61.8 ± 0.5	55.9 ± 0.1		1.190 ± 0.017	4.877 ± 0.229	1.034 ± 0.071
esol	53.0 ± 1.4	73.4 ± 0.3	66.5 ± 0.5	91.6 ± 0.3	69.4 ± 0.4	61.0 ± 0.4	56.8 ± 0.5	87.6 ± 1.7		4.674 ± 0.264	1.084 ± 0.022
freesolv	54.5 ± 0.9	71.4 ± 0.8	73.1 ± 1.6	92.7 ± 0.3	72.7 ± 0.8	63.5 ± 1.3	56.0 ± 2.2	93.0 ± 2.1	1.131 ± 0.016		1.000 ± 0.063
lipo	48.4 ± 5.4	71.4 ± 1.1	68.3 ± 1.4	91.5 ± 0.3	68.5 ± 0.7	62.1 ± 0.4	57.2 ± 0.7	88.1 ± 1.4	1.229 ± 0.035	4.985 ± 0.061	
mean	51.9 ± 3.4	71.0 ± 1.7	67.8 ± 2.0	91.2 ± 1.0	70.4 ± 1.2	61.8 ± 0.8	56.6 ± 0.8	88.0 ± 2.8	1.172 ± 0.040	4.789 ± 0.160	1.050 ± 0.054
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Additional Motif Datasets, 500 samples, Test											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		71.3 ± 1.3	76.7 ± 4.8	61.3 ± 2.4	68.5 ± 0.7	60.0 ± 1.1	57.4 ± 1.9	68.3 ± 3.1	1.291 ± 0.017	3.816 ± 0.685	0.967 ± 0.049
hiv	49.1 ± 2.7		83.0 ± 1.1	61.7 ± 0.5	68.6 ± 0.4	61.3 ± 0.5	57.5 ± 0.9	63.0 ± 0.9	1.263 ± 0.043	4.185 ± 0.063	0.982 ± 0.014
bace	46.8 ± 1.9	67.3 ± 1.4		61.2 ± 0.1	68.6 ± 0.3	60.8 ± 0.4	57.2 ± 0.6	67.0 ± 1.8	1.229 ± 0.039	4.493 ± 0.022	1.094 ± 0.055
bbbp	42.7 ± 4.5	70.0 ± 4.2	76.7 ± 5.9		69.1 ± 0.1	60.2 ± 1.5	54.8 ± 2.3	68.2 ± 4.4	1.272 ± 0.145	3.920 ± 0.515	0.950 ± 0.074
tox21	44.5 ± 2.9	70.9 ± 1.6	81.0 ± 1.6	61.1 ± 1.5		61.1 ± 0.2	57.5 ± 1.6	66.0 ± 1.8	1.241 ± 0.086	4.339 ± 0.112	0.924 ± 0.014
toxcast	50.8 ± 0.7	70.6 ± 0.3	80.1 ± 1.3	61.1 ± 2.3	68.2 ± 0.4		56.5 ± 0.4	62.9 ± 1.1	1.205 ± 0.026	4.332 ± 0.083	0.904 ± 0.021
sider	48.6 ± 0.3	71.4 ± 0.3	80.9 ± 1.9	59.3 ± 0.6	68.4 ± 0.0	61.6 ± 0.2		65.4 ± 3.1	1.271 ± 0.031	4.196 ± 0.047	0.944 ± 0.038
clintox	52.5 ± 3.2	71.1 ± 1.1	81.3 ± 1.6	61.5 ± 1.1	68.4 ± 0.8	60.9 ± 0.2	56.3 ± 0.3		1.265 ± 0.031	4.367 ± 0.119	0.949 ± 0.060
esol	48.6 ± 1.5	69.9 ± 0.7	82.1 ± 1.0	61.4 ± 1.9	66.4 ± 0.8	60.2 ± 0.3	57.0 ± 0.9	66.2 ± 3.6		4.284 ± 0.138	0.989 ± 0.031
freesolv	55.5 ± 1.7	72.6 ± 1.2	81.3 ± 1.2	65.9 ± 1.1	71.7 ± 0.8	62.3 ± 0.7	61.5 ± 0.9	82.4 ± 8.5	1.237 ± 0.016		0.899 ± 0.072
lipo	46.0 ± 2.9	73.1 ± 0.2	78.9 ± 0.3	62.1 ± 0.7	67.4 ± 0.7	61.6 ± 0.1	57.8 ± 0.8	66.2 ± 3.0	1.276 ± 0.013	4.570 ± 0.036	
mean	48.5 ± 3.8	70.8 ± 1.6	80.2 ± 2.1	61.7 ± 1.7	68.5 ± 1.3	61.0 ± 0.7	57.4 ± 1.7	67.6 ± 5.5	1.255 ± 0.026	4.250 ± 0.235	0.960 ± 0.056
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Pretraining Additional Motif Datasets, 500 samples, Validation											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		73.3 ± 5.0	79.6 ± 2.9	74.9 ± 5.1	74.9 ± 4.7	76.7 ± 5.0	77.2 ± 5.3	78.8 ± 4.0	80.9 ± 2.4	79.3 ± 4.1	70.8 ± 7.1
hiv	74.9 ± 0.2		80.5 ± 0.8	74.2 ± 0.7	73.5 ± 1.1	75.1 ± 0.9	78.0 ± 0.4	78.9 ± 0.4	80.2 ± 0.8	79.6 ± 1.5	72.7 ± 0.4
bace	74.9 ± 0.5	73.5 ± 0.1		74.5 ± 1.2	73.6 ± 0.7	75.0 ± 0.9	78.4 ± 1.0	78.1 ± 0.4	81.9 ± 0.6	79.1 ± 0.3	73.1 ± 0.3
bbbp	73.9 ± 5.9	74.3 ± 4.5	80.1 ± 4.7		73.5 ± 4.2	76.1 ± 4.0	76.2 ± 5.1	78.3 ± 3.4	80.4 ± 3.5	78.1 ± 6.2	71.1 ± 5.7
tox21	75.1 ± 0.8	73.9 ± 1.7	81.2 ± 1.0	75.4 ± 1.0		76.1 ± 2.0	78.0 ± 1.8	79.3 ± 0.7	81.3 ± 1.3	80.9 ± 1.9	72.5 ± 1.3
toxcast	76.1 ± 1.3	74.2 ± 1.6	81.4 ± 1.2	74.9 ± 1.0	73.2 ± 1.8		78.3 ± 1.3	80.2 ± 0.7	80.7 ± 1.8	79.1 ± 1.5	73.1 ± 1.2
sider	73.4 ± 0.2	71.8 ± 0.7	77.9 ± 1.1	71.8 ± 0.8	71.1 ± 1.2	73.4 ± 1.1		77.2 ± 0.6	78.9 ± 0.7	78.8 ± 1.3	69.2 ± 0.5
clintox	74.9 ± 1.4	73.8 ± 1.3	80.6 ± 0.6	73.4 ± 1.6	72.6 ± 1.4	75.1 ± 2.0	78.2 ± 1.0		79.6 ± 1.5	79.4 ± 1.4	71.1 ± 1.0
esol	75.3 ± 0.7	74.9 ± 0.6	80.1 ± 0.2	74.3 ± 0.4	74.5 ± 0.8	76.7 ± 1.0	76.5 ± 1.2	79.4 ± 0.7		78.9 ± 3.0	73.7 ± 1.6
freesolv	85.1 ± 1.3	82.8 ± 1.7	87.7 ± 0.4	83.5 ± 1.1	83.5 ± 1.5	85.6 ± 1.1	86.4 ± 0.8	84.5 ± 1.8	85.6 ± 0.9		81.7 ± 1.2
lipo	74.1 ± 2.6	73.4 ± 1.8	80.1 ± 0.7	73.8 ± 1.6	74.1 ± 2.2	76.0 ± 2.7	78.5 ± 2.0	78.7 ± 1.0	79.7 ± 1.1	78.0 ± 2.3	
mean	75.8 ± 3.4	74.6 ± 3.0	80.9 ± 2.6	75.1 ± 3.1	74.5 ± 3.3	76.6 ± 3.3	78.6 ± 2.9	79.3 ± 2.0	80.9 ± 1.9	79.1 ± 0.8	72.9 ± 3.4

Pretraining Additional Motif Datasets, 500 samples, Test											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		74.5 ± 4.8	74.3 ± 2.7	76.8 ± 4.7	74.8 ± 5.0	72.9 ± 4.7	78.4 ± 4.1	72.9 ± 4.7	76.1 ± 3.4	73.8 ± 2.4	74.0 ± 5.3
hiv	74.8 ± 0.3		74.0 ± 1.0	77.1 ± 1.4	74.0 ± 1.0	72.4 ± 0.6	78.9 ± 0.9	71.9 ± 0.6	75.5 ± 1.3	73.2 ± 1.6	75.3 ± 0.7
bace	74.7 ± 0.6	75.0 ± 0.2		77.3 ± 0.9	73.6 ± 0.6	71.8 ± 0.7	77.7 ± 0.0	71.4 ± 0.8	75.3 ± 1.0	73.4 ± 0.1	74.4 ± 0.6
bbbp	73.3 ± 5.9	75.0 ± 4.6	75.3 ± 1.9		73.9 ± 4.6	72.1 ± 4.1	76.9 ± 4.7	72.2 ± 4.7	73.7 ± 4.0	71.1 ± 4.6	74.4 ± 4.7
tox21	74.0 ± 1.0	75.2 ± 1.1	74.1 ± 0.7	77.3 ± 1.0		73.1 ± 1.6	78.4 ± 1.5	73.1 ± 1.3	73.9 ± 2.0	73.6 ± 1.8	74.4 ± 0.8
toxcast	75.1 ± 1.1	75.1 ± 1.0	75.1 ± 0.6	78.6 ± 0.7	74.0 ± 1.8		79.2 ± 0.9	73.8 ± 1.8	75.2 ± 2.2	72.9 ± 2.5	75.6 ± 1.2
sider	72.7 ± 0.4	73.3 ± 0.3	73.5 ± 0.5	75.6 ± 0.5	71.6 ± 1.8	70.1 ± 1.0		70.9 ± 0.3	72.1 ± 0.8	71.8 ± 2.5	72.6 ± 0.8
clintox	74.3 ± 0.9	74.4 ± 1.3	73.3 ± 0.2	77.4 ± 1.3	73.8 ± 1.6	71.8 ± 1.3	79.1 ± 0.3		75.2 ± 1.8	71.6 ± 0.8	74.2 ± 1.2
esol	74.4 ± 0.8	75.9 ± 0.3	75.7 ± 0.5	78.0 ± 0.6	74.4 ± 1.7	72.5 ± 1.0	79.2 ± 0.4	72.9 ± 0.7		72.1 ± 2.0	75.1 ± 0.5
freesolv	82.9 ± 1.4	82.6 ± 1.6	81.6 ± 0.9	84.3 ± 1.4	83.4 ± 1.4	81.8 ± 1.5	86.2 ± 1.6	81.4 ± 0.8	81.9 ± 0.5		85.5 ± 1.1
lipo	73.2 ± 2.4	74.9 ± 1.8	74.9 ± 0.8	77.6 ± 1.5	74.7 ± 2.7	72.3 ± 2.5	79.1 ± 1.0	72.6 ± 1.4	74.3 ± 2.1	70.2 ± 1.1	
mean	74.9 ± 2.9	75.6 ± 2.6	75.2 ± 2.4	78.0 ± 2.3	74.8 ± 3.1	73.1 ± 3.2	79.3 ± 2.5	73.3 ± 3.0	75.3 ± 2.6	72.4 ± 1.2	75.6 ± 3.6

Table C.7: Validation and Test Performance during pretraining with additional motif datasets and 500 samples.

C.4.3 1500 samples

Pretraining Additional Motif Datasets, 1500 samples, Validation											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		74.0 ± 0.5	75.2 ± 1.2	92.8 ± 0.2	71.5 ± 0.1	64.1 ± 0.4	56.1 ± 0.4	94.8 ± 3.8	1.076 ± 0.028	5.431 ± 0.111	0.917 ± 0.051
hiv	53.4 ± 5.2		74.1 ± 0.5	92.0 ± 0.2	71.6 ± 0.4	64.3 ± 0.8	56.3 ± 0.9	95.3 ± 0.7	1.082 ± 0.094	5.592 ± 0.278	0.911 ± 0.041
bace	48.1 ± 4.6	71.8 ± 1.0		91.2 ± 0.6	71.5 ± 1.4	64.0 ± 0.6	56.7 ± 1.9	93.1 ± 2.3	1.163 ± 0.036	5.033 ± 0.364	0.944 ± 0.067
bbbp	51.8 ± 0.9	74.4 ± 0.4	68.5 ± 2.3		72.0 ± 0.3	64.2 ± 0.3	55.5 ± 0.8	96.0 ± 0.2	1.046 ± 0.070	5.604 ± 0.064	0.856 ± 0.022
tox21	53.9 ± 4.0	74.1 ± 1.8	71.4 ± 1.5	92.8 ± 0.6		64.0 ± 0.6	56.9 ± 1.0	94.9 ± 1.4	1.063 ± 0.010	5.513 ± 0.168	0.875 ± 0.017
toxcast	53.3 ± 3.1	72.9 ± 0.3	74.1 ± 1.4	91.9 ± 0.6	71.6 ± 0.6		56.0 ± 0.5	95.7 ± 2.7	1.062 ± 0.034	5.601 ± 0.055	0.889 ± 0.012
sider	59.2 ± 2.2	72.6 ± 0.6	72.6 ± 2.4	92.2 ± 0.8	72.0 ± 0.5	64.5 ± 0.2		93.5 ± 1.7	1.074 ± 0.026	5.520 ± 0.041	0.832 ± 0.024
clintox	53.1 ± 2.7	72.4 ± 0.9	74.6 ± 1.5	92.3 ± 0.4	71.8 ± 0.7	63.6 ± 1.3	57.5 ± 0.7		1.032 ± 0.038	5.430 ± 0.084	0.958 ± 0.032
esol	59.2 ± 3.6	72.8 ± 0.8	73.5 ± 2.6	92.9 ± 0.7	72.8 ± 0.3	62.6 ± 0.7	56.8 ± 0.2	94.5 ± 0.6		5.565 ± 0.076	0.981 ± 0.036
freesolv	72.0 ± 1.1	74.5 ± 1.6	74.6 ± 2.2	92.8 ± 1.2	76.8 ± 0.2	65.4 ± 0.4	61.2 ± 0.1	93.4 ± 1.5	1.019 ± 0.068		0.873 ± 0.025
lipo	56.9 ± 3.4	73.6 ± 0.8	69.8 ± 0.3	91.8 ± 0.7	70.9 ± 1.3	63.7 ± 0.7	55.7 ± 0.4	90.0 ± 5.1	1.125 ± 0.013	5.352 ± 0.094	
mean	56.1 ± 6.5	73.3 ± 0.9	72.8 ± 2.3	92.3 ± 0.6	72.2 ± 1.7	64.0 ± 0.7	56.9 ± 1.6	94.1 ± 1.7	1.074 ± 0.043	5.464 ± 0.174	0.904 ± 0.047
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Additional Motif Datasets, 1500 samples, Test											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		71.0 ± 1.3	82.8 ± 0.7	65.2 ± 1.8	70.5 ± 1.2	61.8 ± 0.2	62.0 ± 0.6	83.1 ± 2.4	1.184 ± 0.020	4.461 ± 0.153	0.853 ± 0.048
hiv	58.6 ± 6.5		81.4 ± 1.7	65.6 ± 1.2	71.4 ± 0.4	62.3 ± 0.5	63.4 ± 1.4	84.0 ± 3.3	1.243 ± 0.083	4.452 ± 0.170	0.836 ± 0.026
bace	52.6 ± 5.5	72.2 ± 0.6		67.4 ± 1.7	69.8 ± 1.5	61.7 ± 0.6	61.2 ± 1.6	84.5 ± 0.5	1.202 ± 0.070	4.613 ± 0.212	0.888 ± 0.058
bbbp	54.6 ± 5.0	71.0 ± 0.8	82.3 ± 2.0		70.2 ± 0.5	61.8 ± 0.1	61.4 ± 0.4	85.2 ± 2.1	1.188 ± 0.082	4.371 ± 0.097	0.816 ± 0.011
tox21	54.6 ± 1.5	70.7 ± 1.0	81.9 ± 0.3	65.0 ± 0.9		62.3 ± 0.3	63.8 ± 1.0	84.8 ± 4.3	1.213 ± 0.017	4.462 ± 0.053	0.825 ± 0.025
toxcast	53.1 ± 2.7	71.9 ± 0.5	78.9 ± 3.9	64.2 ± 0.6	70.6 ± 1.0		62.2 ± 0.8	81.5 ± 3.0	1.212 ± 0.023	4.595 ± 0.063	0.826 ± 0.009
sider	62.1 ± 0.4	72.5 ± 0.4	79.2 ± 1.0	65.2 ± 1.9	71.6 ± 0.4	62.5 ± 0.2		81.8 ± 1.4	1.195 ± 0.020	4.439 ± 0.143	0.781 ± 0.017
clintox	58.9 ± 5.4	71.1 ± 0.5	81.7 ± 0.8	65.4 ± 0.5	70.8 ± 1.5	62.4 ± 0.8	61.7 ± 1.7		1.203 ± 0.030	4.570 ± 0.176	0.876 ± 0.044
esol	61.5 ± 1.0	70.6 ± 0.9	79.8 ± 1.6	65.6 ± 0.7	70.8 ± 1.0	61.8 ± 0.1	60.7 ± 1.1	87.3 ± 2.5		4.514 ± 0.127	0.898 ± 0.051
freesolv	72.1 ± 1.1	73.7 ± 0.8	81.2 ± 2.6	70.0 ± 1.0	74.0 ± 0.5	64.0 ± 0.3	62.6 ± 0.5	89.6 ± 2.3	1.139 ± 0.046		0.821 ± 0.028
lipo	54.8 ± 2.3	72.6 ± 0.5	80.2 ± 1.1	65.6 ± 0.7	70.4 ± 1.4	62.4 ± 0.7	63.1 ± 1.2	79.4 ± 4.2	1.228 ± 0.054	4.510 ± 0.073	
mean	58.3 ± 5.9	71.7 ± 1.0	80.9 ± 1.3	65.9 ± 1.6	71.0 ± 1.2	62.3 ± 0.7	62.2 ± 1.0	84.1 ± 2.9	1.201 ± 0.028	4.499 ± 0.076	0.842 ± 0.036
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Pretraining Additional Motif Datasets, 1500 samples, Validation											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		81.7 ± 1.6	88.1 ± 0.8	84.0 ± 0.1	82.7 ± 1.8	84.7 ± 1.6	85.2 ± 1.2	84.7 ± 0.7	84.7 ± 0.4	86.9 ± 2.6	80.9 ± 0.9
hiv	85.5 ± 2.0		89.3 ± 1.7	85.1 ± 1.0	85.6 ± 1.6	87.1 ± 1.3	87.0 ± 1.6	86.4 ± 1.8	86.1 ± 2.2	86.8 ± 1.0	82.4 ± 1.6
bace	84.1 ± 1.6	82.8 ± 1.1		84.4 ± 1.5	84.5 ± 1.6	86.3 ± 1.1	86.3 ± 1.2	86.1 ± 1.4	85.9 ± 1.0	85.9 ± 2.4	81.6 ± 1.1
bbbp	85.0 ± 0.4	82.8 ± 0.4	88.9 ± 0.4		84.6 ± 0.4	85.8 ± 0.6	85.9 ± 0.7	86.6 ± 0.7	86.4 ± 0.5	87.3 ± 0.7	81.1 ± 0.7
tox21	85.5 ± 1.6	84.3 ± 1.5	89.4 ± 1.3	86.1 ± 1.1		86.9 ± 1.5	87.5 ± 1.6	87.4 ± 1.8	86.5 ± 1.0	87.5 ± 0.7	82.6 ± 1.3
toxcast	83.9 ± 1.4	82.2 ± 0.9	88.1 ± 1.6	84.7 ± 1.0	83.5 ± 1.4		85.5 ± 0.3	84.6 ± 0.5	85.1 ± 0.5	87.4 ± 0.7	81.4 ± 1.2
sider	86.6 ± 1.4	84.3 ± 1.4	90.4 ± 0.9	86.1 ± 1.0	86.2 ± 2.0	87.5 ± 1.6		87.1 ± 1.8	86.4 ± 0.6	88.8 ± 1.1	83.4 ± 1.4
clintox	85.5 ± 2.2	83.4 ± 2.2	88.6 ± 1.6	85.3 ± 1.1	84.2 ± 2.1	86.1 ± 1.8	86.0 ± 1.5		87.1 ± 1.2	87.2 ± 1.5	82.2 ± 2.2
esol	85.0 ± 0.1	83.6 ± 0.4	88.1 ± 1.9	84.3 ± 0.8	84.2 ± 0.4	86.6 ± 0.2	85.5 ± 0.3	85.7 ± 0.4		87.3 ± 1.7	82.3 ± 0.2
freesolv	94.4 ± 0.1	91.0 ± 0.1	94.5 ± 0.3	91.6 ± 0.8	92.7 ± 0.2	93.7 ± 0.2	93.3 ± 0.1	93.7 ± 0.2	91.3 ± 0.5		91.0 ± 0.6
lipo	84.5 ± 1.4	82.2 ± 1.3	88.4 ± 1.4	84.4 ± 2.0	84.4 ± 1.6	85.8 ± 1.6	85.3 ± 1.4	84.6 ± 1.5	85.4 ± 0.4	86.6 ± 1.9	
mean	86.0 ± 3.1	83.8 ± 2.7	89.4 ± 1.9	85.6 ± 2.2	85.3 ± 2.8	87.0 ± 2.5	86.7 ± 2.4	86.7 ± 2.7	86.5 ± 1.8	87.2 ± 0.7	82.9 ± 2.9

Pretraining Additional Motif Datasets, 1500 samples, Test											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		82.6 ± 1.3	79.1 ± 1.2	84.5 ± 1.1	82.9 ± 1.8	81.5 ± 1.6	84.7 ± 0.9	81.2 ± 1.2	82.9 ± 0.3	78.3 ± 0.3	83.1 ± 0.8
hiv	84.7 ± 1.9		82.0 ± 2.8	85.4 ± 2.0	85.7 ± 1.5	83.9 ± 1.7	85.9 ± 1.2	83.2 ± 2.1	83.0 ± 1.4	79.1 ± 1.9	85.4 ± 1.5
bace	83.8 ± 1.8	83.0 ± 1.0		85.2 ± 1.5	83.9 ± 1.5	82.8 ± 1.3	84.7 ± 0.8	81.6 ± 1.6	83.1 ± 0.6	80.8 ± 1.0	83.7 ± 1.6
bbbp	84.2 ± 0.6	83.2 ± 0.2	80.1 ± 0.7		84.6 ± 0.2	83.1 ± 0.2	86.1 ± 0.4	81.1 ± 1.0	81.6 ± 0.6	79.9 ± 0.6	84.4 ± 0.3
tox21	85.1 ± 2.1	84.7 ± 1.3	81.1 ± 3.7	86.1 ± 1.1		84.3 ± 1.6	86.9 ± 1.2	82.5 ± 2.0	83.9 ± 1.0	80.6 ± 0.9	85.6 ± 1.7
toxcast	82.9 ± 1.2	82.5 ± 1.0	79.2 ± 1.6	84.6 ± 0.7	83.6 ± 1.1		85.0 ± 0.7	80.8 ± 0.8	82.4 ± 0.8	79.0 ± 0.3	83.6 ± 0.8
sider	85.9 ± 1.8	84.7 ± 1.0	82.1 ± 3.0	86.6 ± 1.3	86.1 ± 1.7	84.7 ± 1.9		84.2 ± 1.2	83.9 ± 0.5	81.0 ± 1.2	86.5 ± 2.0
clintox	84.8 ± 2.3	83.8 ± 1.9	81.0 ± 3.9	85.6 ± 1.5	84.9 ± 2.0	83.1 ± 2.2	85.4 ± 1.9		84.1 ± 1.5	78.7 ± 2.1	85.4 ± 1.9
esol	84.4 ± 0.1	83.3 ± 0.4	81.2 ± 1.4	85.2 ± 0.3	84.7 ± 0.1	82.8 ± 0.5	86.3 ± 0.4	82.0 ± 0.6		79.5 ± 1.3	84.5 ± 0.2
freesolv	93.8 ± 0.3	91.0 ± 0.2	91.9 ± 0.3	93.3 ± 0.4	92.9 ± 0.3	92.0 ± 0.4	92.6 ± 0.6	90.0 ± 0.3	87.4 ± 1.3		93.7 ± 0.3
lipo	83.5 ± 1.5	82.9 ± 1.3	80.5 ± 1.8	84.4 ± 1.2	83.7 ± 1.6	82.4 ± 1.6	84.4 ± 1.4	81.5 ± 0.9	84.0 ± 0.6	77.1 ± 0.9	
mean	85.3 ± 3.1	84.2 ± 2.5	81.8 ± 3.7	86.1 ± 2.6	85.3 ± 2.8	84.1 ± 2.9	86.2 ± 2.4	82.8 ± 2.7	83.6 ± 1.5	79.4 ± 1.2	85.6 ± 3.0

Table C.8: Validation and Test Performance during pretraining with additional motif datasets and 1500 samples.

C.4.4 3000 samples

Pretraining Additional Motif Datasets, 3000 samples, Validation											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		74.5 ± 0.8	73.4 ± 2.7	92.2 ± 1.1	75.4 ± 0.3	64.1 ± 0.4	59.7 ± 0.5	92.5 ± 2.0	1.000 ± 0.026	5.590 ± 0.084	0.839 ± 0.020
hiv	70.2 ± 2.3		74.5 ± 0.7	92.8 ± 0.1	75.6 ± 1.2	65.8 ± 0.7	60.2 ± 1.2	91.1 ± 2.6	0.982 ± 0.033	5.405 ± 0.218	0.835 ± 0.029
bace	69.9 ± 2.5	74.8 ± 2.1		92.5 ± 0.3	77.2 ± 0.4	64.7 ± 0.4	63.2 ± 1.6	87.8 ± 2.4	0.996 ± 0.012	5.284 ± 0.218	0.862 ± 0.028
bbbp	72.1 ± 0.4	77.4 ± 0.1	75.0 ± 1.0		77.1 ± 0.6	65.3 ± 0.4	62.8 ± 1.0	92.8 ± 2.1	0.970 ± 0.021	5.450 ± 0.067	0.853 ± 0.029
tox21	68.7 ± 4.6	76.0 ± 1.1	71.5 ± 1.1	92.3 ± 1.0		63.4 ± 0.6	62.7 ± 2.2	91.8 ± 0.8	1.018 ± 0.028	5.597 ± 0.154	0.837 ± 0.033
toxcast	68.1 ± 2.2	75.5 ± 2.3	73.8 ± 1.1	92.7 ± 0.9	75.8 ± 0.6		61.7 ± 0.7	94.7 ± 0.7	0.979 ± 0.042	5.380 ± 0.147	0.842 ± 0.029
sider	72.0 ± 4.3	76.6 ± 0.7	74.8 ± 2.2	93.0 ± 0.6	76.7 ± 1.1	65.6 ± 0.3		92.5 ± 2.1	0.967 ± 0.063	5.232 ± 0.108	0.839 ± 0.018
clintox	65.1 ± 4.2	75.8 ± 1.7	75.0 ± 2.1	93.2 ± 0.4	75.6 ± 1.2	64.7 ± 0.2	60.5 ± 0.5		1.008 ± 0.023	5.606 ± 0.073	0.877 ± 0.012
esol	70.2 ± 1.3	74.9 ± 1.7	71.7 ± 0.7	91.7 ± 0.5	75.5 ± 0.8	64.3 ± 0.4	59.4 ± 1.2	95.9 ± 1.1		5.364 ± 0.085	0.937 ± 0.049
freesolv	72.2 ± 1.0	77.3 ± 1.8	74.5 ± 1.1	93.0 ± 1.3	78.3 ± 0.4	66.1 ± 0.8	64.4 ± 1.0	91.4 ± 2.1	0.982 ± 0.015		0.857 ± 0.027
lipo	65.7 ± 3.1	76.5 ± 1.2	71.5 ± 0.6	93.0 ± 0.7	75.0 ± 1.1	65.1 ± 0.4	59.5 ± 1.8	91.1 ± 4.3	1.054 ± 0.033	5.562 ± 0.078	
mean	69.4 ± 2.5	75.9 ± 1.0	73.6 ± 1.5	92.6 ± 0.5	76.2 ± 1.0	64.9 ± 0.8	61.4 ± 1.8	92.2 ± 2.2	0.996 ± 0.026	5.447 ± 0.136	0.858 ± 0.031
no pretrain	75.6 ± 0.9	80.9 ± 2.0	70.9 ± 1.9	93.4 ± 0.3	76.3 ± 0.3	64.0 ± 1.1	60.5 ± 0.9	98.7 ± 0.5	1.234 ± 0.070	1.723 ± 0.141	0.804 ± 0.007

Pretraining Additional Motif Datasets, 3000 samples, Test											
left out dataset ↓	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
muv		70.9 ± 0.5	79.6 ± 1.2	67.5 ± 0.5	73.1 ± 0.4	64.1 ± 0.8	64.2 ± 0.4	79.8 ± 3.4	1.114 ± 0.011	4.835 ± 0.123	0.796 ± 0.009
hiv	71.0 ± 2.2		79.7 ± 1.3	67.9 ± 0.7	74.0 ± 0.8	63.9 ± 0.7	62.7 ± 0.4	86.5 ± 1.7	1.105 ± 0.071	4.789 ± 0.135	0.791 ± 0.030
bace	68.1 ± 3.5	73.2 ± 0.6		69.1 ± 2.0	73.9 ± 0.6	64.0 ± 0.4	63.2 ± 0.5	87.3 ± 1.4	1.093 ± 0.006	4.661 ± 0.052	0.804 ± 0.012
bbbp	73.3 ± 1.9	72.9 ± 2.2	79.7 ± 1.8		75.2 ± 0.5	64.4 ± 0.4	63.0 ± 0.6	86.4 ± 1.2	1.059 ± 0.024	4.554 ± 0.201	0.774 ± 0.002
tox21	69.5 ± 4.5	71.5 ± 1.8	79.9 ± 1.3	67.5 ± 1.1		64.0 ± 0.6	64.0 ± 0.8	82.3 ± 2.2	1.112 ± 0.022	4.727 ± 0.173	0.778 ± 0.032
toxcast	70.2 ± 3.1	71.9 ± 0.4	79.8 ± 0.7	66.6 ± 1.5	73.0 ± 0.8		62.7 ± 0.6	82.3 ± 1.2	1.097 ± 0.028	4.644 ± 0.172	0.791 ± 0.022
sider	71.7 ± 2.5	71.9 ± 0.8	81.3 ± 0.5	64.8 ± 1.3	74.8 ± 0.7	64.9 ± 0.9		84.2 ± 1.4	1.081 ± 0.041	4.328 ± 0.072	0.786 ± 0.015
clintox	69.9 ± 1.2	71.6 ± 0.5	80.3 ± 0.6	66.3 ± 2.2	72.9 ± 0.7	64.1 ± 0.6	64.6 ± 1.0		1.129 ± 0.041	4.747 ± 0.128	0.820 ± 0.033
esol	72.1 ± 2.6	71.0 ± 1.5	81.5 ± 1.4	66.2 ± 0.9	73.7 ± 0.2	63.6 ± 0.3	62.4 ± 0.5	88.7 ± 4.0		4.463 ± 0.174	0.876 ± 0.041
freesolv	75.1 ± 1.8	74.6 ± 0.7	80.7 ± 1.6	71.1 ± 0.8	75.2 ± 0.2	64.9 ± 0.2	62.3 ± 0.8	91.2 ± 1.9	1.076 ± 0.031		0.788 ± 0.035
lipo	68.1 ± 1.3	73.5 ± 0.8	79.9 ± 1.3	67.7 ± 1.7	73.7 ± 0.8	63.5 ± 0.5	63.8 ± 1.5	84.3 ± 1.3	1.143 ± 0.002	4.708 ± 0.094	
mean	70.9 ± 2.2	72.3 ± 1.2	80.2 ± 0.7	67.5 ± 1.7	74.0 ± 0.9	64.1 ± 0.5	63.3 ± 0.8	85.3 ± 3.4	1.101 ± 0.025	4.646 ± 0.156	0.800 ± 0.030
no pretrain	70.7 ± 2.4	75.1 ± 1.0	79.0 ± 1.2	63.2 ± 2.2	73.2 ± 0.2	61.3 ± 0.7	56.7 ± 0.4	84.3 ± 0.9	1.407 ± 0.186	2.634 ± 0.072	0.773 ± 0.013

Pretraining Additional Motif Datasets, 3000 samples, Validation											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		88.6 ± 0.3	93.9 ± 0.5	89.5 ± 0.7	90.6 ± 0.6	91.1 ± 0.4	90.4 ± 0.3	91.4 ± 0.7	89.6 ± 0.6	90.7 ± 1.0	87.6 ± 0.6
hiv	91.4 ± 1.3		93.4 ± 1.4	89.3 ± 1.0	90.5 ± 1.4	91.2 ± 1.2	90.8 ± 0.7	91.3 ± 0.8	89.9 ± 0.8	89.8 ± 2.4	86.9 ± 1.2
bace	93.6 ± 0.7	90.9 ± 0.9		91.9 ± 1.0	92.7 ± 0.7	93.1 ± 0.8	93.1 ± 0.9	92.7 ± 0.5	91.0 ± 0.5	91.6 ± 0.8	90.1 ± 0.5
bbbp	94.8 ± 1.0	91.7 ± 1.1	95.3 ± 0.7		92.7 ± 0.9	93.4 ± 1.1	92.9 ± 1.1	93.4 ± 0.7	92.1 ± 0.4	92.0 ± 0.3	90.8 ± 1.0
tox21	91.5 ± 2.6	89.4 ± 2.4	93.9 ± 1.4	90.3 ± 2.0		91.6 ± 1.8	91.1 ± 1.8	91.5 ± 1.4	90.6 ± 2.0	90.0 ± 1.6	88.3 ± 2.4
toxcast	92.5 ± 1.6	90.0 ± 1.3	93.8 ± 0.5	90.0 ± 1.5	91.3 ± 1.2		91.6 ± 1.2	92.1 ± 1.0	89.5 ± 1.5	89.2 ± 2.8	88.9 ± 1.7
sider	94.3 ± 1.6	91.5 ± 1.2	95.4 ± 1.2	91.8 ± 2.1	92.6 ± 1.6	93.7 ± 1.3		92.9 ± 0.6	91.8 ± 1.2	91.5 ± 1.5	90.9 ± 2.0
clintox	91.7 ± 1.5	88.9 ± 1.1	93.4 ± 0.7	89.6 ± 1.3	90.2 ± 1.2	91.3 ± 1.2	90.3 ± 1.2		90.6 ± 0.4	91.1 ± 0.4	88.1 ± 2.0
esol	93.1 ± 1.7	90.7 ± 1.5	94.6 ± 1.2	90.0 ± 1.2	91.3 ± 1.7	92.5 ± 1.4	91.5 ± 1.5	92.7 ± 1.1		90.9 ± 1.6	90.0 ± 2.3
freesolv	96.7 ± 0.3	93.2 ± 0.3	96.5 ± 0.2	94.5 ± 0.1	95.1 ± 0.3	96.3 ± 0.4	95.1 ± 0.9	95.2 ± 0.2	92.4 ± 0.8		94.4 ± 0.7
lipo	91.3 ± 0.7	88.5 ± 0.7	93.4 ± 0.7	89.4 ± 0.8	90.2 ± 0.6	90.9 ± 0.6	90.6 ± 0.9	90.8 ± 0.5	89.0 ± 0.5	89.6 ± 0.6	
mean	93.1 ± 1.8	90.3 ± 1.5	94.4 ± 1.1	90.6 ± 1.6	91.7 ± 1.6	92.5 ± 1.7	91.7 ± 1.5	92.4 ± 1.3	90.7 ± 1.2	90.6 ± 0.9	89.6 ± 2.2

Pretraining Additional Motif Datasets, 3000 samples, Test											
left out dataset ↓	$\text{MUV}_{\text{motif}}$	$\text{HIV}_{\text{motif}}$	$\text{BACE}_{\text{motif}}$	$\text{BBBP}_{\text{motif}}$	$\text{TOX21}_{\text{motif}}$	$\text{TOXCAST}_{\text{motif}}$	$\text{SIDER}_{\text{motif}}$	$\text{CLINTOX}_{\text{motif}}$	$\text{ESOL}_{\text{motif}}$	$\text{FREESOLV}_{\text{motif}}$	$\text{LIPO}_{\text{motif}}$
muv		88.3 ± 0.1	89.2 ± 0.1	89.9 ± 0.5	90.6 ± 0.5	89.5 ± 0.4	89.6 ± 0.2	88.2 ± 0.7	86.7 ± 1.8	81.2 ± 0.5	91.1 ± 0.4
hiv	90.9 ± 1.5		89.8 ± 0.4	90.2 ± 1.3	90.9 ± 1.3	89.9 ± 1.3	90.5 ± 1.0	87.7 ± 1.2	87.4 ± 1.5	81.9 ± 1.0	91.5 ± 1.6
bace	93.4 ± 0.5	90.4 ± 0.9		92.1 ± 0.4	92.7 ± 0.6	92.0 ± 0.7	91.4 ± 1.0	88.7 ± 0.6	88.5 ± 1.4	82.9 ± 0.3	93.1 ± 0.9
bbbp	94.2 ± 0.9	91.1 ± 1.1	93.3 ± 0.9		93.6 ± 0.9	92.7 ± 1.1	92.5 ± 0.4	90.8 ± 1.4	89.7 ± 1.9	83.9 ± 0.7	93.8 ± 0.8
tox21	91.3 ± 2.6	89.1 ± 2.0	89.5 ± 1.9	90.5 ± 1.7		90.1 ± 2.3	90.3 ± 1.5	88.1 ± 1.7	86.6 ± 1.8	83.0 ± 0.5	91.5 ± 2.4
toxcast	92.7 ± 1.4	89.4 ± 1.0	91.0 ± 1.1	91.0 ± 1.2	91.6 ± 1.1		90.7 ± 1.0	88.5 ± 1.5	88.3 ± 1.1	81.1 ± 2.3	92.3 ± 1.2
sider	94.0 ± 1.6	91.2 ± 1.3	92.4 ± 1.8	92.6 ± 1.3	93.3 ± 1.5	92.4 ± 1.7		90.5 ± 1.1	89.5 ± 1.5	83.2 ± 1.2	94.1 ± 1.3
clintox	91.3 ± 1.4	88.5 ± 1.3	90.0 ± 1.6	90.9 ± 1.2	90.5 ± 1.3	89.3 ± 1.4	89.6 ± 1.1		86.6 ± 1.0	81.0 ± 1.1	91.1 ± 1.7
esol	93.0 ± 1.7	89.9 ± 1.4	91.4 ± 1.5	91.8 ± 1.4	92.2 ± 1.8	91.2 ± 1.7	91.7 ± 1.2	88.8 ± 1.8		82.4 ± 0.8	92.0 ± 1.5
freesolv	96.0 ± 0.3	93.5 ± 0.1	94.6 ± 0.7	95.5 ± 0.2	95.6 ± 0.3	94.8 ± 0.2	95.6 ± 0.3	93.0 ± 0.3	90.4 ± 0.9		95.7 ± 0.3
lipo	91.0 ± 0.5	88.3 ± 0.5	89.3 ± 1.0	89.8 ± 0.1	90.2 ± 0.5	89.1 ± 0.6	89.4 ± 0.2	88.5 ± 0.3	86.8 ± 1.6	80.9 ± 1.9	
mean	92.8 ± 1.7	90.0 ± 1.6	91.0 ± 1.9	91.4 ± 1.7	92.1 ± 1.7	91.1 ± 1.9	91.1 ± 1.9	89.3 ± 1.6	88.0 ± 1.4	82.2 ± 1.1	92.6 ± 1.5

Table C.9: Validation and Test Performance during pretraining with additional motif datasets and 3000 samples.

C.5 Only Motif Datasets

C.5.1 500 samples

Pretraining only on motif datasets, 500 samples, Validation											
	ROC-AUC(%) ↑										
Dataset	MUV_{motif}	HIV_{motif}	$BACE_{motif}$	$BBBP_{motif}$	$TOX21_{motif}$	$TOXCAST_{motif}$	$SIDER_{motif}$	$CLINTOX_{motif}$	$ESOL_{motif}$	$FREESOLV_{motif}$	$LIPO_{motif}$
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	85	85	85	85	85	85	85	85	85	85	85
left out dataset											
muv		96.5 ± 0.1	99.2 ± 0.1	97.2 ± 0.3	98.2 ± 0.3	98.4 ± 0.3	98.0 ± 0.2	98.4 ± 0.1	95.3 ± 0.4	93.8 ± 1.2	97.5 ± 0.2
hiv	97.7 ± 0.0		99.2 ± 0.1	97.0 ± 0.3	98.2 ± 0.1	98.3 ± 0.2	97.8 ± 0.1	98.6 ± 0.2	94.9 ± 0.1	94.2 ± 0.8	98.1 ± 0.4
bace	97.4 ± 0.3	96.5 ± 0.2		97.0 ± 0.1	98.2 ± 0.2	98.5 ± 0.2	98.0 ± 0.3	98.5 ± 0.1	95.3 ± 0.8	93.6 ± 0.8	98.1 ± 0.3
bbbp	97.4 ± 0.2	96.8 ± 0.1	99.2 ± 0.2		98.3 ± 0.1	98.5 ± 0.1	98.0 ± 0.1	98.2 ± 0.1	95.2 ± 0.5	93.4 ± 0.4	98.1 ± 0.3
tox21	97.8 ± 0.4	96.6 ± 0.3	99.2 ± 0.1	97.4 ± 0.2		98.4 ± 0.1	97.9 ± 0.2	98.4 ± 0.1	94.9 ± 0.6	93.9 ± 0.3	98.0 ± 0.4
toxcast	97.7 ± 0.4	96.8 ± 0.1	99.2 ± 0.2	97.3 ± 0.0	98.4 ± 0.2		97.9 ± 0.2	98.6 ± 0.1	94.7 ± 1.5	94.2 ± 0.6	98.1 ± 0.2
sider	97.9 ± 0.6	96.8 ± 0.2	99.3 ± 0.1	97.3 ± 0.3	98.3 ± 0.2	98.6 ± 0.2		98.3 ± 0.1	95.5 ± 0.7	93.8 ± 1.3	98.1 ± 0.3
clintox	97.7 ± 0.1	96.5 ± 0.1	99.2 ± 0.1	97.1 ± 0.4	98.3 ± 0.1	98.6 ± 0.0	97.7 ± 0.3		95.2 ± 0.6	93.2 ± 0.4	98.2 ± 0.3
esol	97.7 ± 0.1	96.9 ± 0.1	99.3 ± 0.0	97.4 ± 0.4	98.4 ± 0.1	98.6 ± 0.1	98.0 ± 0.2	98.6 ± 0.0		94.0 ± 1.0	98.4 ± 0.0
freesolv	97.8 ± 0.2	96.8 ± 0.2	99.3 ± 0.1	97.6 ± 0.1	98.3 ± 0.1	98.4 ± 0.0	98.1 ± 0.1	98.4 ± 0.1	95.1 ± 0.2		98.2 ± 0.1
lipo	97.6 ± 0.3	96.5 ± 0.1	99.3 ± 0.1	97.4 ± 0.3	98.1 ± 0.1	98.4 ± 0.2	97.6 ± 0.1	98.5 ± 0.2	95.1 ± 0.9	94.3 ± 1.3	
mean	97.7 ± 0.2	96.7 ± 0.2	99.2 ± 0.1	97.3 ± 0.2	98.3 ± 0.1	98.5 ± 0.1	97.9 ± 0.2	98.5 ± 0.1	95.1 ± 0.2	93.8 ± 0.4	98.1 ± 0.2

Pretraining only on motif datasets, 500 samples, Test											
	ROC-AUC(%) ↑										
Dataset	MUV_{motif}	HIV_{motif}	$BACE_{motif}$	$BBBP_{motif}$	$TOX21_{motif}$	$TOXCAST_{motif}$	$SIDER_{motif}$	$CLINTOX_{motif}$	$ESOL_{motif}$	$FREESOLV_{motif}$	$LIPO_{motif}$
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	85	85	85	85	85	85	85	85	85	85	85
left out dataset											
muv		96.0 ± 0.3	98.0 ± 0.2	98.0 ± 0.4	98.6 ± 0.3	97.9 ± 0.2	98.5 ± 0.1	96.4 ± 0.1	93.6 ± 0.8	83.0 ± 1.5	97.4 ± 0.5
hiv	97.2 ± 0.1		98.2 ± 0.1	98.1 ± 0.3	98.5 ± 0.1	98.2 ± 0.1	98.6 ± 0.2	96.5 ± 0.3	93.4 ± 0.4	84.0 ± 2.7	97.8 ± 0.4
bace	97.5 ± 0.5	96.3 ± 0.4		98.2 ± 0.1	98.6 ± 0.1	98.0 ± 0.0	98.3 ± 0.1	96.8 ± 0.2	93.9 ± 1.3	84.4 ± 1.5	98.0 ± 0.3
bbbp	97.9 ± 0.2	96.4 ± 0.2	98.1 ± 0.2		98.7 ± 0.1	98.0 ± 0.1	98.2 ± 0.2	95.7 ± 0.6	92.9 ± 0.5	83.1 ± 0.7	98.0 ± 0.1
tox21	98.3 ± 0.1	96.2 ± 0.4	98.2 ± 0.2	97.7 ± 0.2		98.0 ± 0.2	98.4 ± 0.2	96.5 ± 0.3	93.2 ± 0.5	85.1 ± 1.3	98.1 ± 0.3
toxcast	98.0 ± 0.2	96.6 ± 0.1	98.2 ± 0.2	98.2 ± 0.2	98.7 ± 0.1		98.6 ± 0.1	97.0 ± 0.4	92.8 ± 0.3	83.9 ± 1.4	97.9 ± 0.2
sider	98.1 ± 0.3	96.6 ± 0.2	98.1 ± 0.2	97.5 ± 0.3	98.7 ± 0.2	98.2 ± 0.2		96.2 ± 0.2	93.3 ± 0.7	83.8 ± 0.5	97.8 ± 0.2
clintox	98.3 ± 0.1	96.5 ± 0.2	98.1 ± 0.2	97.6 ± 0.1	98.5 ± 0.2	98.2 ± 0.1	98.7 ± 0.1		93.1 ± 1.0	83.9 ± 1.0	98.2 ± 0.3
esol	98.1 ± 0.1	96.4 ± 0.2	98.3 ± 0.2	98.2 ± 0.3	98.7 ± 0.1	98.2 ± 0.2	98.4 ± 0.1	96.7 ± 0.3		83.1 ± 1.2	97.9 ± 0.4
freesolv	98.3 ± 0.1	96.6 ± 0.2	98.3 ± 0.2	97.9 ± 0.4	98.7 ± 0.1	98.1 ± 0.2	98.6 ± 0.1	96.3 ± 0.5	93.5 ± 0.4		97.8 ± 0.1
lipo	98.0 ± 0.3	96.4 ± 0.2	97.9 ± 0.3	97.8 ± 0.4	98.8 ± 0.1	98.3 ± 0.2	98.7 ± 0.2	96.6 ± 0.3	93.0 ± 0.7	84.9 ± 2.4	
mean	98.0 ± 0.4	96.4 ± 0.2	98.1 ± 0.1	97.9 ± 0.3	98.7 ± 0.1	98.1 ± 0.1	98.5 ± 0.2	96.5 ± 0.4	93.3 ± 0.3	83.9 ± 0.7	97.9 ± 0.2

Table C.10: Validation and Test Performance during pretraining only on motif datasets and 500 samples.

C.5.2 3000 samples

Pretraining only on motif datasets, 3000 samples, Validation											
	ROC-AUC(%) ↑										
Dataset	MUV_{motif}	HIV_{motif}	$BACE_{motif}$	$BBBP_{motif}$	$TOX21_{motif}$	$TOXCAST_{motif}$	$SIDER_{motif}$	$CLINTOX_{motif}$	$ESOL_{motif}$	$FREESOLV_{motif}$	$LIPO_{motif}$
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	85	85	85	85	85	85	85	85	85	85	85
left out dataset											
muv		99.2 ± 0.1	99.9 ± 0.1	99.3 ± 0.2	99.8 ± 0.1	99.8 ± 0.1	99.5 ± 0.1	99.8 ± 0.0	96.3 ± 0.3	93.4 ± 0.4	99.9 ± 0.0
hiv	99.6 ± 0.1		99.7 ± 0.0	99.5 ± 0.2	99.9 ± 0.1	99.8 ± 0.0	99.8 ± 0.0	99.7 ± 0.1	96.7 ± 0.6	93.5 ± 0.8	99.8 ± 0.1
bace	99.9 ± 0.0	99.2 ± 0.1		99.5 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.9 ± 0.0	99.8 ± 0.0	96.2 ± 0.2	93.4 ± 0.5	100.0 ± 0.0
bbbp	99.8 ± 0.1	99.2 ± 0.1	99.9 ± 0.0		99.9 ± 0.0	99.9 ± 0.1	99.8 ± 0.0	99.8 ± 0.0	96.4 ± 0.1	94.7 ± 0.3	99.9 ± 0.0
tox21	99.8 ± 0.1	99.1 ± 0.2	99.9 ± 0.1	99.7 ± 0.1		99.9 ± 0.1	99.7 ± 0.1	99.6 ± 0.1	96.5 ± 0.3	94.2 ± 0.6	99.9 ± 0.0
toxcast	99.6 ± 0.2	99.3 ± 0.2	99.9 ± 0.1	99.6 ± 0.1	99.9 ± 0.0		99.6 ± 0.1	99.8 ± 0.1	96.9 ± 0.6	93.3 ± 0.9	99.9 ± 0.0
sider	99.8 ± 0.1	99.3 ± 0.1	99.9 ± 0.1	99.6 ± 0.1	99.9 ± 0.0	99.9 ± 0.0		99.6 ± 0.1	96.7 ± 0.1	93.1 ± 0.3	100.0 ± 0.0
clintox	99.8 ± 0.1	99.2 ± 0.0	99.9 ± 0.0	99.5 ± 0.1	99.9 ± 0.1	99.9 ± 0.0	99.7 ± 0.1		96.3 ± 0.3	94.0 ± 0.4	99.9 ± 0.1
esol	99.8 ± 0.1	99.3 ± 0.1	100.0 ± 0.0	99.6 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.9 ± 0.1	99.8 ± 0.1		94.3 ± 0.6	100.0 ± 0.0
freesolv	99.9 ± 0.1	99.2 ± 0.2	99.9 ± 0.1	99.7 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.8 ± 0.2	99.8 ± 0.0	97.0 ± 0.9		100.0 ± 0.0
lipo	99.8 ± 0.1	99.2 ± 0.1	99.9 ± 0.1	99.6 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.7 ± 0.1	99.8 ± 0.0	96.6 ± 0.7	94.1 ± 0.8	
mean	99.8 ± 0.1	99.2 ± 0.1	99.9 ± 0.1	99.6 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.7 ± 0.1	99.7 ± 0.1	96.6 ± 0.3	93.8 ± 0.5	99.9 ± 0.1

Pretraining only on motif datasets, 3000 samples, Test											
	ROC-AUC(%) ↑										
Dataset	MUV_{motif}	HIV_{motif}	$BACE_{motif}$	$BBBP_{motif}$	$TOX21_{motif}$	$TOXCAST_{motif}$	$SIDER_{motif}$	$CLINTOX_{motif}$	$ESOL_{motif}$	$FREESOLV_{motif}$	$LIPO_{motif}$
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	85	85	85	85	85	85	85	85	85	85	85
left out dataset											
muv		98.8 ± 0.0	99.5 ± 0.0	99.1 ± 0.3	99.8 ± 0.1	99.8 ± 0.1	99.5 ± 0.1	98.1 ± 0.1	94.0 ± 0.4	83.5 ± 0.8	99.8 ± 0.1
hiv	99.2 ± 0.2		99.5 ± 0.2	99.1 ± 0.2	99.8 ± 0.0	99.8 ± 0.1	99.5 ± 0.0	98.4 ± 0.3	93.4 ± 0.8	83.7 ± 0.5	99.4 ± 0.1
bace	99.3 ± 0.2	99.2 ± 0.1		99.2 ± 0.3	99.9 ± 0.0	99.9 ± 0.0	99.5 ± 0.0	98.5 ± 0.1	94.0 ± 0.8	84.5 ± 1.1	99.6 ± 0.1
bbbp	99.5 ± 0.0	99.0 ± 0.1	99.5 ± 0.1		99.9 ± 0.0	99.9 ± 0.0	99.5 ± 0.1	98.4 ± 0.1	93.8 ± 0.5	84.3 ± 0.7	99.9 ± 0.0
tox21	99.5 ± 0.1	98.9 ± 0.1	99.5 ± 0.0	98.9 ± 0.2		99.8 ± 0.1	99.2 ± 0.1	98.7 ± 0.1	93.8 ± 0.2	84.6 ± 0.5	99.9 ± 0.1
toxcast	99.0 ± 0.2	98.9 ± 0.1	99.5 ± 0.1	99.2 ± 0.1	99.8 ± 0.0		99.3 ± 0.1	98.7 ± 0.1	94.0 ± 0.4	84.2 ± 1.4	99.9 ± 0.0
sider	99.1 ± 0.3	98.9 ± 0.2	99.4 ± 0.1	99.0 ± 0.1	99.9 ± 0.0	99.9 ± 0.0		98.6 ± 0.1	93.4 ± 0.5	84.2 ± 0.4	99.7 ± 0.2
clintox	99.6 ± 0.2	98.8 ± 0.1	99.4 ± 0.1	98.9 ± 0.2	99.9 ± 0.0	99.9 ± 0.1	99.6 ± 0.1		93.8 ± 0.4	84.7 ± 1.9	99.8 ± 0.2
esol	99.3 ± 0.4	98.9 ± 0.2	99.6 ± 0.1	99.0 ± 0.0	99.9 ± 0.0	99.9 ± 0.1	99.6 ± 0.1	98.6 ± 0.0		84.5 ± 0.7	99.9 ± 0.1
freesolv	99.2 ± 0.4	99.1 ± 0.1	99.6 ± 0.0	99.1 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.7 ± 0.1	98.7 ± 0.2	93.6 ± 0.2		99.8 ± 0.2
lipo	99.5 ± 0.1	99.0 ± 0.2	99.5 ± 0.0	99.1 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.5 ± 0.1	98.8 ± 0.2	93.6 ± 0.5	83.8 ± 1.2	
mean	99.3 ± 0.2	99.0 ± 0.1	99.5 ± 0.1	99.1 ± 0.1	99.9 ± 0.0	99.9 ± 0.0	99.5 ± 0.1	98.5 ± 0.2	93.7 ± 0.2	84.2 ± 0.4	99.8 ± 0.2

Table C.11: Validation and Test Performance during pretraining only on motif datasets and 3000 samples.

Appendix D

No GNN Finetuning

Here we provide in addition to the test scores, that have been already presented in the results chapter, the validation scores of the "No GNN Finetuning" setup.

No GNN Finetuning, Validation											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
random GNN-weights	62.4 ± 3.9	54.6 ± 1.7	57.8 ± 5.2	71.2 ± 6.2	65.7 ± 1.8	54.1 ± 0.8	52.7 ± 1.3	57.3 ± 12.7	3.747 ± 0.169	7.296 ± 0.092	1.732 ± 0.110
pretrained$_{0.5k}$ (default)	71.5 ± 2.3	69.3 ± 1.4	55.0 ± 4.9	83.5 ± 1.4	70.7 ± 0.5	62.6 ± 0.9	55.7 ± 0.7	79.5 ± 2.4	1.476 ± 0.024	6.250 ± 0.222	1.136 ± 0.027
pretrained$_{3k}$ (default)	71.3 ± 4.7	70.3 ± 1.7	60.0 ± 3.7	83.1 ± 0.1	69.5 ± 0.5	63.0 ± 1.0	55.7 ± 0.5	79.9 ± 2.0	1.405 ± 0.021	6.233 ± 0.163	1.135 ± 0.004
pretrained$_{0.5k}$ (+ motif labels)	73.5 ± 2.2	71.4 ± 1.3	69.6 ± 0.9	91.7 ± 0.4	74.4 ± 0.6	63.4 ± 1.1	59.6 ± 0.8	82.3 ± 5.5	1.798 ± 0.015	6.522 ± 0.135	1.084 ± 0.010
pretrained$_{3k}$ (+ motif labels)	72.9 ± 2.5	72.2 ± 1.2	73.2 ± 1.7	93.0 ± 0.3	76.7 ± 0.2	65.3 ± 0.2	60.8 ± 0.5	85.3 ± 3.6	1.516 ± 0.112	6.303 ± 0.139	1.005 ± 0.001
pretrained$_{0.25k}$ (+ motif datasets)	70.4 ± 2.5	70.0 ± 0.7	58.1 ± 1.0	86.3 ± 0.8	70.2 ± 0.9	62.7 ± 0.5	56.1 ± 0.1	78.3 ± 3.0	1.624 ± 0.035	5.354 ± 0.387	1.138 ± 0.012
pretrained$_{0.5k}$ (+ motif datasets)	70.9 ± 1.9	70.0 ± 2.9	57.8 ± 1.6	83.5 ± 1.6	69.8 ± 0.9	62.0 ± 0.8	55.7 ± 1.0	78.7 ± 1.6	1.434 ± 0.061	5.494 ± 0.080	1.131 ± 0.010
pretrained$_{1.5k}$ (+ motif datasets)	74.8 ± 2.3	68.7 ± 0.8	59.9 ± 3.5	87.2 ± 1.2	70.9 ± 0.8	62.6 ± 0.4	56.5 ± 0.6	82.6 ± 2.9	1.339 ± 0.016	5.806 ± 0.010	1.103 ± 0.014
pretrained$_{3k}$ (+ motif datasets)	75.6 ± 1.1	69.4 ± 0.7	61.4 ± 3.0	89.6 ± 1.0	72.8 ± 1.6	63.5 ± 0.9	57.9 ± 0.8	81.8 ± 1.5	1.349 ± 0.021	5.941 ± 0.066	1.028 ± 0.010
pretrained$_{0.5k}$ (only motif datasets)	72.1 ± 0.5	71.3 ± 1.0	72.5 ± 0.6	91.9 ± 0.3	71.7 ± 0.4	61.6 ± 0.5	59.7 ± 0.5	85.8 ± 2.1	1.962 ± 0.043	6.269 ± 0.234	1.080 ± 0.016
pretrained$_{3k}$ (only motif datasets)	68.5 ± 2.3	71.9 ± 1.8	75.2 ± 1.4	92.3 ± 1.0	69.2 ± 0.7	61.4 ± 0.4	61.5 ± 0.5	84.6 ± 2.1	1.825 ± 0.064	6.290 ± 0.254	1.043 ± 0.004

No GNN Finetuning, Test											
	ROC-AUC(%) ↑								RMSE ↓		
Dataset	MUV	HIV	BACE	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	ESOL	FREESOLV	LIPO
# compounds	93087	41127	1513	2039	7831	8577	1427	1480	1128	642	4200
# tasks	17	1	1	1	12	617	27	2	1	1	1
random GNN-weights	59.5 ± 2.1	51.3 ± 4.3	68.2 ± 4.7	53.7 ± 2.2	64.0 ± 2.3	54.1 ± 0.3	52.1 ± 0.4	41.0 ± 4.1	3.801 ± 0.174	5.679 ± 0.109	1.795 ± 0.121
pretrained$_{0.5k}$ (default)	63.7 ± 1.7	68.6 ± 1.2	67.0 ± 2.5	54.8 ± 0.7	68.1 ± 0.8	61.3 ± 0.3	54.0 ± 0.9	63.7 ± 5.6	1.642 ± 0.018	4.681 ± 0.154	1.067 ± 0.020
pretrained$_{3k}$ (default)	68.0 ± 4.9	69.9 ± 2.1	67.0 ± 2.3	57.2 ± 3.3	68.8 ± 0.6	61.8 ± 0.2	57.1 ± 0.6	62.9 ± 2.2	1.636 ± 0.057	4.330 ± 0.116	1.071 ± 0.011
pretrained$_{0.5k}$ (+ motif labels)	74.5 ± 0.5	68.3 ± 3.4	72.2 ± 0.3	65.2 ± 1.1	71.6 ± 0.7	61.0 ± 0.2	59.3 ± 0.1	52.5 ± 1.9	1.802 ± 0.052	4.746 ± 0.132	0.984 ± 0.004
pretrained$_{3k}$ (+ motif labels)	72.7 ± 0.6	71.5 ± 1.0	73.1 ± 1.9	67.1 ± 1.6	75.0 ± 0.1	63.0 ± 0.8	59.3 ± 0.8	65.2 ± 3.1	1.468 ± 0.040	4.519 ± 0.122	0.952 ± 0.012
pretrained$_{0.25k}$ (+ motif datasets)	63.1 ± 1.5	67.8 ± 1.4	61.6 ± 1.7	56.1 ± 1.0	67.8 ± 0.1	60.4 ± 0.2	55.3 ± 0.5	59.0 ± 0.6	1.683 ± 0.035	3.958 ± 0.368	1.048 ± 0.024
pretrained$_{0.5k}$ (+ motif datasets)	67.6 ± 3.0	69.5 ± 1.9	67.7 ± 0.8	55.6 ± 0.7	67.7 ± 1.1	60.3 ± 0.4	56.8 ± 0.5	58.8 ± 2.6	1.496 ± 0.015	4.005 ± 0.028	1.047 ± 0.015
pretrained$_{1.5k}$ (+ motif datasets)	70.8 ± 0.7	67.0 ± 2.1	66.7 ± 2.4	57.6 ± 1.6	70.6 ± 0.3	61.1 ± 0.5	59.8 ± 0.5	61.4 ± 0.9	1.466 ± 0.092	4.081 ± 0.061	1.014 ± 0.017
pretrained$_{3k}$ (+ motif datasets)	73.7 ± 1.0	67.9 ± 2.3	71.9 ± 1.8	60.2 ± 1.0	72.1 ± 1.4	63.2 ± 0.8	59.6 ± 1.0	61.5 ± 3.6	1.415 ± 0.035	4.263 ± 0.093	0.943 ± 0.014
pretrained$_{0.5k}$ (only motif datasets)	73.0 ± 2.1	66.6 ± 1.2	74.0 ± 0.4	65.4 ± 1.3	70.0 ± 0.7	60.3 ± 0.1	58.4 ± 0.6	60.1 ± 1.0	1.899 ± 0.059	4.578 ± 0.313	1.011 ± 0.021
pretrained$_{3k}$ (only motif datasets)	68.7 ± 1.4	64.3 ± 3.9	73.4 ± 1.4	67.0 ± 0.7	68.7 ± 0.6	59.8 ± 0.2	57.3 ± 0.6	68.8 ± 3.7	1.866 ± 0.055	4.499 ± 0.088	0.988 ± 0.003

Table D.1: Validation and Test performance when finetuning only the decoder for 5 epochs.