1 Submission15320: Exploring Molecular Pretraining Model at Scale

1.1 Pretrain Time Complexity and GPU Resources

We utilized a computational cluster comprising 64 NVIDIA A100 GPUs, each equipped with 80GB of HBM2 memory. The GPUs were interconnected via a high-speed Nvidia Infiniband fabric, offering 400 Gbps bandwidth for inter-GPU communication. The details of each model size are listed in table 1.

Table 1: Training Time of Uni-Mol2 at different scale

Params	Compute Resouce (GPUs)	Training Time (GPU hours)
84M	32	2585.6
164M	32	5120
310M	32	7680
570M	64	13824
1.1B	64	30720

1.2 Additional QM9 Property Experiment Results

We have undertaken additional experiments utilizing the QM9 dataset to assess molecular properties. Our findings indicate that the model's performance consistently improves with an increase in model size. This scalability suggests that employing larger models will further enhance the accuracy and reliability of molecular property predictions, thereby increasing the utility of our approach for practical applications in the field.

Table 2: Mean absolute error(MAE, ↓) results on QM9 Dataset

Model	HOMO / LUMO / GAP	alpha	C_v	mu	R^2	ZPVE	U_0	U	G	H
Uni-Mol2 310M	0.0036(1e-05)	0.315(0.003)	**0.143(0.002)**	0.092(0.0013)	4.672(0.245)	0.0005(1e-05)	6.149(0.161)	6.009(0.217)	5.564(0.124)	6.3547(0.1048)
Uni-Mol2 570M	0.0036(2e-05)	0.315(0.004)	0.147(0.0007)	0.089(0.0015)	4.523(0.080)	0.0005(3e-05)	5.421(0.432)	5.396(0.384)	5.178(0.1842)	5.9046(0.23469)
Uni-Mol2 1.1B	**0.0035(1e-05)**	**0.305(0.003)**	0.144(0.002)	**0.089(0.0004)**	**4.265(0.067)**	**0.0005(8e-05)**	**4.512(0.188)**	**5.774(0.165)**	**3.921(0.076)**	**5.880(0.293)**
Improvment	2.78%	3.17%	-0.6%	3.26%	8.71%	0%	26.6%	3.9%	29.5%	7.5%

1.3 Biogen ADME Dataset

The Biogen ADME dataset focuses on the evaluation of drug metabolism and pharmacokinetics (DMPK) properties, specifically assessing absorption, distribution, metabolism, and excretion (ADME) characteristics of potential drug candidates. Table 3 illustrates the predictive performance of the Uni-Mol2 model regarding these ADME properties.

Table 3: Mean absolute error(MAE, ↓) results on Biogen ADME Dataset

Model	HCLint-1	PERM-1	SOLU-1
Uni-Mol	0.3085(0.003)	0.2886(0.008)	0.3167(0.010)
Uni-Mol2 84M	0.3117(0.007)	0.2853(0.006)	0.325(0.006)
Uni-Mol2 1.1B	**0.3045(0.003)**	**0.3043(0.003)**	**0.3062(0.005)**