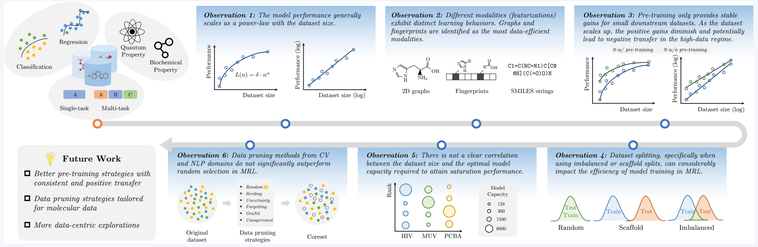
Uncovering Neural Scaling Laws in Molecular Representation Learning

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

[*]Equal Contribution [1]CRIPAC, CASIA [2]UCLA [3]UW [4]Cornell [5]CUHK

TL;DR

We conduct comprehensive experiments from a data-centric perspective to study the learning efficiency of Molecular Representation Learning (MRL) and identify potential avenues for its improvement.



Motivation

Most existing MRL models takes a **model-centric** approach that develops neural architectures and training strategies to improve the expressiveness of the learned molecular representations. However, the influence of varying data scales on the performance of MRL under different learning scenarios is yet to be fully understood.

Drawing upon the successful practice of data-centric AI in the field of NLP and CV, we choose neural scaling laws and data pruning as two essential avenues for exploring how molecular data contributes in MRL.

- Common research objectives in neural scaling laws:
  - Relationship between data quantities and model performance
  - The impact of pre-training
  - Model capacity (parameter size)
- Unique data-oriented challenges in MRL:
  - Molecular data modalities (featurizations)
  - Out-Of-Distribution (OOD) generalizaiton abilities
- Data pruning for MRL:
  - Strategies of sampling representative subsets

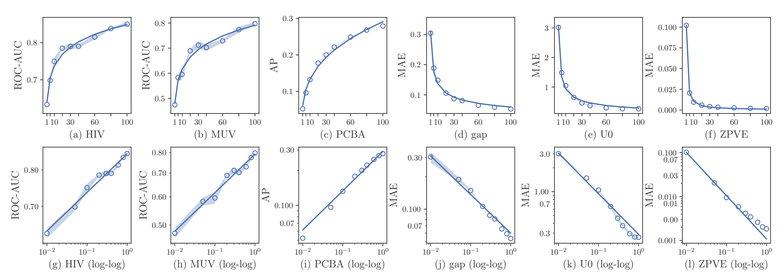
Paper **Code** **Poster**

Design Dimensions

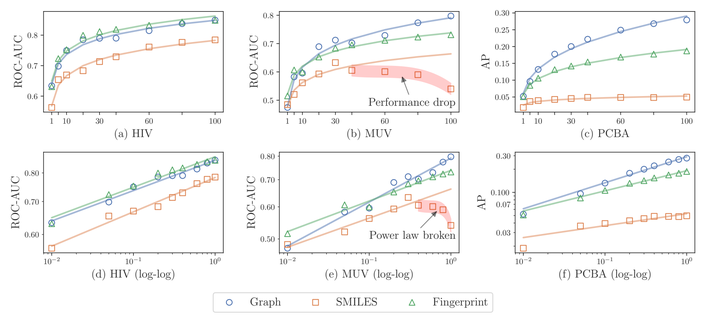
- **Data Modalities (Featurizations)**: 2D topology graphs, 3D geometry graphs, Morgan fingerprints, and SMILES strings
- **Dataset Splitting**: Random splitting, imbalanced splitting, and scaffold splitting
- **The Role of Pretraining**: Training from scratch vs fine-tuning with pre-trained models
- **Model Capacity**: Model parameters determined by the number of layers (depth D) and hidden units (width W)
- **Data Pruning**: Random pruning, Herding, Entropy, Least Confidence, Forgetting, GraNd and k-means

Empirical Results and Observations

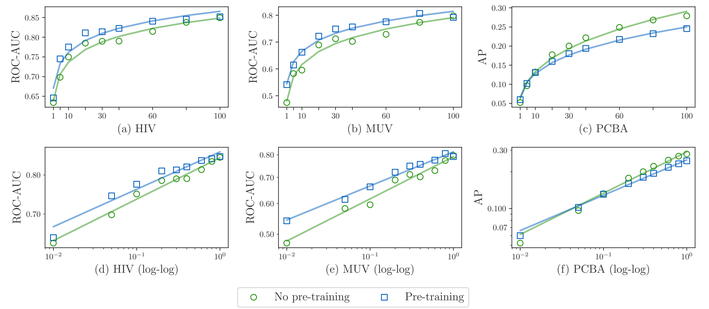
Obs. 1. The model performance on all datasets adheres to the power law relationship as the data quantity varies. This pattern remains consistent in both low- and high-data regimes.



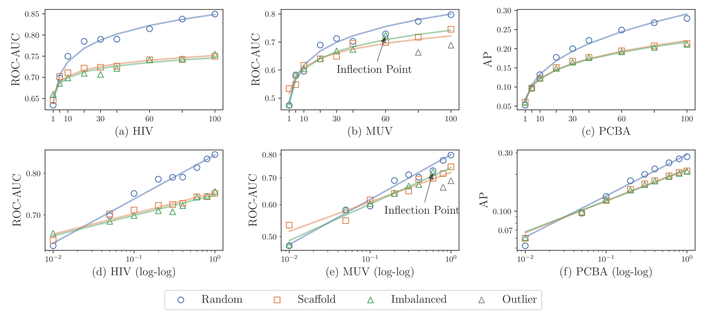
Obs. 2. Different modalities (featurizations) exhibit distinct learning behaviors in MRL. 2D graphs and fingerprints stand out as the most efficient modalities for MRL.



Obs. 3. Pretrained models benefit from better initialization, but they demonstrate reduced data efficiency and could lead to negative transfer.



Obs. 4. While power law remains consistent with different data splitting, uniform setting has a higher data utilization efficiency compared to OOD settings.



Obs. 5. There is not a clear correlation between the dataset size and the optimal model capacity required to attain saturation performance.

Obs. 6. Data pruning methods from CV and NLP domains do not significantly outperform random selection, which highlights the need for developing data pruning strategies specifically tailored to molecular data.

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
k-means	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
k-means	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