Journal of Data-centric Machine Learning Research (2024) Submitted 2/24; Revised 4/24; Published 5/24

Autoregressive activity prediction for low-data drug discovery

Johannes Schimunek[1], **Lukas Friedrich**[2], **Daniel Kuhn**[2], **and Günter Klambauer**[1]
SCHIMUNEK@ML.JKU.AT
[1] *ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning, Johannes Kepler University Linz, Austria*
[2] *Merck Healthcare KGaA, Medicinal Chemistry and Drug Design, 64293 Darmstadt, Germany*

Reviewed on OpenReview: *https://openreview.net/forum?id=eLg88qUVFT*

Abstract

Autoregressive modeling is the main learning paradigm behind the currently so successful large language models (LLM). For sequential tasks, such as generating natural language, autoregressive modeling is a natural choice: the sequence is generated by continuously appending the next sequence token. In this work, we investigate whether the autoregressive modeling paradigm could also be successfully used for molecular activity and property prediction models, which are equivalent to LLMs in molecular sciences. To this end, we formulate autoregressive activity prediction modeling (AR-APM), draw relations to transductive and active learning, and assess the predictive quality of AR-APM models in few-shot learning scenarios. Our experiments show that using an existing few-shot learning system without any other changes, except switching to autoregressive mode for inference, improves ΔAUC-PR up to ∼40%. Code is available here: `https://github.com/ml-jku/autoregressive_activity_prediction`.

Keywords: autoregressive inference, few-shot learning, low-data drug discovery

1 Introduction

Autoregressive modeling and large language models. Autoregressive modeling (Yule, 1927; Whittle, 1951; Box et al., 2015; Radford et al., 2018) is a fundamental approach within the domain of sequence and generative modeling, exemplified prominently by the recent success of large language models (LLMs) (Vaswani et al., 2017; Brown et al., 2020b). LLMs have shown remarkable capabilities at text generation, translation, text summarization, and as conversational agents (Zhao et al., 2023). These LLMs are deep neural networks, usually based on the Transformer architecture (Vaswani et al., 2017), and are trained to solve autoregressive tasks, concretely predicting the correct next tokens given a sequence of previous tokens. Overall, autoregressive modeling has emerged as a cornerstone technique for natural language processing and other areas. However, in other areas, such as the molecular sciences, where the data is not naturally sequential, it is still unclear whether the success of autoregressive modeling can be carried over and how these areas will be impacted. In this work, we investigate whether molecular activity and property prediction

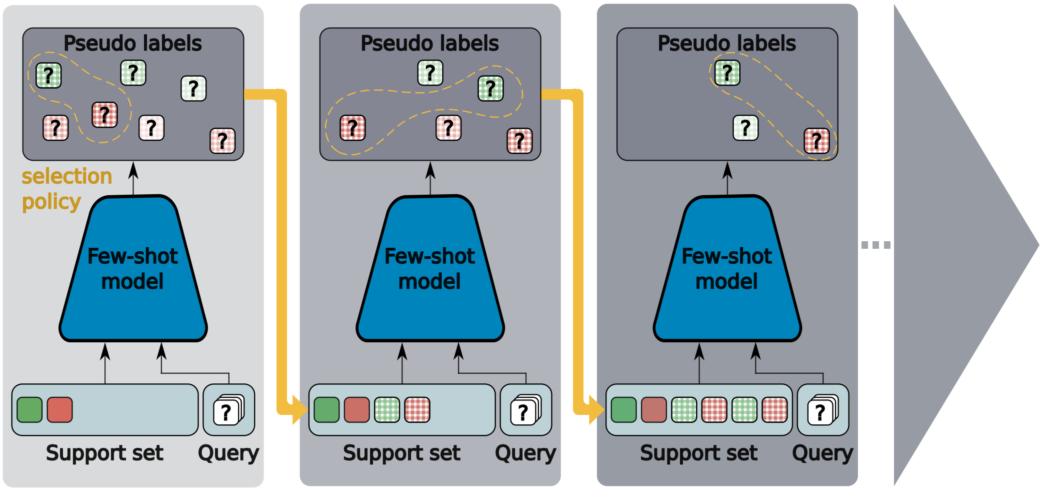

Figure 1: Overview of autoregressive activity prediction (AAP) models. **Left column:** Initially, only 1 active and 1 inactive molecules are known. Based on these two molecules, an embedding-based few-shot learning model predicts the labels of the query set. **Second column:** The top predictions of the query set are added to the support set. Again, the few-shot model predicts the labels of the query set. Thus, the model is conditioned on its own prediction from the step before.

models (Mayr et al., 2018; Yang et al., 2019; Deng et al., 2023), which are the equivalent of language models for molecular sciences, can be improved with the autoregressive modeling paradigm. Concretely, we will propose autoregressive modeling of few-shot learning data in drug discovery.

Low-data drug discovery, few- and zero-shot learning on molecules. Molecular activity and property prediction models play a crucial role in numerous drug discovery projects (Green, 2019; Brown et al., 2020a; Tyrchan et al., 2022; Volkamer et al., 2023). Since seeking suitable drug candidates often faces the low-data obstacle, Deep Learning methods have been designed for, applied to, and evaluated for these low-data scenarios (Stanley et al., 2021; The COVID Moonshot Consortium et al.; Schimunek et al., 2024). While already optimized for these low-data scenarios, recently used few-shot models (Altae-Tran et al., 2017; Guo et al., 2021; Wang et al., 2021; Schimunek et al., 2023; Chen et al., 2022) stick to the initially available measurements provided as a support set.

Contributions. In contrast to the existing few-shot learning modeling, we suggest an autoregressive inference scheme for few-shot drug-discovery models which enables support set augmentation in an iterative fashion by including new pseudo-labeled samples to the support set. We show that this autoregressive inference scheme applied to the SOTA few-shot model MHNfs boosts the model performance in terms of AUC and ΔAUC-PR for very low-data scenarios, i.e. scenarios in which just one active and one inactive molecule is known.

2 Background and related work

Few-shot learning refers to methods that are geared to learning accurate predictive models with in scenarios $\mathcal{T}_*$ where only little data is available. Usually, a few-shot model $g(\cdot; \boldsymbol{w})$ with learnable parameters $\boldsymbol{w}$ is provided with a set of training tasks $\mathcal{D}_{\text{train}} = \{\mathcal{T}_t\}_{t=1}^{T}$ during training, and, during test/inference time, a set of unseen tasks $\mathcal{D}_{\text{eval}} = \{\mathcal{T}_*\}$: $\mathcal{D}_{\text{eval}} \cap \mathcal{D}_{\text{train}} = \varnothing$. Each task $\mathcal{T}$ comprises a set of data points, i.e. pairs of molecular inputs $\boldsymbol{x}$ and associated labels y: $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_K, y_K)\}$. The labels are assumed to be either binary or unknown: $y \in \{0, 1, \square\}$, where $\square$ indicates the unknown class. For test/inference tasks $\mathcal{T}_*$, typically the amount of labeled data $\mathcal{S} \subset \mathcal{T}_*$ is assumed to be limited which could be used to tune model parameters and help to predict molecules $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L\}$ from the query set $\mathcal{Q} \subset \mathcal{T}_*$, $\mathcal{Q} \cap \mathcal{S} = \varnothing$. $\mathcal{S}$ is called the support set. Notably, molecules assigned to the label 0 (1) are considered active (inactive).

Inductive inference. In standard few-shot drug-discovery settings, e.g., as provided in the FS-Mol benchmark (Stanley et al., 2021), query molecules $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L\}$ usually are treated independently. This is called inductive inference:

$$\hat{y}_l = g\left(\boldsymbol{q}_l; \mathcal{A}(\boldsymbol{w}, \mathcal{S}, \mathcal{D}_{\text{train}})\right) \quad \forall\ 1 \leqslant l \leqslant L, \tag{1}$$

where $\mathcal{A}$ is a possibly complex learning algorithm which maps the parameters, the training data and the support set onto new parameters.

Semi-supervised learning and transductive inference. Semi-supervised learning methods extend the support set $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ with unlabeled data $\mathcal{U} = \{(\boldsymbol{x}_{N+1}, \square), \ldots, (\boldsymbol{x}_{N+M}, \square)\}$, which frames few-shot learning as the task of learning from labeled and unlabeled data:

$$\hat{y}_l = g\left(\boldsymbol{q}_l; \mathcal{A}(\boldsymbol{w}, \mathcal{S} \cup \mathcal{U}, \mathcal{D}_{\text{train}})\right) \quad \forall\ 1 \leqslant l \leqslant L. \tag{2}$$

In transductive inference the samples included in this additional unlabeled data set $\mathcal{U}$ are the query molecules $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L\}$.

Pseudo labeling and label propagation. Recent methods (Iscen et al., 2019; Liu et al., 2018; Lazarou et al., 2021; Zhu and Koniusz, 2023) leverage this unlabeled dataset $\mathcal{U}$ by augmenting the support set (iteratively) with pseudo-labeled samples given in $\mathcal{U}$. Label propagation (Zhu and Ghahramani, 2002; Zhou et al., 2003; Liu et al., 2018) is the process of creating pseudo-labels for unlabeled samples by propagating given label information through an nearest-neighbor based graph which includes both labeled and unlabeled samples.

Feature space adaption and embedding propagation. Rodríguez et al. (2020) propose embedding propagation, which is an unsupervised non-parametric regularizer for manifold smoothing in few-shot classification. Embedding propagation leverages interpolations between the extracted features of a neural network based on a similarity graph. Similarly, Hu et al. (2021a) use feature interpolations based on a similarity graph in few-shot settings. Hu et al. (2021b) introduce class-wise feature preprocessing and feature distribution leveraging

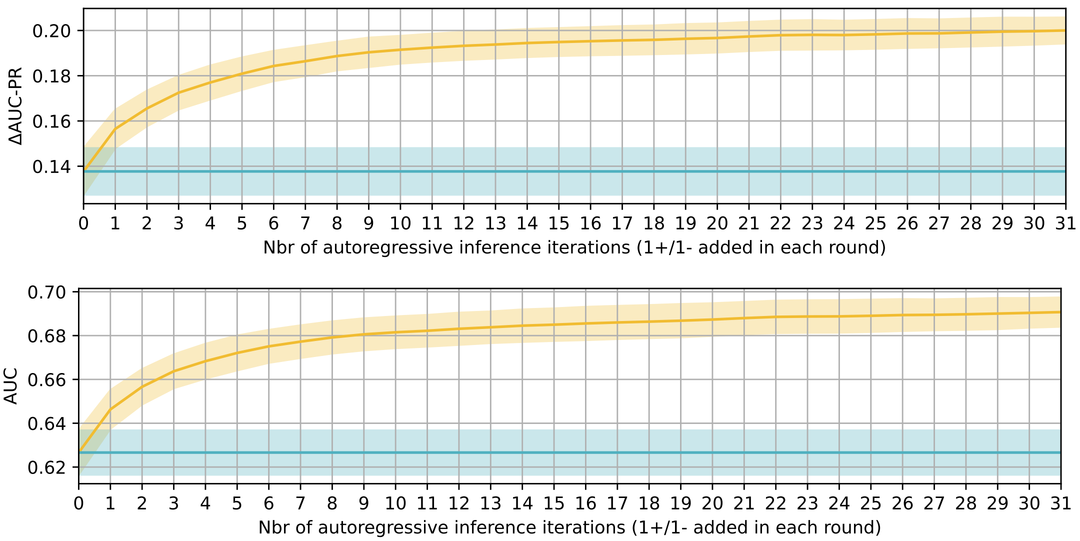

Figure 2: Results of autoregressive inference experiment. The model performance for autoregressive inference mode (yellow) and naive baseline (blue) are shown across inference iterations. The shaded area indicates the standard deviation across experiment reruns.

in few-shot learning.

Few-shot drug discovery. Different types of few-shot approaches have been suggested to few-shot learning in drug-discovery. Some of them build up on meta-learning frameworks (Finn et al., 2017) and use the support set to adapt to the new task within a few update steps (Guo et al., 2021; Wang et al., 2021; Chen et al., 2022). Other models, known as embedding-based methods, compute similarities between query and support set samples and eventually build predictions from a weighted sum over the support set labels (Altae-Tran et al., 2017; Schimunek et al., 2023). Since for embedding-based methods no re-training or fine-tuning (in the sense of a backward pass to adjust parameters) is necessary, these methods are intuitively well suited for an iterative autoregressive inference procedure.

3 Autoregressive activity prediction

We employ autoregressive inference as a strategy to surmount the challenge of the pronounced scarcity of available data. In autoregressive inference mode, a given few-shot model iteratively augments the support set with additional pseudo-labeled samples. Formally, given a set of unlabeled data $\mathcal{U} = \{(\boldsymbol{u}_1, \Box), \ldots, (\boldsymbol{u}_M, \Box)\}$ and a selection policy π, which selects samples to be added to the support set based on the model predictions:

$$\pi\big(\{(\boldsymbol{u}_1, \hat{y}_1), \ldots, (\boldsymbol{u}_M, \hat{y}_M)\}\big) = (\boldsymbol{u}_i, \hat{y}_i),$$

Table 1: Autoregressive inference results on FS-Mol. The first column gives the inference mode, either inductive ("Ind.") or autoregressive ("AR-") of the APM. The backbone of the APM is given in column two. The columns "n+/n−" show the model performance with n active pseudo-labeled and n inactive pseudo-labeled samples added to the support set. Error bars represent the standard deviation across ten experiment reruns. The metrics are averaged across tasks.

Inf. Mode	Backbone	ΔAUC-PR			AUC		
		1+/1−	2+/2−	8+/8−	1+/1−	2+/2−	8+/8−
Ind. APM	MHNfs	$.138^{\pm.010}$	$.138^{\pm.010}$	$.138^{\pm.010}$	$.623^{\pm.010}$	$.623^{\pm.010}$	$.623^{\pm.010}$
AR-APM	MHNfs	$.156^{\pm.009}$	$.177^{\pm.008}$	$.189^{\pm.006}$	$.646^{\pm.009}$	$.657^{\pm.008}$	$.679^{\pm.007}$
Performance Gain		$.019^{\pm.003}$	$.028^{\pm.004}$	$.051^{\pm.008}$	$.020^{\pm.003}$	$.030^{\pm.005}$	$.053^{\pm.007}$

the autoregressive inference procedure is performed in iterations to augment the support set and eventually improve the model performance:

$$\begin{aligned}
(\boldsymbol{u}_1, \hat{y}_1) &= \pi\Big(\big\{(\boldsymbol{u}, \hat{y}) \mid \hat{y} = g\big(U_1; \mathcal{S}\big), \boldsymbol{u} \in U_1\big\}\Big)\\
(\boldsymbol{u}_2, \hat{y}_2) &= \pi\Big(\big\{(\boldsymbol{u}, \hat{y}) \mid \hat{y} = g\big(U_2; \mathcal{S} \cup \{(\boldsymbol{u}_1, \hat{y}_1)\}\big), \boldsymbol{u} \in U_2\big\}\Big)\\
&\dots\\
(\boldsymbol{u}_n, \hat{y}_n) &= \pi\Big(\big\{(\boldsymbol{u}, \hat{y}) \mid \hat{y} = g\big(U_{n-1}; \mathcal{S} \cup \{(\boldsymbol{u}_i, \hat{y}_i)\}_{i=1}^{n-1}\big), \boldsymbol{u} \in U_{n-1}\big\}\Big),
\end{aligned}$$

where $U_i = (\boldsymbol{u}_i, \boldsymbol{u}_{i+1}, \dots \boldsymbol{u}_M)$. Here, we used $g(\cdot; \mathcal{S})$ as a shorthand for $g\left(\cdot; \mathcal{A}(\boldsymbol{w}, \mathcal{S} \cup \mathcal{U}, \mathcal{D}_{\text{train}})\right)$. Also, for simplicity, we assumed the selection policy selects the unlabeled elements sequentially. Note that active learning is similar, but instead of adding the datapoint with the pseudo-label $(\boldsymbol{u}_i, \hat{y}_i)$, the datapoint with the correct label $(\boldsymbol{u}_i, y_i)$ is added to the training or support set.

We choose MHNfs (Schimunek et al., 2023) as the backbone few-shot model since a) MHNfs has already proven to be SOTA on the FS-Mol benchmark experiment, b) MHNfs is an embedding-based method and thus does not require any backward passes to adapt to changing support sets, and c) already includes the idea of feature manifold smoothing (Rodríguez et al., 2020; Hu et al., 2021a). As a selection policy, we choose the candidate with the highest (lowest) few-shot model prediction to be added to the support for the active (inactive) class.

4 Experiments

Data. Recently Stanley et al. (2021) proposed the FS-Mol dataset to benchmark few-shot models. Extracted from ChEMBL27 (Mendez et al., 2019), it consists 5,125 separate assays, 233,786 compounds and 489,133 measurements. The tasks are well-balanced by design which means that the mean ratio of active and inactive molecules per task is 1. The authors provide a training (4,938 tasks), validation (40 tasks), and test split (157 tasks), whereas

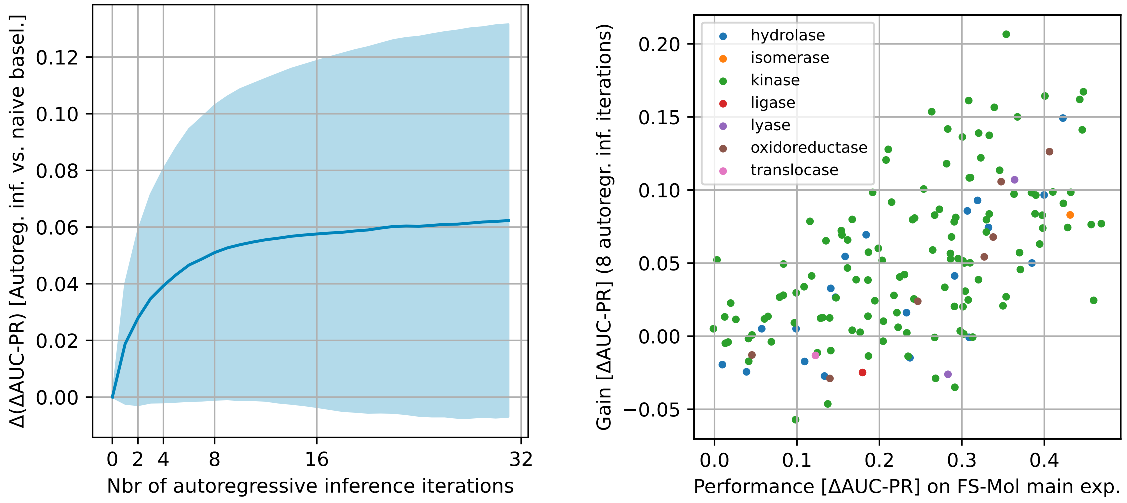

Figure 3: Results of autoregressive inference experiment. Left: The model performance within the autoregressive inference loop (mean over experiment reruns) is shown. The shaded area reports the standard deviation across tasks. Right: For the backbone few-shot model, task-wise the performance value evaluated on the FS-Mol benchmark experiment (Schimunek et al., 2023) is associated with the performance gain in the autoregressive inference procedure.

training, validation, and test tasks build disjoint sets (Stanley et al., 2021). We use the FS-Mol test set data for benchmarking performance gains of autoregressive activity prediction.

Methods compared. We compare MHNfs run in autoregressive inference mode ”AR-APM (MHNfs)” with a naive baseline ”Inductive APM (MHNfs)”. The naive baseline is the MHNfs model which predicts the evaluation set only once just being aware of the initial support set, consisting of one active and one inactive sample. Notably, the naive baseline is the model proposed by Schimunek et al. (2023). Therefore, it already was compared to other competitors in the FS-Mol benchmark experiment.

We run experiments in two slightly different experimental setups, which are a) semi-supervised learning with fixed evaluation set, and b) a transductive learning setting (see Appendix A.3).

Semi-supervised learning setting with fixed evaluation set

This experiment evaluates whether the predictive power of the few-shot classifier increases in autoregressive inference mode. For undistorted performance evaluation we fix the evaluation set which means that samples for performance evaluation and potential support set candidates come from different sets. The evaluation set samples are predicted in inductive inference mode, support set candidates are processed in transductive inference mode.

Experimental setup. For each FS-Mol test task, the available data points are split into three sets a) an initial support set, b) a candidate set, and c) an evaluation set. The initial support set consists of 1 active and 1 inactive molecule. It functions as the initial, first support set the model is provided with during the autoregressive inference procedure. The candidate set consists of 32 active, and 32 inactive data points. Candidate set samples are processed in transductive inference mode which means that for all available candidates activity is predicted jointly and eventually a selection policy π decides for the candidates to augment the support set with. The evaluation set includes all other datapoints the FS-Mol test task provides. It is fixed in terms of it does not change during the autoregressive inference iterations. Evaluation set samples are treated independently which means the few-shot model runs in inductive inference mode for these samples. All available datapoints for a task are distributed randomly. For experiment reruns the samples for initial support set and candidate set are drawn with different seeds, the evaluation set is not changed.

Results. The results in terms of ΔAUC-PR and AUC are presented in Table 1 and Figure 2. The standard deviation is reported across ten experiment reruns, i.e. the comparison of autoregressive inference with naive baseline. In the table, the support set augmentation realized with the autoregressive inference procedure causes performance gains up to 0.051 ± 0.008 for the ΔAUC-PR and 0.053 ± 0.007 for the AUC metric. This means the model performance increases from 0.138 ± 0.010 to 0.189 ± 0.006 for the ΔAUC-PR and from 0.623 ± 0.010 to 0.679 ± 0.007 for the AUC metric without neither having changed any model parameter nor having included any new measurements. Despite this generally found performance boost, Figure 3 shows that, in fact, performance gains are highly task dependent and vary a lot. While some correlation between model performance on a specific task and potential gains using the autoregressive inference scheme seems present, detailed analysis is up to future work. Notably, the FS-Mol main benchmark experiment was performed with support set size 16, while the size of the initial support set for this experiment is 2.

5 Discussion

Our work has introduced the idea of autoregressive activity prediction modeling, while connecting it to the fields of active learning and transductive learning. Our experiments showed that applying this autoregressive inference mode to MHNfs improved both the AUC and the ΔAUC-PR metric. Generalization to another embedding-based few-shot method failed and requires further exploration A.2. Another evaluation in relation to the selection policy A.1 shows that even randomly selecting candidates for the support set helps to improve the model performance. Still, AR-APM (MHNfs), i.e. MHNfs in autoregressive inference mode with suggested selection policy, excels.

Broader Impact Statement

Impact on machine learning and related scientific fields. We believe that with the increasing availability of drug discovery and further improved biotechnologies, the drug discovery process will be made more efficient. Our approach might bridge the gap for scenarios in which data is very scarse.
Impact on society. Should this method prove effective, it could contribute to a faster and more cost-efficient drug discovery process. The COVID-19 pandemic underscored the importance of accelerating the drug discovery timeline to few years or even months. We hope that this work contributes to this effort and eventually leads to safer drugs developed faster.
Consequences of failures of the method. As is usually the case with machine learning techniques, there is a risk that users rely to much on our new approach without reflecting on the outcomes. Human beings would not directly be affected by failure modes since wrong model predictions would rather lead to failed wet-lab and in-vitro experiments than to harmful therapies.
Leveraging of biases in the data and potential discrimination. As for almost all machine learning methods, confounding factors, lab or batch effects, could be used for classification. This might lead to biases in predictions or uneven predictive performance across different drug targets or bioassays.

Acknowledgments and Disclosure of Funding

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids (FFG- 899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo), Software Competence Center Hagenberg GmbH, Borealis AG, TÜV Austria, Frauscher Sensonic, TRUMPF and the NVIDIA Corporation.

We thank Edoardo Cetin and Pieter-Jan Hoedt for their valuable input and help.

References

Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

Nathan Brown, Peter Ertl, Richard Lewis, Torsten Luksch, Daniel Reker, and Nadine Schneider. Artificial intelligence in chemistry and drug design. *Journal of Computer-Aided Molecular Design*, 34:709–715, 2020a.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020b.

Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887*, 2024.

Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395, 2023.

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

Darren VS Green. Using machine learning to inform decisions in drug discovery: an industry perspective. In *Machine learning in chemistry: data-driven algorithms, learning systems, and predictions*, pages 81–101. ACS Publications, 2019.

Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021*, pages 2559–2567, 2021.

Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Graph-based interpolation of feature vectors for accurate few-shot classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8164–8171. IEEE, 2021a.

Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*, pages 487–499. Springer, 2021b.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079, 2019.

Michalis Lazarou, Tania Stathaki, and Yannis Avrithis. Iterative label cleaning for transductive and semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8751–8760, 2021.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24): 5441–5451, 2018.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47 (D1):D930–D940, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 121–138. Springer, 2020.

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

Johannes Schimunek, Philipp Seidl, Katarina Elez, Tim Hempel, Tuan Le, Frank Noé, Simon Olsson, Lluís Raich, Robin Winter, Hatice Gokcan, Filipp Gusev, Evgeny M. Gutkin, Olexandr Isayev, Maria G. Kurnikova, Chamali H. Narangoda, Roman Zubatyuk, Ivan P. Bosko, Konstantin V. Furs, Anna D. Karpenko, Yury V. Kornoushenko, Mikita Shuldau, Artsemi Yushkevich, Mohammed B. Benabderrahmane, Patrick Bousquet-Melou, Ronan Bureau, Beatrice Charton, Bertrand C. Cirou, Gérard Gil, William J. Allen, Suman Sirimulla, Stanley Watowich, Nick Antonopoulos, Nikolaos Epitropakis, Agamemnon Krasoulis, Vassilis Pitsikalis, Stavros Theodorakis, Igor Kozlovskii, Anton Maliutin, Alexander Medvedev, Petr Popov, Mark Zaretckii, Hamid Eghbal-Zadeh, Christina Halmich, Sepp Hochreiter, Andreas Mayr, Peter Ruch, Michael Widrich, Francois Berenger, Ashutosh Kumar, Yoshihiro Yamanishi, Kam Y. J. Zhang, Emmanuel Bengio, Yoshua Bengio, Moksh J. Jain, Maksym Korablyov, Cheng-Hao Liu, Gilles Marcou, Enrico Glaab, Kelly Barnsley, Suhasini M. Iyengar, Mary Jo Ondrechen, V. Joachim

Haupt, Florian Kaiser, Michael Schroeder, Luisa Pugliese, Simone Albani, Christina Athanasiou, Andrea Beccari, Paolo Carloni, Giulia D'Arrigo, Eleonora Gianquinto, Jonas Goßen, Anton Hanke, Benjamin P. Joseph, Daria B. Kokh, Sandra Kovachka, Candida Manelfi, Goutam Mukherjee, Abraham Muñiz-Chicharro, Francesco Musiani, Ariane Nunes-Alves, Giulia Paiardi, Giulia Rossetti, S. Kashif Sadiq, Francesca Spyrakis, Carmine Talarico, Alexandros Tsengenes, Rebecca C. Wade, Conner Copeland, Jeremiah Gaiser, Daniel R. Olson, Amitava Roy, Vishwesh Venkatraman, Travis J. Wheeler, Haribabu Arthanari, Klara Blaschitz, Marco Cespugli, Vedat Durmaz, Konstantin Fackeldey, Patrick D. Fischer, Christoph Gorgulla, Christian Gruber, Karl Gruber, Michael Hetmann, Jamie E. Kinney, Krishna M. Padmanabha Das, Shreya Pandita, Amit Singh, Georg Steinkellner, Guilhem Tesseyre, Gerhard Wagner, Zi-Fu Wang, Ryan J. Yust, Dmitry S. Druzhilovskiy, Dmitry A. Filimonov, Pavel V. Pogodin, Vladimir Poroikov, Anastassia V. Rudik, Leonid A. Stolbov, Alexander V. Veselovsky, Maria De Rosa, Giada De Simone, Maria R. Gulotta, Jessica Lombino, Nedra Mekni, Ugo Perricone, Arturo Casini, Amanda Embree, D. Benjamin Gordon, David Lei, Katelin Pratt, Christopher A. Voigt, Kuang-Yu Chen, Yves Jacob, Tim Krischuns, Pierre Lafaye, Agnès Zettor, M. Luis Rodríguez, Kris M. White, Daren Fearon, Frank Von Delft, Martin A. Walsh, Dragos Horvath, Charles L. Brooks III, Babak Falsafi, Bryan Ford, Adolfo García-Sastre, Sang Yup Lee, Nadia Naffakh, Alexandre Varnek, Günter Klambauer, and Thomas M. Hermans. A community effort in sars-cov-2 drug discovery. *Molecular Informatics*, 43(1):e202300262, 2024. doi: https://doi.org/10.1002/minf.202300262. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202300262.

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

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

The COVID Moonshot Consortium, John Chodera, Alpha Lee, Nir London, and Frank von Delft. Open science discovery of oral non-covalent sars-cov-2 main protease inhibitors.

Christian Tyrchan, Eva Nittinger, Dea Gogishvili, Atanas Patronov, and Thierry Kogej. Approaches using ai in medicinal chemistry. In *Computational and data-driven chemistry using artificial intelligence*, pages 111–159. Elsevier, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andrea Volkamer, Sereina Riniker, Eva Nittinger, Jessica Lanini, Francesca Grisoni, Emma Evertsson, Raquel Rodríguez-Pérez, and Nadine Schneider. Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056, 2023.

Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. *Advances in Neural Information Processing Systems*, 34:17441–17454, 2021.

Peter Whittle. *Hypothesis testing in time series analysis.* PhD thesis, Uppsala University, 1951.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

George Udny Yule. Vii. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646):267–298, 1927.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

Hao Zhu and Piotr Koniusz. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23996–24006, 2023.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *ProQuest Number: INFORMATION TO ALL USERS*, 2002.

Appendix A.

A.1 Details on semi-supervised learning experiment with fixed evaluation set

Random sampling baseline for the semi-supervised learning setting

As a additional baseline we define a ”random selection policy” which ignores the MHNfs predictions. In every iteration, randomly one candidate is chosen to be added to the support set with an active pseudo label and one candidate is chosen to be added with an inactive pseudo label neglecting both the true labels and the few-shot model predictions. Since the candidate set is balanced, this leads to adding candidates with wrong pseudo-label in 50 % of the cases.

Figure A1 shows that running MHNfs in autoregressive inference mode with this random selection policy already improves the model performance in comparison to the naive baseline. We speculate this is due to three reasons a) the data points the experiment is based on were curated by chemists and therefore already include some sort of inductive bias, b) MHNfs, which originally was trained to behalf well for support set sizes around 16, might generally perform better for larger support set sizes, and c) adding noisy data could even help, similar to retrieval-augmented generation systems for which including irrelevant documents can unexpectedly enhance performance (Cuconasu et al., 2024).

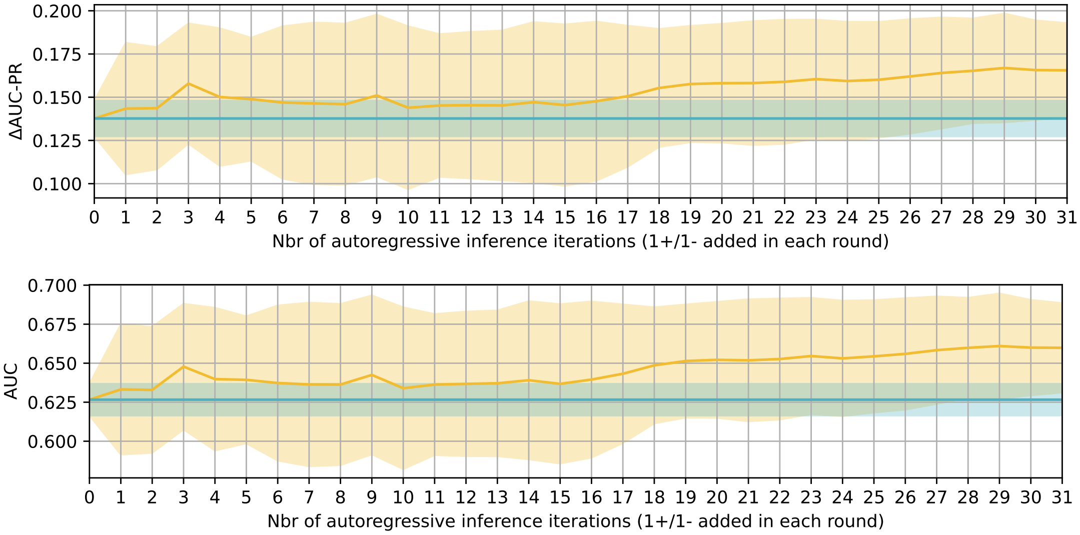

Figure A1: Results of the autoregressive inference experiment with the ”random selection policy”. The model performance for autoregressive inference mode (yellow) and naive baseline (blue) are shown across inference iterations. The shaded area indicates the standard deviation across experiment reruns.

Table A2: Extended autoregressive inference results on FS-Mol. The columns "n+/n−" show the model performance with n active pseudo-labeled and n inactive pseudo-labeled samples added to the support set. Error bars represent the standard deviation across ten experiment reruns. The metrics are averaged across tasks.

Inf. Mode	Backbone	ΔAUC-PR			AUC		
		1+/1−	2+/2−	8+/8−	1+/1−	2+/2−	8+/8−
Ind. APM	ProtoNet	$.138^{\pm.008}$	$.138^{\pm.008}$	$.138^{\pm.008}$	$.617^{\pm.011}$	$.617^{\pm.011}$	$.617^{\pm.011}$
Ind. APM	MHNfs	$.138^{\pm.010}$	$.138^{\pm.010}$	$.138^{\pm.010}$	$.623^{\pm.010}$	$.623^{\pm.010}$	$.623^{\pm.010}$
AR-APM	ProtoNet	$.138^{\pm.010}$	$.138^{\pm.011}$	$.135^{\pm.013}$	$.614^{\pm.011}$	$.615^{\pm.014}$	$.609^{\pm.016}$
AR-APM	MHNfs	$.156^{\pm.009}$	$.177^{\pm.008}$	$.189^{\pm.006}$	$.646^{\pm.009}$	$.657^{\pm.008}$	$.679^{\pm.007}$
AR-APM (random π)	MHNfs	$.143^{\pm.038}$	$.144^{\pm.035}$	$.146^{\pm.047}$	$.633^{\pm.042}$	$.633^{\pm.040}$	$.636^{\pm.052}$

A.2 Generalization to different backbone few-shot model

For this evaluation, we replace the MHNfs backbone model with a Prototypical-Network based few-shot model (Snell et al., 2017). As a similarity measure dot-product distance is used. The model was trained on the FS-Mol training set. Table A1 reports the performance on the FS-Mol main benchmark experiment.

Table A1: Model performance comparison on the FS-Mol benchmark experiment. Standard deviation is reported across tasks.

Model	ΔAUC-PR	AUC
Prototypical Networks	0.218 ± 0.135	0.719 ± 0.131
MHNfs	0.241 ± 0.119	0.739 ± 0.114

Surprisingly, this ProtoNet implementation in autoregressive mode performs worse than its naive baseline (see Figure A2, and A2), which needs further exploration.

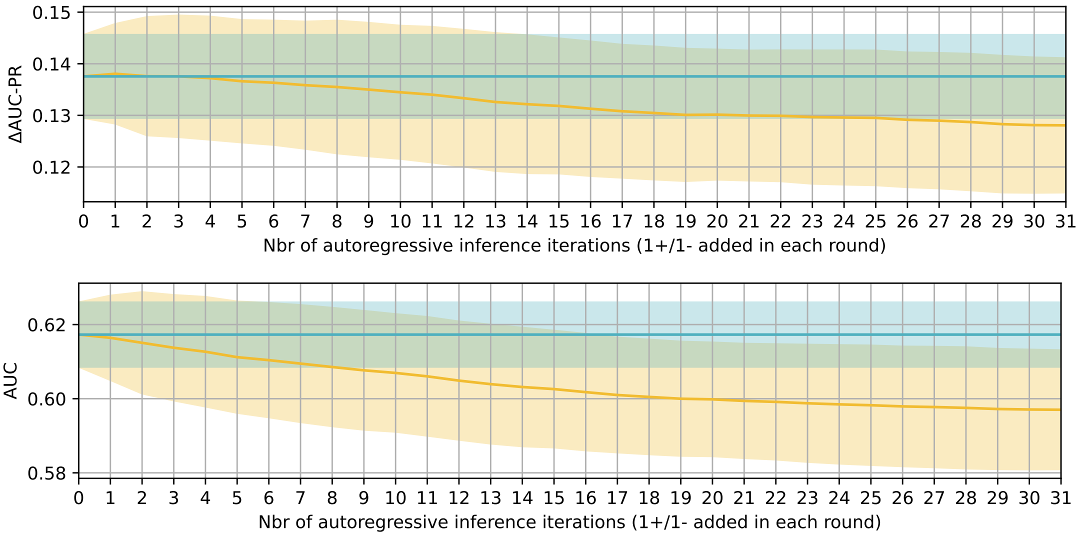

Figure A2: Prototypical Network results of the autoregressive inference experiment. The model performance for autoregressive inference mode (yellow) and naive baseline (blue) are shown across inference iterations. The shaded area indicates the standard deviation across experiment reruns.

A.3 Transductive learning experiment

This setting mimics virtual screenings in which pseudo labels for some of the queried molecules might boost the prediction for others.

Experimental setup. For each FS-Mol test task, the available data points are split into two sets which are an initial support set (analogous to 4) and a query set. The query set takes over both the role of the candidate set and the evaluation set. Notably this setting might be closer to real-world applications but reported performance values are based on test sets in which samples are not i.i.d..

Results. The results in terms of ΔAUC-PR and AUC are presented in Table A3. The standard deviation is reported across ten experiment reruns. The table shows that the support set augmentation realized with the autoregressive inference procedure causes performance gains up to 0.040 ± 0.005 for the ΔAUC-PR and 0.041 ± 0.006 for the AUC metric. This means the model performance increases from 0.130 ± 0.010 to 0.169 ± 0.010 for the ΔAUC-PR and from 0.624 ± 0.011 to 0.665 ± 0.011 for the AUC metric without neither having changed any model parameter nor having included any new measurements.

Table A3: Transductive inference results on FS-Mol. The columns ”n+/n−” show the model performance with n active pseudo-labeled and n inactive pseudo-labeled samples added to the support set. Error bars represent the standard deviation across ten experiment reruns. The metrics are averaged across tasks.

Inf. Mode	Backbone	ΔAUC-PR			AUC		
		1+/1−	2+/2−	8+/8−	1+/1−	2+/2−	8+/8−
Ind. APM	MHNfs	$.130^{\pm.010}$	$.130^{\pm.010}$	$.130^{\pm.010}$	$.624^{\pm.011}$	$.624^{\pm.011}$	$.624^{\pm.011}$
AR-APM	MHNfs	$.146^{\pm.010}$	$.154^{\pm.010}$	$.169^{\pm.010}$	$.641^{\pm.011}$	$.650^{\pm.011}$	$.665^{\pm.011}$
Performance Gain		$.016^{\pm.002}$	$.024^{\pm.003}$	$.040^{\pm.005}$	$.018^{\pm.002}$	$.026^{\pm.004}$	$.041^{\pm.006}$

A.4 Used performance metrics

In this manuscript results are presented in terms of AUC and ΔAUC-PR.

The AUC metric computes the area under the receiver operating characteristic curve (ROC AUC). AUC values are between 0 and 1 and indicate how well active and inactive test molecule predictions are separable. Random classifier achieve AUC scores around 0.5.

The ΔAUC-PR metric computes the area under the precision recall curve and reports the model performance as the difference from a random classifier. It was used by Stanley et al. (2021) in the FSMol few-shot drug-discovery benchmark experiment.