Do Machine Learning Force Fields Overcome Adversarial Attacks?

Declan Chan — Engineering Science 2T9, U of T
AutoDIAL Research Group
Department of Materials Science & Engineering
UnERD 2026

UNIVERSITY OF TORONTO

Why Do We Use Machine Learning Force Fields?

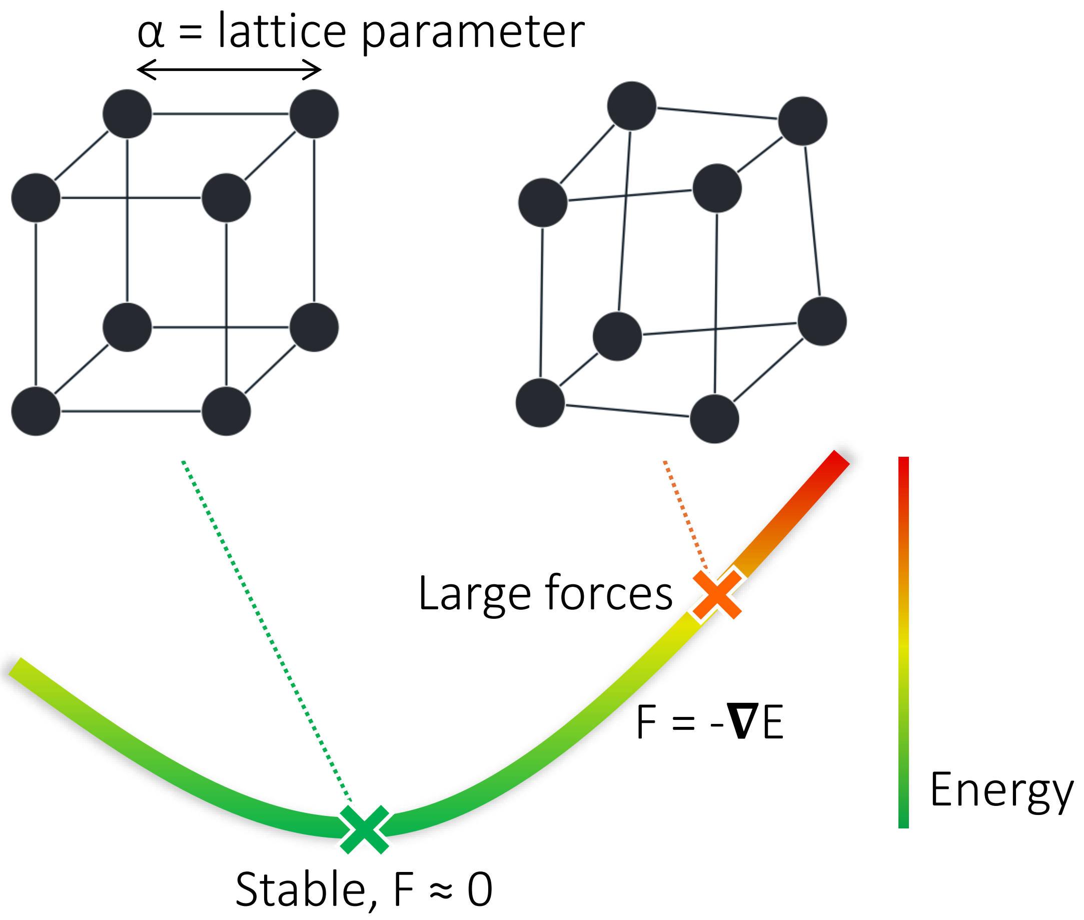
Screen many atomic structures for materials discovery

→ **D**ensity **F**unctional **T**heory

To save time, cost & resources

→ **M**achine **L**earning **F**orce **F**ields

MLFFs learn the DFT PES to predict energies & forces efficiently.



What Happens When MLFFs Leave Their Comfort Zone?

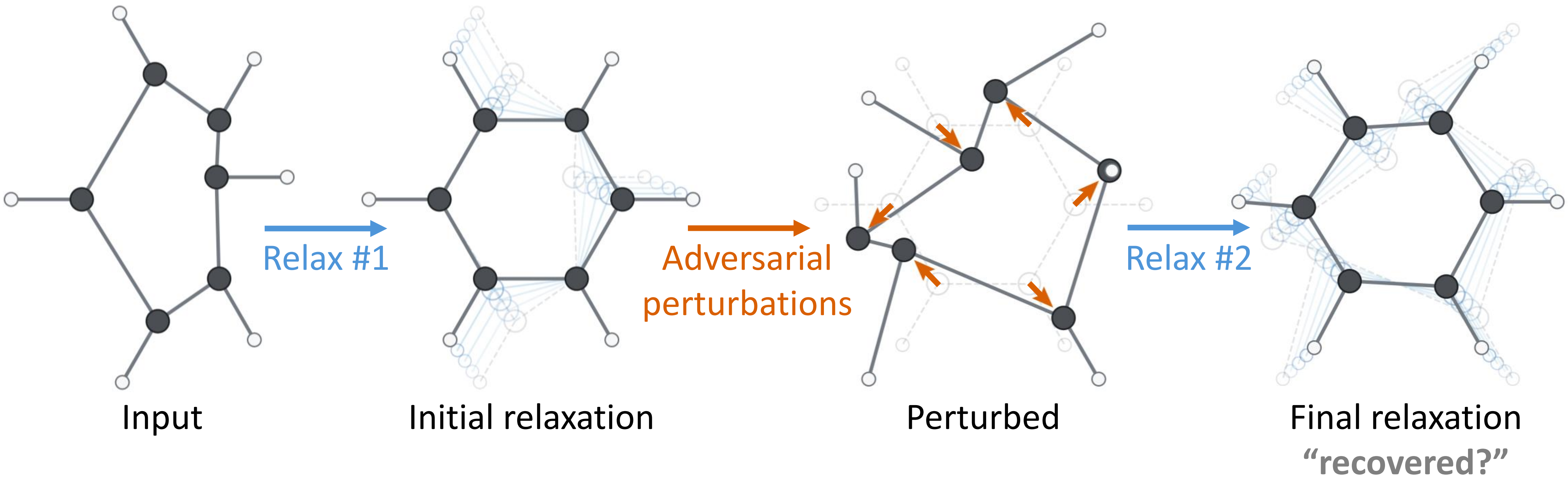
Risk → Out-of-distribution structures may lead to unreliable predictions.

Conventional benchmarks use static errors against DFT but miss:
unphysical artifacts and structural stability.

Low errors can still lead to unstable simulations: *Fu, X., et al. (2023). Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. Transactions on Machine Learning Research.*

Assessing MLFF Robustness Using Perturbations

How do MLFFs relax structures from adversarial perturbations?

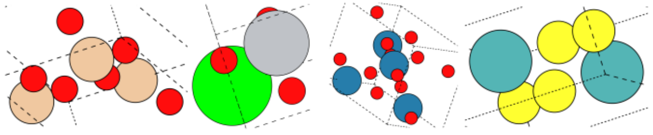


Workflow for relaxation and perturbations: *Dale, A. S., Wan, H., & Chan, D. (2026). MLFF Attack: A library for attacking MLFF models. GitHub. https://github.com/Trustworthy-AI-Tools-for-Science/mlff_attack*

Experimental Design: 20 Structures and 2 MLFFs

__20 Inorganic Crystals__

From *Materials Project*:



SiO_2	$SrTiO_3$	ReO_3	MoS_2
mp-6922	mp-5229	mp-190	mp-2815

Initial relaxation ↓

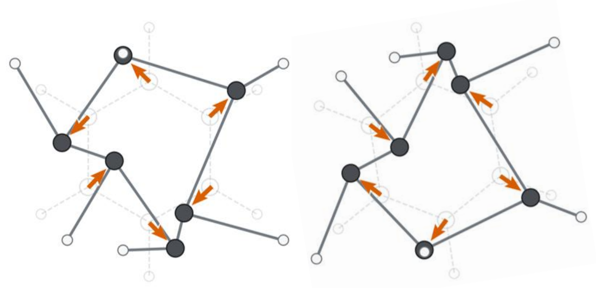
__MLFFs__

Model 1: **UMA**—universal models for atoms

Model 2: **MACE-MH**—multi-head message passing

Perturb structures →

__Perturbations__



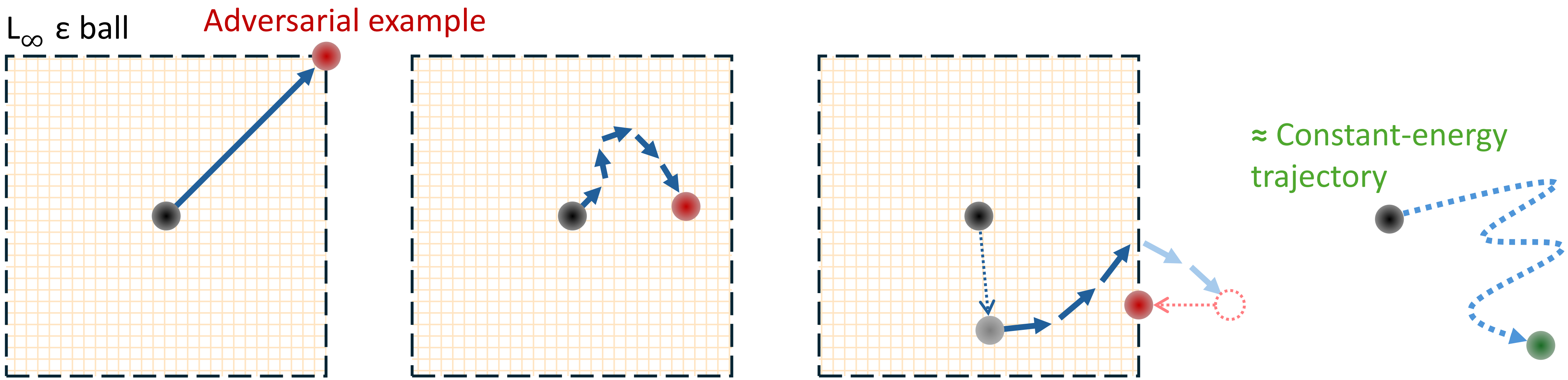
Final relaxation ↑

__MLFFs__ ≈ __DFT__

Final recovered structure = initial structure?

Measure changes in energy, forces, topology

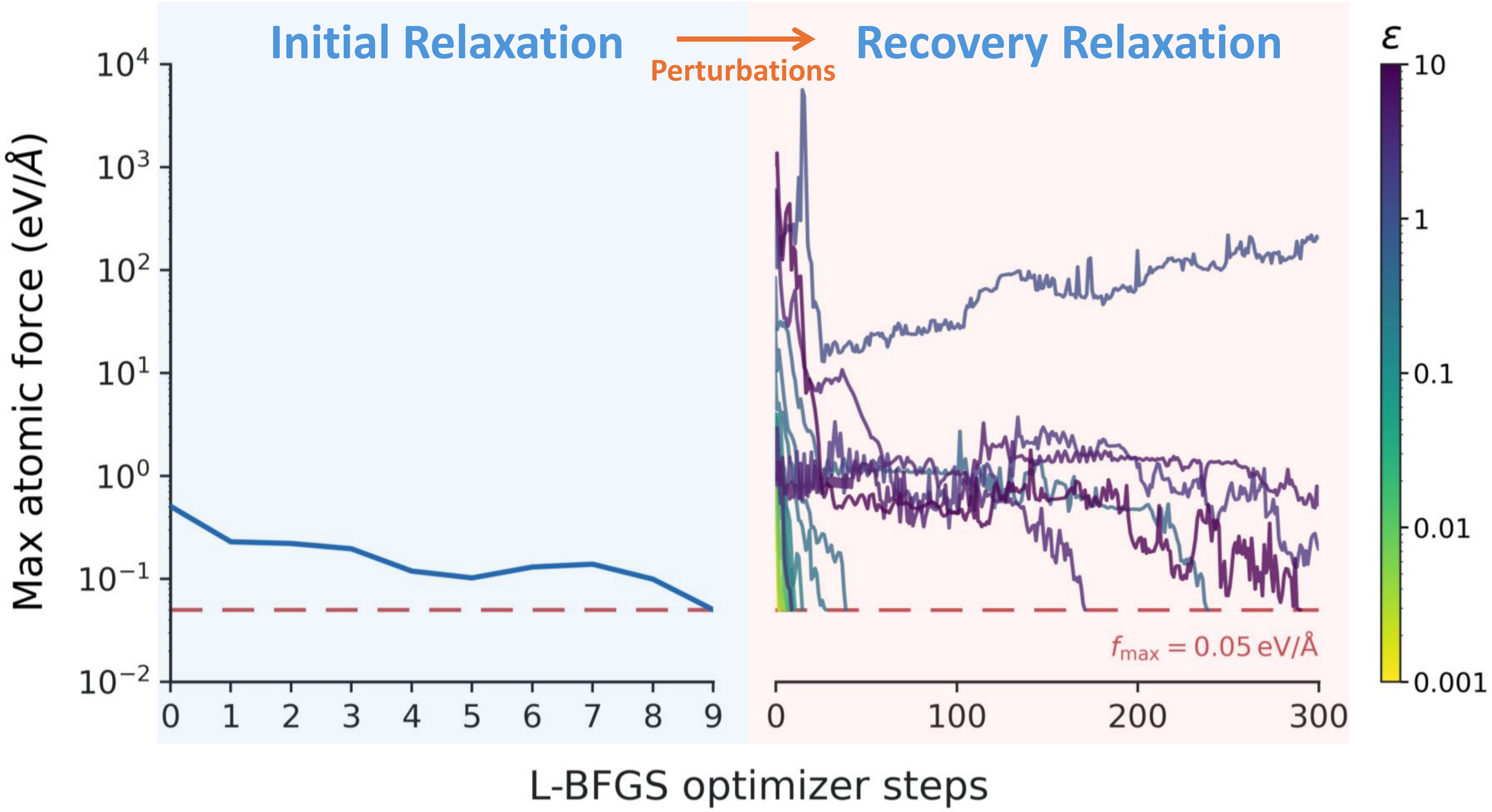
Robustness Tests Through Adversarial Perturbations and Contour Exploration



FGSM	**I-FGSM**	**PGD**	**Contour exploration**
Fast Gradient Sign Method	Iterative Fast Gradient Sign Method	Projected Gradient Descent after random start	Non-adversarial baseline

Adversarial robustness tests: *Madry, A., et al. (2017). Towards deep learning models resistant to adversarial attacks. arXiv. https://arxiv.org/abs/1706.06083*

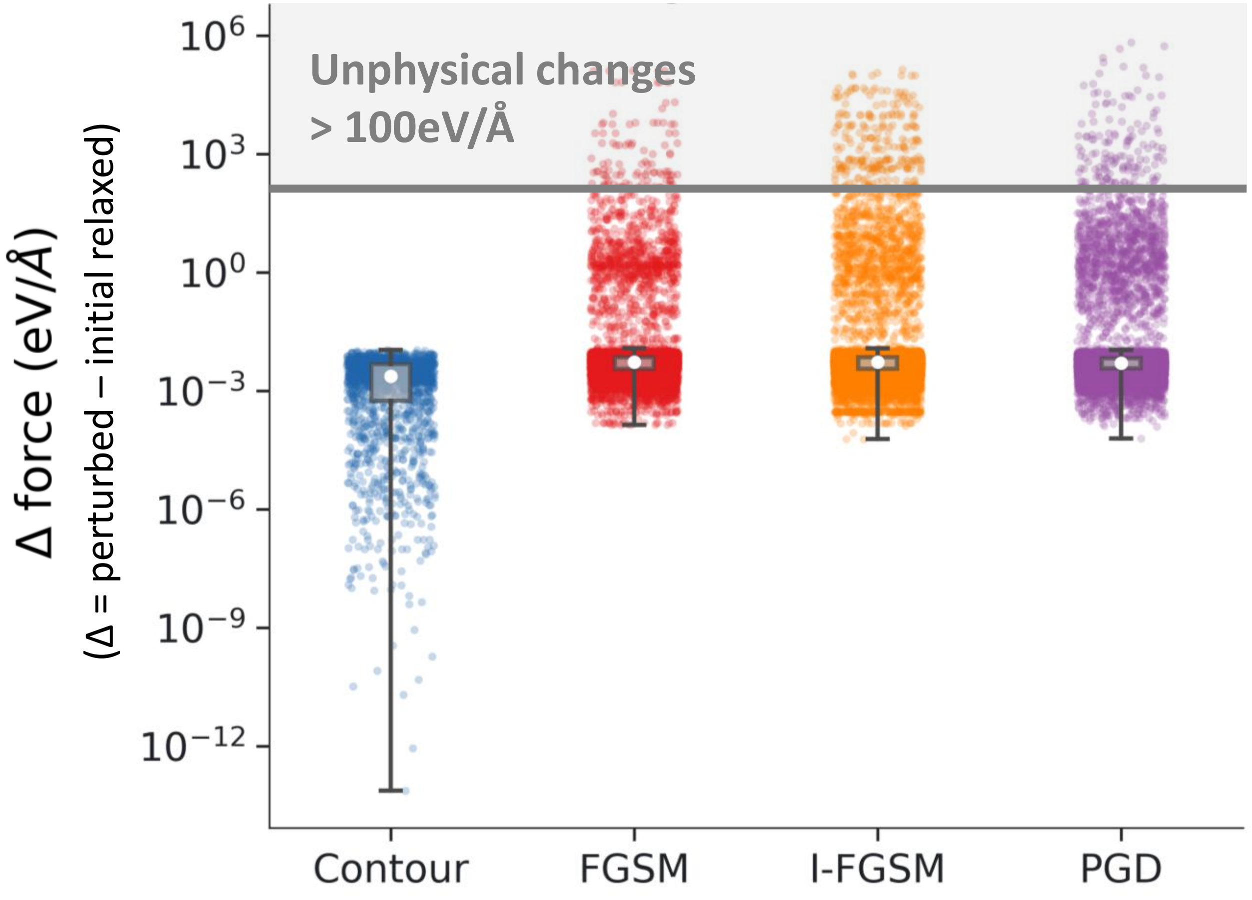
Testing MLFF Relaxation Recovery After Perturbations



Experiments repeated across:	
dtype	Seed
float32	42
	43
	44
float64	45
	46

Ran experiments on HPC with 20 structures × 3 attacks × 21 ε's = 1260 cases per model:
Digital Research Alliance of Canada. Fir Cluster.

Adversarial Gradients Expose Unphysical MLFF Predictions



Convergence rate:

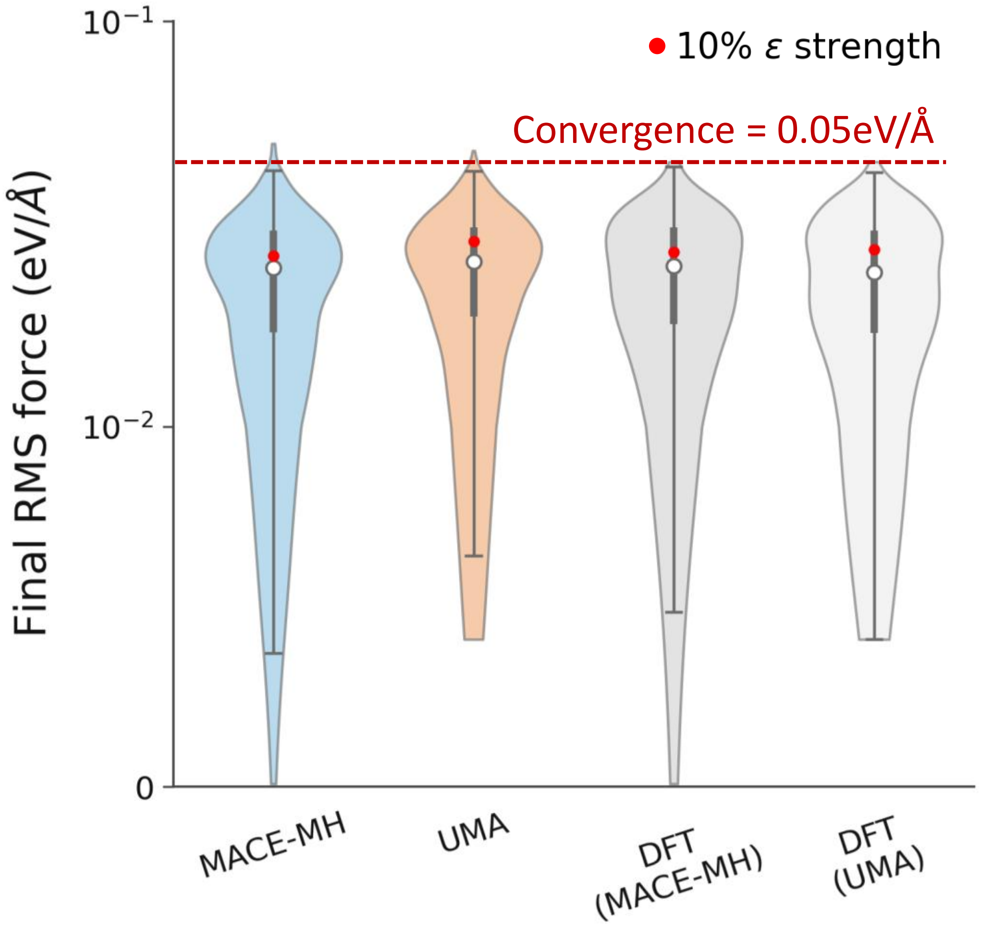
94% adversarial

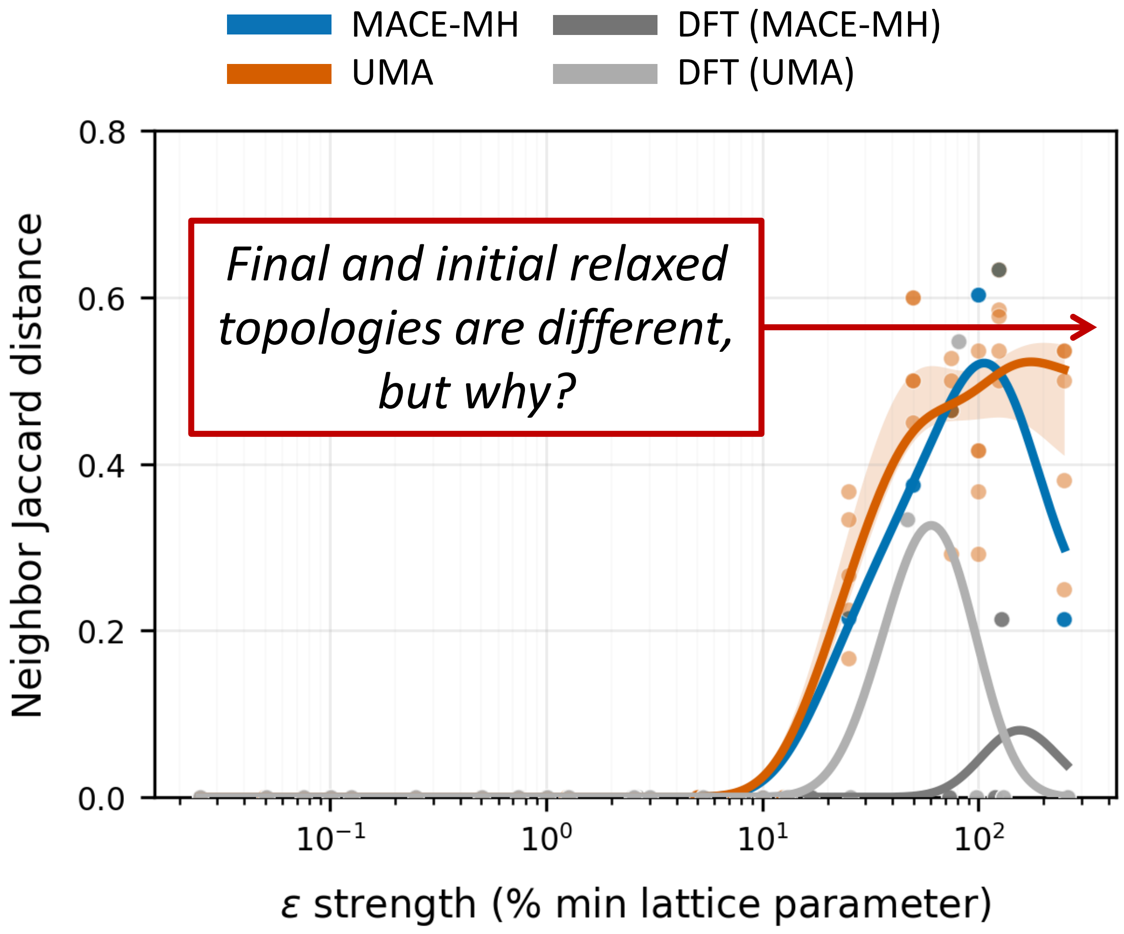
100% non-adversarial

After perturbations, relaxations can converge to low forces—but do they recover structurally?

> 100eV/Å are unphysical and cause simulation failures: *Mannan, S., et al. (2025). UniFFBench: Evaluating universal machine learning force fields against experimental measurements. arXiv. https://arxiv.org/abs/2508.05762*

Although Forces Converge, Topology Fails to Recover > 10% ε

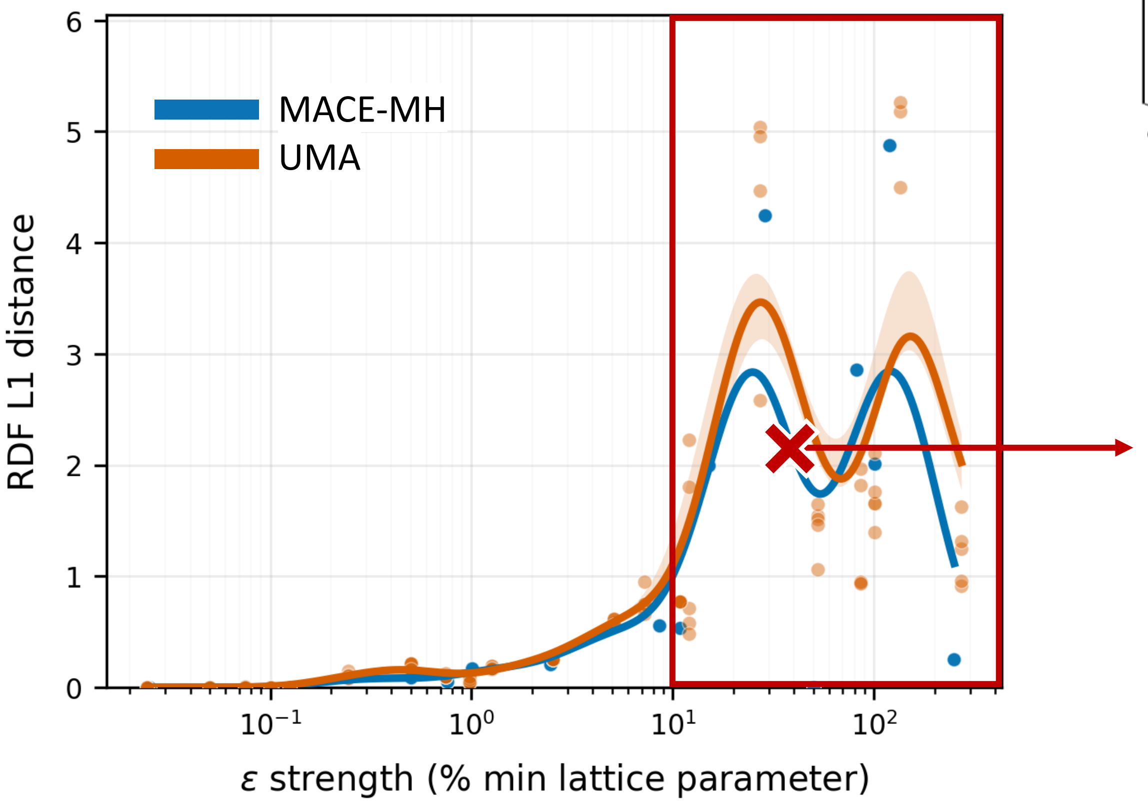




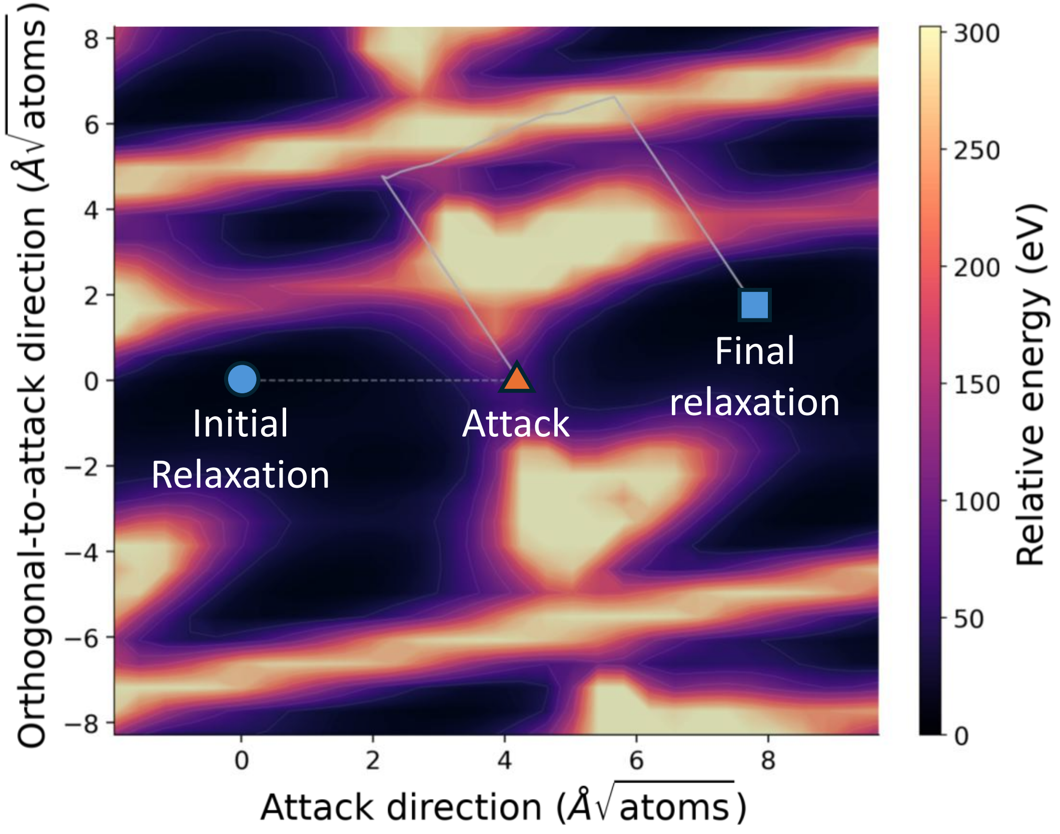
DFT (MACE-MH) = Initial relaxation by MACE-MH & final relaxation by DFT
DFT (UMA) = Initial relaxation by UMA & final relaxation by DFT

Structural Recovery Fails Because MLFFs Relax into Different Basins

Geometric Differences in MLFF Recovery and DFT



Example MLFF PES for ε > 10% (hBN)



2D PES: *Heidenreich, H. (2025). Implementing the Müller-Brown potential in PyTorch. https://hunterheidenreich.com/posts/muller-brown-in-pytorch/*

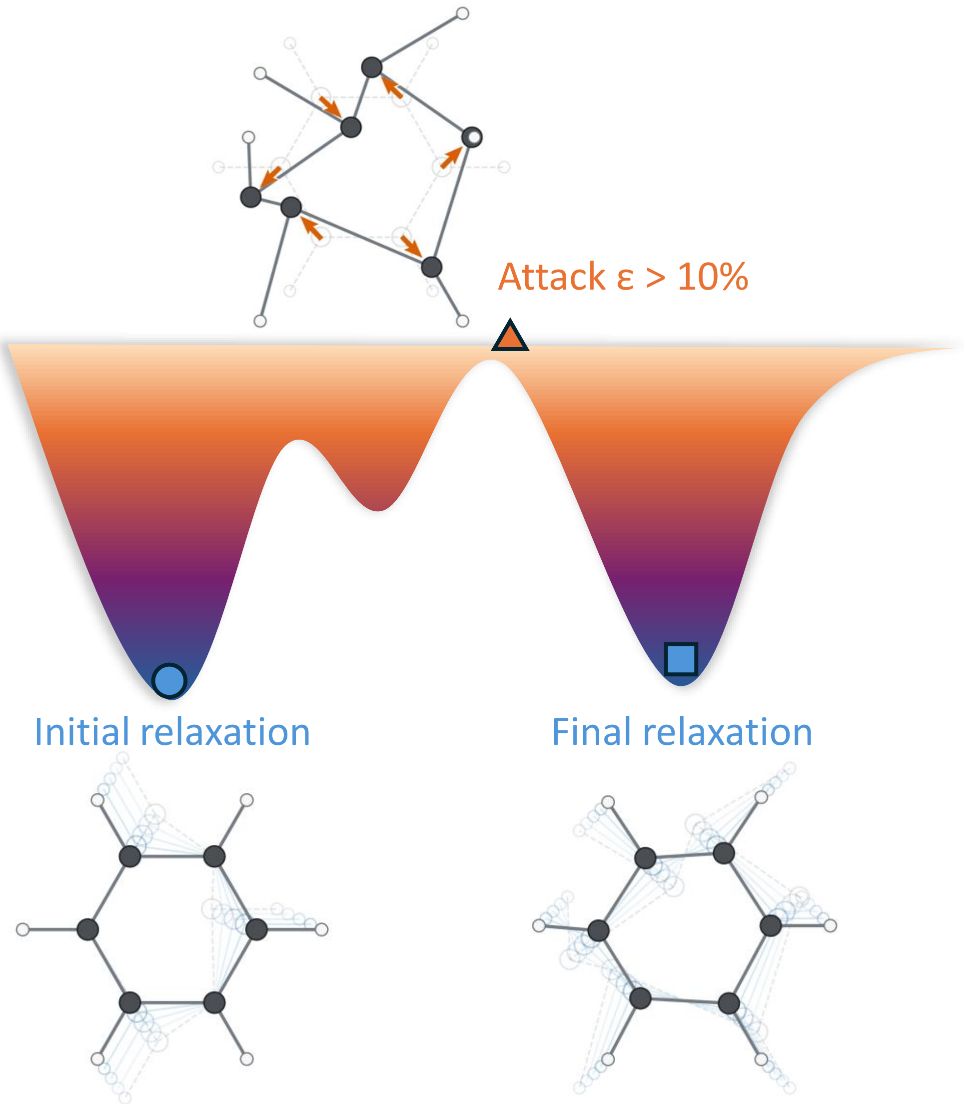
Towards More Reliable MLFFs for Materials Discovery

Takeaways:

- Low forces ≠ structural recovery
- Adversarial gradients redirect relaxations
- 10% perturbation threshold observed in the dataset.

Ongoing:

- More MLFFs & structures
- Dynamic stability with phonons



Thank You

Acknowledgements

- Professor Jason Hattrick-Simpers
- Dr. Ashley Dale, Dr. Dan McHaffie, and Dr. Hao Wan
- AutoDIAL Research Group
- Department of Materials Science & Engineering
- Digital Research Alliance of Canada

UNIVERSITY OF TORONTO