Ekin Dogus Cubuk, Samuel S. Schoenholz

Google Research, Brain Team, {cubuk, schsam}@google.com

Adversarial Forces of Physical Models

- Simulating quantum mechanics is expensive.
- A hierarchy of approximate models are commonly used in biology, chemistry, and materials science.
- Such ML approximations are usually assessed based on their average-case performance.
- We show that there is a well defined sense of adversarial direction that governs the worst-case behavior for these models.
- Unlike in other contexts, where adversarial examples are scarce absent malicious intervention, in physical systems we show that the laws of physics can naturally lead the model to move in adversarial directions.
- Surprisingly, we find that these adversarial directions can exist even for traditional, analytic force fields such as the BKS potential.

Adversarial directions of physical models

Whether a model is trained with this loss or not, it is an important error measure for the validity of an approximate energy model:

$$\mathcal{L} = \sum_j \left[E^{\text{Acc}}(\vec{R}_j) - E^{\text{App}}(\vec{R}_j)\right]^2$$

where E^{Acc} is the ground truth energy, and E^{App} is the approximate energy. Then it follows that adversarial direction that maximizes this error measure is:

$$\begin{aligned}\vec{A}(\vec{R}_j) \equiv \nabla_{\vec{R}_j}\mathcal{L} &= \nabla_{\vec{R}_j}\left(E^{\text{Acc}}(\vec{R}_j) - E^{\text{App}}(\vec{R}_j)\right)^2 \\ &= \underbrace{2\left(E^{\text{Acc}}(\vec{R}_j) - E^{\text{App}}(\vec{R}_j)\right)}_{\text{scalar}}\left(\vec{F}^{\text{App}}(\vec{R}_j) - \vec{F}^{\text{Acc}}(\vec{R}_j)\right)\end{aligned}$$

which might point in the same direction as F^{App} if F^{Acc} is small (as would be the case near the local minima of the ground truth energy landscape). Since thermalized and quasi-thermalized systems are exponentially more likely to be close to their local minima, such physical models might move in the same direction as their adversarial direction as defined above. Next, we will empirically measure the cosine angle between the model force and the adversarial direction:

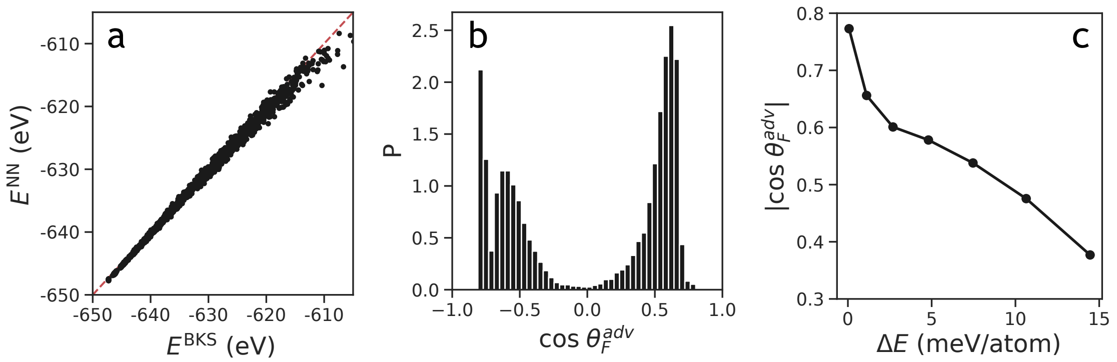
$$\cos\theta_F^{adv} = \frac{\vec{A}(\vec{R}) \cdot \vec{F}^{\text{App}}(\vec{R})}{||\vec{A}(\vec{R})||\ ||\vec{F}^{\text{App}}(\vec{R})||}$$

We start by showing that moving the atoms in this adversarial direction does increase the model error compared to a random direction (where ε is total distortion magnitude in Å^2):

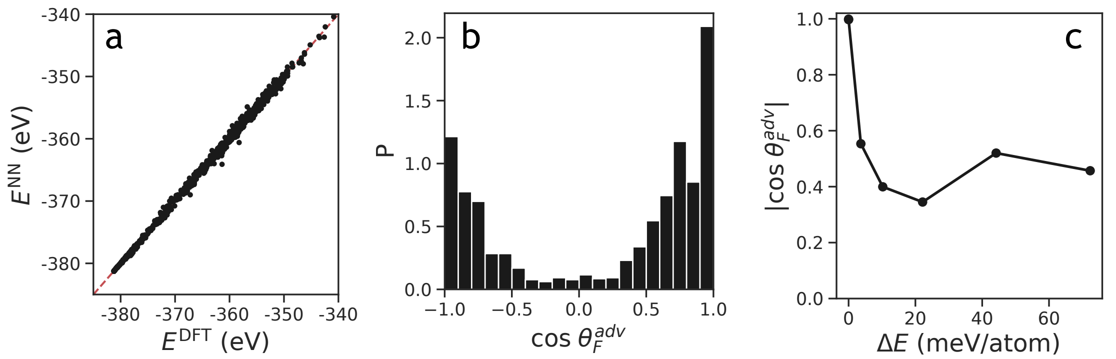
	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$
Random direction	5.1	5.3	5.9	6.8
Adversarial direction	5.1	15.4	31.9	42.0

For a variety of physical models, their force is correlated with their adversarial direction:

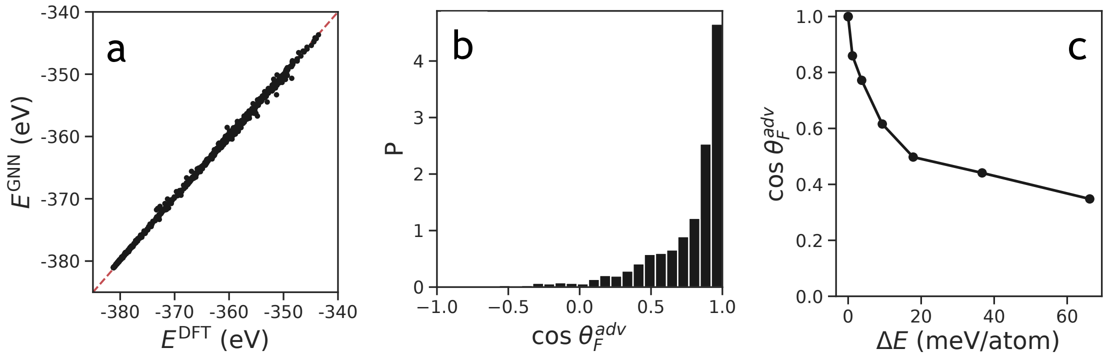
Behler-Parrinello neural network approximating BKS (analytic physics model)



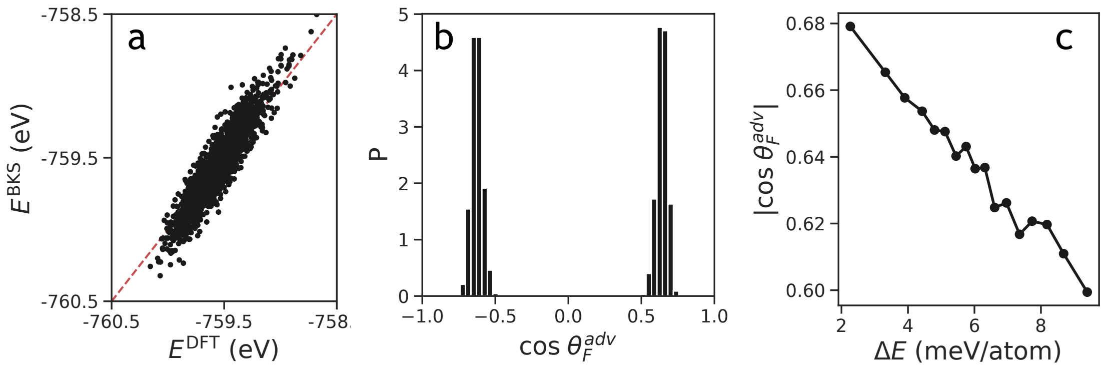
Behler-Parrinello neural network approximating DFT



Graph neural network approximating DFT



BKS (analytic physics model) approximating DFT

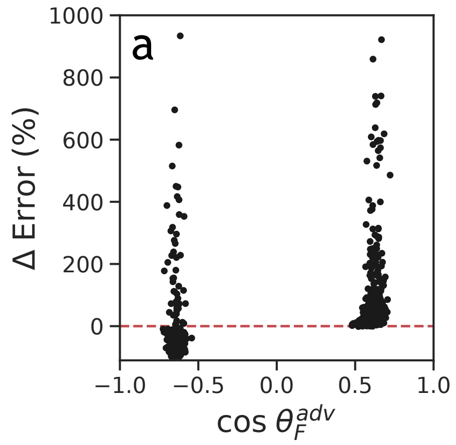


a)
ground truth energy
vs
predicted energy

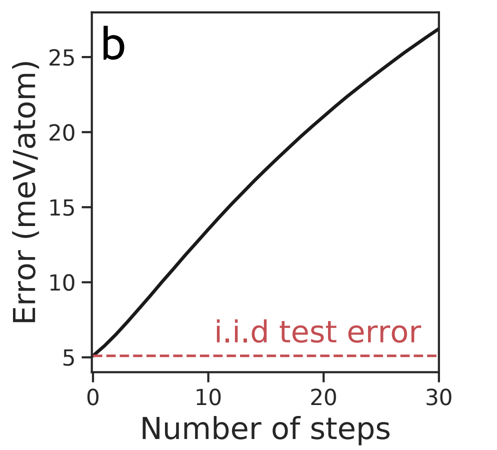
b)
histogram of
$\cos\theta_F^{adv}$

c)
$\cos\theta_F^{adv}$
vs
energy above
local minimum

How does this effect the use of force fields?
Let's look at how error grows with structural relaxation (gradient descent in energy landscape)



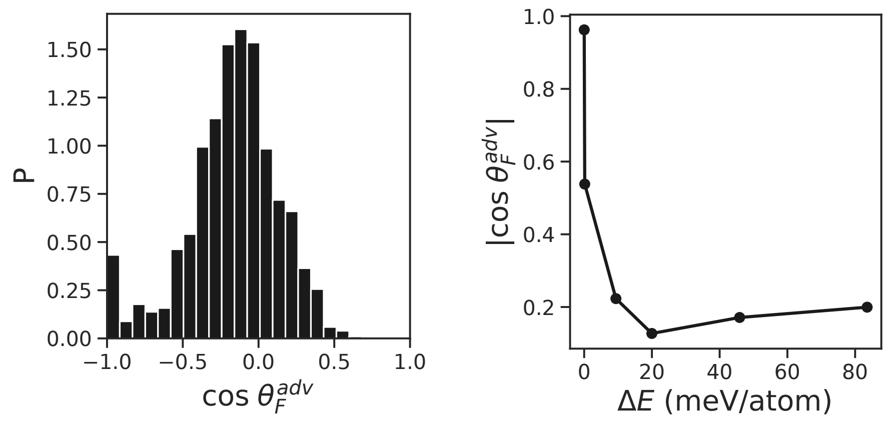
Error after the first step, as a function of
$\cos\theta_F^{adv}$



Error grows with more steps taken with the energy model

Potential solutions

1) Adversarial training
2) Augmenting training by taking a step in F^{App} (similar to (1))
3) Augmenting training by taking a step in F^{Acc} (similar to (2))
4) Related to above, training on E^{Acc} *and* F^{Acc}, which is commonly used already. We see that it does reduce the correlation between adversarial directions and model force significantly:



Graph neural network approximating DFT, trained on energies *and* forces. Overlap is still high for only the lowest energy configurations. Further work is needed if these adversarial directions pose a significant obstacle to the use of force fields.