May the Force be with You: Unified Pre-Training for 3D Molecular Conformations

Rui Feng*, Qi Zhu**, Huan Tran*, Binghong Chen*, Aubrey Toland*, Rampi Ramprasad*, Chao Zhang*

* Georgia Institute of Technology, Georgia, United States

** University of Illinois Urbana-Champaign, Illinois, United States

{rfeng, huan.tran, binghong, artoland, rampi.ramprasad, chaozhang}@gatech.edu, qi.zhu.ckc@gmail.com

Introduction

- We develop a **pre-training method** for **3D molecular conformations**.
- Common pre-training strategy: self-supervised de-noising, such as Pre-training-via-Denoising (Zaidi et al., 2022) and UniMol (Zhou et al., 2022).
- De-noising can be thought of an approximation to learning atomic forces. Forces are defined as:
  $F = -\nabla_x E(x)$, x: atomic positions, E: potential energy.
- **However,** this assumption would only be true for *equilibrium* data, i.e. 3D molecular conformations at zero potential energy.
  - A large amount of non-equilibrium data during simulations and optimizations do not fit this description;
  - The approximation is not necessarily accurate and lacks physical information.
- **Furthermore,** existing machine learning for molecules predominantly focus on **extensive training on a single domain**, limiting practical usability and encouraging overfitting.
- Extension of pre-training to more available data, both equilibrium and off-equilibrium, is largely unexplored.
- **We incorporate both equilibrium and off-equilibrium data in a unified representation learner** by a force-centric training paradigm.

Our Contributions

- Introduced a novel force-centric pretraining paradigm for molecular conformations, unifying equilibrium and off-equilibrium data.
- Developed a model that enhances molecular dynamics (MD) simulations, achieving high accuracy and efficient simulation.
- Provided a diverse set of DFT simulation data for polymers, aiding in the study of polymer properties and molecular forces modeling.

ET-OREO Pre-training Objective

$$\mathbb{E}_{x\sim\mathcal{E}}\Big[\underbrace{\|\nabla_{\mathbf{r}_x}\Phi(\mathbf{r}_x)\|_2^2}_{\text{zero-force regularization}} + \underbrace{\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}\left[\|\nabla_{\mathbf{r}_x}\Phi(\mathbf{r}_x-\varepsilon)-\varepsilon\|_2^2\right]}_{\text{de-noising equilibrium}}\Big] + \mathbb{E}_{x\subset\mathcal{F}}\Big[\underbrace{\|F(\mathbf{r}_x)-\nabla_{\mathbf{r}_x}\Phi(\mathbf{r}_x)\|_2^2}_{\text{forces optimization}}\Big]$$

- **Assume:** Equivariant Transformer Φ, coordinates $\boldsymbol{r_x} \in \boldsymbol{R}^{n_{atoms}\times 3}$.
- **Non-equilibrium** ⟹ High energy ⟹ DFT Forces
- **Equilibrium** ⟹ Low energy ⟹ Zero Forces
- **Perturbed equilibrium** ⟹ High energy ⟹ Approximate Forces
- **Why?**
  - Pre-train the model with the physics-informed interatomic relations by forces;
  - Unify the training objective for all data with one physical principle;
  - De-noising objective helps explore the potential energy surface.

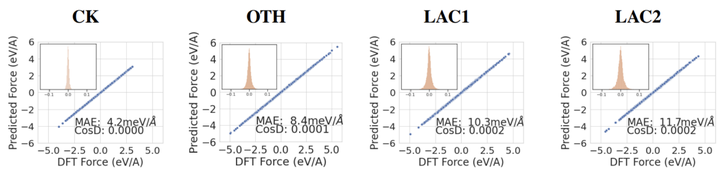
Experiment Results

- **ET-OREO** consistently outperforms baseline models in terms of force accuracy, molecular dynamics simulation accuracy, and simulation robustness.

Molecule	Metric	DimeNet	GemNet-T	GemNet-dT	NequIP	TorchMDNet	ET-ORO	ET-OREO
Aspirin	Force (↓)	10.0	3.3	5.1	2.3	7.4	4.2	**1.0**
	Stability (↑)	$54_{(12)}$	$72_{(50)}$	$192_{(132)}$	$300_{(0)}$	$102_{(45)}$	$94_{(42)}$	$300_{(0)}$
	$h(r)$ (↓)	$0.04_{(0.00)}$	$0.04_{(0.02)}$	$0.04_{(0.01)}$	$0.02_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.02_{(0.00)}$
Ethanol	Force	4.2	2.1	1.7	1.3	5.6	3.1	**1.0**
	Stability	$26_{(10)}$	$169_{(98)}$	$300_{(0)}$	$300_{(0)}$	$121_{(34)}$	$300_{(0)}$	$300_{(0)}$
	$h(r)$	$0.15_{(0.03)}$	$0.10_{(0.02)}$	$0.09_{(0.00)}$	$0.08_{(0.00)}$	$0.12_{(0.01)}$	$0.10_{(0.00)}$	$0.03_{(0.00)}$
Naphthalene	Force	5.7	1.5	1.9	1.1	3.3	2.0	**0.9**
	Stability	$85_{(68)}$	$8_{(2)}$	$25_{(10)}$	$300_{(0)}$	$50_{(20)}$	$25_{(9)}$	$300_{(0)}$
	$h(r)$	$0.10_{(0.01)}$	$0.13_{(0.00)}$	$0.12_{(0.01)}$	$0.12_{(0.01)}$	$0.12_{(0.00)}$	$0.11_{(0.00)}$	$0.03_{(0.00)}$
Salicylic Acid	Force	9.6	4.0	4.0	1.6	4.7	2.5	**0.9**
	Stability	$73_{(82)}$	$26_{(24)}$	$94_{(109)}$	$300_{(0)}$	$60_{(69)}$	$94_{(58)}$	$300_{(0)}$
	$h(r)$	$0.06_{(0.02)}$	$0.08_{(0.04)}$	$0.07_{(0.03)}$	$0.03_{(0.00)}$	$0.06_{(0.02)}$	$0.05_{(0.01)}$	$0.02_{(0.00)}$

Table 2: Simulation results on MD17. For all results, force MAE is reported in the unit of [meV/Å], and stability is reported in the unit of [ps]. The distribution of interatomic distances $h(r)$ MAE is unitless. FPS stands for frames per second. For all metrics (↓) indicates the lower the better, and (↑) indicates the higher the better. The first group of methods is taken from [8]. The second group of methods is our new baselines, including TorchMDNet [8], ET-ORO, and ET-OREO. These models share the same architecture and have the same FPS.

- **ET-OREO** maintains high force accuracy in molecular dynamics simulations.



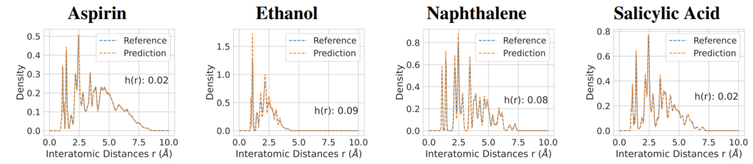
(a) Forces correlation of ET-OREO to DFT forces during simulations.

Dataset	# Conformations	Equilibrium	Off-equilibrium
PCQM4Mv2	3,378,606	✓	✗
ANI1x	4,956,005	✓	✓
MD17	3,611,115	✗	✓
poly24	3,851,540	✗	✓
Total	15,718,279	✓	✓

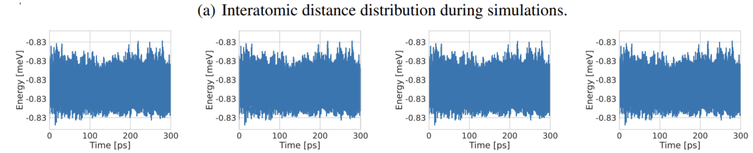
Table 1: Datasets used in our model pre-training process.

Nice properties of atomic forces:

- They are physically well-defined observable, i.e., the force acting on an atom is determined solely and uniquely from its local environment, defined as the real-space distribution of its neighboring atoms;
- They are generalizable across various molecules in the sense that atoms from different molecules that have the same local environment should experience the same atomic forces;
- They can unify equilibrium and off-equilibrium data, characterizing the whole landscape of the potential energy surface;
- Empirically, forces across different datasets calculated different ab initio methods have relatively similar distributions, among other chemical properties.



(a) Interatomic distance distribution during simulations.



(b) Potential energy curve during simulations.

ET-OREO improves property prediction on QM9 by ~30%, on par with NoisyNodes.

	TorchMDNet	NoisyNode	ET-OREO
$\varepsilon_{\text{HOMO}}$	20.3	**15.6**	16.8
$\varepsilon_{\text{LUMO}}$	17.5	**13.2**	14.5
$\Delta\varepsilon$	36.1	**24.5**	26.4

Table 4: Fine-tuning on HOMO-LUMO properties on QM9. Metrics are MAE in meV.

Off-equilibrium data helps more with simulation and optimization than property prediction.