Supplementary Materials: Graph DiT-UQ

Additional Methods

PPO Implementation Details

Our PPO implementation uses: - Learning rate: 3e-4 - Clip ratio: 0.2 - Value function coefficient: 0.5 - Entropy coefficient: 0.01

Multi-Objective Reward Function

$R = \lambda\_QED \cdot r\_QED + \lambda\_dock \cdot r\_dock + \lambda\_SA \cdot r\_SA + \beta \cdot \sqrt{u}$

Where: - $\lambda\_QED = 0.4$, $\lambda\_dock = 0.4$, $\lambda\_SA = 0.2$ - β controls uncertainty exploration strength

Additional Results

Hyperparameter Sensitivity

Parameter	Value Range	Best Value
β (uncertainty)	0.05-0.2	0.2
Learning rate	1e-4 to 1e-3	3e-4
Batch size	32-128	64

Training Convergence

- Convergence achieved in 20 iterations
- No overfitting observed
- Stable reward improvement throughout training

Software Dependencies

- PyTorch 2.0+
- RDKit 2023.9.5
- PyTorch-Geometric 2.4.0
- Weights & Biases for logging

Computational Resources

- GPU: NVIDIA RTX 4090 (24GB)
- Training time: ~30 minutes
- Memory usage: <8GB
- Carbon footprint: 0.14 μg CO_2 per 10k molecules