# Using Domain-specific Fingerprints Generated Through Neural Networks to Enhance Ligand-based Virtual Screening.

**Janosch Menke**
Institute of Pharmaceutical and Medicinal Chemistry
Westfälische Wilhelms-Universität Münster
Corrensstraße 48, 48149 Münster

**Oliver Koch***
Institute of Pharmaceutical and Medicinal Chemistry
Westfälische Wilhelms-Universität Münster
Corrensstraße 48, 48149 Münster

Center for Multiscale Theory and Computation
Westfälische Wilhelms-Universität Münster
Corrensstraße 40, 48149 Münster

oliver.koch@uni-muenster.de

## Abstract

Molecular fingerprints are essential for different cheminformatics approaches like similarity-based virtual screening. In this work, the concept of neural (network) fingerprints in the context of similarity search is introduced in which the activation of the last hidden layer of a trained neural network represents the molecular fingerprint. The neural fingerprint performance of five different neural network architectures was analyzed and compared to the well-established Extended Connectivity Fingerprint (ECFP) and an autoencoder-based fingerprint. This is done using a published compound dataset with known bioactivity on 160 different kinase targets. We expect neural networks to combine information about the molecular space of already known bioactive compounds together with the information on the molecular structure of the query and by doing so enrich the fingerprint. The results show that indeed neural fingerprints can greatly improve the performance of similarity searches. Most importantly, it could be shown that the neural fingerprint performs well even for kinase targets that were not included in the training. Surprisingly, while Graph Neural Networks (GNNs) are thought to offer an advantageous alternative, the best performing neural fingerprints were based on traditional fully connected layers using the ECFP4 as input. The best performing kinase-specific neural fingerprint will be provided for public use.

## Introduction

A backbone of cheminformatics and computer-aided drug design is the ability to translate molecules into a computer-readable format. While formats such as SMILES[1], InCHI Key[2], and several file formats represent molecules in their entirety, other encoding strategies are better suited for computational methods. *Molecular fingerprints* convert molecules into numeric vectors of fixed length. Instead of encoding the complete molecule, this representation captures structural characteristics and chemical properties. The MACCS 166 key, for example, encodes the presence of predefined substructures into a bit vector of length 166 (Figure 1a).[3] *Hashed Fingerprints* such as the Extended Connectivity Fingerprint (ECFP) do not rely on such substructure dictionaries but use a hash function to record information for each atoms neighborhood up to a prespecified diameter (Figure 1b).[4]

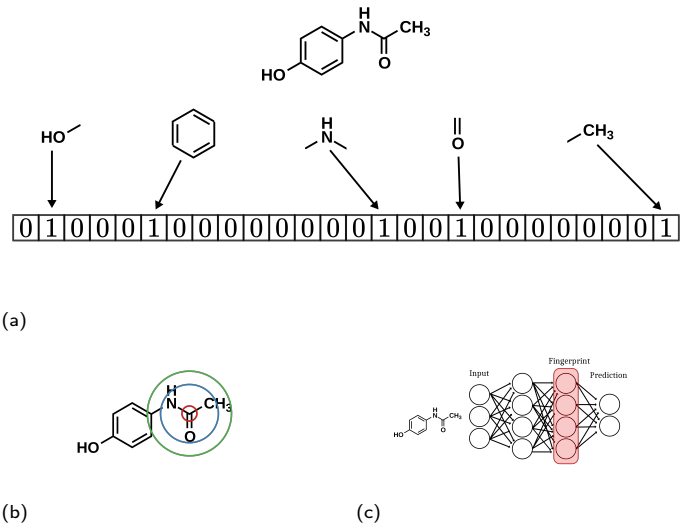



Figure 1 Examples of different fingerprint generation strategies. (a) In dictionary-based Fingerprint each bit corresponds to a substructure. (b) The ECFP encodes substructures based on the neighborhood of each atom. (c) Neural Fingerprints are generated through the training of neural networks

A central dogma in drug design is that similar molecules have similar properties and should bind to the same drug target.[5] Fingerprints can be used to compare molecules based on their similarity which is applied in fingerprint-based virtual screening. They are used to analyze virtual compound libraries for the identification of similar molecules with presumably similar bioactivity.[6] The choice of fingerprint is crucial to the success of such a similarity search, as different fingerprints focus on different chemical domains and properties.[7]

Machine learning (ML) based approaches offer an alternative approach for identifying molecules with bioactivity of interest. Rather than searching for similar molecules, machine learning models are trained to predict the activities of molecules based on their fingerprints.[8–11] This bypasses the need for similarity search but these approaches still rely, at its core, on precalculated fingerprints. A new class of ML algorithms, called Graph Neural Networks (GNN) are thought to overcome the calculation of fingerprints.[12] These networks can handle molecules directly and learn how to encode them during the training process. This allows GNNs to change the way molecules are transformed depending on the task they are being trained for. The hope is that GNNs can learn and encode structural information relevant for a specific task, something that traditional fingerprints cannot do.

One of the earliest usages of Graph Neural Networks in the context of molecular fingerprints was described by Duvenaud et. al.[13] They introduced a GNN that mimics the ECFP. Later research expanded on the idea with more complex models but still relate to the concept of chemical fingerprints.[12,14] These papers explicitly state the connection to molecular fingerprints, but virtually any Graph Neural Network (technically any neural network) that has been trained on molecules has the same capabilities. These neural fingerprints are simply the activation of a specific hidden layer in the network. As Kearnes[12] states: "[...] graph convolutions and other graph-based approaches purposefully blur the distinction between molecular features and predictive models".

Studies, so far, focused on the possible improvements of prediction achieved through the use of GNNs.[9,15–17] A few studies evaluate and compare the quality of fingerprints that can be obtained from such networks. In a paper by Hirohara et. al.[18] the chemical space covered by their fingerprint is compared to the one of the ECFP. In the already mentioned paper by Duvenaud[13], the similarity between the ECFP and the GNN fingerprint is assessed. In a study by Winter et. al.[19] a similarity search using a neural fingerprint based on an autoencoder is conducted. However, those fingerprints are generated through unsupervised training. While unsupervised models are trained to reconstruct or translate molecular representations[20–22], supervised models are trained for a specific task, like predicting the activity of molecules. This allows unsupervised models to be used across many domains, but they cannot encode any domain-specific/task-specific information like the supervised models. So far, the similarity search performance of fingerprints extracted from supervised models has been neglected. Therefore, the approach presented in this paper is focusing on the idea to use and evaluate (supervised) neural network fingerprints for similarity-based virtual screening.

The here proposed idea aims to combine the learned structural features of a trained model with the specific structural information of a query compound so that relevant information beyond target affinity is included. Typical structural features of a class of ligands, for example, kinase inhibitors, are learned by the trained model and incorporated into the fingerprint. Beyond ligand-class features, molecules similar to the query are expected to share the additional properties and leading to a lower false-positive rate. Choosing an appropriate query molecule with desired properties increases the chance of finding hits that are not only active on the same target but also share desired properties. In contrast, many application of Machine Learning focus solely on target or property prediction for a given molecule.[23] The disadvantage is, that such prediction models can only take into account properties they are being trained for. For that reason, similarity search remains an important and powerful tool in the field of computer-aided drug discovery.

In this study, we evaluate the performance of fingerprints extracted from neural networks. We investigate whether fingerprint-based virtual screening can benefit from (implicit) incorporation of additional target information. Different neural networks were trained on a large kinase dataset with known active kinase inhibitors and inactive molecules.[24] The activations of the last hidden layer are used as a neural fingerprint for a fingerprint-based virtual screening (see Figure 1c). The expectation was that molecules sharing similar bioactivity and similar structural features will also have similar activations in the final hidden layers. These fingerprints are expected to carry information on the molec-

ular space of already known kinase inhibitors together with the information on the molecular structure of the query. This specific information should elevate the performance of the fingerprint in comparison to traditional fingerprints. The initial assumption was that many ATP-competitive kinase inhibitors share similar structural elements, e.g. the hinge binder[25] and such information would be learned implicitly across many different kinase targets.

This method of extracting the activations of a hidden layer to generate fingerprints is not limited to GNNs since the activation can also be extracted from neural networks that use a traditional fingerprint as input. Deeper in the network activation should also contain implicit information on that target. Thus, such a process could transform static precomputed fingerprints to fingerprints which adapt to the task, similar to a GNN. Therefore, graph-neural networks and simpler feed-forward neural networks are trained to generate neural fingerprints which are compared to the ECFP4 with regards to performance in similarity search.

## Experimental Section

Two different architectures were used to generate and access the neural network fingerprints (see Figure 2). After training, the activations of the last hidden layer are used as the neural network fingerprint for similarity comparison of different molecules. The Graph Convolution based approach takes the molecular structure as input and encodes this into a vector that is fed into the fully-connected layers. The alternative approach directly uses the precalculated ECFP4 fingerprint as input for the fully connected layer.

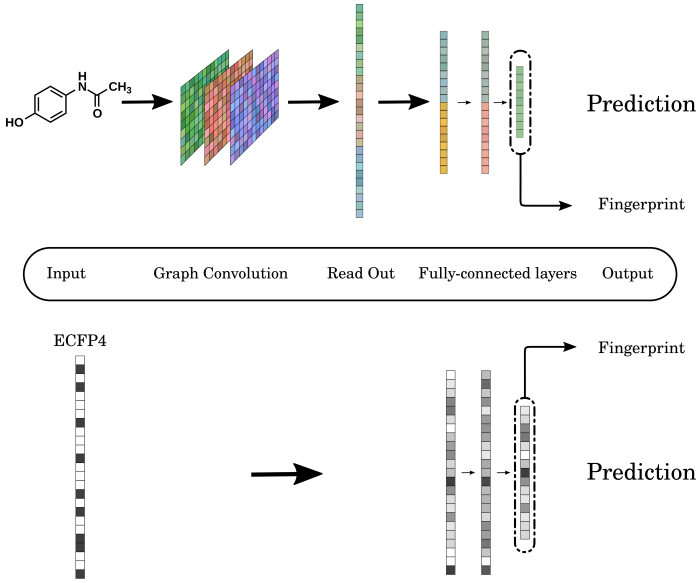



Figure 2 Architecture of Neural Networks.
*Top:* GNN-based Model using the molecular structure as input which is encoded and fed into a network of fully-connected layers.
*Bottom:* MLP-based architecture using a precalculated ECFP4 fingerprint as input for the fully-connected layer.

## Graph Convolution Network

The Graph Neural Network implemented in this paper can be separated into three distinct parts: the (1) graph convolution layer, the (2) read-out layer, and the (3) fully-connected layers.

*Graph Convolution*

The basic idea of most graph neural networks is that the values of nodes are updated by the aggregated information of its neighboring nodes/atoms[16] (see Figure 3). After updating the values, each node carries information on itself and also its direct neighbors. If the process is repeated a second time, each node will not only carry its direct neighborhood information but also the information of nodes two hops away. Repeatedly updating nodes by their neighbors grows the size of the neighborhood of which the nodes carry information. A similar process is used in the generation of the widely applied Extended Connectivity Fingerprint (ECFP). The ECFP uses a hash-function to aggregate the information while GNNs can make use of learnable weight matrices.[13] This difference allows GNNs to learn how to encode the molecule depending on the task. Many different variants of GNNs exist and differ in the way information is aggregated and passed through the graph.[16,26]

The Graph Neural Network architecture that we chose is an extension of the Graph Convolution Network (GCN) by Kipf & Welling (2017), here the nodes are updated by the mean of its neighborhood:

$$X^{l+1} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X^l W_1^l + X^l W_2^l)$$

The addition of the skip connection $W_2^l X^l$ was introduced by Cangea et. al.[27] and allows the model to carry information on smaller substructures deeper into the model. Here $X^l$ are the activation's of layer $l$. $W_1$ and $W_2$ are trainable weight matrices, $\hat{A}$ is the adjacency matrix with added self-loops ($\hat{A} = A + I$) and $\hat{D}$ is the degree matrix of $\hat{A}$. $\sigma$ represents the ReLU activation function.

The GCN requires two matrices for each molecule as input. The first one is the adjacency matrix $A$ of size $n \times n$ where $n$ is the number of atoms in the molecule. The adjacency matrix contains information on which atoms are connected via a bond. The second matrix is the feature matrix $X$ of size $n \times f$ where $f$ is the number of features. The second matrix, the feature matrix, contains the feature information for each atom in the molecule. Both matrices were calculated using RDKit[28] (version: 2009.03.04). The included features for the GCN are shown in Table 1.

Table 1 Features used in the GCN

| Feature |
|---|
| Atom Type |
| Degree |
| Hybridization |
| Formal Charge |
| Num. Implicit Hydrogen Atoms |
| In aromatic ring |

*Read-Out Layer*

After the input has passed through several graph convolutions the activations of the nodes have to be pooled to create a single latent vector for the graph. This can be done through a read-out layer, which performs a global pooling step on the activation of the graph convolution. To ensure that information on small substructures is not lost during the propagation through the convolution layers, the pooling is not only performed on the activations of the last layer but the activation's after each convolution are pooled. We slightly alter the read-out function from Cangea et. al.[27], where a global mean pooling step is combined with a global max pooling step. The pooled vector $p^l$ is calculated as follows:

$$\vec{p}^l = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^l \,||\, \max_{i=1}^{n}\vec{x}_i^l$$

where $\vec{x}_i^l$ is the activation of the $i$th node of the $l$th convolution and $||$ represents the concatenation. Different to Cangea, where

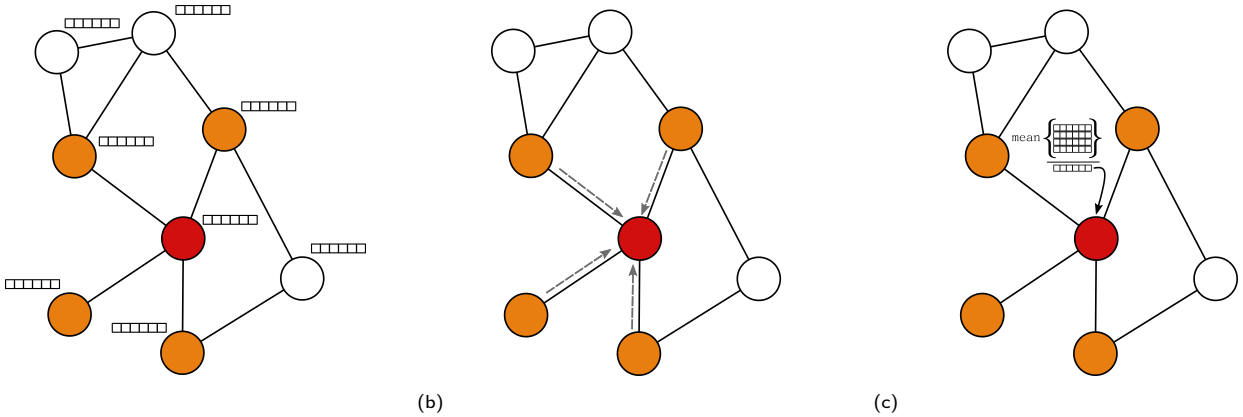



Figure 3 Each node/atom has a vector storing atom properties assigned to it (*a*). To update a node (red) the neighbouring nodes information is collected (*b*). Lastly the central node is updated by calculating the columns-wise mean (depending on the GNN used) over all neighbouring node vectors (*c*).

the pooled vectors $\vec{p} = \sum_{l=1}^{L} p^l$ are summed up across all the convolution layers $L$, we chose to concatenate the pooled vectors.

$$\vec{p} = \vec{p}^1 || \vec{p}^2 || ... || \vec{p}^{L-1} || \vec{p}^L$$

This process of pooling is similar to the way the ECFP encodes substructures as it not only includes the information of the last aggregation iteration but also earlier ones.

*Fully Connected Layers*

In the final step, the vector $\vec{p}$ is passed through a network of fully connected layers, with subsequent dropout and batch normalization, to make the final predictions (see Figure 4). The last hidden layer is the layer from which the neural fingerprint is extracted.

### Multi-Layer Perceptron Architecture.

The architecture of the MLP is identical to the fully connected layers of the GCN. But rather than receiving the input from the Graph Convolutions the MLP uses the ECFP as input. We chose the ECFP4 as the input fingerprint because it is one of the best-performing ones across many different tasks.[29] The ECFP4 of length 1024 for each molecule was calculated using RDkit. Because the GCN and ECFP have similar encoding strategies, the MLP can loosely be thought of as a GCN alternative in which the encoding of the molecule is not learnable. As in the GCN, the neural fingerprint is extracted from the last hidden layer.

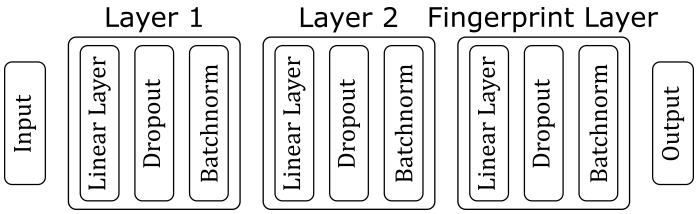



Figure 4 Detailed overview of the fully-connected layers. Each Linear Layer is followed up with a Dropout and Batchnormlayer. The last layer is used to extract the Fingerprint.

### Dataset and processing

For the training and testing of the models, a dataset on kinases activities was used. The dataset was created and provided on request by Pogodin et. al.[24] based on ChEMBL bioactivity data.[30] It contains 55,594 ligands with activities measured on 160 different human-protein kinases. All of the ligands are classified as ATP-competitive and were either scored as active or inactive depending on their inhibition rate. For more information on the data curation, we refer to the original paper. The dataset contains many ligands that were only tested a few kinases, which leads to the issue of molecules missing activity information on many kinases. We chose to consider ligands which did not have a recorded activity on a kinase as inactive on that specific kinases. The final dataset was made up out of the SMILES strings and the activities on all 160 kinases for each ligand. We used 5-fold cross-validation with a random split of 80%/10%/10% for the training, test, and external validation set. The external set was not only used to evaluate the performance of the model but also for the evaluation of the similarity search. All statistics reported are thus the mean across the five external validation sets. In a second experiment, we used a 60%/20%/20% split to evaluate the effect of decreased training set size. To investigate the ability of the fingerprint to generalize to out-of-sample targets, two additional experiments were conducted. First, to investigate the effect on leaving out inhibitors for specific kinases during the training process was investigated. For ten kinases we trained models ones including its inhibitors and ones excluding them and compared the change in performance afterward. In the second experiment, we kept all inhibitors in the training set but removed the target information for 64 (40%) kinases. Thus, those models were not trained to predict activities on 160 targets but only on 96.

### Model Training

Overall, five different models were created based on the GCN and MLP architectures. We wanted to investigate, whether training models to predict general kinase inhibition is already sufficient for an adequate fingerprint. For this, we trained generic models which only predicted whether a given molecule is a kinase inhibitor or not (single-task prediction). This was done for the MLP (MLP generic) and the GCN (GCN generic). The second kind of model we trained was a multitask model predicting which specific kinase the molecules are inhibiting. This leads to the four models: MLP generic and multitarget (MT), and GCN generic and multitarget (MT). The last model (Freeze MT) is based on a GCN with fixed weights of the convolution layers. When freezing these layers, the information is still aggregated as it passes through the network, but the weights of graph convolutions cannot be changed, and the encoding of molecules remains fixed. The random weights were

sampled from a normal distribution and the activations were normalized before being passed through the fully connected layers.

All models were created and trained with PyTorch (1.3.1)[31]. We used the Adam algorithm[32] for optimization. Additionally, we used Early Stopping with a patience of five based on the validation loss to prevent overfitting. The final models were evaluated based on the performance on the external validation set. Facebook's Open-source library *Ax*[33] was used for Bayesian hyper-parameter optimization. For each model, 100 trials were run. Each trial was evaluated based on the mean loss across the validation sets. Details on which parameters were optimized are presented in the Supporting Information. The final models were evaluated on the external validation set.

## Similarity Search

The created external validation set was the basis for the evaluation of the fingerprint-based virtual screening, where for each active molecule a similarity search was performed against all other molecules in this external validation set. The molecules were ranked afterward from highest to lowest similarity. This ranking was used to evaluate the performance of the similarity search. For the performance assessment, the AUC and different enrichment factors were calculated based on the different trained neural network architecture. For comparison, a baseline similarity search was performed using the standard ECFP4 (without any neural network). To assess the similarity, the Tanimoto coefficient was used for the ECFP4 and the cosine similarity for the neural fingerprints, since it has been proven that these measures are comparable.[34]

Additionally, we wanted to investigate how well neural fingerprints from an unsupervised model perform. For this, we chose to use the "continuous and data-driven molecular descriptors" (CDDD) proposed by Winter et. al.[19] The pretrained model provided in the original paper was used to generate a fingerprint of length 512. For 1730 of the molecules we were not able to produce a CDDD due to issues with the SMILES strings. Thus, the similarity search for the CDDD was only performed for molecules on which we were able to generate the CDDD. For evaluation of the similarity search we followed the protocol proposed by Riniker et. al.[35], with the slight adaptation that we used every active in the external validation set as a query.

## Performance assessment

### Area under the Curve (AUC).

The area under the Curve (AUC) measures the area underneath the Receiver Operating Characteristics (ROC) curve and provides an aggregated measure of model performance in classification problems ($> 0.5$ = better than random). The ROC curve is calculated using the true positive rate (sensitivity = $\frac{\text{True Positive}}{\text{False Positive+True Negative}}$) and false positive rate ($\frac{\text{False Positive}}{\text{False Positives+True Negatives}}$). The Curve is constructed by shifting the threshold for the binary classification. For each threshold, the false positive rate is plotted against the true positive rate. For a more detailed explanation, we refer to Sonego et. al.[36] The AUC was used for both, the assessment of the model performance and the fingerprint-based virtual screening approach.

### Enrichment factor.

The enrichment factor (EF) describes the enrichment of active kinase inhibitors as opposed to the number of inactive molecules. It is an alternative measure and evaluates how well algorithms perform for the best-ranked items. It is of great importance as in reality fingerprints are used to screen through millions of compounds. As resources are limited only the top-ranked compounds are can be considered for further evaluation. The EF for $x\%$ of the screened dataset is calculated based on the number of true actives in the top $x\%$ relative to the number of true actives in the complete dataset. $EF_{x\%} = \frac{N_{x\%}^{active}}{N_{x\%}^{total}} / \frac{N_{100\%}^{active}}{N_{100\%}^{total}}$. The $EF_{1\%} = 5$ indicates that five times more actives were are found in the Top 1% relative to the complete set of molecules. For the similarity search, we evaluated both the $EF_{1\%}$ and $EF_{2.5\%}$.

### Average Similarity.

The average similarity is calculated based on the ECFP4 similarity of the top 1% of the active hits in comparison to the query molecule. This is an indicator of the structural diversity of the identified hits relative to the ECFP4.

# Results & Discussion

The first step in creating the neural network fingerprints is to train and evaluate the prediction models. The initial idea was to learn the generic features that are most common amongst kinase inhibitors, e.g. a hinge binding motif. The generic models were implemented to only use information about general kinase activity or inactivity. As these models might run the risk of overgeneralizing, meaning that the neural fingerprint extracted from the network might only capture kinases specific information and "ignores" much of the information that describes the structure of the molecule. Hence, models for multitask predictions were trained that predict the specific kinase target activity. In general, a GCN with subsequent fully connected layers is in many ways similar to a multi-layer perceptron (MLP) that uses an ECFP fingerprint as input. Therefore, two different models using the ECFP4 fingerprint (MLP generic/ MLP MT) were also trained as a comparison to the GCN models. The big difference between both is that the GCN can learn how to encode a molecular structure while the ECFPs encoding remains fixed. One could also think about the MLP in combination with the ECFP as a "GNN" in which the convolution layers are not differentiable. Vice versa, freezing convolutional layers, mimics an MLP using an ECFP as input, since the information is still aggregated and passed through the network. However, as the weights of graph convolutions cannot be changed the encoding of molecules remains fixed. We can compare the performance of the Freeze MT to the performance of the GCN MT to assess the effects that the learnable feature encoding of the GCN has on the fingerprint.

### Model Performance

Table 2 displays the model performance with varying fingerprint sizes. All models perform well at predicting activities. It can be seen that the performance of the MLPs is slightly better than the performance of the GCNs. It is not uncommon that graph-agnostic models can outperform some GNNs[9,15,37,38] and performance differences depend on the exact choice of model and task at hand. For the MT models, the performance improves noticeably through an increasing fingerprint size while for the generic models the performance decrease slightly. Overall, the MLP generic performs better than the GCN generic and the MLP MT performs better than both the GCN MT and Freeze MT model.

Table 2 Average AUC of the trained Models. AUC for MT models is the average across all targets.

| | AUC (SD) | | |
|---|---|---|---|
| Fingerprint Size | 64 | 254 | 1024 |
| MLP generic | 0.9222 (0.004) | 0.9207 (0.003) | 0.9186 (0.004) |
| GCN generic | 0.9044 (0.004) | 0.8984 (0.007) | 0.8961 (0.008) |
| MLP MT | 0.9128 (0.009) | 0.9264 (0.007) | 0.9275 (0.013) |
| GCN MT | 0.9035 (0.012) | 0.9130 (0.013) | 0.9124 (0.012) |
| Freeze MT | 0.8889 (0.012) | 0.9098 (0.005) | 0.9130 (0.005) |

**Fingerprint Distribution**

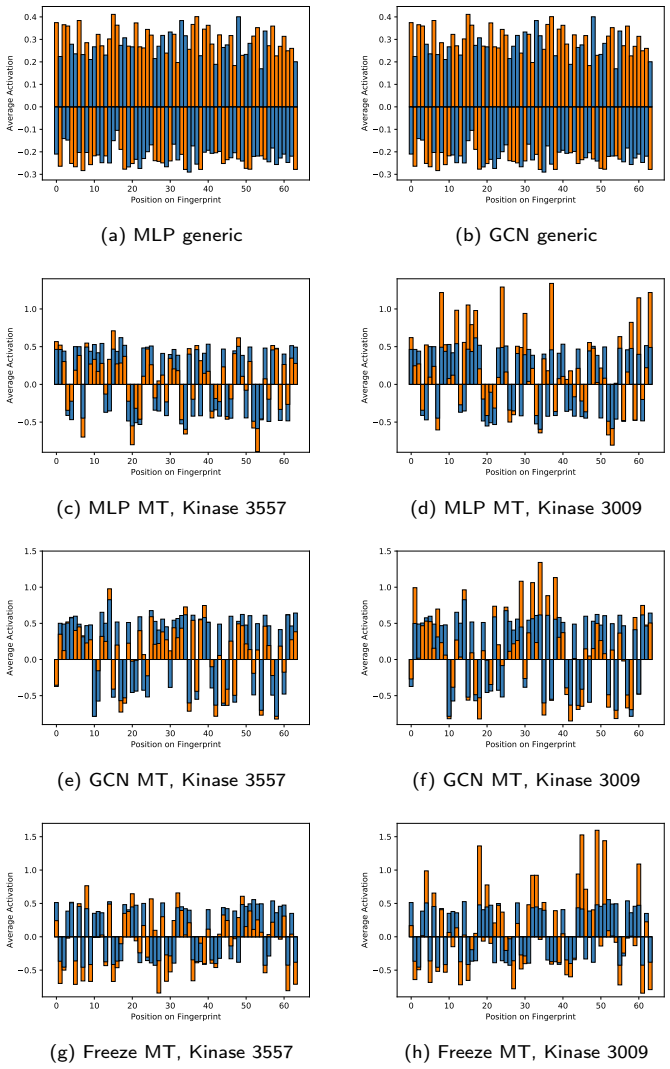
In Figure 5 the average values for each position of the neural fingerprints are shown and average activations of actives are compared to inactives. The generic models show a strong separation between active and inactive molecules. The separation is somewhat greater for the MLP than for the GCN. For the MT models, the fingerprint generated for two specific kinase targets are compared. Here the means of active and inactive molecules are closer together. The strong separation like in the generic models is lost, but smaller differences can still be identified, especially when comparing two different kinases. These results are not surprising as generic models can more easily separate actives from inactives, while multitask models have to be more specific with regards to their prediction. This resulting in more nuanced fingerprints that are capable of discriminating between different kinase targets in similarity-based searches.

**Similarity Search**

Figure 6 shows the results of the similarity search. The AUC describes the performance of the fingerprint across the whole external validation set. The size of the ECFP4 is kept at length 1024 and the size of the CDDD at 512. Overall, the AUC for all fingerprints is relatively similar. Here, the ECFP4 is the worst performing fingerprint. The best performing fingerprint is generated by the Freeze Model, closely followed by the MLP generic. The performance of some models differs depending on the fingerprint size. The MLP MT improves with increasing size while the performance of GCN generic worsens as its size increases.

Further Figure 6 shows the enrichment factor at 1% and 2.5%. The fingerprints based on the MLP MT model outperform all other fingerprints by a great margin. Its enrichment factor is two times larger than that of the ECFP4 and CDDD. The next best fingerprint is based on the Freeze MT model which outperforms the ECFP4 as well. The GCN MT initially performs better than the ECFP4 fingerprint, but the performance decrease with increasing fingerprint size. Another important finding is that the generic models perform worse than ECFP4. This indicates that a generic kinase-like property is not sufficient for a successful similarity search. The CDDD performs worse than the ECFP4 with regards to enrichment but performs better than the generic models.

Lastly, the analysis of the average similarity highlights another useful property of the neural fingerprint. It evaluates the average ECFP4 similarity of the actives to query in the top-ranked (1%). The neural fingerprints all have a lower average similarity than the ECFP4 based similarity search, especially the MLP MT and Freeze MT differ noticeably from the traditional ECFP4. These findings indicate that a different chemical space is learned and identified using neural fingerprints. It also points to the fact that the ECFP4 had less discriminatory power for this particular chemical space, and as a result identified fewer actives as top-ranked hits. Besides the higher enrichment, this is an additional benefit of herein proposed fingerprint. Importantly, the MLP MT and Freeze MT fingerprint




(a) MLP generic (b) GCN generic

(c) MLP MT, Kinase 3557 (d) MLP MT, Kinase 3009

(e) GCN MT, Kinase 3557 (f) GCN MT, Kinase 3009

(g) Freeze MT, Kinase 3557 (h) Freeze MT, Kinase 3009

Figure 5 Mean Activation per position on the vector for a fingerprint size of 64, *orange*: active, *blue*: inactive

clearly outperform the ECFP4 with regards to similarity search. Proving that the initial idea of combining knowledge on learned kinase inhibitors in a neural network fingerprint is a promising and successful approach. Somewhat surprising, graph convolutions seem to be useful only when the weights are frozen during learning. Further, we could show that neural networks trained on a predictive task were able to generate better fingerprints than networks that were only trained to reconstruct molecules (CDDD).

Next to the average performance across targets, we also evaluate how often each fingerprint performs best on each target. In Figure 7 the mean rank of each fingerprint is shown. The results do not differ much to the mean performance. The Freeze MT is ranked the highest for the AUC. The ECFP4 is ranked the lowest. When it comes enrichment the MLP MT is ranked on average the best. Lastly the Freeze MT and MLP MT are also the lowest rank fingerprint based on their average similarity. These results demonstrate that the neural fingerprints not only provide better enrichment and a better AUC but they do so consistently across most targets.

We also tested the performance while using less data during training in the form of a 60%/20%/20% split. This did not lead to a significantly different performance of the fingerprints (details

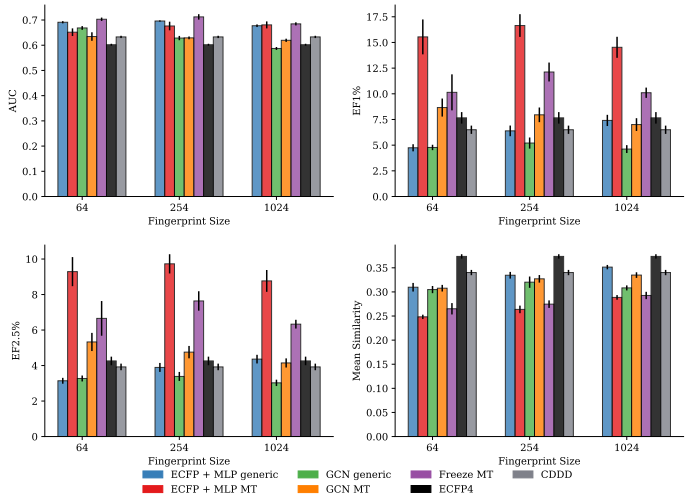



Figure 6 Average performance statistics of neural fingerprints in similarity search in comparison to the ECFP4 fingerprint. AUC: Area under the curve, EF: enrichment factor at the top 1%/2.5% of the dataset, The mean ECFP4 based similarity of the query and the corresponding hit.

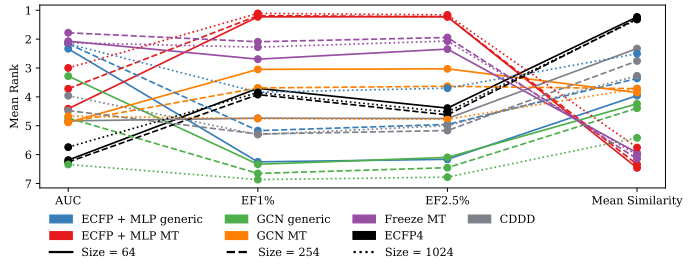



Figure 7 Average ranking of the fingerprints per performance measure

in the Supporting Information). The overall performance is similar, which means that the MLP MT-based fingerprint clearly outperforms the ECFP4 and CDDD. The second-best model, as in the 80%/10%/10% split, was the Freeze MT. This shows that equal performance can be achieved with less training data.

### Out-of-sample Performance

The next analysis deals with the question, if generic kinase properties can be learned and if they can be helpful in neural fingerprint-based screening. First, a single kinase and its inhibitors were removed from the training set. In the second experiment, we removed multiple kinases at the same time but kept the inhibitors in the training set. Table 3 shows the effect of not having trained with inhibitors for a specific target. We compare the performance of the fingerprint for a particular kinase, ones when trained with the active molecules in the training set, and ones when they are removed from the training. The GCN generic is effected the least from removing inhibitors specific to a kinase target from training. The effects are more noticeable for the MLP generic where an average performance loss of 5% to 10% can be seen. For both MT models the removal of inhibitors from training has a strong effect resulting in an average reduction of up to 20%. Lastly, the difference between training with vs without specific inhibitors shows a large variance across the different targets. An indication that the effects of excluding inhibitors from training differ between the kinase targets.

For one specific kinase we observed interesting behavior (see

Table 3 Average relative difference for specific kinase targets between models excluding and including the inhibitors during training. (Standard Deviation)

| | MLP generic | MLP MT | GCN generic | GCN MT |
|---|---|---|---|---|
| AUC | -0.066 (0.067) | -0.125 (0.112) | -0.003 (0.021) | -0.071 (0.054) |
| EF1% | -0.072 (0.135) | -0.2 (0.267) | 0.067 (0.068) | -0.206 (0.241) |
| EF2.5% | -0.098 (0.114) | -0.201 (0.287) | 0.08 (0.076) | -0.16 (0.226) |
| Mean Sim. | -0.039 (0.066) | 0.17 (0.238) | 0.0 (0.05) | 0.117 (0.159) |

Table 4). The MLP generic performance in similarity search increased greatly when removing its inhibitors from the training set. We found that the MLP generic trained without the kinase inhibitors for CHEMBL4899 led to a model that considered all inhibitors for that kinase as inactives. Thus, while the model got worse at making predictions for those inhibitors the fingerprint improved. A possible explanation is that the ECFP4 of the inhibitors for CHEMBL4899 differs from those of other kinase inhibitors. For the model to correctly classify those inhibitors, their ECFP4 is transformed to match the activations of other inhibitors. This allows the model to make better predictions but in return, the neural fingerprints for those inhibitors carry less unique molecular information decreasing the similarity search performance. This effect is not seen in the GCN generic, most likely because it can adapt the encoding of inhibitors during training, and does not rely on the ECFP4.

In Figure 8 the effects of leaving out multiple kinases during the training process are shown. Here, 40% of the kinase targets were left out during training but the inhibitors themselves were kept in the training set. The performance of the generic models is not influenced by the decrease in target information. This was expected, as the generic models are not trained with specific target information. In contrast, removing target information decreases the MT models AUC, especially for the Freeze MT and MLP MT model. The GCN MT performance also decreases but not as strongly. For a fingerprint of size 64, the performance of the MLP MT is worse than the one of the ECFP4. With increasing fingerprint size all models perform again better than the ECFP4 but the effects of training with fewer targets remain noticeable.

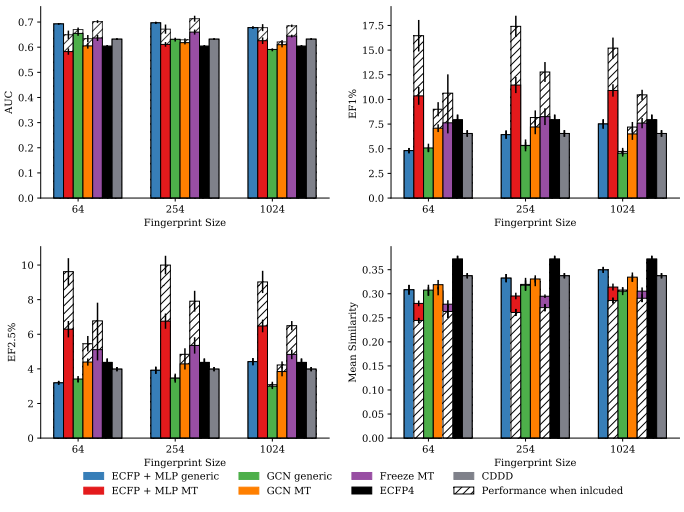



Figure 8 Effects of leaving out target information for specific kinases on the similarity search. Average performance when target information is removed (*colored*) and when target information is included during training (*striped*)

.

Table 4 Difference in performance of the similarity search when training with vs training without the inhibitors active on CHEMBL4899

| | Average Performance when trained without/with | | | | |
|---|---|---|---|---|---|
| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP |
| AUC | 0.873/0.65 | 0.81/0.944 | 0.786/0.845 | 0.767/0.765 | 0.773/0.773 |
| EF1% | 61.377/21.662 | 64.084/64.987 | 29.786/31.591 | 42.422/51.448 | 64.987/64.987 |
| EF2.5% | 26.545/9.455 | 26.182/29.455 | 14.545/17.091 | 22.182/21.455 | 26.182/26.182 |
| Mean Sim | 0.446/0.47 | 0.446/0.446 | 0.46/0.478 | 0.467/0.458 | 0.446/0.446 |

The effects of removing kinases are even stronger for the enrichment. Here the Freeze MT and MLP MT performance is decreased by up to a third. Similar to the AUC the reduction is less noticeable for the GCN MT. While the MLP MT still performs better than the ECFP4, the decrease in performance lets the Freeze MT and GCN MT model perform worse than the ECFP4. These results show that the performance of the fingerprints depends greatly on the targets they have been trained for. But while the performance decrease, the MLP MT neural fingerprints can still outperform the ECFP4 and CDDD even when the model was never trained to predict activities on a particular kinase. This allows the neural fingerprints to be used for virtual screening even when the desired kinase target is not included in the dataset used by us.

Figure 9 further supports this point. While the MLP MT model clearly performs better on targets which were included in the training set, it still provides the best performance on most targets not included during training.

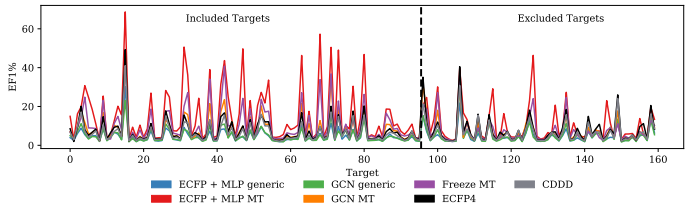



Figure 9 Average Enrichment per target for a fingerprint of length 64. The first 96 targets were included in the training, the last 64 targets were excluded

.

These results demonstrate that training neural networks can create fingerprints that are better performing than traditional fingerprints. The choice of task and architecture, however, is quite important. Fingerprints obtained through multitask training lead to higher enrichment than models trained on a generic task. MLP models also produced better neural fingerprints than the GCN. This is surprising as the model performance (Table 2) is quite similar between those models. An explanation for this discrepancy can be found in the pronounced difference in enrichment between the Freeze MT and GCN MT. The GCN models can adapt the feature generation to the task. In contrast, the Freeze MT is unable to change its feature generation because the weights of the convolution layers were frozen. The additional learnable weights allow the GCN to transform the input more specific to the task causing the model to ”overfit” to the task. Through this process, later activations carry more task-relevant information, and much unique molecular information is lost. This molecule-specific information is retained in the MLP MT and Freeze MT allowing for a more diverse fingerprint which in return leads to better performances. It can be expected that the quality of neural fingerprints also depends on the dataset used. When active and inactive molecules are increasingly similar, the neural networks are required to encode the feature space more nuanced. While when actives and inactives are less similar the networks only have to encode major differences to be successful in target prediction. A more difficult task could force the networks to produce a more diverse and better-performing fingerprint. We believe that the GCN with the additional layers would profit most from a more difficult task.

Most importantly the lack of performance could be due to the inherent limitations of many GNNs including the architecture used in our analysis. It was shown that simple linear graph models that do not aggregate the neighboring atoms perform on par with the GCN. This hints to the potential inability of the GCN to capture more sophisticated substructures.[39] In a different paper introducing the Graph Isomorphism Network (GIN), the authors point to the fact that the GCN and comparable networks have issues distinguishing certain simple graph structures.[40] This lack of so-called expressiveness has gained more attention in the last few years. It was shown that many GNNs have strong limits with regards to differentiating graphs and their substructures.[41,42] Closely related is the issue of *oversmoothing*.[43,44] As the input passes through the network, the features related to each node become increasingly similar to each other. This leads to further difficulties distinguishing nodes and the features they represent. Overall these factors contribute to the fact that many GNNs are not able to encode molecules efficiently and explain the low performance of the GCN fingerprint.

New, more expressive architectures have been proposed such as the aforementioned GIN, the SGN[45], and others.[42,46] While it was shown that more expressive architectures do not automatically lead to better model performance[38] they bear the potential to create fingerprints better than those of the GCN and pose an interesting direction for future research.

## Conclusions

In this work, several important findings are presented and discussed. First of all, we introduced the novel idea of using the activations from neural networks trained for target predictions as fingerprints for similarity search. The proof-of-concept analysis using a dataset of active and non-active kinase inhibitors showed the usefulness of neural fingerprints in ligand-based virtual screening. When choosing the right model the neural fingerprints can outperform traditional ones like the ECFP4. Somewhat surprising, the most successful architecture is based on a multi-layer perceptron with the ECFP as input trained for multitask classification.

Our initial hypothesis was, that Graph Neural Networks would perform better due to their ability to adapt the encoding of molecules. However, the Graph Convolution Network was not able to produce fingerprints that matched the performance of standard multi-layer perceptrons. Most interestingly, this seems to be due to the additional learning capabilities. When the weights of the convolutional layers are frozen, the performance of the GCN fingerprints massively increases. This indicates that the non-frozen model "overtrains" to the task. Beyond that, the lack of expressive power inherent to many Graph Neural Networks is made out to be a reason for the weak performance. Our results, in combina-

tion with findings from recent literature, question the usefulness of many current GNNs for the here proposed approach.

In contrast, using the MLP to generate fingerprints provides consistently higher enrichment than the ECFP4 and other neural fingerprints. Additionally, the hits in the top 1% found by the fingerprint were less similar than those of the ECFP4. Thus, a different chemical space was learned and identified. Even when this model was not trained to predict a subset of kinases, the fingerprint performed better on the excluded targets compared to the ECFP4. Therefore, this fingerprint is suitable for ATP-competitive inhibitors not included in our training set. While kinase inhibitors are a rather homogeneous group of inhibitors we believe that, given enough data, our approach can be extended to other ligand domains.

The pretrained model together with a script to generate the most successful kinase-specific neural fingerprint will be provided to the community to enhance the search for new kinase inhibitors.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, priority program „Algorithms for Big Data", SPP 1736, Grant No. KO 4689/2-2) and funding of the research training group "GRK 2515: Chemical biology of ion channels (Chembion)" by the DFG is gratefully acknowledged.

## References

[1] Weininger, D. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

[2] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics* **2015**, *7*, 23.

[3] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for use in drug Ddscovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.

[4] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[5] Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley, 1990.

[6] Willett, P. Similarity-based Virtual Screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

[7] Koutsoukas, A.; Paricharak, S.; Galloway, W. R.; Spring, D. R.; IJzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *Journal of Chemical Information and Modeling* **2014**, *54*, 230–242.

[8] Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science* **2018**, *9*, 5441–5451.

[9] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

[10] Yang, M.; Tao, B.; Chen, C.; Jia, W.; Sun, S.; Zhang, T.; Wang, X. Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 nhibitors. *Journal of Chemical Information and Modeling* **2019**, *59*, 5002–5012.

[11] Rodriguez-Perez, R.; Bajorath, J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. *Acs Omega* **2019**, *4*, 4367–4375.

[12] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-aided Molecular Design* **2016**, *30*, 595–608.

[13] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. $29^{th}$ Annual Conference on Neural Information Processing Systems, NeurIPS. 2015; pp 2224–2232.

[14] Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of Chemical Information and Modeling* **2017**, *57*, 1757–1772.

[15] Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

[16] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. $34^{th}$ International Conference on Machine Learning, ICML. 2017; pp 1263–1272.

[17] Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *Journal of Cheminformatics* **2020**, *12*, 1.

[18] Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* **2018**, *19*, 526.

[19] Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science* **2019**, *10*, 1692–1701.

[20] Xu, Z.; Wang, S.; Zhu, F.; Huang, J. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. $8^th$ ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2017; pp 285–294.

[21] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **2018**, *4*, 268–276.

[22] Bjerrum, E. J.; Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **2018**, *8*, 131.

[23] Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: quo vadis? *Journal of Chemical Information and Modeling* **2012**, *52*, 1413–1437.

[24] Pogodin, P. V.; Lagunin, A. A.; Rudik, A. V.; Filimonov, D. A.; Druzhilovskiy, D. S.; Nicklaus, M. C.; Poroikov, V. V. How to achieve better results using PASS-based virtual screening: Case study for kinase inhibitors. *Frontiers in Chemistry* **2018**, *6*, 133.

[25] Xing, L.; Klug-Mcleod, J.; Rai, B.; Lunney, E. A. Kinase hinge binding scaffolds and their hydrogen bond patterns. *Bioorganic & Medicinal Chemistry* **2015**, *23*, 6520–6527.

[26] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, 1–21.

[27] Cangea, C.; Veličković, P.; Jovanović, N.; Kipf, T.; Liò, P. Towards sparse hierarchical graph classifiers. *Workshop on Relational Representation Learning (R2L)* at the $32^{nd}$ Annual Conference on Neural Information Processing Systems, NeurIPS. 2018, arXiv:1811.01287v1.

[28] Landrum, G., et al. RDKit: Open-source cheminformatics. 2006.

[29] O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics* **2016**, *8*, 1–14.

[30] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B., et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.

[31] Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. $33^{nd}$ Annual Conference on Neural Information Processing Systems, NeurIPS. 2019; pp 8024–8035.

[32] Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. $3^{rd}$ International Conference on Learning Representations ICLR. 2015.

[33] Bakshy, E.; Dworkin, L.; Karrer, B.; Kashin, K.; Letham, B.; Murthy, A.; Singh, S. AE: A domain-agnostic platform for adaptive experimentation. *Workshop on System for ML* at the $32^{nd}$ Annual Conference on Neural Information Processing Systems, NeurIPS. 2018.

[34] Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7*, 20.

[35] Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* **2013**, *5*, 26.

[36] Sonego, P.; Kocsor, A.; Pongor, S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Briefings in Bioinformatics* **2008**, *9*, 198–209.

[37] Errica, F.; Podda, M.; Bacciu, D.; Micheli, A. A Fair Comparison of Graph Neural Networks for Graph Classification. $8^th$ International Conference on Learning Representations ICLR. 2020.

[38] Dwivedi, V. P.; Joshi, C. K.; Laurent, T.; Bengio, Y.; Bresson, X. Benchmarking Graph Neural Networks. *arXiv preprint arXiv:2003.00982v1* **2020**,

[39] Chen, T.; Bian, S.; Sun, Y. [preprint] Are powerful graph neural nets necessary? a dissection on graph classification. *arXiv preprint arXiv:1905.04579v3* **2019**,

[40] Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? $7^th$ International Conference on Learning Representations, ICLR. 2019.

[41] Chen, Z.; Chen, L.; Villar, S.; Bruna, J. [preprint] Can graph neural networks count substructures? *arXiv preprint arXiv:2002.04025v3* **2020**,

[42] Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. $33^{rd}$ Conference on Artificial Intelligence, AAAI. 2019; pp 4602–4609.

[43] Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; Sun, X. Measuring and Relieving the Over-Smoothing Problem for Graph Neural Networks from the Topological View. $34th$ Conference on Artificial Intelligence, AAAI. 2020; pp 3438–3445.

[44] Zhao, L.; Akoglu, L. PairNorm: Tackling Oversmoothing in GNNs. $8^th$ International Conference on Learning Representations, ICLR. 2020, arXiv:1909.12223v2.

[45] Bouritsas, G.; Frasca, F.; Zafeiriou, S.; Bronstein, M. M. Improving Graph Neural Network Expressivity via Subgraph Isomorphism Counting. *Graph Representation Learning and Beyond (GRL+) Workshop* at the $37^{th}$ International Conference on Machine Learning, ICML. 2020, arXiv:2006.09252.

[46] Maron, H.; Ben-Hamu, H.; Serviansky, H.; Lipman, Y. Provably Powerful Graph Networks. $33^{nd}$ Annual Conference on Neural Information Processing Systems, NeurIPS. 2019; pp 2153–2164.

# Supporting Information

## Results 80%-10%-10% Split

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.3325 (1.4917) | 4.4112 (2.2493) | 3.2812 (1.3001) | 4.8825 (1.2012) | 2.07 (1.1051) | 6.19 (0.9525) | 4.8325 (1.2131) |
| EF1% | 6.2513 (0.7871) | 1.2288 (0.6583) | 6.335 (0.9034) | 3.0481 (0.8396) | 2.6944 (1.1837) | 3.705 (0.9579) | 4.7375 (0.9264) |
| EF2.5% | 6.1638 (0.9686) | 1.2244 (0.6024) | 6.0962 (1.0765) | 3.0288 (0.7837) | 2.3463 (0.9832) | 4.3894 (1.0444) | 4.7513 (1.0132) |
| Mean Sim. | 3.9562 (1.4002) | 6.4588 (0.8467) | 4.24 (1.2817) | 3.8525 (1.0199) | 5.9362 (0.8852) | 1.23 (0.6003) | 2.3262 (0.8542) |

Table 1: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6911 (0.0462) | 0.6517 (0.106) | 0.6685 (0.0465) | 0.6345 (0.0673) | 0.703 (0.0708) | 0.6018 (0.0501) | 0.6329 (0.0497) |
| EF1% | 4.7443 (3.9901) | 15.5406 (13.7324) | 4.7762 (4.0505) | 8.661 (7.9419) | 10.1418 (9.2473) | 7.6572 (7.1812) | 6.5001 (5.7875) |
| EF2.5% | 3.1413 (1.818) | 9.29 (6.6907) | 3.2707 (2.0137) | 5.3285 (3.827) | 6.6565 (4.7716) | 4.2594 (3.0614) | 3.922 (2.6154) |
| Mean Sim. | 0.3097 (0.0693) | 0.2483 (0.0452) | 0.3049 (0.0679) | 0.3077 (0.0539) | 0.265 (0.0539) | 0.3737 (0.0458) | 0.3402 (0.0558) |

Table 2: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.1612 (1.2744) | 3.72 (2.1011) | 4.7312 (1.291) | 4.875 (1.0629) | 1.78 (0.8032) | 6.2575 (0.9487) | 4.475 (1.2772) |
| EF1% | 5.17 (1.1068) | 1.1875 (0.5947) | 6.6525 (0.7792) | 3.6906 (1.029) | 2.0912 (0.7727) | 3.9219 (0.9747) | 5.2862 (1.0911) |
| EF2.5% | 4.955 (1.2694) | 1.2238 (0.5982) | 6.4537 (0.9632) | 3.6331 (0.9841) | 1.9431 (0.5923) | 4.6237 (1.1625) | 5.1675 (1.1578) |
| Mean Sim | 3.3613 (1.2412) | 6.3362 (0.9222) | 4.3925 (1.4182) | 3.7156 (1.1256) | 6.145 (0.7672) | 1.2919 (0.7283) | 2.7575 (1.0686) |

Table 3: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.696 (0.0462) | 0.6762 (0.1056) | 0.6284 (0.0488) | 0.6291 (0.0582) | 0.7123 (0.0776) | 0.6018 (0.0501) | 0.6329 (0.0497) |
| EF1% | 6.3907 (5.5934) | 16.646 (14.7808) | 5.205 (4.6032) | 7.9529 (7.3027) | 12.1225 (10.646) | 7.6572 (7.1812) | 6.5001 (5.7875) |
| EF2.5% | 3.8938 (2.4448) | 9.7286 (7.1847) | 3.3866 (2.2746) | 4.7556 (3.4193) | 7.6383 (5.3841) | 4.2594 (3.0614) | 3.922 (2.6154) |
| Mean Sim. | 0.3346 (0.0624) | 0.2637 (0.0407) | 0.3202 (0.0638) | 0.3271 (0.0531) | 0.2746 (0.0474) | 0.3737 (0.0458) | 0.3402 (0.0558) |

Table 4: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.1638 (1.1922) | 2.9988 (1.9531) | 6.3413 (0.9423) | 4.67 (1.0932) | 2.1237 (0.8293) | 5.7413 (0.9396) | 3.9613 (1.2848) |
| EF1% | 3.8394 (1.0715) | 1.1081 (0.4157) | 6.8662 (0.5125) | 4.7525 (1.1395) | 2.2781 (0.8988) | 3.8519 (1.0828) | 5.3037 (1.0286) |
| EF2.5% | 3.6988 (1.0779) | 1.16 (0.5003) | 6.78 (0.6998) | 4.7756 (1.1465) | 2.0662 (0.6964) | 4.5206 (1.151) | 4.9987 (1.1428) |
| Mean Sim | 2.505 (0.9213) | 5.7544 (1.3608) | 5.42 (1.5015) | 3.7175 (1.0846) | 6.0125 (0.8031) | 1.3194 (0.7689) | 3.2712 (1.129) |

Table 5: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6773 (0.0496) | 0.6806 (0.0981) | 0.5871 (0.0437) | 0.6195 (0.0497) | 0.6846 (0.0684) | 0.6018 (0.0501) | 0.6329 (0.0497) |
| EF1% | 7.4082 (6.5602) | 14.5275 (12.5815) | 4.6182 (4.2315) | 7.0045 (6.3773) | 10.1031 (8.8352) | 7.6572 (7.1812) | 6.5001 (5.7875) |
| EF2.5% | 4.3613 (2.8499) | 8.7676 (6.393) | 3.0244 (2.0595) | 4.146 (2.9518) | 6.3307 (4.4012) | 4.2594 (3.0614) | 3.922 (2.6154) |
| Mean Sim. | 0.3514 (0.0529) | 0.2886 (0.0401) | 0.3083 (0.0644) | 0.3349 (0.056) | 0.2926 (0.0482) | 0.3737 (0.0458) | 0.3402 (0.0558) |

Table 6: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

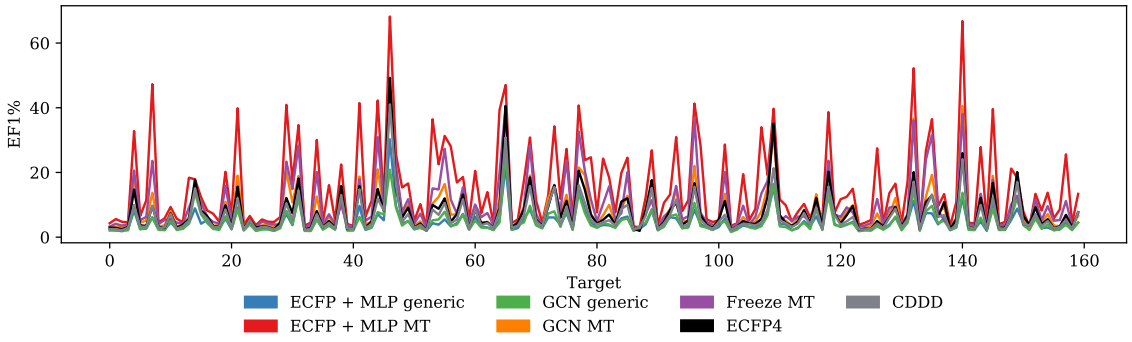



Figure 1: Mean EF1% per Target for a fingerprint of size 64. The Performance of each fingerprint is shown for each target. The length of the ECFP4 and CDDD are 1024 and 512 respectively

## Results 60%-20%-20% Split

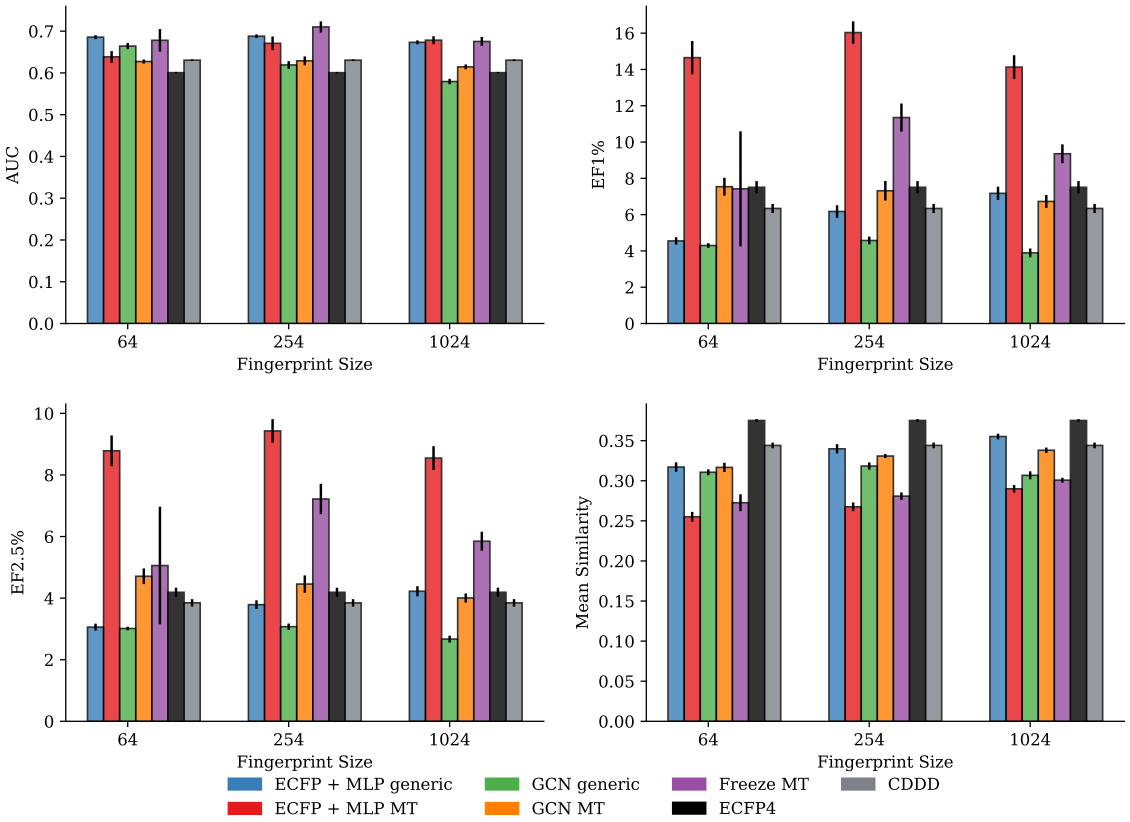



Figure 2: Average Mean Similarity Search Performance of the Models trained on only 60% of the data

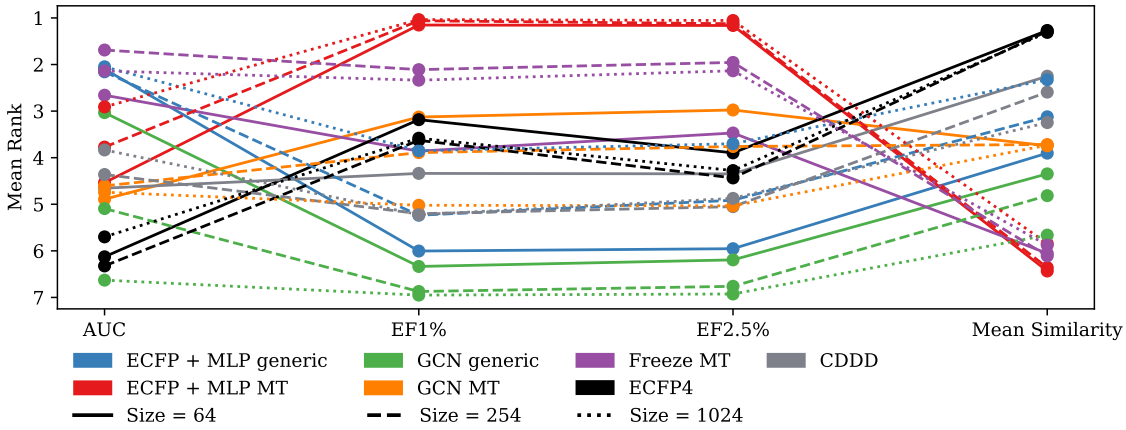



Figure 3: Average Ranking of each fingerprint on each performance measure

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.0988 (1.3737) | 4.5362 (2.4427) | 3.0325 (1.269) | 4.8913 (1.1709) | 2.6588 (1.2756) | 6.1275 (0.9044) | 4.655 (1.1224) |
| EF1% | 6.0038 (0.795) | 1.1525 (0.5496) | 6.335 (0.7824) | 3.1275 (0.8898) | 3.86 (1.2156) | 3.1838 (0.9222) | 4.3375 (0.8461) |
| EF2.5% | 5.9538 (0.9352) | 1.1625 (0.5124) | 6.1938 (0.9207) | 2.9762 (0.8462) | 3.47 (1.2361) | 3.8963 (1.1368) | 4.3475 (0.9499) |
| Mean Sim | 3.9025 (1.3762) | 6.4325 (0.8741) | 4.3475 (1.2012) | 3.7438 (1.0154) | 6.0538 (0.7862) | 1.2688 (0.749) | 2.2513 (0.7299) |

Table 7: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6855 (0.0402) | 0.6384 (0.1044) | 0.6642 (0.0405) | 0.6272 (0.056) | 0.678 (0.0581) | 0.6006 (0.0469) | 0.6304 (0.0446) |
| EF1% | 4.5499 (3.4524) | 14.6507 (12.8999) | 4.2975 (3.157) | 7.5395 (6.7117) | 7.4207 (6.4479) | 7.5073 (7.0001) | 6.3391 (5.5114) |
| EF2.5% | 3.0584 (1.6135) | 8.7841 (6.2785) | 3.0141 (1.6379) | 4.71 (3.2937) | 5.0577 (3.377) | 4.1895 (2.9862) | 3.8459 (2.5159) |
| Mean Sim. | 0.3171 (0.0724) | 0.2551 (0.0446) | 0.3106 (0.0682) | 0.3166 (0.0557) | 0.2726 (0.0557) | 0.3752 (0.0406) | 0.3441 (0.0538) |

Table 8: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.1562 (1.1551) | 3.7725 (2.1922) | 5.0925 (1.1804) | 4.605 (1.0127) | 1.6888 (0.7619) | 6.3237 (0.8721) | 4.3612 (1.2125) |
| EF1% | 5.2356 (0.926) | 1.0606 (0.2894) | 6.8725 (0.5103) | 3.8912 (0.9855) | 2.11 (0.6176) | 3.6275 (0.8875) | 5.2025 (0.9239) |
| EF2.5% | 4.9138 (1.1465) | 1.1244 (0.3676) | 6.7613 (0.6699) | 3.76 (1.0239) | 1.955 (0.4701) | 4.4344 (1.1102) | 5.0513 (1.0285) |
| Mean Sim | 3.1225 (1.0726) | 6.3594 (0.9435) | 4.815 (1.4503) | 3.7175 (1.0246) | 6.1088 (0.6422) | 1.2844 (0.7407) | 2.5925 (0.8892) |

Table 9: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6878 (0.0435) | 0.6708 (0.1013) | 0.619 (0.0397) | 0.6288 (0.0499) | 0.71 (0.0726) | 0.6006 (0.0469) | 0.6304 (0.0446) |
| EF1% | 6.1729 (5.4781) | 16.034 (14.0027) | 4.5741 (3.8649) | 7.3141 (6.382) | 11.3483 (9.9383) | 7.5073 (7.0001) | 6.3391 (5.5114) |
| EF2.5% | 3.787 (2.3921) | 9.4294 (6.7965) | 3.0726 (1.8998) | 4.4558 (3.0791) | 7.2183 (5.0166) | 4.1895 (2.9862) | 3.8459 (2.5159) |
| Mean Sim. | 0.3398 (0.0603) | 0.2675 (0.0385) | 0.3183 (0.0644) | 0.3308 (0.0535) | 0.2808 (0.0448) | 0.3752 (0.0406) | 0.3441 (0.0538) |

Table 10: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.0488 (1.0188) | 2.91 (1.8793) | 6.6287 (0.725) | 4.7412 (1.0023) | 2.1375 (0.8275) | 5.7 (0.8592) | 3.8338 (1.1589) |
| EF1% | 3.8512 (1.0036) | 1.0356 (0.2144) | 6.95 (0.3242) | 5.0212 (1.0482) | 2.3362 (0.9103) | 3.5819 (0.9709) | 5.2238 (0.9398) |
| EF2.5% | 3.7006 (0.9896) | 1.055 (0.2445) | 6.9262 (0.3883) | 5.0338 (1.1001) | 2.1344 (0.6555) | 4.2762 (1.0954) | 4.8737 (1.062) |
| Mean Sim | 2.3263 (0.7986) | 5.8444 (1.2799) | 5.6625 (1.5031) | 3.7325 (0.9439) | 5.8775 (0.6779) | 1.3106 (0.8092) | 3.2462 (0.982) |

Table 11: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6732 (0.0459) | 0.6783 (0.094) | 0.5793 (0.0342) | 0.6144 (0.0449) | 0.6752 (0.0642) | 0.6006 (0.0469) | 0.6304 (0.0446) |
| EF1% | 7.1723 (6.4152) | 14.1281 (12.1811) | 3.8915 (3.1077) | 6.7252 (6.1995) | 9.3536 (8.2474) | 7.5073 (7.0001) | 6.3391 (5.5114) |
| EF2.5% | 4.2225 (2.7573) | 8.5476 (6.1526) | 2.6693 (1.543) | 4.0046 (2.8733) | 5.8472 (4.0239) | 4.1895 (2.9862) | 3.8459 (2.5159) |
| Mean Sim. | 0.3552 (0.0491) | 0.2899 (0.0369) | 0.3067 (0.0653) | 0.338 (0.0537) | 0.3007 (0.0472) | 0.3752 (0.0406) | 0.3441 (0.0538) |

Table 12: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

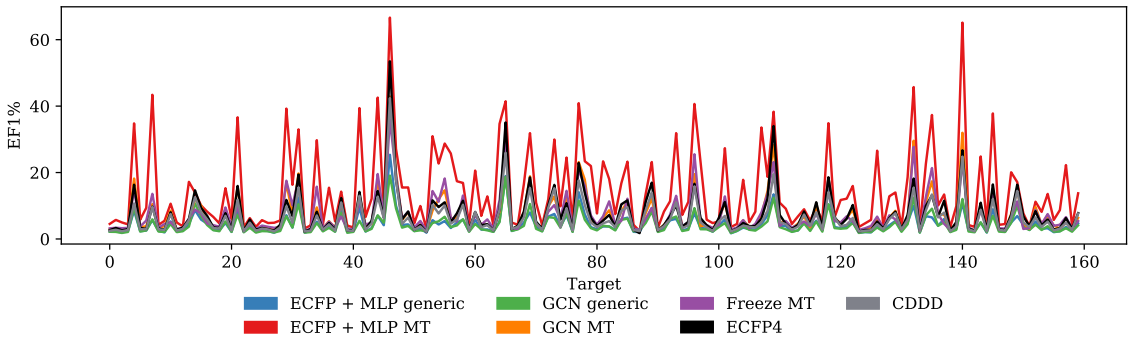



Figure 4: Mean EF1% per Target for a fingerprint of size 64. The Performance of each fingerprint is shown for each target. The length of the ECFP4 and CDDD are 1024 and 512 respectively.

# Out-of-Sample Analysis

## Single Kinase Removed

| Inhibitors Excluded/Inhibitors Included | | | | | |
|---|---|---|---|---|---|
| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
| AUC | 0.609/0.641 | 0.54/0.568 | 0.58/0.584 | 0.544/0.565 | 0.595 |
| EF1% | 1.84/1.89 | 6.353/6.303 | 3.866/3.33 | 6.258/6.204 | 5.782 |
| EF2.5% | 1.899/1.818 | 4.021/3.442 | 2.699/2.324 | 3.462/3.411 | 3.62 |
| Mean Sim | 0.227/0.257 | 0.3/0.292 | 0.279/0.265 | 0.275/0.283 | 0.36 |

Table 13: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL2801

| Inhibitors Excluded/Inhibitors Included | | | | | |
|---|---|---|---|---|---|
| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
| AUC | 0.614/0.606 | 0.525/0.533 | 0.566/0.566 | 0.533/0.569 | 0.554 |
| EF1% | 1.614/1.467 | 2.645/2.508 | 1.693/1.585 | 2.358/2.606 | 2.313 |
| EF2.5% | 1.447/1.404 | 2.115/1.987 | 1.536/1.468 | 1.972/2.05 | 1.76 |
| Mean Sim | 0.185/0.216 | 0.262/0.266 | 0.246/0.234 | 0.243/0.242 | 0.341 |

Table 14: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL3829

| Inhibitors Excluded/Inhibitors Included | | | | | |
|---|---|---|---|---|---|
| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
| AUC | 0.724/0.731 | 0.561/0.584 | 0.602/0.606 | 0.565/0.587 | 0.599 |
| EF1% | 2.346/2.523 | 4.335/4.17 | 2.126/2.178 | 3.811/3.963 | 3.494 |
| EF2.5% | 2.188/2.253 | 3.297/3.374 | 1.946/1.84 | 3.159/3.15 | 2.561 |
| Mean Sim | 0.225/0.238 | 0.27/0.268 | 0.232/0.253 | 0.273/0.271 | 0.332 |

Table 15: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL3357

| Inhibitors Excluded/Inhibitors Included | | | | | |
|---|---|---|---|---|---|
| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
| AUC | 0.634/0.64 | 0.522/0.541 | 0.59/0.593 | 0.545/0.578 | 0.56 |
| EF1% | 1.716/1.536 | 3.353/3.483 | 2.137/2.132 | 2.725/2.825 | 2.622 |
| EF2.5% | 1.467/1.484 | 2.592/2.601 | 1.783/1.783 | 2.32/2.374 | 1.857 |
| Mean Sim | 0.241/0.265 | 0.249/0.264 | 0.262/0.259 | 0.255/0.249 | 0.343 |

Table 16: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL4482

Inhibitors Excluded/Inhibitors Included

| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
|---|---|---|---|---|---|
| AUC | 0.655/0.671 | 0.598/0.605 | 0.592/0.593 | 0.555/0.582 | 0.575 |
| EF1% | 2.201/2.326 | 5.3/5.666 | 2.633/2.522 | 3.794/3.855 | 3.575 |
| EF2.5% | 1.952/2.044 | 3.725/4.119 | 2.108/2.003 | 2.862/2.976 | 2.495 |
| Mean Sim | 0.284/0.28 | 0.286/0.277 | 0.275/0.27 | 0.304/0.293 | 0.356 |

Table 17: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL4225

Inhibitors Excluded/Inhibitors Included

| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
|---|---|---|---|---|---|
| AUC | 0.612/0.719 | 0.549/0.746 | 0.662/0.657 | 0.556/0.681 | 0.607 |
| EF1% | 6.83/7.494 | 19.381/37.622 | 10.935/9.988 | 14.936/23.525 | 19.503 |
| EF2.5% | 3.522/4.607 | 8.742/18.668 | 5.926/5.175 | 7.812/10.868 | 8.369 |
| Mean Sim | 0.433/0.463 | 0.426/0.294 | 0.45/0.436 | 0.446/0.368 | 0.452 |

Table 18: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL3910

Inhibitors Excluded/Inhibitors Included

| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
|---|---|---|---|---|---|
| AUC | 0.626/0.764 | 0.767/0.925 | 0.702/0.722 | 0.697/0.817 | 0.78 |
| EF1% | 9.65/12.592 | 20.437/60.175 | 13.495/13.027 | 13.201/43.651 | 28.366 |
| EF2.5% | 5.073/6.813 | 11.22/29.534 | 7.48/7.934 | 6.85/23.15 | 15.165 |
| Mean Sim | 0.485/0.471 | 0.425/0.291 | 0.465/0.448 | 0.472/0.331 | 0.402 |

Table 19: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL4439

Inhibitors Excluded/Inhibitors Included

| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
|---|---|---|---|---|---|
| AUC | 0.632/0.662 | 0.594/0.773 | 0.619/0.593 | 0.661/0.661 | 0.596 |
| EF1% | 10.393/15.623 | 25.88/28.766 | 14.95/14.55 | 23.586/26.333 | 28.156 |
| EF2.5% | 5.739/7.469 | 11.354/13.455 | 7.955/7.011 | 11.101/11.61 | 12.009 |
| Mean Sim | 0.466/0.441 | 0.413/0.403 | 0.444/0.478 | 0.43/0.419 | 0.413 |

Table 20: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL3142

Inhibitors Excluded/Inhibitors Included

| | MLP generic | MLP MT | GCN generic | GCN MT | ECFP4 |
|---|---|---|---|---|---|
| AUC | 0.641/0.747 | 0.613/0.886 | 0.727/0.752 | 0.69/0.736 | 0.619 |
| EF1% | 10.873/11.515 | 21.048/47.671 | 20.939/17.509 | 20.3/41.446 | 21.298 |
| EF2.5% | 5.557/6.527 | 9.296/23.819 | 11.528/9.739 | 12.274/18.26 | 9.201 |
| Mean Sim | 0.439/0.432 | 0.421/0.265 | 0.351/0.367 | 0.401/0.298 | 0.415 |

Table 21: Difference in performance of the similarity search when training with vs. training without the inhibitors active on CHEMBL1907601

# Results - Removing multiple Kinases from Training Set

The CHEMBL-IDs for the Kinase which were either removed or included in the training set.

| Kinases included in the training set | |
|---|---|
| CHEMBL2695 | CHEMBL4708 |
| CHEMBL3797 | CHEMBL5314 |
| CHEMBL1907605 | CHEMBL3905 |
| CHEMBL3863 | CHEMBL5147 |
| CHEMBL4630 | CHEMBL5331 |
| CHEMBL2147 | CHEMBL299 |
| CHEMBL1844 | CHEMBL2094128 |
| CHEMBL2534 | CHEMBL2250 |
| CHEMBL3829 | CHEMBL2595 |
| CHEMBL1824 | CHEMBL3116 |
| CHEMBL2068 | CHEMBL4036 |
| CHEMBL3529 | CHEMBL3650 |
| CHEMBL3009 | CHEMBL1913 |
| CHEMBL2938 | CHEMBL2185 |
| CHEMBL262 | CHEMBL5543 |
| CHEMBL4899 | CHEMBL3024 |
| CHEMBL4045 | CHEMBL5518 |
| CHEMBL2094127 | CHEMBL3961 |
| CHEMBL2468 | CHEMBL4898 |
| CHEMBL279 | CHEMBL2345 |
| CHEMBL4204 | CHEMBL1991 |
| CHEMBL1868 | CHEMBL3788 |
| CHEMBL3616 | CHEMBL1936 |
| CHEMBL4223 | CHEMBL1907601 |
| CHEMBL2426 | CHEMBL4482 |
| CHEMBL3982 | CHEMBL2959 |
| CHEMBL2111389 | CHEMBL3778 |
| CHEMBL3582 | CHEMBL4237 |
| CHEMBL4501 | CHEMBL1974 |
| CHEMBL4203 | CHEMBL3983 |
| CHEMBL2850 | CHEMBL1841 |
| CHEMBL5145 | CHEMBL4128 |
| CHEMBL1957 | CHEMBL1907600 |
| CHEMBL2815 | CHEMBL2041 |
| CHEMBL1075104 | CHEMBL1955 |
| CHEMBL2971 | CHEMBL1075167 |
| CHEMBL3831 | CHEMBL308 |
| CHEMBL2828 | CHEMBL5818 |
| CHEMBL2431 | CHEMBL260 |
| CHEMBL5330 | CHEMBL4897 |
| CHEMBL2793 | CHEMBL4895 |
| CHEMBL2835 | CHEMBL3234 |
| CHEMBL4282 | CHEMBL2007 |
| CHEMBL2276 | CHEMBL5650 |
| CHEMBL3629 | CHEMBL2292 |
| CHEMBL4273 | CHEMBL2527 |
| CHEMBL258 | CHEMBL3835 |
| CHEMBL2208 | CHEMBL2543 |

| Kinases excluded from the training set | |
|---|---|
| CHEMBL3142 | CHEMBL5568 |
| CHEMBL2094126 | CHEMBL5600 |
| CHEMBL5719 | CHEMBL301 |
| CHEMBL3553 | CHEMBL4247 |
| CHEMBL4816 | CHEMBL4179 |
| CHEMBL4722 | CHEMBL4101 |
| CHEMBL6166 | CHEMBL267 |
| CHEMBL2742 | CHEMBL3045 |
| CHEMBL4040 | CHEMBL5627 |
| CHEMBL5608 | CHEMBL3055 |
| CHEMBL5491 | CHEMBL2973 |
| CHEMBL2637 | CHEMBL3836 |
| CHEMBL203 | CHEMBL4525 |
| CHEMBL3231 | CHEMBL3587 |
| CHEMBL5261 | CHEMBL1862 |
| CHEMBL1906 | CHEMBL5251 |
| CHEMBL4576 | CHEMBL5408 |
| CHEMBL3357 | CHEMBL4900 |
| CHEMBL4601 | CHEMBL5407 |
| CHEMBL4523 | CHEMBL4224 |
| CHEMBL4454 | CHEMBL2689 |
| CHEMBL331 | CHEMBL4439 |
| CHEMBL2148 | CHEMBL4575 |
| CHEMBL2801 | CHEMBL4225 |
| CHEMBL2073 | CHEMBL3935 |
| CHEMBL4202 | CHEMBL5469 |
| CHEMBL2599 | CHEMBL5794 |
| CHEMBL3032 | CHEMBL2749 |
| CHEMBL4599 | CHEMBL3717 |
| CHEMBL2996 | CHEMBL4852 |
| CHEMBL3476 | CHEMBL3920 |
| CHEMBL4578 | CHEMBL1981 |

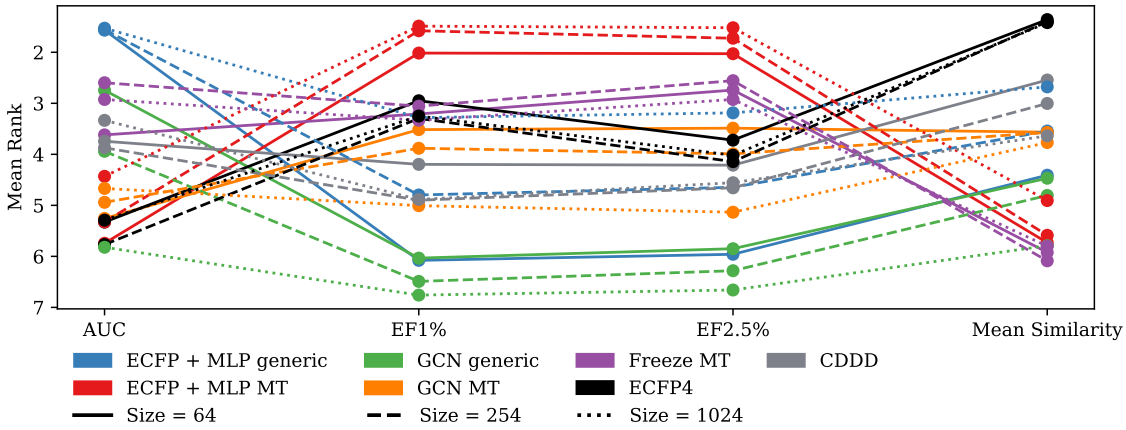



Figure 5: Average ranking of each fingerprint on excluded Targets, when models were trained without the target information on the excluded targets

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.3062 (1.4452) | 4.4188 (2.2439) | 3.2406 (1.2969) | 4.9125 (1.1849) | 2.025 (0.9946) | 6.2281 (0.9128) | 4.8688 (1.1273) |
| EF1% | 6.2531 (0.8079) | 1.1688 (0.51) | 6.2844 (0.8888) | 3.0594 (0.8485) | 2.6875 (1.1336) | 3.7438 (0.9265) | 4.8031 (0.9213) |
| EF2.5% | 6.1656 (0.9366) | 1.175 (0.5103) | 6.0625 (1.064) | 3.0156 (0.7145) | 2.3688 (0.9444) | 4.4531 (1.0493) | 4.7594 (1.0282) |
| Mean Sim. | 3.9281 (1.3748) | 6.5406 (0.7212) | 4.1 (1.262) | 3.85 (1.001) | 5.9812 (0.8177) | 1.2406 (0.6441) | 2.3594 (0.8815) |

Table 22: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 1.5656 (1.1727) | 5.7438 (1.9681) | 2.7438 (1.1668) | 5.2562 (1.3009) | 3.6188 (1.3553) | 5.3281 (1.1448) | 3.7438 (1.3254) |
| EF1% | 6.0781 (1.0196) | 2.0156 (1.579) | 6.0344 (1.2358) | 3.5156 (1.16) | 3.2094 (1.6359) | 2.95 (1.4042) | 4.1969 (1.3837) |
| EF2.5% | 5.9594 (1.1491) | 2.0281 (1.6668) | 5.85 (1.3534) | 3.4875 (1.2608) | 2.7406 (1.5003) | 3.7219 (1.6196) | 4.2125 (1.3931) |
| Mean Sim. | 4.4094 (1.51) | 5.7312 (1.5144) | 4.4688 (1.3736) | 3.5656 (1.1631) | 5.9219 (1.0261) | 1.3594 (0.9672) | 2.5438 (1.13) |

Table 23: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6926 (0.0434) | 0.6494 (0.1054) | 0.6702 (0.048) | 0.6339 (0.066) | 0.7017 (0.068) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 4.8079 (3.9544) | 16.4581 (14.2163) | 4.9364 (4.4316) | 9.0027 (8.4472) | 10.6325 (10.1414) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 3.1925 (1.8638) | 9.6177 (6.7945) | 3.348 (2.2345) | 5.4527 (4.0701) | 6.7742 (4.9788) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3083 (0.0712) | 0.2445 (0.0441) | 0.3046 (0.0681) | 0.3057 (0.0531) | 0.2632 (0.0512) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 24: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 64 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6926 (0.0434) | 0.5825 (0.0815) | 0.6548 (0.0505) | 0.6051 (0.0587) | 0.6366 (0.0603) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 4.8079 (3.9544) | 10.3592 (9.4776) | 5.0793 (4.741) | 7.0841 (6.4387) | 7.6259 (6.9301) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 3.1925 (1.8638) | 6.2937 (4.8609) | 3.4132 (2.3914) | 4.3942 (3.0889) | 5.1112 (3.7599) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3083 (0.0712) | 0.2799 (0.0693) | 0.3076 (0.0666) | 0.3189 (0.0622) | 0.2783 (0.0609) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 25: Average Mean(SD) of the Similarity Search Performance Measures with a Fingerprint size of 64 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.1719 (1.272) | 3.8344 (2.1354) | 4.6688 (1.3058) | 4.85 (1.0175) | 1.675 (0.6516) | 6.2469 (0.9448) | 4.5531 (1.2078) |
| EF1% | 5.1625 (1.116) | 1.175 (0.5068) | 6.6406 (0.8168) | 3.7375 (0.9426) | 2.0188 (0.6771) | 3.925 (1.0303) | 5.3406 (1.0373) |
| EF2.5% | 4.9594 (1.2853) | 1.225 (0.5847) | 6.425 (0.9704) | 3.6938 (0.9691) | 1.8938 (0.4834) | 4.6531 (1.1985) | 5.15 (1.144) |
| Mean Sim. | 3.3188 (1.2253) | 6.4 (0.8517) | 4.2938 (1.4087) | 3.7375 (1.0499) | 6.1531 (0.7677) | 1.2969 (0.7224) | 2.8 (1.1247) |

Table 26: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 1.5562 (1.1183) | 5.325 (2.1344) | 3.9406 (1.3805) | 4.9406 (1.1715) | 2.5969 (1.1997) | 5.7688 (1.1313) | 3.8719 (1.395) |
| EF1% | 4.7969 (1.3227) | 1.5781 (1.1646) | 6.4906 (1.0485) | 3.8812 (1.215) | 3.0562 (1.6825) | 3.2969 (1.4964) | 4.9 (1.412) |
| EF2.5% | 4.6469 (1.531) | 1.725 (1.3525) | 6.2812 (1.1613) | 3.9906 (1.1955) | 2.5562 (1.4711) | 4.1438 (1.7338) | 4.6562 (1.4899) |
| Mean Sim. | 3.5438 (1.3338) | 5.5875 (1.5054) | 4.8062 (1.5871) | 3.5719 (1.1636) | 6.0875 (1.0455) | 1.4 (1.0223) | 3.0031 (1.4074) |

Table 27: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6969 (0.0442) | 0.6718 (0.1038) | 0.6301 (0.0482) | 0.6294 (0.0593) | 0.7133 (0.0774) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 6.4273 (5.4834) | 17.4081 (14.927) | 5.3393 (4.8846) | 8.1632 (7.7244) | 12.7796 (11.2379) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 3.9184 (2.4352) | 9.9987 (7.1789) | 3.4637 (2.4133) | 4.8412 (3.6451) | 7.9113 (5.5725) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3325 (0.0613) | 0.2611 (0.0394) | 0.3194 (0.0638) | 0.3249 (0.053) | 0.2712 (0.0442) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 28: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 254 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6969 (0.0442) | 0.6111 (0.081) | 0.6315 (0.0472) | 0.618 (0.0533) | 0.66 (0.0595) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 6.4273 (5.4834) | 11.4568 (10.2393) | 5.3209 (5.0244) | 7.2046 (6.4526) | 8.253 (7.1874) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 3.9184 (2.4352) | 6.7569 (5.0808) | 3.4619 (2.4589) | 4.2877 (3.007) | 5.3542 (3.7783) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3325 (0.0613) | 0.2953 (0.059) | 0.3177 (0.065) | 0.3307 (0.0579) | 0.2948 (0.0583) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 29: Average Mean(SD) of the Similarity Search Performance Measures with a Fingerprint size of 254 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 2.1562 (1.166) | 3.0312 (2.0248) | 6.3219 (0.9692) | 4.6281 (1.0587) | 2.0781 (0.7499) | 5.7594 (0.9082) | 4.025 (1.2417) |
| EF1% | 3.9344 (1.0658) | 1.0969 (0.3052) | 6.85 (0.5275) | 4.6688 (1.1308) | 2.2031 (0.7983) | 3.9 (1.0932) | 5.3469 (1.0267) |
| EF2.5% | 3.7594 (1.0548) | 1.1719 (0.4957) | 6.7438 (0.7228) | 4.7406 (1.154) | 2.0188 (0.5879) | 4.5719 (1.2103) | 4.9938 (1.133) |
| Mean Sim. | 2.4781 (0.8921) | 5.7875 (1.3292) | 5.3125 (1.5694) | 3.7594 (1.0825) | 6.0375 (0.74) | 1.3188 (0.7645) | 3.3062 (1.1746) |

Table 30: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 1.5281 (0.977) | 4.4312 (2.2064) | 5.8219 (1.3009) | 4.6688 (1.2516) | 2.925 (1.1767) | 5.2906 (1.1954) | 3.3344 (1.3872) |
| EF1% | 3.2812 (1.1483) | 1.4875 (1.0985) | 6.7594 (0.774) | 5.0062 (1.1536) | 3.3312 (1.5931) | 3.25 (1.3389) | 4.8844 (1.297) |
| EF2.5% | 3.1875 (1.1619) | 1.5188 (1.1216) | 6.6594 (0.8827) | 5.1344 (1.2287) | 2.925 (1.4722) | 4.0188 (1.4992) | 4.5562 (1.3325) |
| Mean Sim. | 2.6781 (1.0784) | 4.9062 (1.5425) | 5.7906 (1.6173) | 3.7688 (1.2748) | 5.8 (1.1038) | 1.4188 (1.0192) | 3.6375 (1.3264) |

Table 31: Average Rank (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6779 (0.049) | 0.6777 (0.0961) | 0.589 (0.0415) | 0.621 (0.0512) | 0.6848 (0.0687) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 7.5112 (6.6055) | 15.2025 (13.0442) | 4.726 (4.5044) | 7.197 (6.6415) | 10.4608 (9.4026) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 4.4164 (2.9585) | 9.0239 (6.5004) | 3.0841 (2.1521) | 4.225 (3.0779) | 6.4953 (4.6746) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3496 (0.0508) | 0.2859 (0.0391) | 0.3077 (0.0645) | 0.332 (0.0561) | 0.2906 (0.0469) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 32: Average Mean (SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024 when included during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

| | MLP generic | MLP MT | GCN generic | GCN MT | Freeze MT | ECFP4 | CDDD |
|---|---|---|---|---|---|---|---|
| AUC | 0.6779 (0.049) | 0.6258 (0.075) | 0.591 (0.0415) | 0.6099 (0.0486) | 0.6443 (0.0576) | 0.604 (0.0503) | 0.632 (0.0502) |
| EF1% | 7.5112 (6.6055) | 10.9072 (9.3638) | 4.5517 (4.1088) | 6.501 (6.0679) | 7.5994 (6.4745) | 7.9292 (7.6529) | 6.5469 (5.8295) |
| EF2.5% | 4.4164 (2.9585) | 6.4804 (4.6945) | 2.9974 (1.9874) | 3.8456 (2.7433) | 4.8364 (3.3044) | 4.3726 (3.2898) | 3.9855 (2.7332) |
| Mean Sim. | 0.3496 (0.0508) | 0.3139 (0.0535) | 0.3053 (0.066) | 0.3345 (0.0593) | 0.3054 (0.0574) | 0.372 (0.0417) | 0.3374 (0.0562) |

Table 33: Average Mean(SD) of the Similarity Search Performance Measures with a Fingerprint size of 1024 when excluded during training. The Standard Deviation presented here is the average Standard Deviation across the targets, not the 5-folds of the Cross-Validation.

# Hyperparameters

| Parameter | Range | Optimal MLP (gen/MT) | GCN (gen/MT) |
|---|---|---|---|
| Learning Rate | $10^{-6}$–0.05 (log-scale) | $5.9 \cdot 10^{-5}$/ $4.6 \cdot 10^{-5}$ | $1.7 \cdot 10^{-4}$/ $1.8 \cdot 10^{-4}$ |
| Dropout | 0.0–0.4 | 0.21/0.33 | 0.2/0.22 |
| N. of Convolution layers (GCN only) | $[1, 2, 3]$ | - | 3/3 |
| Size of Embedding (GCN only) | $[50, 100, 200]$ | - | 200/200 |
| N. of Hidden Linear Layers | $[1, 2, 3, 4]$ | 2/2 | 1/1 |
| Size of Linear Layer | $[254, 512, 1024]$ | 1024/1024 | -/- |
| Fingerprint Size | $[64, 254, 1024]$ | 64/1024 | 64/1024 |

Table 34: Hyperparameters and their Ranges.