Neural Similarity Search for Drug Target Prediction

Nicolas Madella-Mella - k12021216
Department of Machine Learning
Johannes Kepler University Linz
Austria, Linz 4040
Github-Projekt
nicolas.madellamella@gmail.com

June 2022

Abstract

Building models for molecular properties and activities with minimal measurements is a crucial challenge in drug discovery. Typically drug discovery projects begin with one or a few known active molecules and then face the task of identifying promising candidates for further screening. Few-shot learning methods have been introduced to computer-aided drug design to enhance this critical phase of the process. It has been shown that effective few-shot learning methods and architectures can significantly reduce time and costs associated with in-vitro testing, thereby accelerating the drug development process. In this work the Neural Similarity Search was reimplemented, showing that a fully-connected Deep Neural Network using scaled exponential units in a Siamese fashion and paired with Similarity Search is suitable to tackle the problem of low measurements. For this the SIDER dataset was used to evaluate the model's performance. The Neural Similarity Search not only significantly outperformed a Random Forest baseline, but also achieved superior results compared to state-of-the-art benchmarks on the SIDER dataset.

1 The story summary

The focus of this report is on utilizing few-shot learning methods to address the challenge of building predictive models with limited data in drug discovery. Initially, a descriptor-based fully-connected Deep Neural Network (DNN) with scaled exponential units (Klambauer et al., 2017) was employed, configured in a Siamese Network architecture (Koch et al., 2015). This approach was chosen based on the framework suggested by Schimunek et al. (2021). The performance of this Neural Network model was assessed against a Random Forest baseline, using metrics such as the Area Under the Curve (AUC) and the normalized Area Under the Precision-Recall Curve (ΔAUC-PR) as suggested in (Stanley

et al., 2021). The evaluation results demonstrated that the neural variant significantly outperformed the Random Forest baseline, indicating the effectiveness of embedding-based few-shot learning methods paired with this type of model architecture and Similarity Search. The superiority is evident even with a minimal number of measurements, as exemplified by the SIDER dataset (Kuhn et al., 2016). Additionally, the Neural Similarity Search model not only exceeded the Random Forest baseline but also achieved superior results compared to state-of-the-art benchmarks.

2 Introduction

Enhancing human health, fighting diseases, and addressing future pandemics, as demonstrated by recent experiences with the COVID pandemic, have shown that discovering new drugs quickly and efficiently is essential. But the drug discovery process is time-consuming and cost-intensive (Arrowsmith, 2011). In order to reduce costly and time-consuming in vitro experiments and accelerate drug development, machine learning algorithms are used to predict the binding affinity of molecules towards target proteins, which allows to narrow down the number of suitable candidates.

Specifically Deep Learning models were used to reach those goals, with the disadvantage of them being very data hungry (Marcus, 2018) and therefore need very large amounts of measurements for accurate activity prediction. So to reach high predictive performance for Deep Learning-based binding activity prediction, hundreds or thousands of measurements per task are needed. Well performing models for activity prediction tasks of ChEMBL have been trained with an average of 3,621 activity points per task (Mayr et al., 2018). The ExCAPE-DB dataset offers an average of 42,501 measurements per task (Sun et al., 2017). Wu et al. (2018) provided a large-scale benchmark for molecular machine learning, including prediction models for the SIDER dataset (Kuhn et al., 2016) with an average of 5,187 data points, which is also the dataset being focused on in this work, Tox21 (Huang et al., 2016; Mayr et al., 2016) with an average of 9,031 data points, and ClinTox (Wu et al., 2018) with 1,491 measurements per task. However, for many drug targets, the available measurements are very limited (Stanley et al., 2021; Altae-Tran et al., 2017; Waring et al., 2015) due to the resources required for in-vitro experiments. Therefore, methods that can utilize few measurements to build valuable property and activity prediction models are necessary.

The challenge of training models with limited data is addressed by the machine learning fields of meta-learning and few-shot learning (Schmidhuber, 1987; Bengio et al., 1990; Hochreiter et al., 2001; Miller et al., 2000; Bendre et al., 2020; Y. Wang et al., 2020). Few-shot learning, initially developed for image datasets (Bendre et al., 2020; Y. Wang et al., 2020), is now being applied to drug discovery (Adler et al., 2020; Stanley et al., 2021; Altae-Tran et al., 2017).

These methods are categorized into four main approaches (Bendre et al., 2020; Y. Wang et al., 2020):

a Data-augmentation-based approaches generate diverse data points from the limited available samples (T. Chen et al., 2020; Zhao et al., 2019; Antoniou and Storkey, 2019).

b Embedding-based and nearest neighbor approaches create meaningful representations for low-shot predictions, such as Matching Networks using attention mechanisms (Vinyals et al., 2016) and Prototypical Networks creating class prototypes (Snell et al., 2017).

c Optimization-based or fine-tuning methods, like MAML, adapt initial weights to new tasks with few optimization steps (Finn et al., 2017).

d Semantic-based approaches use additional semantic information for new task learning (Schwartz et al., 2022; Li et al., 2019).

In drug discovery, Adler et al. (2020) proposed a cross-domain few-shot learning method using representation fusion. Nguyen et al. (2020) and Guo et al. (2021) explored GNNs combined with meta-learning. Altae-Tran et al. (2017) developed Iterative Refinement Long Short-Term Memory for molecule embeddings. Stanley et al. (2021) created a benchmark dataset for few-shot learning in drug discovery, but connections to classic cheminformatics techniques like Similarity Search (Cereto-Massagué et al., 2015) remain unclear.

Schimunek et al. (2021) therefore presented a framework for embedding-based few-shot learning methods integrating Similarity Search, as well as Deep Learning techniques. This work aims at reimplementing this approach, train it in a few shot manner using the SIDER Datset (Kuhn et al., 2016), which contains a low amount of measurements and compared it to a Random Forest Baseline.

3 Methods

3.1 Description of datasets used:

For this work's experiments the SIDER dataset was used (Kuhn et al., 2016). It contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. The available information include side effect frequency, drug and side effect classifications as well as links to further information, for example drug–target relations. SIDER consists of activity measurements from 1427 molecules on 27 tasks, either being inactive or active. As there are ni missing values, this leads up to 38529 measurements overall.

3.2 Descriptors and Preprocessing:

In this preprocessing pipeline, molecular data encoded as SMILES strings is systematically transformed into a numerical format suitable for multi-task learning. The process begins by converting SMILES strings into RDKit Mol objects,

which accurately represent the molecular structures. Extended Connectivity Fingerprints (ECFPs) (Rogers and Hahn, 2010) are then generated for each molecule, capturing structural features in a high-dimensional vector format. Additionally, RDKit descriptors are computed to quantify various molecular properties. A crucial step involves applying an empirical cumulative distribution function (ECDF) to the descriptors, which normalizes them by transforming raw values into quantiles. This method mitigates the impact of outliers and enhances the consistency of the data across features. Unlike traditional approaches where the ECDF is fitted solely on the training set, here it is applied across all molecules in the dataset. This strategy is justified by the taskwise splitting of the data, where each molecule of the SIDER Dataset (Kuhn et al., 2016) appears in training, validation, and test sets, but with different task labels. By applying the ECDF globally, we ensure consistent feature representation across tasks. The final step involves concatenating the ECFPs and quantile-transformed descriptors, followed by standardization using sklearn's (Pedregosa et al., 2011) StandardScaler to ensure all features are on a uniform scale. This preprocessing strategy effectively prepares the data for robust and consistent model training, ensuring that molecular features are uniformly processed across all tasks.

3.3 Model architecture:

As introduced by Schimunek et al. (2021), a descriptor-based fully-connected Deep Neural Network with scaled exponential units (Klambauer et al., 2017) as the molecule and memory encoder in a Siamese Network fashion (Koch et al., 2015) was implemented. In a first step the representations for the query molecule and support set molecules are encoded by this fully-connected network and transformed by the Layer Normalization operation (J. L. Ba et al., 2016). This can be seen as a function h_ω with learnable parameters ω to compute the feature representations for molecules

$$h_m ol(x) = h_\omega(x)$$

where h_{mol} represents the embedding of molecule x.

Applying this function to the query molecule, as well to the as positive and negative support set molecules respectively, gives the representations q, p_i, n_j, for $i \in \{1, ..., P\}$ and $j \in \{1, ..., N\}$ where $N = P = 8$, being the number of support set molecules used in this work's experiments.

The resulting embeddings are then compared using the dot product to measure similarity. The dot product similarities between the query molecule and each of its corresponding support set molecules are averaged separately for the positive and negative support sets.

$$q_p = \frac{1}{|P|} \sum_{i=1}^{|P|} (q \cdot p_i)$$

$$q_n = \frac{1}{|N|} \sum_{j=1}^{|N|} (q \cdot n_j)$$

Finally, the raw prediction values are multiplied by a constant c, being $1/\sqrt{D}$, where D is the embedding dimension and scaled to a range between zero and one using the sigmoid function.

$$\hat{y} = \sigma\left(c\left(q_n - q_p\right)\right)$$

This process is also illustrated in Figure 1. During the experiments also the cosine similarity was tested, but no benefits were gained from it, so the dot product was being kept as the similarity comparison in the association function.

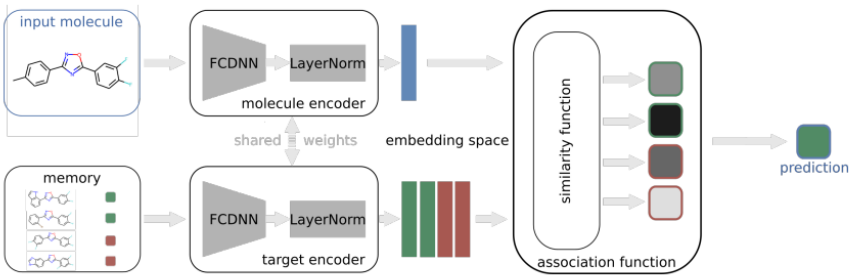

Figure 1: Overview neural search variant

First attempts at implementing the method, unknowingly the prototypical network (Snell et al., 2017) was mimicked. The key idea of this network is that each class can be represented by a single prototype in an embedding space. This prototype is the mean of the class's support set in the embedding space. A neural network is used to learn a non-linear mapping from input data to this space. Classification of a query point is then done by finding the nearest class prototype.

3.4 Evaluation

The performance of the Neural Similarity Search variant (Schimunek et al., 2021) and the Random Forest baseline are assessed using the Area Under the Curve (AUC) and as proposed by Stanley et al. (2021) the normalized Area Under the Precision-Recall Curve (ΔAUC-PR). Each task of the Sider dataset (Kuhn et al., 2016) is evaluated individually, with the results averaged to provide a comprehensive measure across all tasks. The ΔAUC-PR metric quantifies the model's improvement over random guessing by subtracting the AUC-PR of a random classifier, based on the proportion of positive samples. A positive ΔAUC-PR indicates superior performance, whereas a near-zero or negative value suggests performance comparable to or worse than random guessing.

The comparison of the Neural Similarity Search with the two benchmark methods from (W. Chen et al., 2022) is limited to the AUC metric, as ΔAUC-PR was not reported for those methods.

4 Experiments

4.1 Methods compared

The primary objective of this work is to compare the performance of the Neural Similarity Search model (Schimunek et al., 2021) against a Random Forest baseline. Additionally, a secondary aim was to surpass benchmark results reported in (W. Chen et al., 2022), specifically the performance of the Siamese network (Koch et al., 2015), which closely aligns with the architecture used in this study, and the PAR (Property-aware Relation network) model (Y. Wang et al., 2021), which previously demonstrated superior results. The PAR model employs a graph-based molecular encoder to generate initial embeddings, which are subsequently refined into property-aware embeddings via a specialized function that learns an adaptive relation graph, effectively capturing molecular relationships and propagating labels. Initial experiments, conducted using the parameters provided in (Schimunek et al., 2021), yielded promising results. Furthermore, a hyperparameter search script was developed and executed within a defined search space. However, this exploration did not significantly improve upon the evaluations already achieved.

4.2 Support sets and Data split

The dataset was divided into training, validation, and test sets, with 60%, 20%, and 20% of the tasks allocated to each split, respectively. During training, for each of the 1,427 molecules, eight active and eight inactive molecules were randomly selected (excluding the original molecule) to form the support set for the training process. In the validation and test sets, a fixed support set of eight active and eight inactive molecules was utilized for each of the remaining 1,411 molecules in the SIDER dataset (Kuhn et al., 2016).

4.3 Hardware Specifications

The experiments were conducted using the following hardware setup:

- **Processor**: AMD Ryzen 5 5600 (3.5 GHz, 6 cores)
- **Memory**: 32 GB DDR4 RAM
- **GPU**: NVIDIA GeForce RTX 4060 TI, 16 GB
- **Operating System**: Microsoft Windos 11 Pro

4.4 Training

The Random Forest baseline consists of 1000 decision trees. Data shuffling was enabled to enhance model robustness against data order effects. The hyperparameters used for the Neural Similarity Search Model (Schimunek et al., 2021) are stated below in Table 1. The model was trained using the Adam optimizer (Kingma and J. Ba, 2014) to minimize binary cross-entropy loss.
To monitor and manage the training statistics of the Neural Similarity Search variant TensorBoard from the TensorFlow framework (Abadi et al., 2015) was utilized. For both the training and validation sets, the area under the receiver

operating characteristic curve (AUC) as well as the normalized Area Under the Precision-Recall Curve (ΔAUC-PR) was recorded. Additionally, the training and validation loss (binary cross-entropy-loss) were logged and the model with the best performance was consistently saved. Early stopping was used to ensure that the training stops when the model is no longer improving, thus avoiding overfitting and unnecessary computations. AThe stopping criterion used was ΔAUC-PR, as it provided the best results. To facilitate training across multiple seeds and to average the evaluation results for both the Neural Similarity Search variant and the Random Forest baseline, a separate script was developed and utilized.

Table 1: Hyperparameters used in Training

Hyperparameter	**Value**
Learning Rate	0.001
Weight Decay	0
Batch Size	32
Input Dimension	2248
Hidden Dimension	1124
Output Dimension	1124
Number of Layers	2
Dropout Probability (p)	0.1

5 Results

In this section, the performance of the trained methods, including the Neural Similarity Search variant (Schimunek et al., 2021) and the Random Forest baseline, are compared. Additionally, benchmark results reported in (W. Chen et al., 2022) are incorporated, specifically those of a Siamese network (Koch et al., 2015), which shares architectural similarities with the Neural Similarity Search variant and the Property-aware Relation Network (PAR) (Y. Wang et al., 2021), which achieved the best performance on the SIDER dataset (Kuhn et al., 2016) in terms of AUC. A visualization of the training's impact on the learned embeddings is also presented.

Table 2: Average Performance of Models Across All Tasks (%)

Model	**AUC**	**ΔAUC-PR**
Neural Similarity Search	**77.97**	**17.79**
Random Forest Baseline	55.11	1.65

Table 2 presents the average performance of the Neural Similarity Search and the Random Forest baseline model across all tasks, measured using AUC and ΔAUC-PR. The Neural Similarity Search model significantly outperforms the Random Forest baseline in both metrics.

Table 3: Performance Comparison Across Benchmarks (%)

Model	**AUC**
Neural Similarity Search	**77.97**
PAR	74.68
Siamese Network	71.10
Random Forest Baseline	55.11

Table 3 compares the AUC performance of the Neural Similarity Search model with several benchmark models, including PAR, Siamese Network and the initial Random Forest baseline. The Neural Similarity Search model demonstrates superior AUC performance compared to all benchmarks.

Table 4: AUC Performance Across Multiple Seeds (%)

Model	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Neural Similarity Search	**76.15**	**81.67**	**76.60**	**81.06**	**76.25**	**83.38**	**75.30**	**76.05**	**75.01**	**78.20**
Random Forest Baseline	53.91	54.81	55.98	55.00	55.21	55.29	56.10	54.23	53.81	56.73

Table 4 details the AUC performance of the Neural Similarity Search and Random Forest models across ten different seeds, representing varied support sets. The Neural Similarity Search consistently outperforms the Random Forest across all seed variations.

Table 5: ΔAUC-PR Performance Across Multiple Seeds (%)

Model	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Neural Similarity Search	**15.94**	**16.17**	**16.03**	**26.29**	**15.80**	**18.81**	**19.83**	**16.35**	**15.76**	**16.90**
Random Forest Baseline	0.89	1.25	2.43	2.00	1.78	1.00	2.54	1.12	1.57	1.96

Table 5 displays the ΔAUC-PR performance of the Neural Similarity Search and Random Forest models across ten seeds, reflecting different support sets. The Neural Similarity Search model consistently achieves higher ΔAUC-PR values compared to the Random Forest.

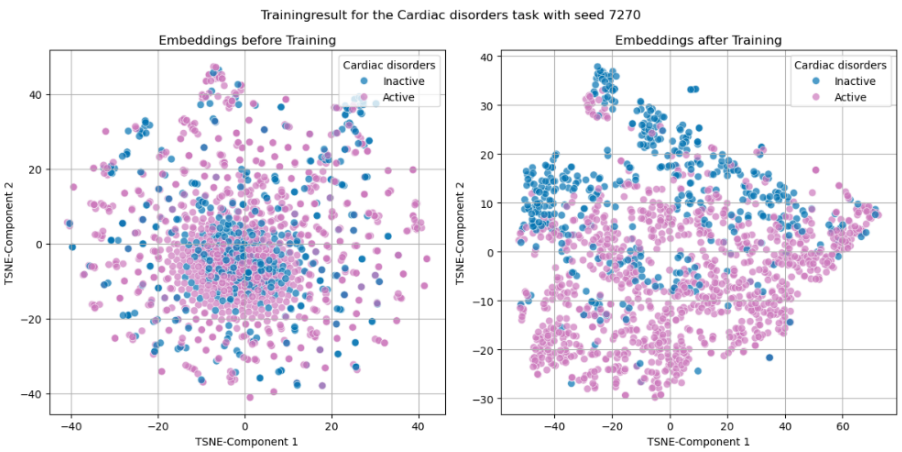

Figure 2: Comparison of Embeddings before and after Training

Figure 2 visualizes the feature embeddings before and after training, illustrating the improvements in clustering and separability as a result of training. The post-training embeddings exhibit clearer groupings, highlighting the impact of the training process on feature representations.

6 Conclusion and Outlook

In this work, the Neural Similarity Search framework (Schimunek et al., 2021) for drug target prediction was reimplemented and was compared against a Random Forest baseline. By applying the model to the SIDER dataset (Kuhn et al., 2016), which contains a relatively small number of samples, it demonstrated the efficacy of few-shot learning methods integrating Similarity Search (Cereto-Massagué et al., 2015) in the drug discovery process. The results show that the Neural Similarity Search model significantly outperformed the Random Forest baseline, achieving higher AUC and ΔAUC-PR scores.

This indicates that embedding-based methods, particularly in a Siamese network configuration (Koch et al., 2015), are more capable of capturing the underlying molecular patterns with limited data compared to traditional machine learning methods. The model's robust performance across different seeds suggests that it generalizes well, even with variations in the support set, reinforcing the potential of neural architectures for drug discovery when faced with data scarcity.

In comparison to state-of-the-art benchmarks, the Neural Similarity Search model surpassed the performance of both the Siamese Network (Koch et al., 2015) and the PAR model (Y. Wang et al., 2021) reported in W. Chen et al. (2022), highlighting the effectiveness of integrating similarity search mechanisms with deep learning approaches. This success emphasizes the importance of adopting flexible neural architectures that can adapt to the specific demands of drug discovery tasks.

Future work will focus on optimizing the training process with better stopping criteria, as well as incorporating an additional Transformer encoder after the fully-connected network encoder, as proposed by Fifty et al. (2023), building on the existing architecture's ability to effectively process molecular embeddings. This extended model would allow for deeper contextualization of the relationships between support set molecules and the query molecule.
By adding another Transformer encoder, the aim is to further enhance the model's capacity to capture complex patterns and dependencies within molecular property prediction tasks. This could enable improved generalization across diverse datasets and boost the model's performance in few-shot learning scenarios, potentially surpassing current results.

References

Klambauer, Günter et al. (2017). "Self-normalizing neural networks". In: *Advances in neural information processing systems* 30.

Koch, Gregory, Richard Zemel, Ruslan Salakhutdinov, et al. (2015). "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. 1. Lille.

Schimunek, Johannes et al. (2021). "A generalized framework for embedding-based few-shot learning methods in drug discovery". In.

Stanley, Megan et al. (2021). "Fs-mol: A few-shot learning dataset of molecules". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kuhn, Michael et al. (2016). "The SIDER database of drugs and side effects". In: *Nucleic acids research* 44.D1, pp. D1075–D1079.

Arrowsmith, John (2011). "Phase II failures: 2008-2010". In: *Nature reviews Drug discovery* 10.5.

Marcus, Gary (2018). "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631*.

Mayr, Andreas et al. (2018). "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL". In: *Chemical science* 9.24, pp. 5441–5451.

Sun, Jiangming et al. (2017). "ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics". In: *Journal of cheminformatics* 9, pp. 1–9.

Wu, Zhenqin et al. (2018). "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical science* 9.2, pp. 513–530.

Huang, Ruili et al. (2016). "Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization". In: *Nature communications* 7.1, p. 10425.

Mayr, Andreas et al. (2016). "DeepTox: toxicity prediction using deep learning". In: *Frontiers in Environmental Science* 3, p. 80.

Altae-Tran, Han et al. (2017). "Low data drug discovery with one-shot learning". In: *ACS central science* 3.4, pp. 283–293.

Waring, Michael J et al. (2015). "An analysis of the attrition of drug candidates from four major pharmaceutical companies". In: *Nature reviews Drug discovery* 14.7, pp. 475–486.

Schmidhuber, Jürgen (1987). *Evolutionary Principles in Self-referential Learning: On Learning how to Learn: the Meta-meta-meta...-hook*.

Bengio, Yoshua, Samy Bengio, and Jocelyn Cloutier (1990). *Learning a synaptic learning rule*. Citeseer.

Hochreiter, Sepp, A Steven Younger, and Peter R Conwell (2001). "Learning to learn using gradient descent". In: *Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11*. Springer, pp. 87–94.

Miller, Erik G, Nicholas E Matsakis, and Paul A Viola (2000). "Learning from one example through shared densities on transforms". In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. Vol. 1. IEEE, pp. 464–471.

Bendre, Nihar, Hugo Terashima Marín, and Peyman Najafirad (2020). "Learning from few samples: A survey". In: *arXiv preprint arXiv:2007.15484*.

Wang, Yaqing et al. (2020). "Generalizing from a few examples: A survey on few-shot learning". In: *ACM computing surveys (csur)* 53.3, pp. 1–34.

Adler, Thomas et al. (2020). "Cross-domain few-shot learning by representation fusion". In.

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR, pp. 1597–1607.

Zhao, Amy et al. (2019). "Data augmentation using learned transformations for one-shot medical image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8553.

Antoniou, Antreas and Amos Storkey (2019). "Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation". In: *arXiv preprint arXiv:1902.09884*.

Vinyals, Oriol et al. (2016). "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29.

Snell, Jake, Kevin Swersky, and Richard Zemel (2017). "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR, pp. 1126–1135.

Schwartz, Eli et al. (2022). "Baby steps towards few-shot learning with multiple semantics". In: *Pattern Recognition Letters* 160, pp. 142–147.

Li, Aoxue et al. (2019). "Large-scale few-shot learning: Knowledge transfer with class hierarchy". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7212–7220.

Nguyen, Cuong Q, Constantine Kreatsoulas, and Kim M Branson (2020). "Meta-learning GNN initializations for low-resource molecular property prediction". In: *arXiv preprint arXiv:2003.05996*.

Guo, Zhichun et al. (2021). "Few-shot graph learning for molecular property prediction". In: *Proceedings of the web conference 2021*, pp. 2559–2567.

Cereto-Massagué, Adrià et al. (2015). "Molecular fingerprint similarity search in virtual screening". In: *Methods* 71, pp. 58–63.

Rogers, David and Mathew Hahn (2010). "Extended-connectivity fingerprints". In: *Journal of chemical information and modeling* 50.5, pp. 742–754.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. URL: `http://jmlr.org/papers/v12/pedregosa11a.html`.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer normalization". In: *arXiv preprint arXiv:1607.06450*.

Chen, Wenlin, Austin Tripp, and José Miguel Hernández-Lobato (2022). "Meta-learning adaptive deep kernel gaussian processes for molecular property prediction". In: *arXiv preprint arXiv:2205.02708*.

Wang, Yaqing et al. (2021). "Property-aware relation networks for few-shot molecular property prediction". In: *Advances in Neural Information Processing Systems* 34, pp. 17441–17454.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Abadi, Martín et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*.

Fifty, Christopher, Jure Leskovec, and Sebastian Thrun (2023). "In-context learning for few-shot molecular property prediction". In: *arXiv preprint arXiv:2310.08863*.