Empirical Observations on Parameter Scaling in Chemical Language Models

Artur Safrastyan [1] **Hrant Khachatrian** [2]

Abstract

Recent studies have shown that increasing parameter count improves the performance in graph-based architectures for molecular property prediction. In this work, we try to examine whether similar scaling behavior can be observed in autoregressive chemical language models trained on SMILES-based datasets. Experiments were conducted on three LLaMA-based models (170M, 380M, and 1.3B parameters) on diverse downstream tasks, including multi-task ADME regression from Polaris, single-task PXR induction prediction, and large-scale binding classification on a DNA-encoded library dataset (BELKA). The results show that larger models can achieve better performance in some settings, particularly in regression tasks, though this behavior is not consistent across all the benchmarks. The PXR induction task showed consistent performance gain with increased model size. On the other hand, the models showed task-dependent variability on the Polaris benchmark. In the BELKA dataset, increased model size improved performance on the public test set but failed to generalize consistently with the private and more out-of-distribution set. The findings in this study suggest that while scaling benefits can extend to chemical language models, they are highly dependent on both task characteristics and the fine-tuning strategies used. We hope this work motivates further research using larger and more diverse model variants to better understand scaling trends in chemical language models.

1. Introduction

The use of transformer-based architectures (Vaswani et al., 2017) has greatly impacted cheminformatics and molecular machine learning (Ross et al., 2022; Guevorguian et al., 2024; Chithrananda et al., 2020). With the growing popularity of transformer-based models and their increasing size and complexity, the need for a scaling law that dictates model performance based on the number of parameters, dataset size, and computational resources has emerged. While these neural scaling laws have been established in natural language processing (Kaplan et al., 2020; Hoffmann et al., 2022), their extension to chemistry remains an area of active research. Recent empirical studies have shown that graph-based architectures (Kipf & Welling, 2016) demonstrate consistent scaling behavior when applied to molecular data (Sypetkowski et al., 2024).

The most commonly used molecular representation system for these language models is the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988). These models treat atoms and their bonds like words in a sentence, learning context-dependent relationships without relying on traditional molecular fingerprints. There exists a gap in literature regarding whether the established scaling law principles observed in natural language processing and graph neural networks can transfer to an autoregressive chemical model.

This work is an empirical study of autoregressive chemical language models and does not attempt to establish a universal chemical scaling law. The models that have been evaluated across several parameter scales are LLaMA 3-based (Touvron et al., 2023; Grattafiori et al., 2024) family of architectures called Chemlactica (Guevorguian et al., 2024; Fahradyan et al., 2024). The main goal is to examine whether increasing model size improves performance when these pre-trained models are fine-tuned on downstream chemical tasks. The study was focused on the 170M, 380M, and 1.3B parameter variants of Chemlactica, all pre-trained on the same Pubchem corpus of approximately 110 million molecules (Kim et al., 2016).

Within the scope of this research, experiments were conducted over diverse downstream property prediction benchmarks, specifically the Polaris ADME multi-task regression dataset (Polaris Hub, 2024), the OpenADMET PXR induction single-task dataset (OpenADMET, 2024), and the combinatorial BELKA binding classification challenge (Kaggle, 2024). The results presented are not meant to es-

[1]American University of Armenia, Yerevan, Armenia [2]Yerevan State University, Yerevan, Armenia. Correspondence to: Artur Safrastyan <artur_safrastyan@aua.am>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

tablish a universal or strictly monotonic scaling law, nor to suggest that larger molecular models are consistently outperforming in all settings. Instead, this work emphasizes the task-dependent nature of scaling. It analyzes how increasing model size affects performance over various downstream chemical prediction tasks, while experimenting with different fine-tuning strategies.

2. Models And Tasks

A family of autoregressive LLaMA-based chemical language models was used within this study, spanning three parameter scales: 170M, 380M, and 1.3B. For architectural consistency, the larger models both have a fixed depth of 16 transformer layers, with the increased capacity achieved through higher embedding dimensionality. All three variants share a common pretraining foundation and have been trained on a large-scale molecular corpus of approximately 110 million compounds. The data was structured using a template system that uses paired tags to delimit molecular descriptors. Tokenization was handled by the base Llama 3 tokenizer, augmented with chemistry-specific markers to prevent fragmentation of structural descriptors. After including all opening and closing tags as distinct special tokens, the final dataset was compressed into 40 billion tokens.

The empirical research focuses on downstream fine-tuning across three benchmarks to examine how effectively pretrained representations transfer to task-specific settings. Three datasets have been considered, with each introducing distinct challenges.

Table 1. Overview of evaluated LLaMA-based molecular architectures and downstream property prediction tasks.

Model	Params	Layers	Dim.
170M	169.3M	8	768
380M	383.0M	16	1024
1.3B	1.24B	16	2048

Dataset	Task	Train	Test
Polaris ADME	Multi-task Reg.	2.8K	704
PXR Induction	Single Reg.	4.1K	513
BELKA	Binary Class.	∼295M	∼1.67M

The Polaris ADME benchmark is formulated as a multi-task regression problem that requires the model to predict six drug properties related to absorption, distribution, metabolism, and excretion simultaneously. The OpenADMET PXR induction dataset is treated as a single-task regression problem, with the objective of predicting ligand-induced activation of the human Pregnane X Receptor. Both the PXR induction and Polaris ADME regression datasets are rather small, with only 3-4.5 thousand samples.

Finally, the BELKA dataset introduces a large-scale binary classification problem derived from DNA-encoded library screening. The task involves predicting binding interactions between small molecules and protein targets under extreme class imbalance, with positive samples accounting for less than 0.5% of the data. To evaluate true out-of-distribution generalization and prevent data leakage, a scaffold (building block) splitting strategy was implemented for this dataset. A random subset representing 3% of the constituent molecular building blocks was isolated, and any molecule containing these held-out blocks was exclusively included in the validation and test sets. This design forced the model to predict binding affinities for entirely novel combinatorial chemical structures simulating real world drug testing.

3. Methodology

During supervised fine-tuning, the entire pre-trained backbone was initially frozen before selectively unfreezing the last transformer layers to mitigate catastrophic forgetting (Luo et al., 2023). A key technical detail is the unfreezing of the normalization layers, which allow the model's statistics to adapt to a new dataset. Furthermore, all training hyperparameters for these architectures were optimized utilizing Bayesian search strategies. For the Polaris multi-task regression dataset, the final two layers were unfrozen across all variants, leaving approximately 10% of the parameters trainable. A similar proportional scaling approach was used for the large-scale BELKA binding dataset, where 6, 8, and 7 layers were unfrozen for the 170M, 380M, and 1.3B models, respectively, to maintain a constant trainable parameter ratio of 31%-34%. In contrast, for the PXR induction challenge, the number of unfrozen layers was not fixed proportionally, and instead, it was tuned independently for each model variant to discover the optimal regularization configuration for each size. The final numbers of unfrozen layers for the PXR induction regression task were 3, 10, and 10, respectively.

To stabilize fine-tuning, the models used differential learning rates across different parameter groups. The parameters of the pre-trained LLaMA backbone and the added classification head are separated into two groups and passed to the optimizer independently. This allows each part of the network to be updated at a different rate during training.

The learning rate of the backbone was one to three orders of magnitude smaller than the learning rate of the added head layer. With this setup, the classification head can adapt quickly to the downstream task, while the backbone only makes small, controlled updates.

The choice of a pooling mechanism for aggregating the token embeddings was dependent on the downstream task. For the Polaris and PXR regression datasets, global average pooling (mean pooling) was used. The choice of a pooling

strategy was made after testing several other methods and mean pooling that outperformed all of them. For the Polaris regression task, a shared projection head was attached to this pooled representation, consisting of a two-layer multi-layer perceptron (MLP) with an intermediate hidden dimension of 256 and a dropout layer with a probability of 0.10 applied to reduce overfitting. For the BELKA classification task last-token pooling yielded the best results. The hidden state corresponding to the final [END_SMILES] delimiter token was extracted, as it is the only token containing the full context of the entire autoregressive molecular sequence.

The models were trained using the AdamW optimizer and a learning rate schedule with an initial linear warmup followed by cosine annealing was also used.

The BELKA dataset introduces an additional challenge with its extreme class imbalance. To address this, standard binary cross-entropy was replaced with Focal Loss combined with label smoothing. γ was set to 2.0. Comprehensive details regarding the empirical hyperparameter search spaces, fixed training configurations, and the specific layer unfreezing strategies utilized across all evaluated benchmarks are provided in Appendices A, B, and C.

4. Results

Full comparative results across all evaluated scales, alongside the established state-of-the-art external benchmarks for each respective dataset, are detailed in Table 2.

In the case of the Polaris multi-task ADME benchmark, a general trend emerged showing that larger models perform better, with the 1.3B parameter variant achieving the highest average performance. Specifically, the largest model yielded an approximately 20% relative improvement in the coefficient of determination (R^2) compared to the 170M baseline. Furthermore, it consistently reduced both Mean Squared Error (MSE) and Mean Absolute Error (MAE) across most individual endpoints.

The single-task PXR induction challenge demonstrated the most consistent and monotonic scaling behavior among the evaluated benchmarks. As model capacity increased from 170M to 1.3B parameters, the predictive performance improved steadily. The R^2 metric increased by approximately 45%, while the MAE and Relative Absolute Error (RAE) decreased by roughly 12% and 11%, respectively.

Evaluation on the highly imbalanced BELKA dataset revealed conflicting trends between the evaluation splits. On the public test set, Average Precision (AP) performance showed a clear positive correlation with model size, ranging from 0.2196 for the 170M model to 0.2549 for the 1.3B model. However, on the private test set, the 380M intermediate-sized model achieved the highest AP of 0.1716, outperforming the 1.3B variant's score of 0.1436.

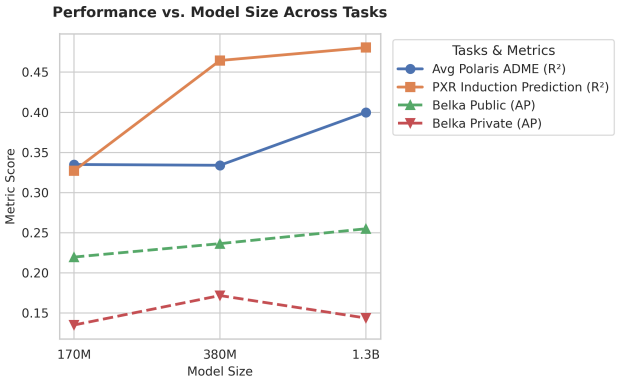

Figure 1. Metric vs Model size across evaluated tasks.

5. Discussion and Limitations

The empirical evaluation demonstrates that increasing the number of model parameters can improve performance for downstream chemical property prediction, though this effectiveness is fundamentally task-dependent. The most consistent scaling behavior was observed in the PXR induction task, while the Polaris benchmark showed improvements that were not strictly uniform across all tasks.

The BELKA binding classification results show limitations in out-of-distribution generalization. Although the 1.3B model achieved the highest performance on the public test set, it was outperformed by the 380M model on the private split, suggesting that severe distribution changes can negate the advantages of increased model capacity.

Although model size appears to be beneficial in certain cases, particularly for regression tasks, the mixed downstream evidence is insufficient to support consistent scaling trend claims for these autoregressive models. Several factors weaken strict causal interpretation. In particular, the need to apply different fine-tuning strategies and hyperparameter configurations across model sizes complicates the isolation of parameter count as the sole driver of performance differences. For example, mean pooling produced the best results for the PXR induction and Polaris tasks, whereas last-token pooling performed better on the BELKA dataset. Additionally, performance for the 1.3B model on BELKA improved when an additional two-layer MLP with a 1024-dimensional hidden layer and GELU activation was appended to the backbone. This suggests that a direct projection from high 2048-dimensional embeddings may create a representational bottleneck.

Another limitation of the research is that, due to the substantial computational cost required for the BELKA dataset,

Table 2. Compact performance results across all downstream tasks. Best results among the evaluated chemical language models are highlighted in bold. The SOTA column represents the current state-of-the-art external benchmarks. Arrows (↓) indicate metrics where lower values represent better performance.

Metric	170M	380M	1.3B	SOTA
Polaris ADME (Average) (Müller et al., 2026)				
R^2	0.335	0.334	**0.400**	0.4760
Pearson	0.601	0.591	**0.638**	0.6982
Spearman ρ	0.583	0.558	**0.613**	0.7034
Explained Variance	0.356	0.338	**0.405**	0.4824
MSE (↓)	0.338	0.337	**0.306**	0.2648
MAE (↓)	0.444	0.448	**0.416**	0.3863
PXR Induction (DiscoveryBytes, 2024)				
MAE (↓)	0.5949	0.5408	**0.5261**	0.4268
RAE (↓)	0.7468	0.6791	**0.6606**	0.5633
R^2	0.3272	0.4646	**0.4807**	0.6033
Spearman ρ	0.6594	0.7115	**0.7304**	0.8317
BELKA Binding Classification (Shlepov, 2026)				
Public AP	0.2196	0.2364	**0.2549**	0.41886
Private AP	0.1349	**0.1716**	0.1436	0.30619

multi-epoch training regimes were not feasible, possibly hindering true evaluation. Additionally, all experiments use only a single random initialization seed. As a result, the variability associated with training could not be systematically evaluated, which makes it difficult to determine whether the observed scaling behaviors can be reproduced across different runs. The absence of confidence intervals also limits the ability to draw strong statistical conclusions about the true nature of the improvements associated with scaling autoregressive chemical language models. Future work must address these limitations.

Impact Statement

This paper presents work aimed at advancing the fields of molecular machine learning and computational chemistry. By investigating the scaling behavior of autoregressive chemical language models across diverse downstream prediction tasks, this research contributes to a broader understanding of how model capacity, fine-tuning strategies, and dataset characteristics influence molecular representation learning and generalization in cheminformatics applications.

References

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

DiscoveryBytes. Openadmet pxr blind challenge: Activity prediction, 2024. URL https://github.com/discoverybytes/openadmet-pxr-blind-challenge/tree/main/activity-prediction.

Fahradyan, T., Geikyan, F., Guevorguian, P., and Khachatrian, H. Towards scaling laws for language model powered evolutionary algorithms: Case study on molecular optimization, 2024. URL https://openreview.net/forum?id=jVAeX6dgDI.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guevorguian, P., Bedrosian, M., Fahradyan, T., Chilingaryan, G., Khachatrian, H., and Aghajanyan, A. Small molecule optimization with large language models. *arXiv preprint arXiv:2407.18897*, 2024.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Kaggle. Leash-belka competition leaderboard, 2024. URL https://www.kaggle.com/competitions/leash-BELKA/leaderboard.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al.

Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Müller, A. T., Atz, K., Reutlinger, M., and Zorn, N. Combining graph attention and recurrent neural networks in a variational autoencoder for molecular representation learning and drug design, 2026. URL https://openreview.net/forum?id=7WYcOGds6R.

OpenADMET. Openadmet pxr challenge train/test dataset, 2024. URL https://huggingface.co/datasets/openadmet/pxr-challenge-train-test.

Polaris Hub. Polaris hub: Adme-fang-r-1, 2024. URL https://polarishub.io/benchmarks/polaris/adme-fang-r-1.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Shlepov, V. 1st place solution updated, 2026. URL https://www.kaggle.com/competitions/leash-BELKA/writeups/victor-shlepov-1st-place-solution-updated.

Sypetkowski, M., Wenkel, F., Poursafaei, F., Dickson, N., Suri, K., Fradkin, P., and Beaini, D. On the scalability of gnns for molecular graphs. *arXiv preprint arXiv:2404.11568*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

A. Polaris Multi-Task Regression Configurations

This section details the parameters used during the multi-task optimization phase on the Polaris ADME regression benchmark. Table 3 defines the shared hyperparameter values applied uniformly across all evaluated scales, while Table 4 records the specific learning rates discovered via the independent optimization grids for each distinct architecture.

Table 3. Fixed training configuration for the Polaris multi-task regression benchmark.

Hyperparameter	**Value**
Batch size	32
Dropout	0.10
Hidden size (MLP)	256
Unfrozen layers	2
Warmup ratio	0.10
Weight decay	0.001

Table 4. Optimized learning rates per model size for the Polaris multi-task regression benchmark.

Model Size	**Learning Rate**
170M	1.05×10^{-3}
380M	2.02×10^{-3}
1.3B	8.20×10^{-5}

B. PXR Induction Task Configurations

This section catalogs the hyperparameter profiles assigned to the single-task PXR induction problem. The broad search spaces mapped out for the optimization runs are summarized in Table 5, and the final configurations chosen are recorded in Table 6.

Table 5. Hyperparameter search space for the PXR induction task across model variants.

Hyperparameter	**170M**	**380M**	**1.3B**
Learning rate	$[10^{-5}, 10^{-2}]$	$[2 \times 10^{-5}, 5 \times 10^{-4}]$	$[1 \times 10^{-6}, 5 \times 10^{-5}]$
Batch size	$\{32, 64\}$	$\{32, 64\}$	$\{32, 64\}$
Dropout	$\{0.1, 0.2, 0.3, 0.4\}$	$\{0.2, 0.25, 0.3, 0.35, 0.4\}$	$\{0.25, 0.3, 0.35, 0.4\}$
Warmup ratio	$\{0.1, 0.2, 0.3, 0.4, 0.5\}$	$\{0.1, 0.2, 0.3, 0.4, 0.5\}$	$\{0.1, 0.2, 0.3, 0.4, 0.5\}$
Weight decay	$\{0.01, 0.05, 0.1, 0.15, 0.2\}$	$\{0.05, 0.1, 0.15, 0.2, 0.25\}$	$\{0.1, 0.15, 0.2, 0.3\}$
Hidden size	$\{256, 512\}$	$\{256, 512\}$	$\{256, 512, 1024\}$
Unfrozen layers	$\{1, \ldots, 8\}$	$\{2, \ldots, 16\}$	$\{2, \ldots, 16\}$

Table 6. Chosen hyperparameters for the PXR induction task per model size.

Parameter	**170M**	**380M**	**1.3B**
Batch size	64	64	32
Dropout	0.20	0.25	0.30
Hidden size	256	512	1024
Learning rate	1.33×10^{-4}	9.53×10^{-5}	3.20×10^{-5}
Unfrozen layers	3	10	10
Warmup ratio	0.30	0.50	0.50
Weight decay	0.15	0.25	0.10

C. BELKA Binding Classification Configurations

This section outlines the training setup used for the BELKA dataset. Table 7 details the specific hyperparameters, loss function configurations, and layer unfreezing strategies applied to each model size.

Table 7. Final hyperparameter configurations for the BELKA binding classification experiments.

Hyperparameter	**170M**	**380M**	**1.3B**
Backbone learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
Head learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch size	64	64	64
Epochs	1	1	1
Focal loss γ	2	2	2
Dropout	0.25	0.25	0.10
Weight decay	0.05	0.10	0.001
Warmup ratio	0.10	0.05	0.35
Unfrozen layers	6	8	7
MLP hidden size	–	–	1024