Uncovering Neural Scaling Law in Molecular Representation Learning

Dingshuo Chen [*1] **Yanqiao Zhu** [*2] **Jieyu Zhang** [*3] **Yuanqi Du** [4]
Zhixun Li [5] **Qiang Liu** [1] **Shu Wu** [1] **Liang Wang** [1]

Abstract

Molecular Representation Learning (MRL) has demonstrated great potential in a variety of tasks such as virtual screening for drug and materials discovery. Despite the widespread interests in advancing model-centric techniques, how the quantity and quality of molecular data affect the learned representations remains an open question in this field. In this paper, we investigate the neural scaling behaviors of MRL from a data-centric perspective across four dimensions, including (1) data modality, (2) data distribution, (3) pre-training involvement, and (4) model capacity. Our empirical studies confirm that the performance of MRL exhibits a power-law relationship with data quantity across the aforementioned four dimensions. Moreover, our fine-grained analysis uncovers potential angles that can be explored to improve the learning efficiency. To seek the possibility to beat the scaling law, we adapt seven popular data pruning strategies to molecular data and benchmark their performance. Drawing from our experimental findings, we underscore the importance of data-centric MRL and discuss their potential for future research.

1. Introduction

The research enthusiasm for Molecular Representation Learning (MRL) is steadily increasing, attributed to its potential in expediting the drug and materials discovery process compared with conventional *in vitro* and *in vivo* experiments(Liu et al., 2017; Shen & Nicolaou, 2019; Wieder et al., 2020). Given a specific featurization (modality) of molecules, the goal of MRL is to learn a continuous vector representation encapsulating rich chemical semantics and possessing high expressiveness to solve downstream tasks(Xu et al., 2018; Gilmer et al., 2017; Rogers & Hahn, 2010; Corso et al., 2020; Li et al., 2020; Beaini et al., 2021; Bodnar et al., 2021; Ying et al., 2021; Yang et al., 2021; Bouritsas et al., 2022; Yu & Gao, 2022).

A trend in the field has been developing neural architectures and training strategies to improve the expressiveness of the learned representations. However, the influence of varying data scales on the performance of MRL under different circumstances is yet to be fully understood. To this end, we draw attention to the following questions: *What is the neural scaling behavior of molecular representation learning? Does it agree with previously discovered scaling laws (such as power-law) in other domains?* To the best of our knowledge, our study is the first to approach MRL from a data-centric perspective, making preliminary strikes in this particular aspect. Through this work, we aim to provide valuable insights that will be instrumental in shaping future explorations in the field.

Beyond common research objects in neural scaling law research such as the impact of pre-training and model parameter size, MRL encounters unique data-oriented challenges including modality selection(Morgan, 1965; Weininger, 1988) and out-of-distribution shift(Hu et al., 2019). In order to conduct a comprehensive study, we investigate the impact of various dimensions on MRL from a data-centric perspective. In particular, we leverage the neural scaling law as a cornerstone, from which we have identified several important scientific questions and systematically explored them as follow:

What kind of scaling law is demonstrated between performance and data quantity? We conduct extensive experiments on four large-scale molecular property prediction datasets. These datasets contain classification and regression tasks, both in single-task and multi-task settings, focusing on properties ranging from quantum mechanical properties to macroscopic influences on human body. The experimental

[*]Equal contribution [1]Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing, China [2]Department of Computer Science, University of California, Los Angeles, United States [3]The Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, United States [4]Department of Computer Science, Cornell University, Ithaca, United States [5]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hongkong, China. Correspondence to: Shu Wu <shu.wu@nlpr.ia.ac.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

results indicate that the model performance generally follows a power-law relationship with data quantity. Compared with the neural scaling law in Natural Language Processing (NLP) and Computer Vision (CV) domain, there is no apparent training bottleneck in the low-data and high-data regimes.

How do different molecular modalities influence the scaling law? The selection of appropriate modalities in MRL has always been an open question. In our investigation, we specifically choose three commonly used modalities (graphs, SMILES strings(Weininger, 1988), and Morgan Fingerprints(Morgan, 1965)). Our experiments, conducted on three classification tasks, aim to discern the effect of these modalities on the performance of MRL. We find that different modalities exhibit distinct learning behaviors in MRL. The graph modality is identified as the most efficient choice for MRL, exhibiting the largest power-law exponent, while fingerprint delivers competitive results. In comparison, the SMILES modality demonstrates a low cost-efficiency with the same data increment.

Is positive transfer consistently observed in scaling law with a pre-trained model? Prior studies generally suggest that molecular pre-training can consistently provide positive performance transfer to downstream tasks. We cast doubt upon this conclusion and investigate the impact of graph-based pre-training on the learning behaviors in downstream tasks. Experimental results demonstrate that in the low-data regime, pre-training indeed leads to beneficial improvements. However, the power-law exponent of pre-training is smaller than that of training from scratch. We suppose that pre-training only provides stable gains when the downstream dataset is small. As the dataset scales up, this positive gains seem to diminish and potentially even leading to negative transfer in the high-data regime.

What influence does data distribution exert on the scaling law? MoleculeNet (Wu et al., 2018) offers scaffold split settings that cater to the practical needs of drug development. Compared to the uniform distribution of random split, scaffold split challenges the model capability for complete out-of-distribution extrapolation. Building upon this, we propose an imbalanced split, which better aligns with real-world requirements compared to existing split settings. Experimental results reveal that random split exhibits the highest power-law exponent, while both imbalanced split and scaffold split demonstrate significantly lower learning efficiency compared to the uniform distribution. These findings underscores that the variations in data distribution can significantly influence the learning behaviors and that imbalanced and scaffold distributions present heightened challenges for MRL.

How does the model capacity affect the scaling law? The size of model parameters stands as another crucial factor impacting the performance of MRL models. Within this context, we concentrate on the widely adopted Graph Isomorphism Network (GIN)(Xu et al., 2018), examining the impact of varying parameter sizes on scaling laws. In general, the power law relationship between model performance and data scale appears consistent, irrespective of the size of model parameters. However, the model capacity does affect training efficiency. Interestingly, there is no apparent relationship between training efficiency and dataset scale across different tasks. For instance, the GIN model reaches its performance peak with a small model capacity on the MUV dataset of a moderate data scale.

Can a curated subset from full dataset yield comparable or even superior results? In the field of CV, the utility of data pruning has been explored due to the computational burden imposed by increasingly large models and massive amounts of data. In the MRL domain, however, the existence of data redundancy and the potential of pruning strategies to alleviate computational burden remain largely unexplored. To address this gap, we benchmark seven data pruning strategies originally proposed for image data on three classification tasks and adapt them to the MRL domain. The results show that existing data pruning methods do not significantly outperform random selection, which highlights the need for the development of data pruning strategies specifically tailored to molecular data.

2. Experiment Setup

2.1. Overall workflow

We follow a consistent research procedure for the six scientific questions under investigation. All experiments are conducted based on the neural scaling law between model performance and data quantity, with the aim of exploring the influence of different factors on the scaling law. Specifically, we divide the complete dataset into nine proportional subsets: $[1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 60\%, 80\%, 100\%]$, and for each ratio, we randomly select five seeds and report the mean result to evaluate the performance. Subsequently, we employ the least squares method to estimate the parameters and fit the performance variation curve, which is visually presented through plotted graphs.

2.2. Datasets and tasks

To reveal the general trends in the neural scaling law of MRL, we opt for datasets from MoleculeNet (Wu et al., 2018) considering three main perspectives: *task type* (classification and regression), *task setting* (single-task and multi-task settings) and *property category* (biophysics and quantum mechanics). Given the potential issues of over-fitting and spurious correlations that may arise when using limited samples, we focus on relatively large-scale datasets (with a

minimum of 40K molecules) for empirical analysis. Please relegate more details about datasets and tasks to Appendix B

2.3. Modalities

Molecular modalities translate chemical information of molecules into representations that can be understood by machine learning algorithms. Concretely, we consider the following four molecular modalities which are widely adopted in the field: (1) **2D topology graph** models atoms and bonds as nodes and edges respectively; (2) **3D geometry** incorporate Cartesian coordinates of atoms in their representation to depict how atoms are positioned relative to each other in the 3D space; (3) **Morgan fingerprint**(Morgan, 1965) encode molecule into fixed-length bit vector which enables mapping of certain structures of the molecule within certain radius of organic molecule bonds; (4) **SMILES string**(Weininger, 1988) is a concise technique that represents chemical structures in a linear notation using ASCII characters, with explicitly depicting information about atoms, bonds, rings, connectivity, aromaticity, and stereochemistry.

2.4. Model and training details

In the following, we will elaborate on the model and training details used in our experiments. Since our empirical analysis cover multiple dimensions, we will only present the general experimental details here. Specific experimental details for each dimension will be discussed in the corresponding subsections of Section 3. Unless otherwise specified, the experimental settings will remain consistent with the description in this section.

Models. Since our experiments involve four different data modalities, each modality is modeled using its corresponding encoders.

- For the graph modality, we utilize the Graph Isomorphism Network (GIN)(Xu et al., 2018) as the encoder. To ensure the generalizability of our research findings, we adopt the commonly recognized experimental settings proposed by Hu et .al(Hu et al., 2019), with 5 layers, 300 hidden units in each of layer and 50% dropout ratio.
- For the 3D geometry modality, we employ the classical SchNet model(Schütt et al., 2017) as the encoder. In this case, we set the hidden dimension and the number of filters in continuous-filter convolution to 128. The interatomic distances are measured with 50 radial basis functions, and we stack 6 interaction layers in the SchNet architecture.
- For the fingerprint modality, we use RDKit(Landrum et al., 2022) to generate 1024-bit molecular fingerprints with radius R = 2, which is roughly equivalent to the ECFP4 scheme(Rogers & Hahn, 2010). We adopt a one-layer Transformer model(Vaswani et al., 2017) with 8 attention heads for modeling. The bit embedding dimension is set to 64, and the hidden dimension is set to 300.
- For the SMILES modality, we employ the same model architecture as the fingerprint modality to ensure a fair comparison. The difference lies in the dictionary dimension, which will be discussed in Section 3.2.

Training details. We follow the experimental settings proposed by Hu et al.(Hu et al., 2019) and used random split for dataset partitioning. For classification tasks, we adopted an 80%/10%/10% split ratio for training/validation/test set, while for regression tasks, we used a split of 110K/10K/10K. All model parameters are initialized using Glorot initialization(Glorot & Bengio, 2010) and trained using the Adam optimizer(Kingma & Ba, 2014). The batch size for all training processes is set to 256. For classification tasks, we set the learning rate to 0.001 and do not utilize a scheduler. For regression tasks, we follow the original experimental settings of SchNet, setting the learning rate to 5×10^{-4} and employing a cosine annealing scheduler.

Evaluation protocols. For HIV and MUV tasks, we report the performance in terms of the Area Under the ROC-Curve (ROC-AUC), while reporting the Average Precision for PCBA performance, where higher values indicate better performance. For quantum property prediction tasks, we measure the performance in Mean Absolute Error (MAE), where lower values are better.

3. Empirical Results and Observations

In this section, we systematically organize the experimental results pertaining to the aforementioned scientific questions. Firstly, we present the neural scaling law between model performance and data quantity across all datasets. Then we demonstrate the differences in the impacts of different settings across four dimensions: data modality, data distribution, pretraining intervention, and model parameter size. In the last subsection, we delve into the applicability of existing data pruning strategies, originally designed for image data, within the molecular domain.

3.1. General neural scaling law

The early studies of both classical learning theory and neural scaling laws on other domains(Amari et al., 1992; Hestness et al., 2017) tell us that the test performance $L(n)$ increases polynomially with the training data size n shown as follow:

$$L(n) = \delta \cdot n^{\alpha} \tag{1}$$

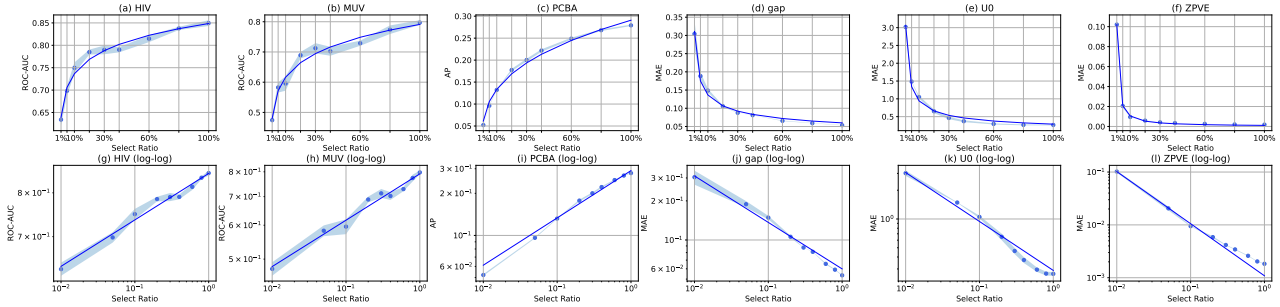

Figure 1. The neural scaling law of molecular representation learning. *Up row:* The effect of scaling up sample size (linear scale) on different tasks. *Bottom row:* The effect of scaling up sample size (log scale) on different tasks.

where δ represents a constant and α represents the exponent of the power law. We investigate whether MRL adheres to this power law relationship. In the selected MRL tasks, both classification and regression tasks are included, covering single- and multi-task scenarios. Starting with the most common setting of graph modality with random split, Figure 1 illustrates the variation in model performance with respect to the size of the data quantity across all datasets. The first row displays the curve changes in linear coordinates, while the second row depicts the curve changes with log scaling. A linear trend in the variation of second-row performance implies compliance with the power law.

Observation (1): It is observed that the model performance on all datasets adheres to the power law relationship as the data quantity varies. Unlike previous findings in the NLP and CV domains(Hestness et al., 2017), there is no obvious performance plateau observed in both small- and high-data regimes, which suggests that the performance of supervised MRL is highly predictable and consistently improves with increasing data quantity. This consistent relationship is also observed with other modalities as shown in Figure 2.

3.2. The effect of modality on the scaling law

The selection of modalities in MRL has been a subject of ongoing debate. The mainstream modalities in MRL currently include graph, SMILES, and molecular fingerprint. However, due to the lack of fair comparisons among these modalities under the same setting, we compare the learning behaviors of different molecular modalities from the perspective of scaling law while controlling other influencing factors as much as possible.

However, due to the variations in widely adopted encoders used for different modalities in the field, there exists a trade-off between expressiveness and uniformity in encoder selection. In other words, using widely-adopted models can reflect the expressive limits of modalities but may introduce unfair comparisons. On the other hand, using the same model may result in under-utilization of modality-specific information. Therefore, we carefully select top-performing models for each modality and ensure a fair model capacity as much as possible: We adopt 5-layer GIN for graph modality and 1-layer transformer for SMILES and fingerprint but with different vocabulary sizes: 2 for fingerprints (binary vector) and 7924 for SMILES (following ChemBERTa(Chithrananda et al., 2020)). Regarding other model details, we adopt the settings as described in Section 2.4 for the corresponding modality. Note that using different numbers of Transformer layers does not affect our conclusion below. Please refer to the Appendix C.2 for the scaling law of Transformer with different numbers of layers.

The results presented in Figure 2 mainly reveal two findings. (1) The graph modality exhibits the superior or near-optimal exponential improvements across all classification tasks, indicating its better learning efficiency and greater potential for gains with the same amount of data increment. (2) The performance of the SMILES modality is consistently worse than the other two modalities, even exhibiting counter-intuitive performance degradation on the `MUV` dataset. We further change the number of model layers and confirm that the performance degradation is a common phenomenon, rather than being attributed to limited model capacity. Figure 6 in Appendix C shows the visualization of the neural scaling law of single-property performance in the MUV task. It demonstrates that performance degradation is evident in most properties, while a few still exhibit stable improvement.

Observation (2): In general, different modalities exhibit distinct learning behaviors in MRL. The graph modality stands out as the most efficient candidate for MRL, while fingerprint also deliver competitive results. The use of SMILES offers the least cost-effectiveness in terms of performance gains. However, it is worth noting that some researches have shown that the language models pre-trained on the large-scale dataset with SMILES modality exhibit remarkable performance in the downstream tasks(Ross et al., 2022).

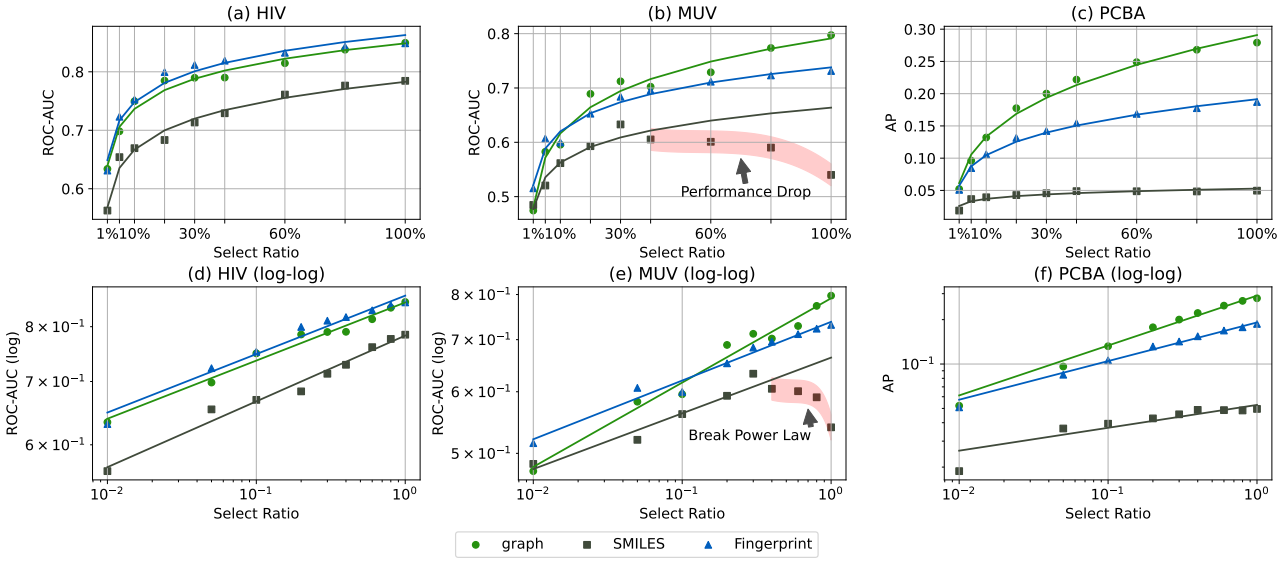

Figure 2. The effect of different molecular modalities. The scatter plot and line chart represent test performance and corresponding fitted curve, respectively.

We will leave this performance gap between pre-trained and train-from-scratch models for future research.

3.3. The effect of pre-training on the scaling law

Molecular pre-training is another important research topic in MRL. Previous studies have generally suggested that pre-training on molecular graph modalities can lead to stable and effective improvements in downstream tasks. However, we raise questions regarding this acknowledged conclusion and investigate how pre-training affects the learning behaviors of MRL, which further facilitates future exploration towards consistent and positive transfer training.

We employ the the pretraining strategy in GraphMAE(Hou et al., 2022), which involved masking the atom type of partial atoms in the molecules, re-masking the encoded atom representations from the backbone model, and eventually using a decoder model to reconstruct the original atom features. The results in Figure 3 illustrate the gains achieved by pre-training on downstream tasks in MRL. In comparison to the model trained from scratch, the pre-trained one still exhibit a power law with data increasing but with a higher intercept and a smaller exponent. Additionally, on the PCBA dataset, a curve intersection occurs at a data scale of 40K, and the pre-trained model's performance noticeably deteriorates compared to the non-pretrained model.

Observation (3): From empirical results, we draw the following conclusion: Pre-training only provides stable gains when the downstream dataset is small, and this positive gains diminish as the dataset scale increases. To be specific, the positive transfer is only observed in the low-data regime, while the negative transfer occurs when the downstream data quantity reaches a certain scale. Furthermore, this impact does not diminish with increasing data quantity. For instance, in PCBA, the performance difference between the two models continues to increase after the intersection point. This suggests the existence of *parameter ossification* (Hernandez et al., 2021) in the pre-trained model, which suggest that pre-training can ossify the model weights so that they do not adapt as well to the fine-tuning distribution in the high-data regime. Given this phenomenon, it requires careful consideration when utilizing pre-training for MRL.

3.4. The effect of distribution on the scaling law

There are two widely adopted data splitting approaches in prior studies on MRL: random split and scaffold split(Wu et al., 2018), representing uniform and out-of-distribution settings, respectively. However, in practical drug development scenarios, it is more common for the test set to contain molecular structures that are scarce in the training set rather than completely absent. To this end, we propose a modified data splitting method that extracts a portion of samples (5% in our experiment) from the test and validation sets and includes them in the training set, creating a more imbalanced data distribution. This approach serves as a trade-off between random split and scaffold split, better aligning with practical requirements.

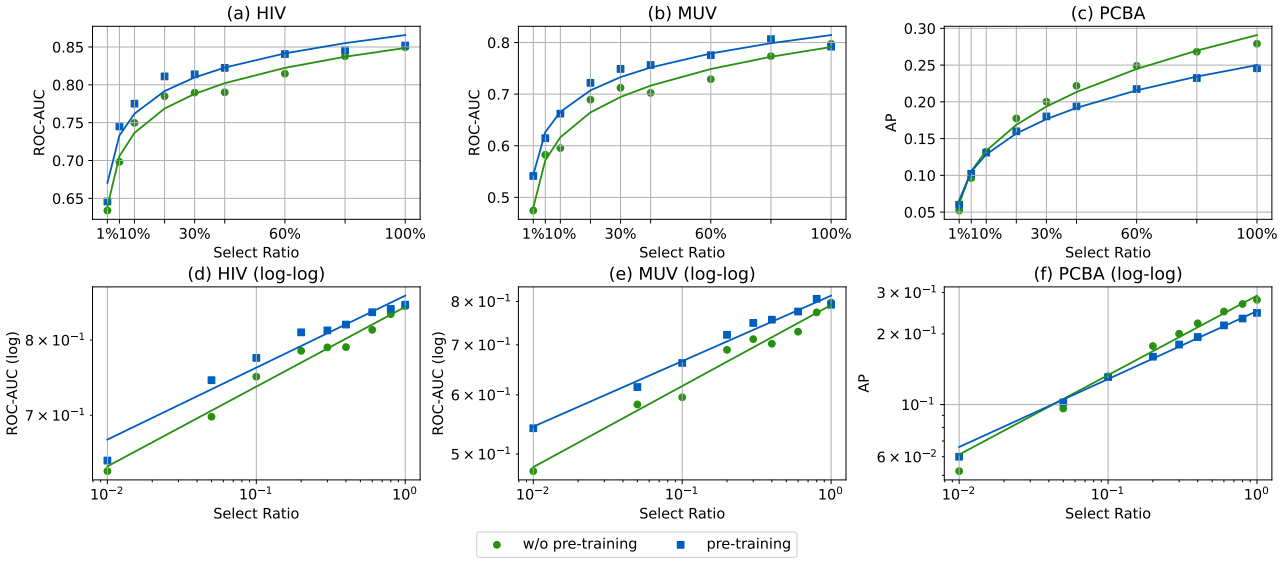

Figure 3. The effect of pre-training on power law.

Observation (4): The results in Figure 4 demonstrate that the model performance also follows a power law with data quantity across different distribution settings. Specifically, the highest exponent is observed in the random split, while the scaffold and imbalanced split exhibit similar learning behaviors. Specifically, imbalanced splitting of the `MUV` dataset exhibits outliers that deviate from the power law at data proportions of 80% and 100%, but not in cases of random or scaffold splitting.

3.5. The effect of model capacity on the scaling law

The capacity is another key factor influencing MRL, primarily determined by the number of layers (depth D) and hidden units (width W) in the model. In our experimental investigation, we ignore the parameters of the embedding layer and output layer and define the model capacity as $D \times W$. We incrementally select four experimental configurations: $[64 \times 2, 100 \times 3, 300 \times 5, 600 \times 10]$. The performance of the four model capacities across different datasets as data varies is shown in Figure 5.

The experimental results demonstrate that the optimal model capacity varies for different tasks, but distinct model capacity does not break the scaling relationship of power law . For `HIV` and `PCBA` datasets, the model with a capacity of 1.5K achieved the best performance. On the contrary, the smallest model with a capacity of 128 achieved optimal performance on `MUV` dataset. Interestingly, on the `PCBA` dataset with an overall data scale of 400K, there are significant differences in power law exponents among different model capacities. The smaller capacity models exhibite a noticeable performance bottleneck as the data scale increase, whereas this phenomenon is not evident in `HIV` and `MUV` datasets.

Observation (5): In general, the power law relationship between model performance and data scale remains unchanged regardless of the chosen model parameters. However, the model capacity could impact training efficiency by increasing or decreasing the power-law exponent. It is worth noting that there is no discernible relationship between training efficiency and dataset size across different tasks. Thus, we call for careful modal capacity selection according to the characteristics of the task, as optimal performance for some tasks may lie in smaller capacity models.

3.6. Data pruning strategies

The great advances in MRL have primarily been driven by highly expressive models and ever-bigger dataset, which imposes substantial computational and storage burdens. In recent years, much attention has been devoted to the design of models, while the exploration from the data perspective have been relatively overlooked. Specifically, there has been a limited focus on strategies to enhance training efficiency and uncover the representation capacity of smaller-scale datasets from a data-centric standpoint. Hence, we shift our attention to *data redundancy* problem in MRL, which is crucial for network training and parameter tuning efficiency. Specifically, we benchmark seven data pruning strategies originally designed for image data and adapt them for the molecular domain: Herding(Chen et al., 2012),

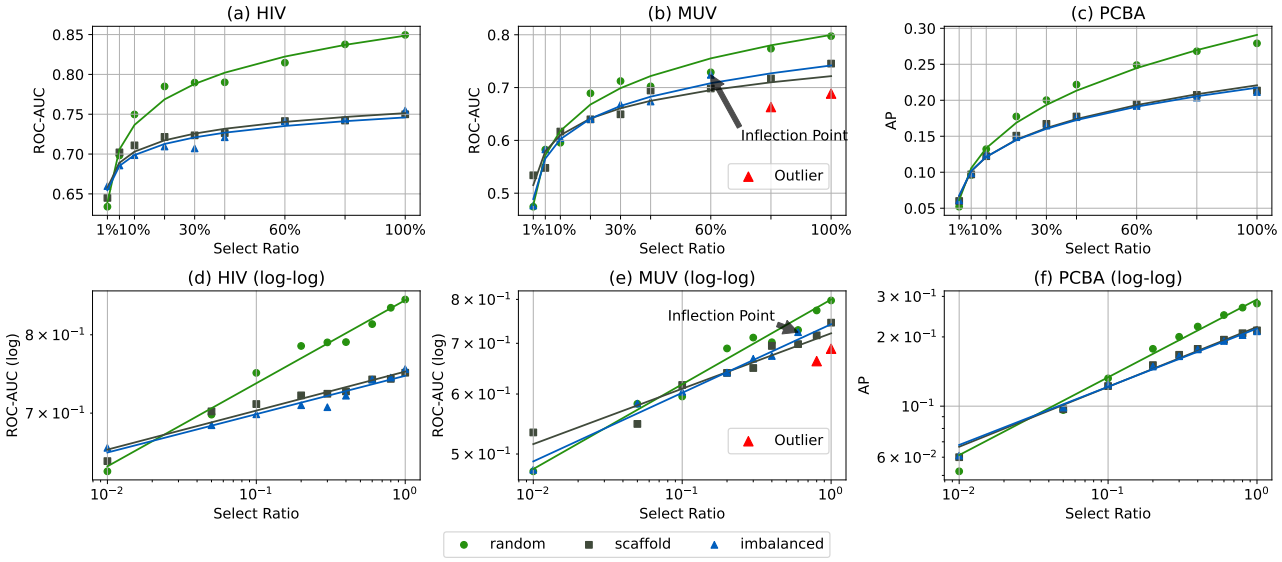

Figure 4. The effect of distribution shift on power law. Outliers that deviate significantly from the fitted curve are marked in red, and arrows indicate the occurrence of inflection point.

Table 1. Performance of data pruning strategies on HIV dataset in terms of ROC-AUC (%, ↑). We highlight the performing results which is higher / lower than random pruning under significance testing (the p-value is lower than 5%).

Uniform	1%	5%	10%	20%	30%	40%	60%	80%
Random	63.4±2.8	69.8±2.2	75.0±2.7	78.5±1.2	79.0±2.2	79.0±1.3	81.5±1.7	83.8±0.8
Herding	60.2±3.9	63.3±3.8	64.7±5.0	69.5±5.4	71.8±7.0	75.8±6.6	80.0±2.8	82.6±1.2
Entropy	67.9±2.2	71.1±3.7	74.2±1.6	76.2±1.2	77.0±2.0	79.2±1.8	81.4±1.9	83.2±1.4
Least Confidence	66.2±4.0	70.4±2.1	72.8±3.9	76.7±2.3	78.0±1.0	81.0±1.4	81.6±1.6	83.3±0.6
Forgetting	67.7±1.2	75.2±1.3	75.1±1.9	76.2±1.7	80.0±1.8	79.8±1.6	82.8±1.0	83.7±1.4
GraNd	66.2±4.0	69.3±2.6	73.6±2.0	78.1±1.1	78.1±1.6	78.6±1.0	82.3±0.8	83.2±1.4
Kmeans	63.8±4.8	64.4±3.4	65.7±1.8	68.1±1.6	71.5±1.4	72.5±3.5	79.2±0.5	82.3±2.2
Imbalanced	1%	5%	10%	20%	30%	40%	60%	80%
Random	66.6±1.7	68.6±3.1	69.9±3.7	70.9±1.4	70.7±4.1	72.1±3.0	74.1±1.1	74.4±1.1
Herding	57.1±3.0	63.0±3.8	64.9±3.6	65.8±5.9	67.3±6.0	72.6±1.8	73.3±2.2	73.7±0.6
Entropy	67.7±7.5	71.5±2.8	70.1±1.1	71.2±2.1	73.2±2.3	71.7±2.6	74.7±1.3	74.8±1.0
Least Confidence	66.8±5.2	71.4±1.0	71.3±2.6	71.8±2.7	69.5±2.8	73.7±3.4	73.4±2.6	73.8±1.8
Forgetting	66.1±3.1	69.7±5.8	70.2±3.6	71.9±1.9	71.6±1.8	71.4±2.0	73.9±1.4	74.2±2.3
GraNd	62.7±4.5	71.0±2.6	69.2±3.6	73.1±1.9	70.0±3.4	72.9±3.0	74.4±1.8	75.9±1.2
Kmeans	67.9±1.8	65.4±3.2	65.0±1.9	67.1±4.3	69.1±4.0	68.5±4.3	72.8±1.2	74.4±1.7

Entropy(Lewis & Gale, 1994), Least Confidence(Lewis & Gale, 1994), Forgetting(Toneva et al., 2018), GraNd(Paul et al., 2021), K-means(Sorscher et al., 2022) and we additionally include random pruning as a baseline method. Please refer to Appendix C.3 for detailed description of adopted data pruning strategies.

Regarding the experimental setup, we employ the graph modality as the subject of our study and conduct data pruning evaluation on three classification datasets. We incrementally select subsets of the complete dataset using various data pruning strategies at eight different proportions: $[1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 60\%, 80\%]$. Additionally, we perform comparative experiments on two distribution settings to observe if there are significant differences in the performance of data pruning strategies for different distributions. The randomly selected results serve as the baseline for comparison. Table 4 demonstrates the experimental results on the `HIV` dataset, where different colors denote results with a p-value less than 5% in the significance testing (t-test) compared with random pruning. Other results of the `MUV` and `PCBA` datasets are shown in Appendix C.3.

Observation (6): From the empirical performance, none of the baselines show a significant advantage over random pruning in most data proportions, regardless of whether the data distribution is uniform or imbalanced. However, in the case of the MUV dataset with an imbalanced data distribution, multiple data pruning methods such as random, GraNd, and Least Confidence exhibit a performance decline, while other methods do not. In the context of PCBA, the performance differences among various strategies are relatively small. This indicates that larger data scales can actually narrow the difference of existing data pruning approaches. Thus, we highlight the need for the development of data pruning strategies specifically tailored to molecular data to better enhance MRL.

4. Limitations and Future Work

Limitations. We conduct our experiments using widely adopted model architectures in the field, such as GIN for the graph modality, SchNet for the 3D geometry modality,

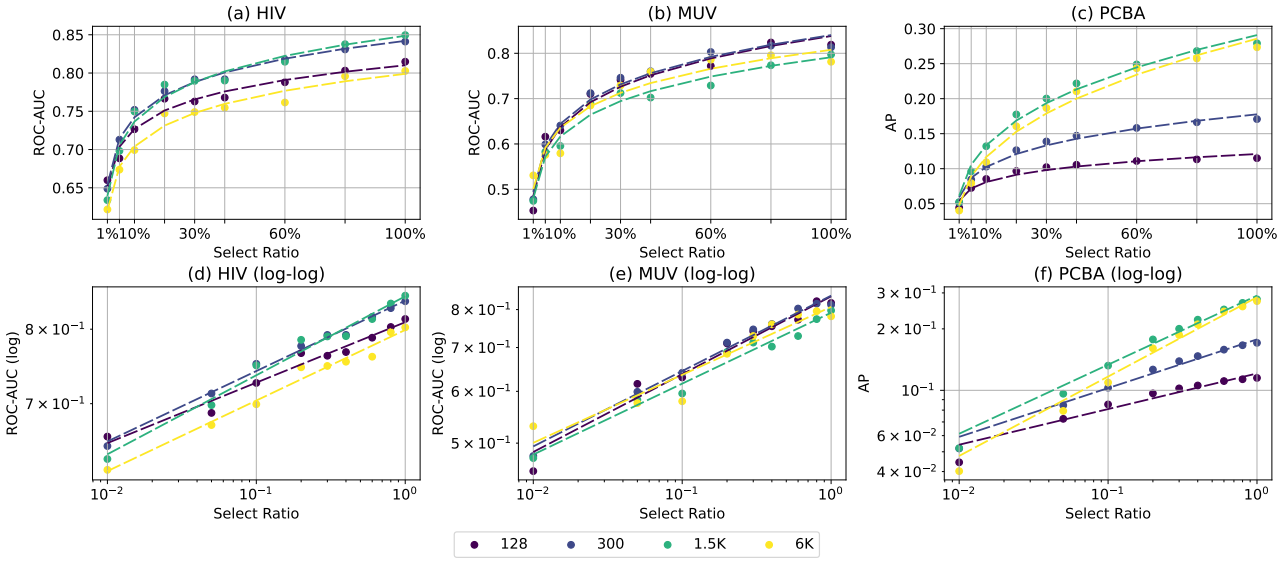

Figure 5. The effect of the amount of model parameters.

and Transformer for SMILES and Fingerprint. However, the rapid development of molecular representation learning in recent years has led to the emergence of many models with enhanced expressiveness. The neural scaling law on these models is still to be explored. Moreover, our focus is primarily on investigating the impact of various dimensions on supervised molecular representation learning. Despite we consider the effect of pre-training, which is another mainstream research area in molecular representation learning, we do not explore the neural scaling behavior between data quantity in pre-training and the corresponding performance on downstream task performance.

Future Work. (1) *Data pruning strategy for molecular data.* As observed in our experiment in Section 3.6, compared to the effectiveness of data pruning strategies in the field of computer vision, these methods do not yield significant results in the molecular domain, and the reasons for this remain to be investigated. Additionally, the design of efficient pruning strategies specifically tailored to molecular data is a promising direction which remain unexplored. (2) *Pre-training strategy with consistent and positive transfer.* How pre-training strategies can consistently and effectively improve downstream tasks is also an open question. We provide a new perspective for molecular pre-training research in MRL: to address the issues of parameter ossification and alleviate the decline in learning efficiency of pre-trained models.

5. Conclusion

We investigate the neural scaling behavior in molecular representation learning to explore how quantity and quality of molecular data affect the performance. Our research confirms that the performance of molecular representation learning follows a power-law relationship with data quantity. Additionally, the experimental results across multiple dimensions demonstrate that the modality, distribution, pre-training intervention, and model parameter size all influence the learning behaviors of molecular data. Specifically, the graph modality and uniformly distributed random split exhibits higher learning efficiency in the studied datasets, and the positive transfer from pre-training diminishes as the data quantity increases. Moreover, the optimal model parameter size is highly correlated with task requirements. We further adapt seven data pruning strategies to molecular data and benchmark their performance. Surprisingly, none of them with simple adaptations can beat the random pruning baseline in MRL. Based on our experimental findings, we raise several key considerations for molecular representation learning, particularly from a data-centric perspective, providing valuable insights for future research endeavors.

References

AIDS Antiviral Screen Data. URL https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data.

Amari, S.-i., Fujita, N., and Shinomoto, S. Four types of learning curves. *Neural Computation*, 4(4):605–618, 1992.

Banko, M. and Brill, E. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pp. 26–33, 2001.

Beaini, D., Passaro, S., Létourneau, V., Hamilton, W., Corso, G., and Liò, P. Directional graph networks. In *International Conference on Machine Learning*, pp. 748–758. PMLR, 2021.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Bodnar, C., Frasca, F., Otter, N., Wang, Y., Lio, P., Montufar, G. F., and Bronstein, M. Weisfeiler and lehman go cellular: Cw networks. *Advances in Neural Information Processing Systems*, 34:2625–2640, 2021.

Bouritsas, G., Frasca, F., Zafeiriou, S., and Bronstein, M. M. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):657–668, 2022.

Caballero, E., Gupta, K., Rish, I., and Krueger, D. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

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

Corso, G., Cavalleri, L., Beaini, D., Liò, P., and Veličković, P. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33: 13260–13271, 2020.

Du, W., Zhang, H., Du, Y., Meng, Q., Chen, W., Zheng, N., Shao, B., and Liu, T.-Y. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pp. 5583–5608. PMLR, 2022.

Du, W., Du, Y., Wang, L., Feng, D., Wang, G., Ji, S., Gomes, C., and Ma, Z.-M. A new perspective on building efficient and expressive 3d equivariant graph neural networks. *arXiv preprint arXiv:2304.04757*, 2023.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Gasteiger, J., Giri, S., Margraf, J. T., and Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020a.

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020b.

Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Guo, C., Zhao, B., and Bai, Y. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*, pp. 181–195. Springer, 2022.

Haussler, D. Quantifying inductive bias: Ai learning algorithms and valiant's learning framework. *Artificial intelligence*, 36(2):177–221, 1988.

Haussler, D., Seung, H. S., Kearns, M., and Tishby, N. Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the seventh annual conference on Computational learning theory*, pp. 76–87, 1994.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

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.

Kimber, T. B., Chen, Y., and Volkamer, A. Deep learning in virtual screening: recent applications and developments. *International Journal of Molecular Sciences*, 22(9):4435, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klug, T. and Heckel, R. Scaling laws for deep learning based image reconstruction. *arXiv preprint arXiv:2209.13435*, 2022.

Landrum, G., Tosco, P., Kelley, B., Ric, sriniker, gedeck, Vianello, R., NadineSchneider, Kawashima, E., Dalke, A., N, D., Cosgrove, D., Cole, B., Swain, M., Turk, S., AlexanderSavelyev, Jones, G., Vaucher, A., Wójcikowski, M., Take, I., Probst, D., Ujihara, K., Scalfani, V. F., guillaume godin, Pahl, A., Berenger, F., JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_03_2 (q1 2022) release, 2022.

Landrum, G. et al. Rdkit: Open-source cheminformatics software. 2016. *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit*, 149(150):650, 2016.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

Li, G., Xiong, C., Thabet, A., and Ghanem, B. Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4541–4549, 2022.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021a.

Liu, Y., Zhao, T., Ju, W., and Shi, S. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B., and Ji, S. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021b.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Neumann, O. and Gros, C. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Rohrer, S. G. and Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.*, 49(2): 169–184, 2009.

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.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

Shen, J. and Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36, 2019.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., et al. Pubchem's bioassay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

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.

Weisfeiler, B. and Leman, A. A Reduction of a Graph to a Canonical Form and an Algebra Arising During This Reduction. *Nauchno-Technicheskaya Informatsia*, 2(9): 12–16, 1968.

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., and Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yang, S., Li, Z., Song, G., and Cai, L. Deep molecular representation learning via fusing physical and chemical information. *Advances in Neural Information Processing Systems*, 34:16346–16357, 2021.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.

Yu, Z. and Gao, H. Molecular representation learning via heterogeneous motif graph neural networks. In *International Conference on Machine Learning*, pp. 25581–25594. PMLR, 2022.

Zhu, J., Xia, Y., Wu, L., Xie, S., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th*

ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2626–2636, 2022.

Zhu, Y., Chen, D., Du, Y., Wang, Y., Liu, Q., and Wu, S. Featurizations matter: A multiview contrastive learning approach to molecular pretraining. In *ICML 2022 2nd AI for Science Workshop*.

A. Implementation Details

A.1. Modality Encoders

In this section, we introduce the detailed implementation of used modality encoders. We denote the representation for node (atom) v_i as $\boldsymbol{h}_i$ and the representation at the graph (molecule) level as $\boldsymbol{z}$.

Embedding 2D graphs. Graph Isomorphism Network (GIN) (Xu et al., 2018) is a simple and effective model to learn discriminative graph representations, which is proved to have the same representational power as the Weisfeiler-Lehman test (Weisfeiler & Leman, 1968). Recall that each molecule is represented as $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{E})$, where $\boldsymbol{A}$ is the adjacency matrix, $\boldsymbol{X}$ and $\boldsymbol{E}$ are features for atoms and bonds respectively. The layer-wise propagation rule of GIN can be written as:

$$\boldsymbol{h}_i^{(k+1)} = f_{\text{atom}}^{(k+1)}\left(\boldsymbol{h}_i^{(k)} + \sum_{j \in \mathcal{N}(i)} \left(\boldsymbol{h}_j^{(k)} + f_{\text{bond}}^{(k+1)}(\boldsymbol{E}_{ij})\right)\right), \tag{2}$$

where the input features $\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i$, $\mathcal{N}(i)$ is the neighborhood set of atom v_i, and f_{atom}, f_{bond} are two MultiLayer Perceptron (MLP) layers for transforming atoms and bonds features, respectively. By stacking K layers, we can incorporate K-hop neighborhood information into each center atom in the molecular graph. Then, we take the output of the last layer as the atom representations and further use the mean pooling to get the graph-level molecular representation:

$$\boldsymbol{z}^{\text{2D}} = \frac{1}{N}\sum_{i \in \mathcal{V}} \boldsymbol{h}_i^{(K)}. \tag{3}$$

Embedding 3D graphs. We use the SchNet (Schütt et al., 2017) as the encoder for the 3D geometry graphs. SchNet models message passing in the 3D space as continuous-filter convolutions, which is composed of a series of hidden layers, given as follows:

$$\boldsymbol{h}_i^{(k+1)} = f_{\text{MLP}}\left(\sum_{j=1}^{N} f_{\text{FG}}(\boldsymbol{h}_j^{(t)}, \boldsymbol{r}_i, \boldsymbol{r}_j)\right) + \boldsymbol{h}_i^{(t)}, \tag{4}$$

where the input $\boldsymbol{h}_i^{(0)} = \boldsymbol{a}_i$ is an embedding dependent on the type of atom v_i, $f_{\text{FG}}(\cdot)$ denotes the filter-generating network. To ensure rotational invariance of a predicted property, the message passing function is restricted to depend only on rotationally invariant inputs such as distances, which satisfying the energy properties of rotational equivariance by construction. Moreover, SchNet adopts radial basis functions to avoid highly correlated filters. The filter-generating network is defined as follow:

$$f_{\text{FG}}(\boldsymbol{x}_j, \boldsymbol{r}_i, \boldsymbol{r}_j) = \boldsymbol{x}_j \cdot e_k(\boldsymbol{r}_i - \boldsymbol{r}_j) = \boldsymbol{x}_j \cdot \exp(-\gamma \| \|\boldsymbol{r}_i - \boldsymbol{r}_j\|_2 - \mu\|_2^2). \tag{5}$$

Similarly, for non-quantum properties prediction concerned in this work, we take the average of the node representations as the 3D molecular embedding:

$$\boldsymbol{z}^{\text{3D}} = \frac{1}{N}\sum_{i \in \mathcal{V}} \boldsymbol{h}_i^{(K)}, \tag{6}$$

where K is the number of hidden layers.

Embedding fingerprints & SMILES strings. Due to the discrete and extremely sparse nature of fingerprint vectors, we first transform all F binary feature fields into a dense embedding matrix $\boldsymbol{F}^{fp} \in \mathbb{R}^{F^{fp} \times D_{\text{F}}}$ via embedding lookup, while we transform SMILES tokens in the same way via another embedding lookup $\boldsymbol{F}^{sm} \in \mathbb{R}^{F^{sm} \times D_{\text{F}}}$. Then, we introduce a positional embedding matrix $\boldsymbol{P} \in \mathbb{R}^{F \times D_{\text{F}}}$ to capture the positional relationship among bits in the fingerprint vector, which is defined as:

$$\boldsymbol{P}_{p,2i} = \sin(p/10000^{2i/D_F}), \tag{7}$$

$$\boldsymbol{P}_{p,2i+1} = \cos(p/10000^{2i/D_F}), \tag{8}$$

where p denotes the corresponding bit position and i is corresponds to the i-th embedding dimension. The positional embedding matrix will be added to the transformed embedding matrix:

$$\boldsymbol{F} = \boldsymbol{F}' + \boldsymbol{P}. \tag{9}$$

Thereafter, we use a multihead Transformer (Vaswani et al., 2017) to model the interaction among those feature fields. Specifically, we first transform each feature into a new embedding space as:

$$\boldsymbol{Q}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\mathrm{Q}}^{(h)}, \tag{10}$$

$$\boldsymbol{K}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\mathrm{K}}^{(h)}, \tag{11}$$

$$\boldsymbol{V}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\mathrm{V}}^{(h)}, \tag{12}$$

where the three linear transformation matrices $\boldsymbol{W}_{\mathrm{Q}}^{(h)}, \boldsymbol{W}_{\mathrm{K}}^{(h)}, \boldsymbol{W}_{\mathrm{V}}^{(h)} \in \mathbb{R}^{D_{\mathrm{F}} \times D/H}$ parameterize the query, key, and value transformations for the h-th attention head, respectively. Following that, we compute the attention scores among all feature pairs and then linearly combine the value matrix from all H attention heads:

$$\boldsymbol{W}_{\mathrm{A}}^{(h)} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}^{(h)}(\boldsymbol{K}^{(h)})^{\top}}{\sqrt{D_{\mathrm{H}}}}\right), \tag{13}$$

$$\widehat{\boldsymbol{Z}} = \left[\boldsymbol{W}_{\mathrm{A}}^{(1)}\boldsymbol{V}^{(1)}; \boldsymbol{W}_{\mathrm{A}}^{(2)}\boldsymbol{V}^{(2)}; \ldots; \boldsymbol{W}_{\mathrm{A}}^{(H)}\boldsymbol{V}^{(H)}\right], \tag{14}$$

Finally, we perform sum pooling on the resulting embedding matrix $\widehat{\boldsymbol{Z}} \in \mathbb{R}^{F \times D_{\mathrm{F}}}$ and use a linear model f_{LIN} to obtain the final fingerprint or SMILES string embedding $\boldsymbol{z} \in \mathbb{R}^{D}$:

$$\boldsymbol{z} = f_{\mathrm{LIN}}\left(\sum_{d=1}^{D_{\mathrm{F}}} \widehat{\boldsymbol{Z}}_d\right). \tag{15}$$

A.2. Computing infrastructures

Software infrastructures. All of the experiments are implemented in Python 3.7, with the following supporting libraries: PyTorch 1.10.2 (Paszke et al., 2019), PyG 2.0.3 (Fey & Lenssen, 2019), RDKit 2022.03.1 (Landrum et al., 2016) and HuggingFace's Transformers 4.17.0 (Wolf et al., 2019).

Hardware infrastructures. We conduct all experiments on a computer server with 8 NVIDIA GeForce RTX 3090 GPUs (with 24GB memory each) and 256 AMD EPYC 7742 CPUs.

A.3. Code availability

The source code of our empirical implementation can be accessed at `https://github.com/Data-reindeer/NSL_MRL`.

B. Datasets and Tasks

In the following, we will elaborate on the adopted datasets and the statistics are summarized in Table 2.

Table 2. Statistics of datasets used in experiments.

	Dataset	Data Type	#Molecules	Avg. #atoms	Avg. #bonds	#Tasks	Avg. degree
Classification	MUV	SMILES	93,087	24.23	26.28	17	2.17
	HIV	SMILES	41,127	25.51	27.47	1	2.15
	PCBA	SMILES	437,929	25.96	28.09	92	2.16
Regression	QM9-ϵ_{gap}	SMILES, 3D	130,831	18.03	18.65	1	2.07
	QM9-U0	SMILES, 3D	130,831	18.03	18.65	1	2.07
	QM9-ZPVE	SMILES, 3D	130,831	18.03	18.65	1	2.07

Datasets. We consider four datasets ranging from molecular-level properties to macroscopic influences on human body for experimental investigation: `HIV` (AID), `MUV` (Rohrer & Baumann, 2009), `PCBA` (Wang et al., 2012) and `QM9` (Ruddigkeit et al., 2012).

- HIV dataset (AIDS Antiviral Screen) was developed by the Drug Therapeutics Program (DTP) (AID), which is designed to evaluate the ability of molecular compounds to inhibit HIV replication.
- Maximum Unbiased Validation (MUV) group was selected from PubChem BioAssay via a refined nearest neighbor analysis approach, which is specifically designed for validation of virtual screening techniques (Rohrer & Baumann, 2009).
- PubChem BioAssay (PCBA) is a database consisting of biological activities of small molecules generated by high-throughput screening (Wang et al., 2012).
- QM9 dataset is a comprehensive dataset that provides geometric, energetic, electronic and thermodynamic properties for a subset of GDB-17 database, comprising 134 thousand stable organic molecules with up to nine heavy atoms (Ruddigkeit et al., 2012). In our experiments, we delete 3,054 uncharacterized molecules which failed the geometry consistency check (Ramakrishnan et al., 2014). We include the ϵ_{gap}, U0, and ZPVE in our experiment, which cover properties related to electronic structure, stability, and thermodynamics. These properties collectively capture important aspects of molecular behavior and can effectively represent various energetic and structural characteristics within the QM9 dataset.

C. Additional Experimental Results

C.1. Single-property performance of MUV dataset with SMILES modality

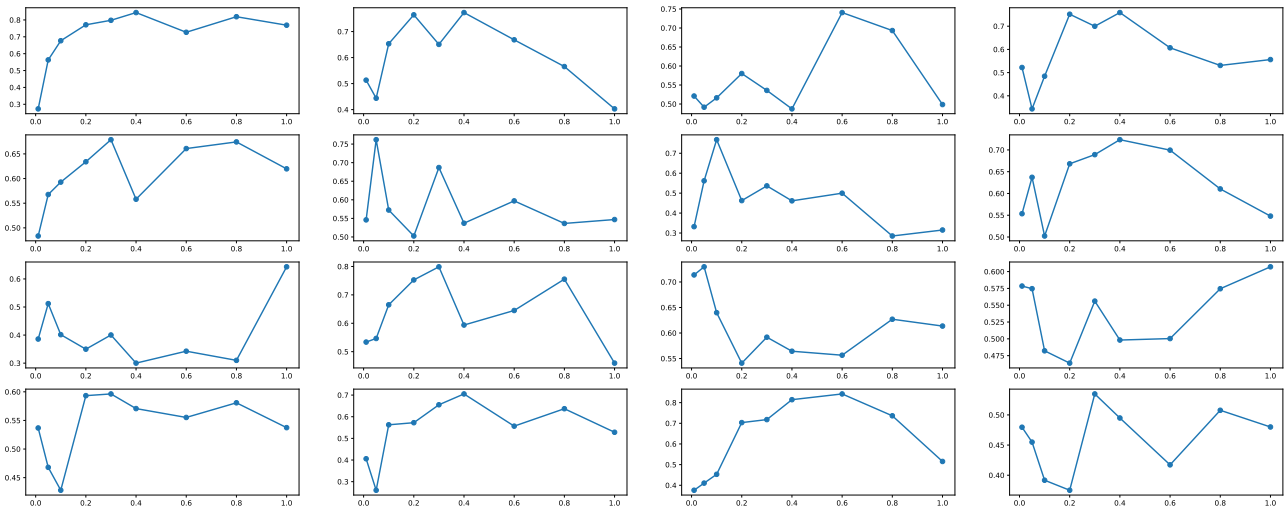

Figure 6. The neural scaling law of single-property performance (ROC-AUC) of MUV dataset with SMILES modality.

In order to gain a more detailed understanding of the performance degradation phenomenon in the multitask scenario of MUV, we specifically demonstrate the neural scaling behavior of the SMILES modalities in single-property ROC-AUC, as shown in Figure 6. Despite some ascending behavior, the bulk of properties exhibit varying degrees of performance drop in the last few proportions, which accounts for the overall performance drop in the multi-task setting.

C.2. Results of different-layers Transformer with SMILES modality.

Figure 7 presents the scaling behavior of transformers with different numbers of layers in terms of their performance. It can be observed that RoBERTa achieves the overall best performance, followed by a 1-layer Transformer, and the worst performance is exhibited by a 3-layer one. In contrast to fingerprint, the SMILES modality could experience a performance drop on some datasets (HIV and MUV) in the high-data regime. Additionally, the varying numbers of Transformer layers do not affect our conclusions in Section 3.2 regarding modality comparison, as even the superior RoBERTa model overall does not surpass the performance of graph and fingerprint modalities.

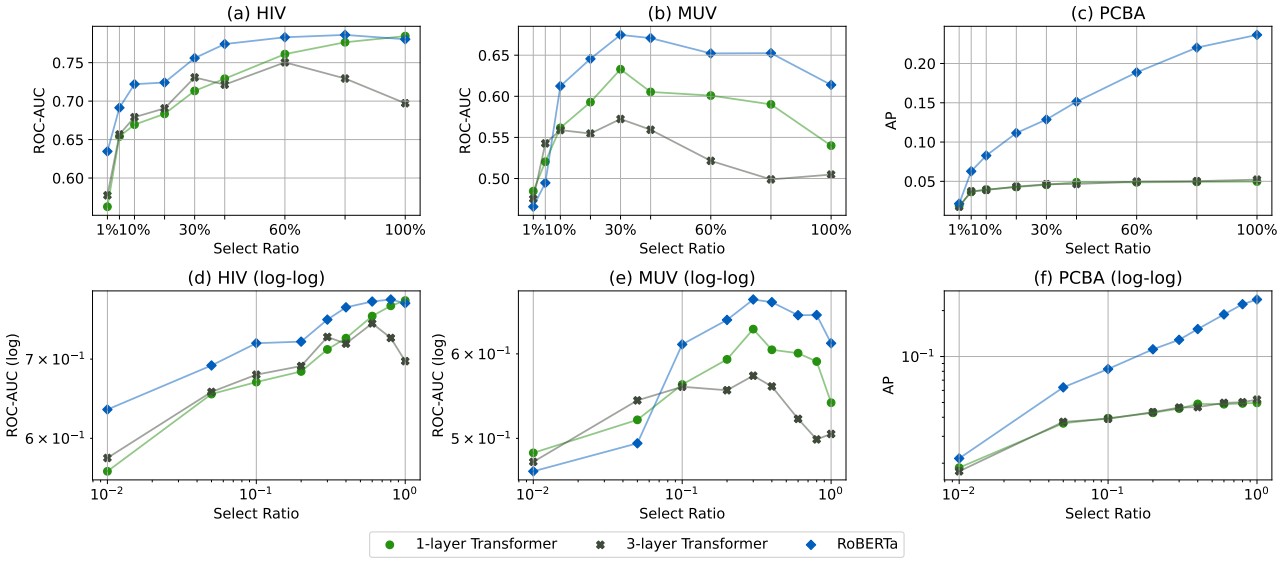

Figure 7. The neural scaling law of different-layer model (Transformer) performance with SMILES modality.

C.3. Data pruning strategies and additional results

Problem Statement. Consider a learning scenario where we have a large training set denoted as $\mathcal{T} = (\boldsymbol{x_i}, y_i)_{i=1}^{|\mathcal{T}|}$, consisting of input-output pairs $(\boldsymbol{x_i}, y_i)$, where $\boldsymbol{x_i} \in \mathcal{X}$ represents the input and $y_i \in \mathcal{Y}$ denotes the ground-truth label corresponding to x_i. Here, $\mathcal{X}$ and $\mathcal{Y}$ refer to the input and output spaces, respectively. The objective of data pruning is to identify a subset $\mathcal{S} \subset \mathcal{T}$, satisfying the constraint $|\mathcal{S}| < |\mathcal{T}|$, that captures the most informative instances. This subset, when used to train a model denoted as $\boldsymbol{\theta}^{\mathcal{S}}$, should yield a similar or better generalization performance to that of the model $\boldsymbol{\theta}^{\mathcal{T}}$, which is trained on the entire training set $\mathcal{T}$.

Data pruning strategies. In our data pruning experiments, we implement a total of seven data pruning (or coreset selection) strategies: Herding(Chen et al., 2012), Entropy(Lewis & Gale, 1994), Least Confidence(Lewis & Gale, 1994), Forgetting(Toneva et al., 2018), GraNd(Paul et al., 2021), K-means(Sorscher et al., 2022) and we additionally include random pruning as a baseline method. These seven strategies are widely used in the field of CV(Guo et al., 2022) and have potential in mitigating the issue of data redundancy in large-scale datasets, thereby saving computational and storage resources. Here we provide a brief overview to each of them.

- **Herding**(Chen et al., 2012) operates by selecting data points in the feature space based on the distance between the coreset center and the original center. It follows an incremental and greedy approach, adding one sample at a time to the coreset in order to minimize the distance between the centers.

- **Entropy** and **Least Confidence**(Lewis & Gale, 1994) iteratively select samples with lower entropy and least confidence, respectively. These methods identify informative samples by considering that lower uncertainty can provide more information gain, thereby benefiting model training and reducing data redundancy.

- **Forgetting**(Toneva et al., 2018) calculates the frequency of forgetting that occurs during the training process, which refers to the number of times the samples correctly classified in the previous epoch are misclassified in the current epoch. Those unforgettable samples, exhibiting robust performance across epochs, have minimal impact on model performance when removed.

- **GraNd**(Paul et al., 2021) measures the average impact of each sample on the reduction of training loss during the initial epochs. Training samples are more important if they contribute more to the error or loss when training neural networks.

- ***k*-means** (Sorscher et al., 2022) employs the application of k-means clustering in the latent space to define the difficulty of each data point based on its Euclidean distance to its nearest cluster centroid. Simple samples (with low difficulty) are considered for removal to reduce data redundancy. It is noteworthy that this method, unlike the aforementioned approaches, does not require any label information or training and can be directly applied to the dataset.

Table 3. Performance of data pruning strategies on MUV dataset in terms of ROC-AUC (%, ↑). We highlight the performing results which is higher / lower than random pruning under significance testing (the p-value is lower than 5%).

Uniform	1%	5%	10%	20%	30%	40%	60%	80%
Random	47.4±3.5	58.3±3.4	59.6±4.7	68.9±2.1	71.2±3.4	70.2±3.0	72.9±1.7	77.4±2.8
Herding	48.4±4.7	51.9±4.8	53.1±10.3	61.5±4.9	68.5±7.3	65.1±5.6	71.7±6.4	78.7±5.0
Entropy	48.9±5.5	58.5±5.1	62.8±4.0	63.8±4.2	68.7±3.4	71.0±3.7	75.0±2.5	79.0±3.8
Least Confidence	52.0±7.9	57.4±5.7	61.1±3.3	66.3±2.1	67.6±4.8	70.0±3.6	75.1±2.5	78.5±2.2
Forgetting	47.1±2.1	58.6±2.1	60.9±5.2	63.3±1.8	67.1±3.0	70.5±2.8	74.5±2.6	76.9±3.8
GraNd	47.3±3.4	52.2±7.1	64.7±5.1	65.9±3.1	64.4±5.9	71.4±4.4	72.5±2.5	76.2±3.2
k-means	49.9±4.7	60.5±7.8	65.5±3.7	68.0±4.2	65.5±3.7	67.1±2.8	72.1±1.4	76.9±4.0
Imbalanced	1%	5%	10%	20%	30%	40%	60%	80%
Random	47.7±3.5	58.3±3.4	60.4±4.7	64.1±2.1	66.8±3.4	67.4±3.0	72.4±1.7	66.3±2.8
Herding	50.2±4.7	55.8±4.8	59.3±10.3	58.8±4.9	64.5±7.3	65.6±5.6	65.8±6.4	69.6±5.0
Entropy	47.7±5.5	57.5±5.1	61.0±4.0	68.0±4.2	64.4±3.4	67.2±3.7	68.1±2.5	69.5±3.8
Least Confidence	50.4±7.9	52.9±5.7	60.5±3.3	64.2±2.1	65.5±4.8	68.3±3.6	69.4±2.5	66.4±2.2
Forgetting	49.3±2.1	51.4±2.1	58.7±5.2	64.0±1.8	65.7±3.0	66.5±2.8	67.7±2.6	68.8±3.8
GraNd	52.0±3.4	55.3±7.1	63.8±5.1	63.4±3.1	67.0±5.9	68.6±4.4	70.3±2.5	69.0±3.2
k-means	52.7±4.7	56.0±7.8	58.5±3.7	61.0±4.2	63.4±3.7	63.4±2.8	67.1±1.4	70.0±4.0

Table 4. Performance of data pruning strategies on PCBA dataset in terms of Average Precision (%, ↑). We highlight the performing results which is higher / lower than random pruning under significance testing (the p-value is lower than 5%).

Uniform	1%	5%	10%	20%	30%	40%	60%	80%
Random	5.2±0.1	9.6±0.2	13.2±0.2	17.7±0.5	20.0±0.6	22.2±0.6	24.9±0.5	26.8±0.4
Herding	3.7±1.5	9.7±0.6	10.2±3.6	15.8±3.7	20.7±0.7	22.6±0.6	25.4±0.8	24.5±3.4
Entropy	5.4±0.2	10.0±0.4	13.7±0.6	17.6±0.1	20.2±0.6	22.1±0.4	25.0±0.3	26.6±0.4
Least Confidence	5.3±0.1	9.7±0.3	13.7±0.5	17.6±0.4	20.3±0.3	22.0±0.5	25.0±0.3	26.5±0.2
Forgetting	5.5±0.2	9.8±0.4	13.5±0.4	17.5±0.3	20.4±0.3	22.1±0.2	24.7±0.6	26.5±0.2
GraNd	5.5±0.3	9.7±0.3	13.3±0.4	17.5±0.4	20.3±0.5	22.1±0.5	24.9±0.3	26.8±0.3
k-means	4.1±0.2	8.1±0.3	11.6±0.3	16.5±0.2	20.4±0.4	22.6±0.3	24.9±0.4	26.3±0.1
Imbalanced	1%	5%	10%	20%	30%	40%	60%	80%
Random	6.1±0.2	9.9±0.2	12.4±0.5	14.9±0.3	16.5±0.2	17.7±0.2	19.2±0.1	20.4±0.2
Herding	4.7±0.6	8.5±0.5	11.5±0.3	13.4±1.6	14.7±2.8	16.5±2.1	16.8±3.6	17.9±3.1
Entropy	6.0±0.3	9.9±0.4	12.3±0.4	15.1±0.5	16.8±0.5	17.8±0.2	19.3±0.4	20.5±0.3
Least Confidence	6.0±0.3	9.9±0.2	12.3±0.4	15.3±0.3	17.0±0.5	17.9±0.3	19.4±0.3	20.5±0.1
Forgetting	5.7±0.3	9.9±0.2	12.2±0.3	14.7±0.2	15.8±0.3	17.8±0.2	19.3±0.5	20.3±0.3
GraNd	5.9±0.4	9.7±0.3	12.2±0.5	15.0±0.6	16.8±0.3	17.8±0.4	19.4±0.2	20.7±0.5
k-means	4.6±0.3	8.1±0.3	10.7±0.3	13.9±0.2	16.4±0.4	17.2±0.3	19.6±0.5	20.4±0.1

D. Related Work

The following section provides a more broad literature review across the spectrum of molecular representation learning and neural scaling law.

D.1. Molecular representation learning

The past decade has seen remarkable success in the application of deep learning in a variety of biochemical tasks, spanning from virtual screening (Kimber et al., 2021) to molecular property prediction (Gilmer et al., 2017). Within this context, molecular representation learning (MRL) serves as a pivotal link between the molecular modalities and the target tasks, efficiently capturing and encoding rich chemical semantic information into vector representations.

One of the mainstream research approaches in MRL is based on *2D topology graphs*. The advancements in Graph Neural Networks (GNNs) have enabled the application of more powerful GNN models in the field of molecular chemistry(Xu et al., 2018; Corso et al., 2020; Bodnar et al., 2021; Li et al., 2020; Beaini et al., 2021; Bouritsas et al., 2022), which has proven effective in enhancing the discriminability between representations and capturing underlying chemical semantics. The study of the expressive power of GNNs using the Weisfeiler-Lehman graph isomorphism test has been widely applied in MRL. GIN (Xu et al., 2018), as one of the most representative works, develops a simple and effective architecture based on a multi-perceptron layer (MLP) that has been proven to be as powerful as the WL test. Some works propose improvements in the expressive power of GNNs to address issues related to long-range interactions(Bodnar et al., 2021; Beaini et al., 2021), higher-order structures (Li et al., 2020; Corso et al., 2020) and substructure recognition(Bouritsas et al., 2022) from different perspectives. Unlike traditional message passing mechanisms, Graphormer (Ying et al., 2021) have explored the direct application of Transformers (Vaswani et al., 2017) to graph representation with tailor-made positional encoding. A few research (Yang et al., 2021; Li et al., 2022) focus more on the ad-hoc model design for biochemical tasks, incorporating constraints based on molecular physics and chemical properties.

The second group of MRL is based on *3D geometry*. Given the Cartesian coordinates of molecular conformations, the main objective of these methods is to learn a molecular representation that adheres to fundamental quantum-mechanical principles by incorporating equilibrium constraints for atomistic systems. SchNet (Schütt et al., 2017) introduces continuous-filter convolutional layers for modeling quantum interactions within molecules, while Message Passing Neural Networks (MPNNs) (Gilmer et al., 2017; Liu et al., 2021b) are designed based on the message passing mechanism, where appropriate message and update functions are employed to provide a useful inductive bias and capture different types of 3D information, including distance, angle, and torsion. Subsequent advancements (Gasteiger et al., 2020b; Satorras et al., 2021; Schütt et al., 2021; Gasteiger et al., 2021; Du et al., 2022; Gasteiger et al., 2020a; Du et al., 2023) have focused on further improvements in addressing translation, rotation, and reflection equivariance (E(n)), as well as permutation equivariance, to enhance the molecular property prediction and enable accurate molecular dynamics simulations.

In the advancements of supervised MRL, there has been limited progress in the model designs specifically tailored to the *SMILES string* and *fingerprint* modalities. It is worth noting that the early benchmark models proposed in MoleculeNet (Wu et al., 2018) have maintained their competitiveness over time. With the rise of pre-training research paradigms, there has been promising progress in recent years towards pre-training approaches based on these two modalities (Wang et al., 2019; Ross et al., 2022; Chithrananda et al., 2020; Zhu et al.) as well as the former two. By employing contrastive (Wang et al., 2022; Liu et al., 2021a; Fang et al., 2022; Li et al., 2022) and generative (Hou et al., 2022; Liu et al., 2021a; Zhu et al., 2022) self-supervised strategies, molecular pre-training approaches guide the model training and subsequently facilitate positive transfer to downstream tasks. However, as mentioned in Section 3.3, existing molecular pre-training still suffer from issues such as parameter ossification (Hernandez et al., 2021), necessitating further exploration for more data-efficient and training-efficient models.

D.2. Neural scaling law

The study of neural scaling law can be traced back to early theoretical analyses of bounding generalization error (Haussler, 1988; Ehrenfeucht et al., 1989; Blumer et al., 1989; Haussler et al., 1994). These works, based on assumptions about model capacity and data volume, reveal power-law relationships between the bounds of model generalization error and the amount of data. However, the conclusions drawn from these theoretical studies often yield loose or even vacuous bounds, leading to a disconnection between the theoretical findings and the empirical results of generalization error.

Early follow-on research have investigated empirical generalization error scaling, which represents an initial attempt at

exploring the neural scaling law. Bango and Bill (Banko & Brill, 2001) conduct experiments on a language modeling problem called confusion set disambiguation, using subsets of a large-scale text corpus containing billions of words. Their findings suggest a power-law relationship between the average disambiguation validation error and the size of the training data. Similarly, Sun et al. (Sun et al., 2017) demonstrate that the accuracy of image classification models improves with larger data sizes and conclude that the accuracy increases logarithmically based on the volume of the training data size.

Hestness et al. (Hestness et al., 2017) empirically validate that model accuracy improves as a power-law as growing training sets in various domains, which exhibit consistent learning behavior across model architectures, optimizers and loss functions. However, there exists generalization error plateau in small data region and irreducible error region. With a broader coverage, Michael et al. (Kaplan et al., 2020) present findings that consistently show the scaling behavior of language model log-likelihood loss in relation to non-embedding parameter count, dataset size, and optimized training computation. They leverage these relationships to derive insights into compute scaling, the extent of overfitting, early stopping step, and data requirements in the training of large language models.

In recent years, several investigations of neural scaling laws specific to particular tasks have been conducted (Cherti et al., 2022; Neumann & Gros, 2022; Caballero et al., 2022; Sorscher et al., 2022). Unlike previous research, while power-law relationships hold within specific ranges of data size or model parameter count, certain tasks exhibit unique and uncommon learning behaviors. For instance, only marginal performance gains are expected beyond a few thousand examples in image reconstruction (Klug & Heckel, 2022).