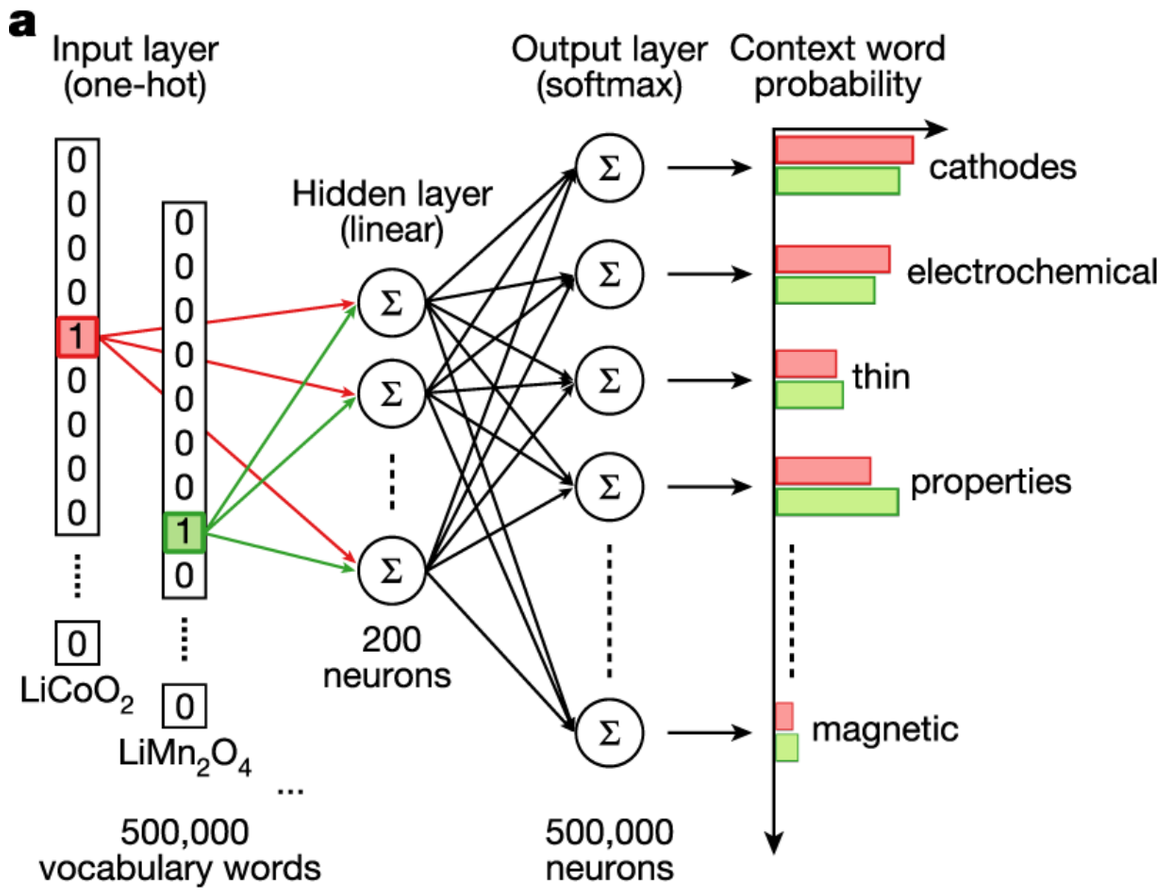
Applications of ML in Chemistry

Key points:

- Get some ideas for what ML can be used in chemistry
- Advanced and recent examples

Word Embeddings: Learning on > 1 Million of Abstracts

single layer NN is trained to predict all context words for the given target word.



for similar words the context words are the same

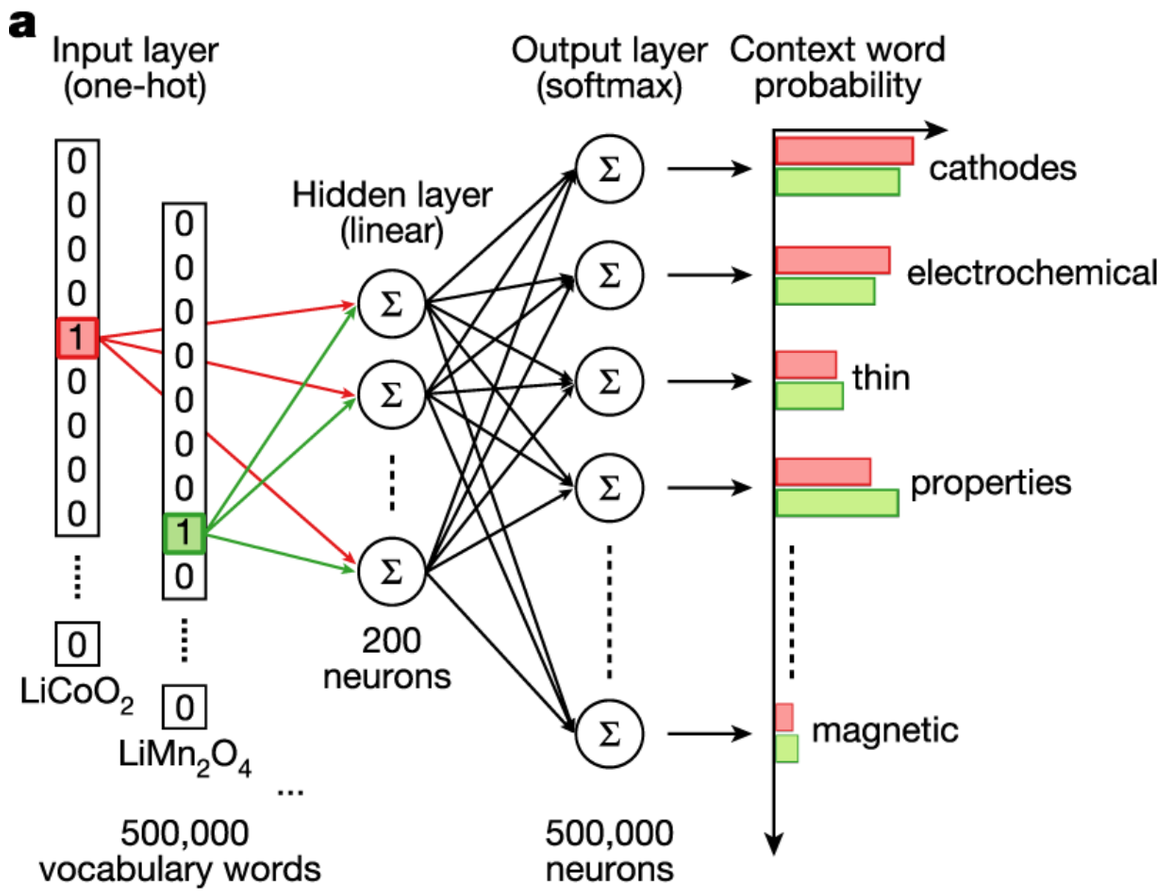
Tshitoyan, V.; *et al*.. *Nature* **2019**, *571* (7763), 95.

Perspective: Isayev, O. *Nature* **2019**, *571* (7763), 42–43.

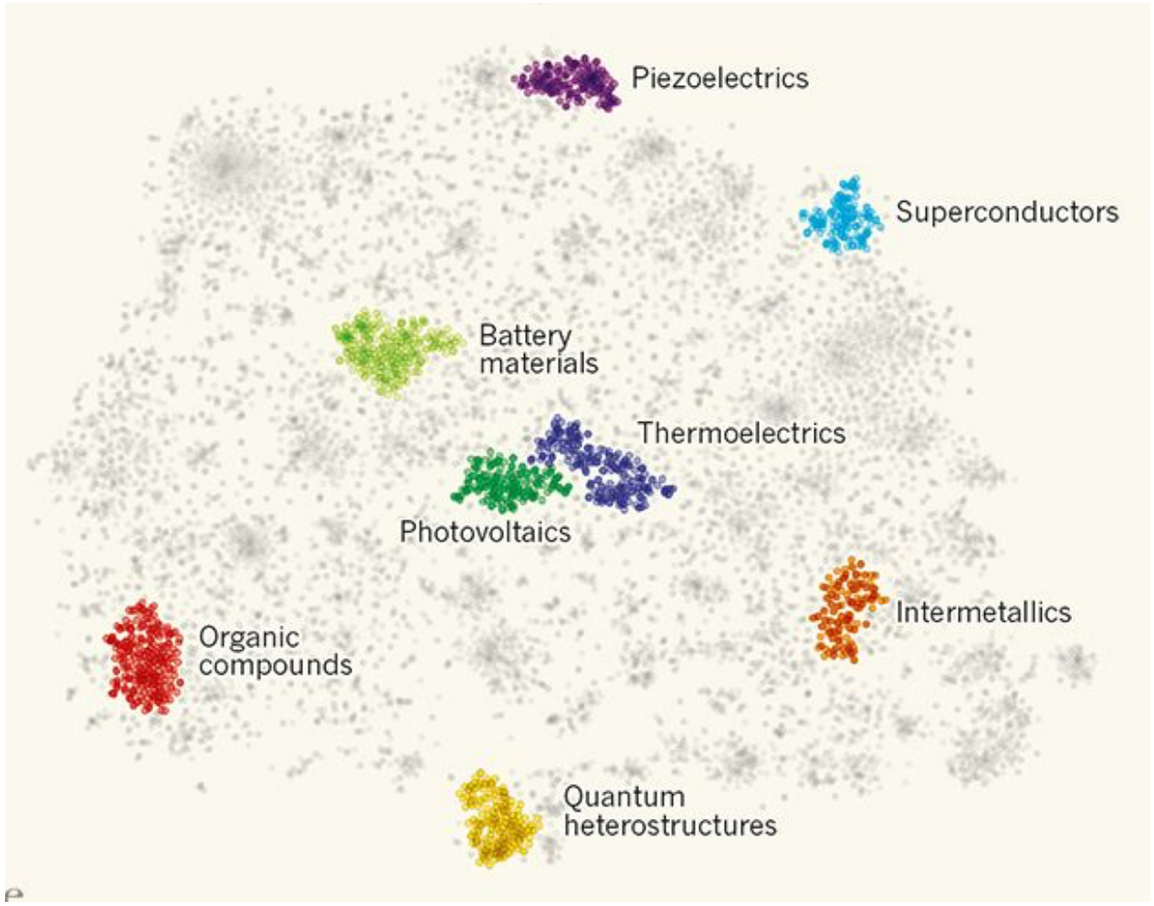
Word Embeddings: Learning on > 1 Million of Abstracts

ferromagnetic−NiFe + IrMn ≈ antiferromagnetic

single layer NN is trained to predict all context words for the given target word.



Projection of embeddings onto two dimensions (t-SNE).



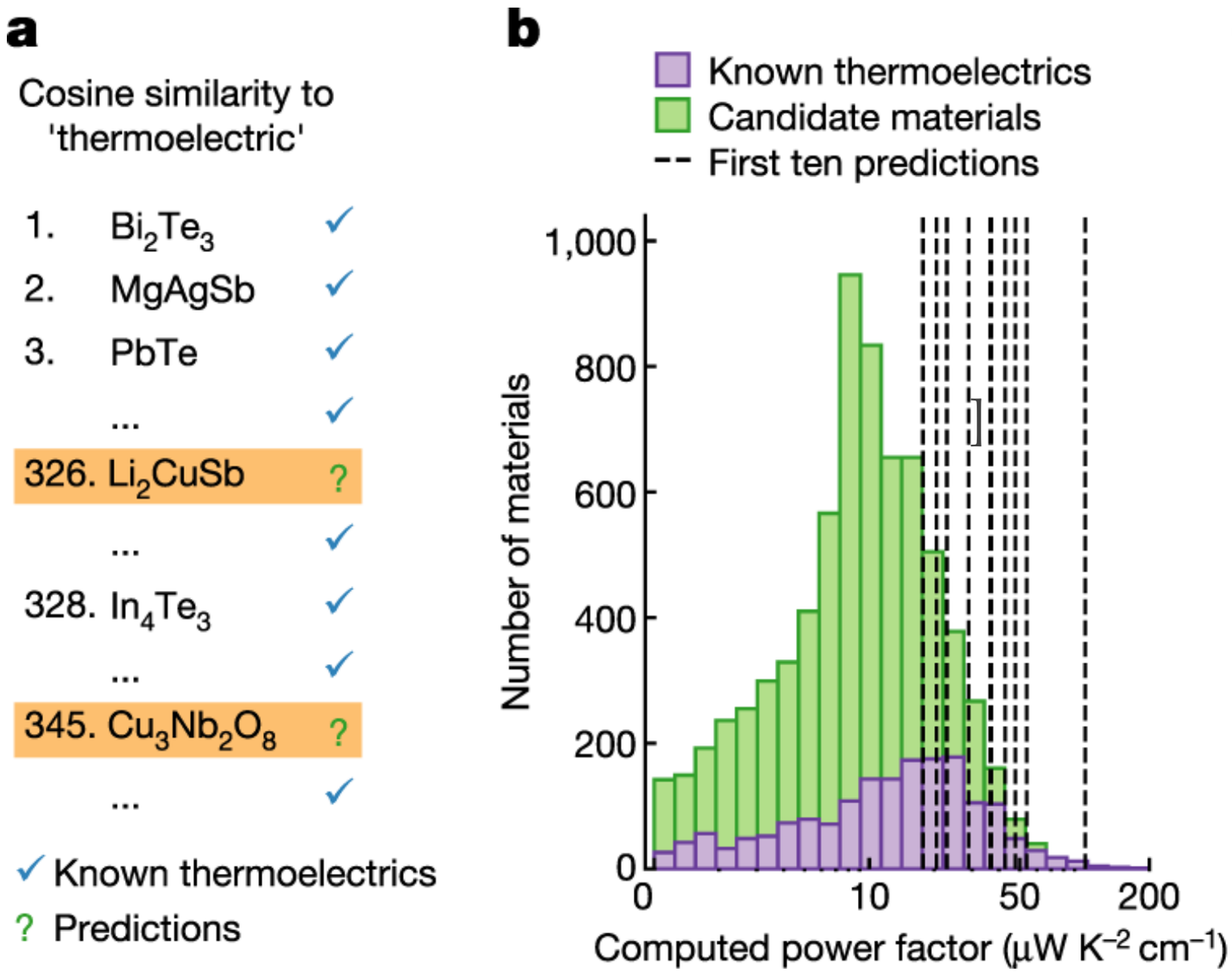
for similar words the context words are the same

Tshitoyan, V.; *et al*.. *Nature* **2019**, *571* (7763), 95.

Perspective: Isayev, O. *Nature* **2019**, *571* (7763), 42–43.

Word Embeddings: Learning on > 1 Million of Abstracts

Embeddings can also be used for predictions.



- Top ten predictions even slightly higher than known average
- Better rank correlation with experiments than DFT
- Training data is important: model trained on all Wikipedia articles performs worse

Tshitoyan, V.; *et al*.. *Nature* **2019**, *571* (7763), 95.

ML for MolSim

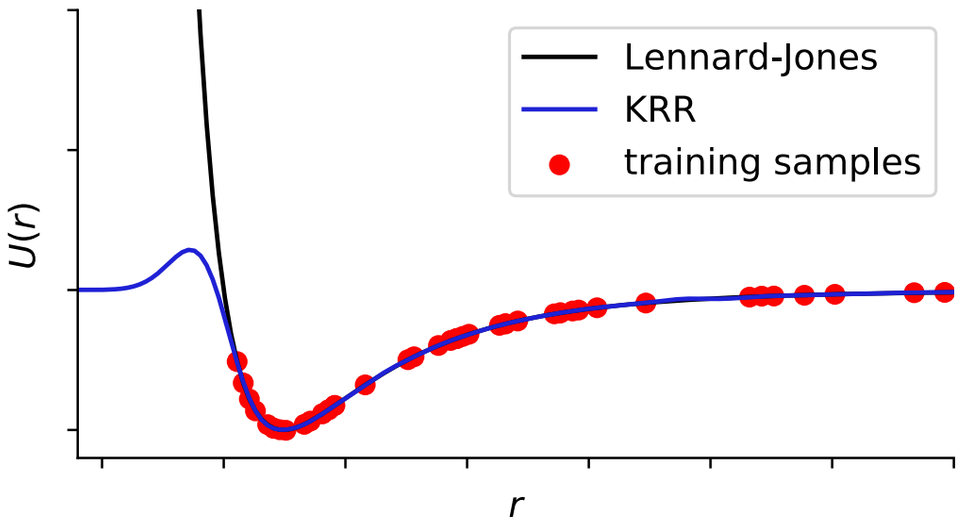
- More efficient MD with ML
- More efficient sampling with Boltzmann generators

GPR for Active Learning of $U(\mathbf{X})$

Bayesian version of KRR: Gaussian Process Regression (GPR)

KRR fails if there is no data …

… but cannot warn us



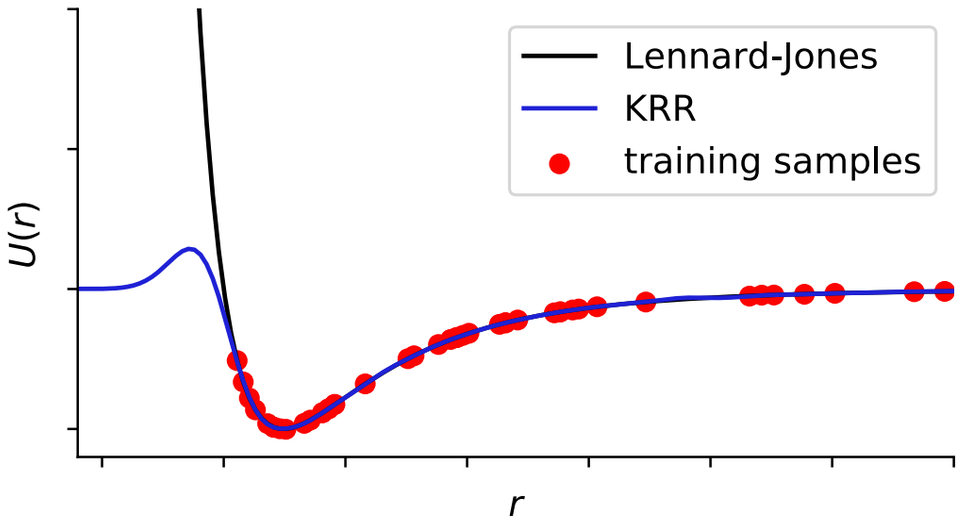
Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003,*

GPR for Active Learning of $U(\mathbf{X})$

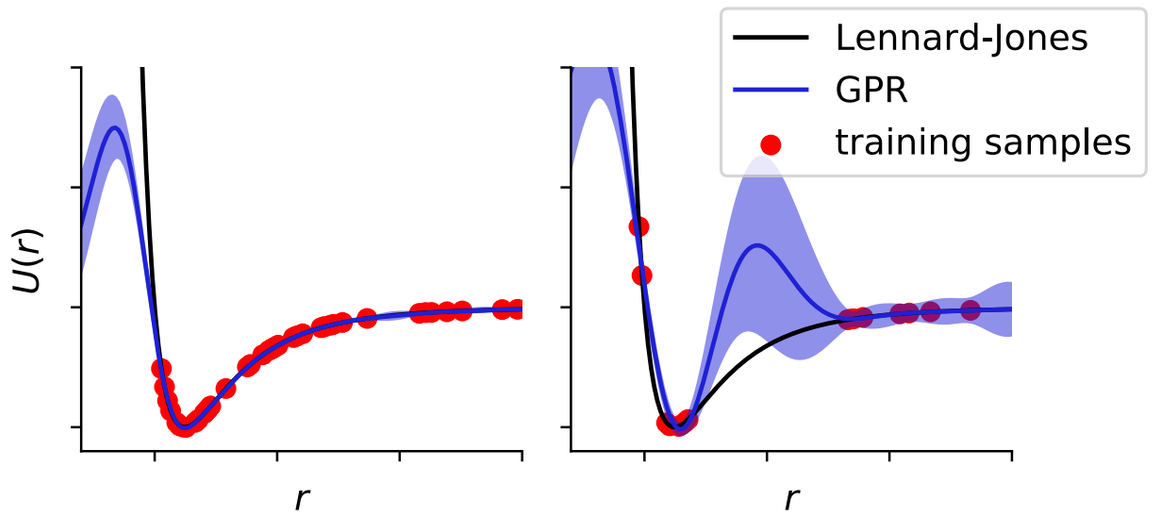
Bayesian version of KRR: Gaussian Process Regression (GPR)

KRR fails if there is no data …

… but cannot warn us

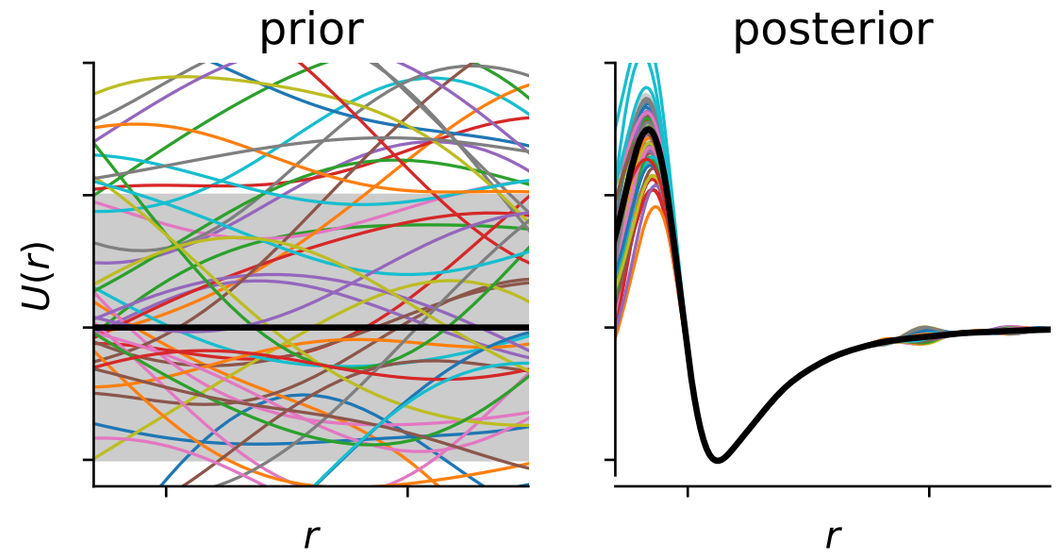


GPR gives uncertainty estimate



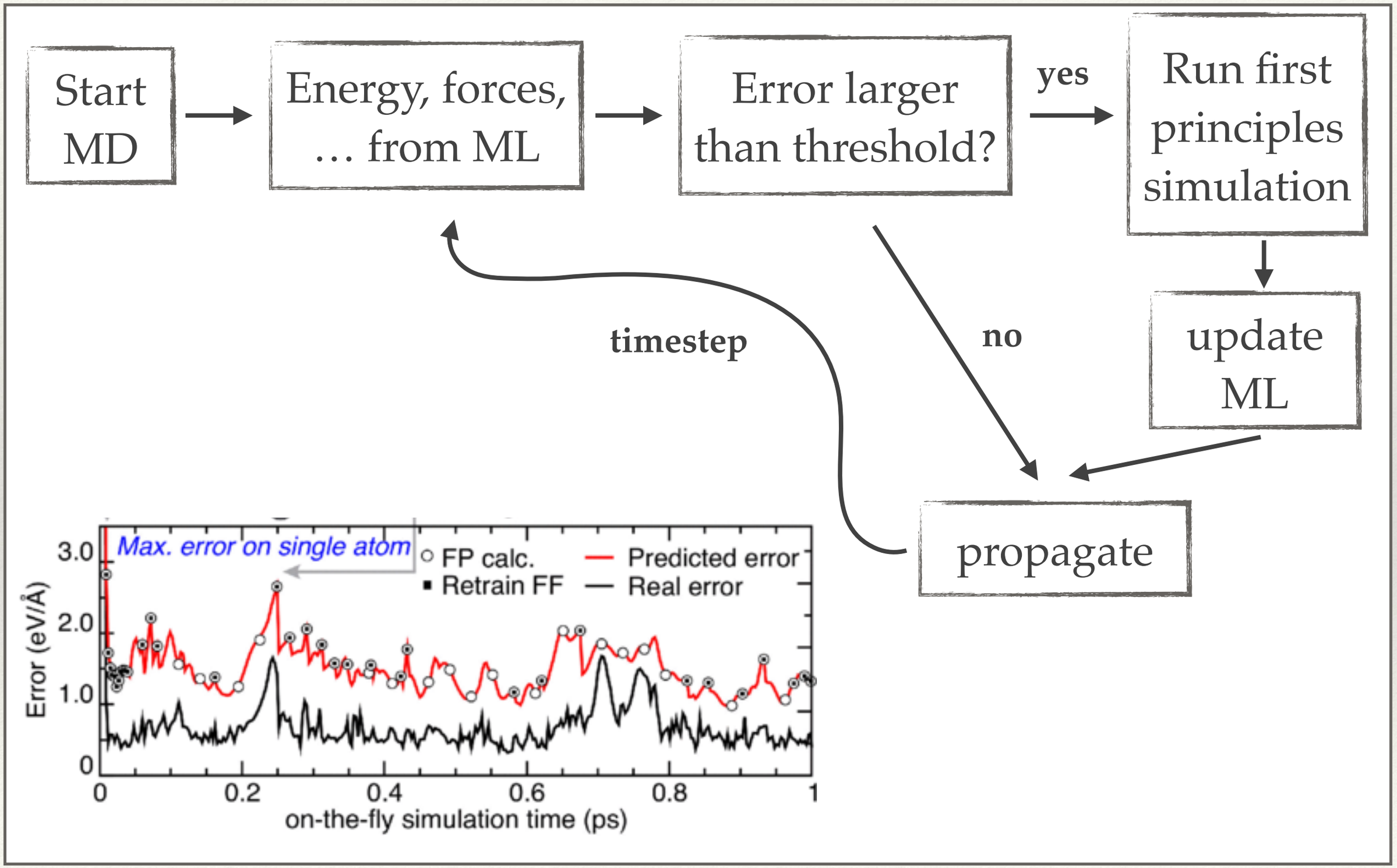
$$\underbrace{P(\theta|D)}_{\text{posterior}} \propto \underbrace{P(D|\theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

GPR uses data to update a prior distribution of functions to a posterior distribution



Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003,*

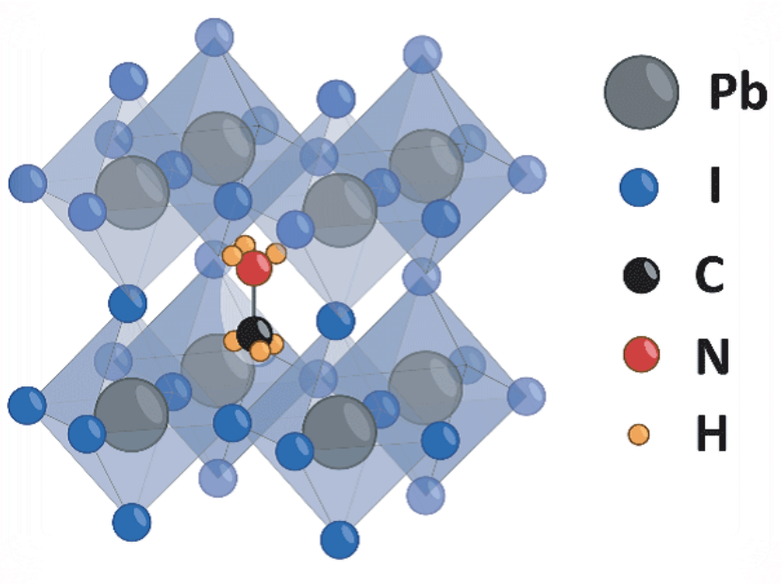
GPR for Active Learning of *U(X)*: Skip 99% of FP Calculations



Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. *Phys. Rev. Lett.* **2019**, *122* (22), 225701.

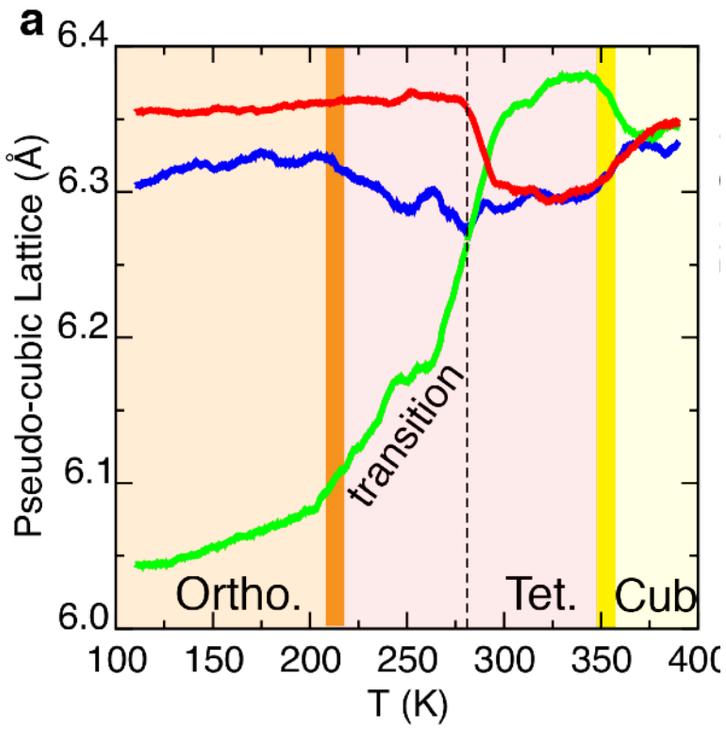
Implemented in VASP6.0.

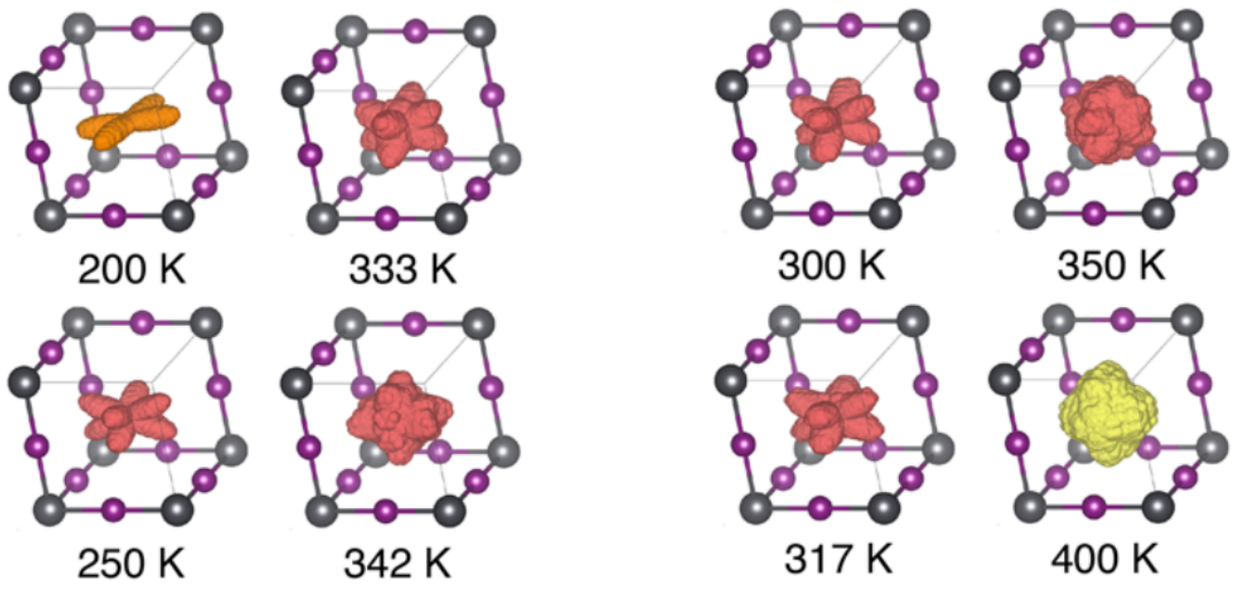
GPR for Active Learning of $U(\mathbf{X})$



Methyl ammonium lead halide perovskites

- Slow rotational dynamics
- Entropy driven phase transition
- Existing force-fields are not accurate



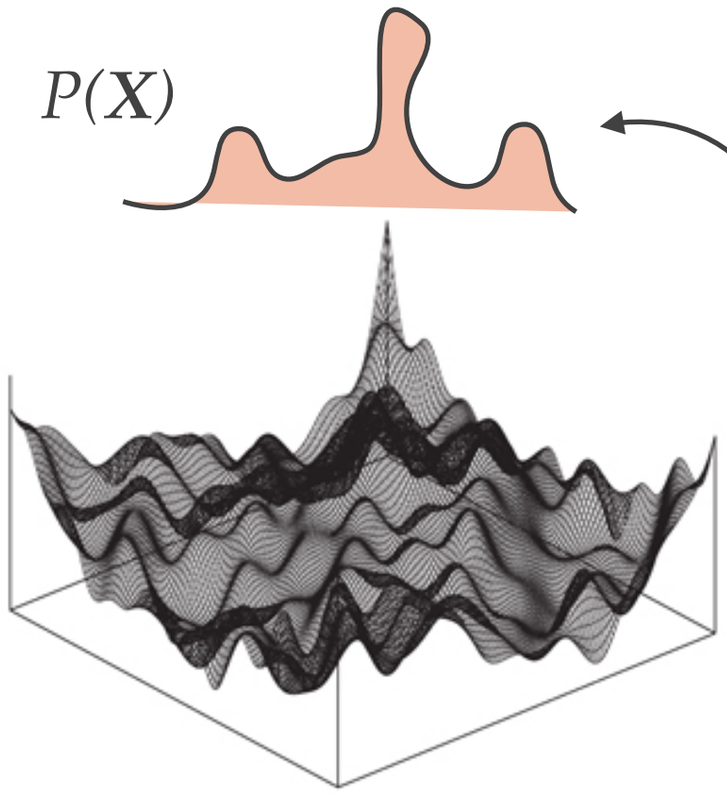


Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. *Phys. Rev. Lett.* **2019**, *122* (22), 225701.

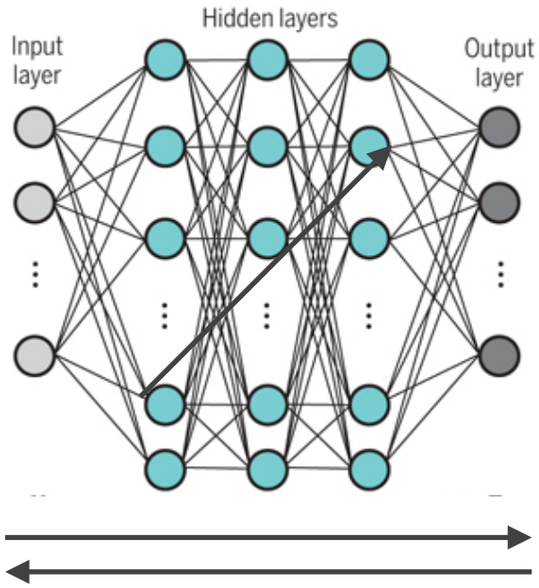
Implemented in VASP6.0.

Boltzmann Generators: A New Approach for Sampling of Microstates in One Shot (Statistically Independent)

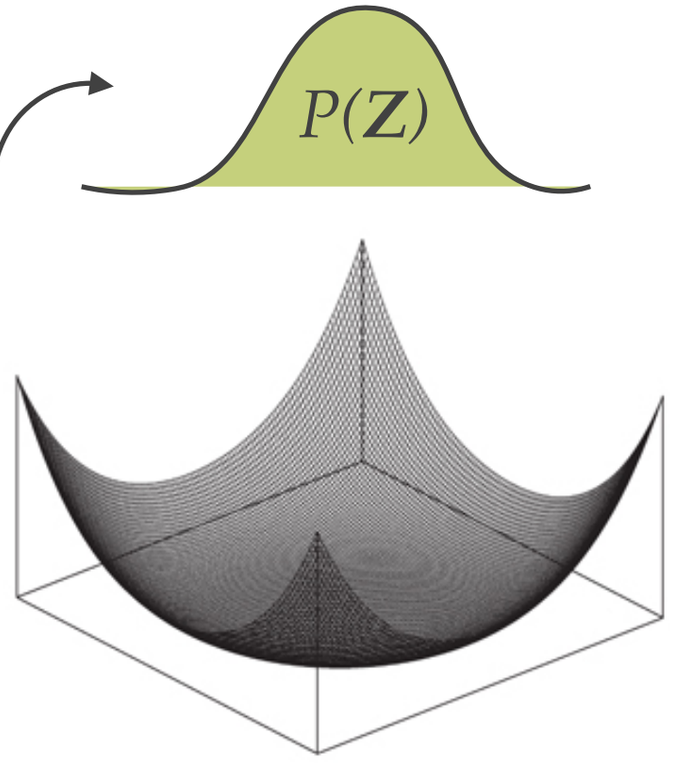
Molecular Simulations: $U(\boldsymbol{X})$ given, need approach to sample $P(\boldsymbol{X})$



X



Z



$-kT\lg(P(\mathbf{X}))$

Landscapes are often rugged, hard to sample

Invertible neural network does the invertible mapping

$-kT\lg(P(\mathbf{Z}))$

… this is why we bias some simulations to have low energy states close to each other (flat sampling)

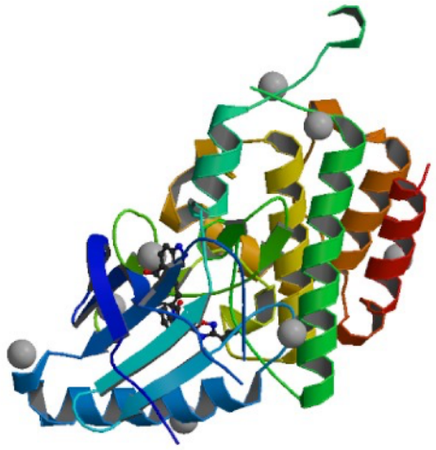
Noé, F.; Olsson, S.; Köhler, J.; Wu, H. *Science* **2019**, *365* (6457),

Perspective: Tuckerman, M. E. *Science* **2019**, *365* (6457), 982–983.

ML for Inverse Design

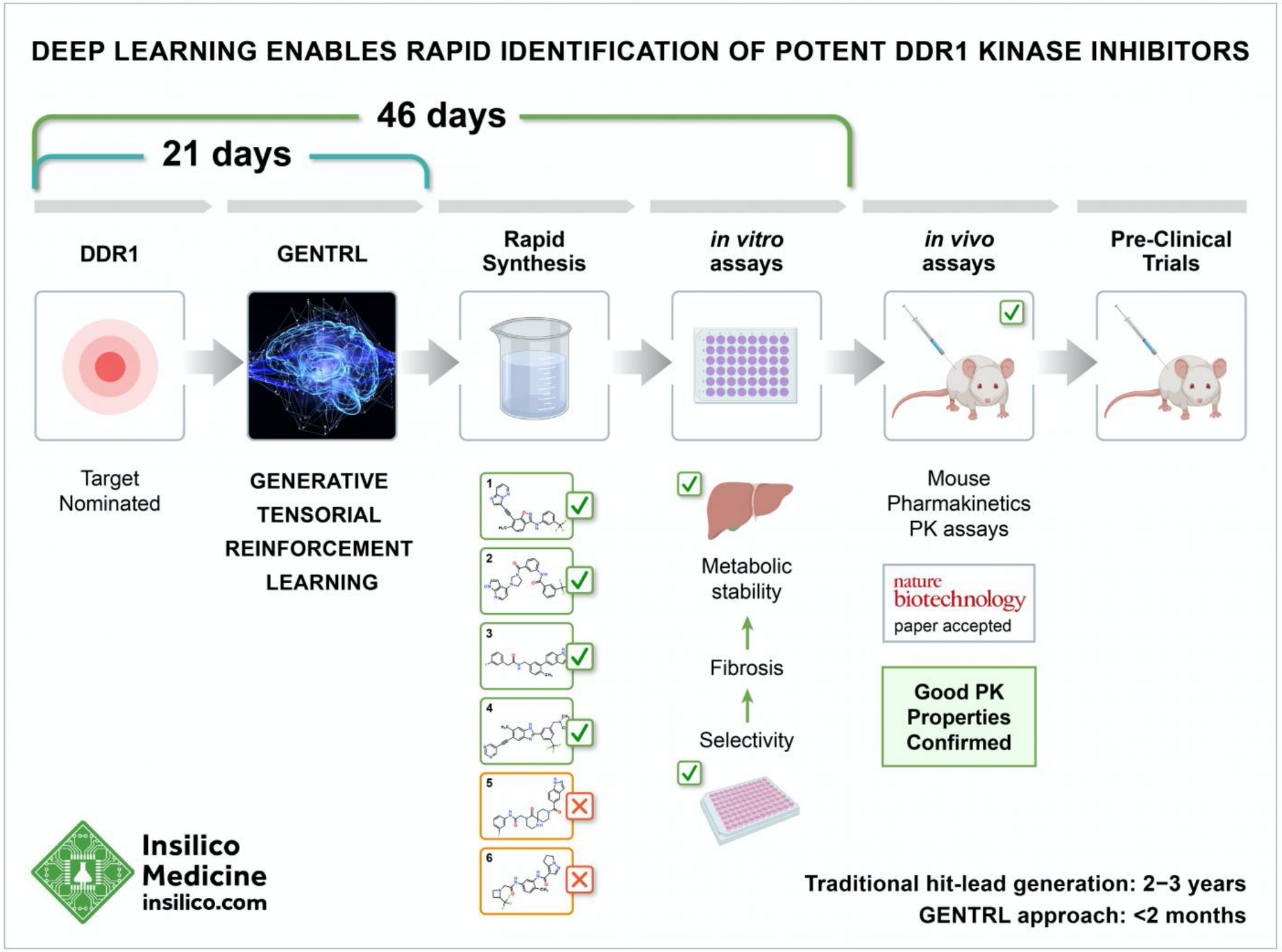
- Reinforcement learning for drug discovery
- Latent representation can be used for inverse design

21 Days to Drug Candidate: Using Reinforcement Learning to Expedite Drug Discovery



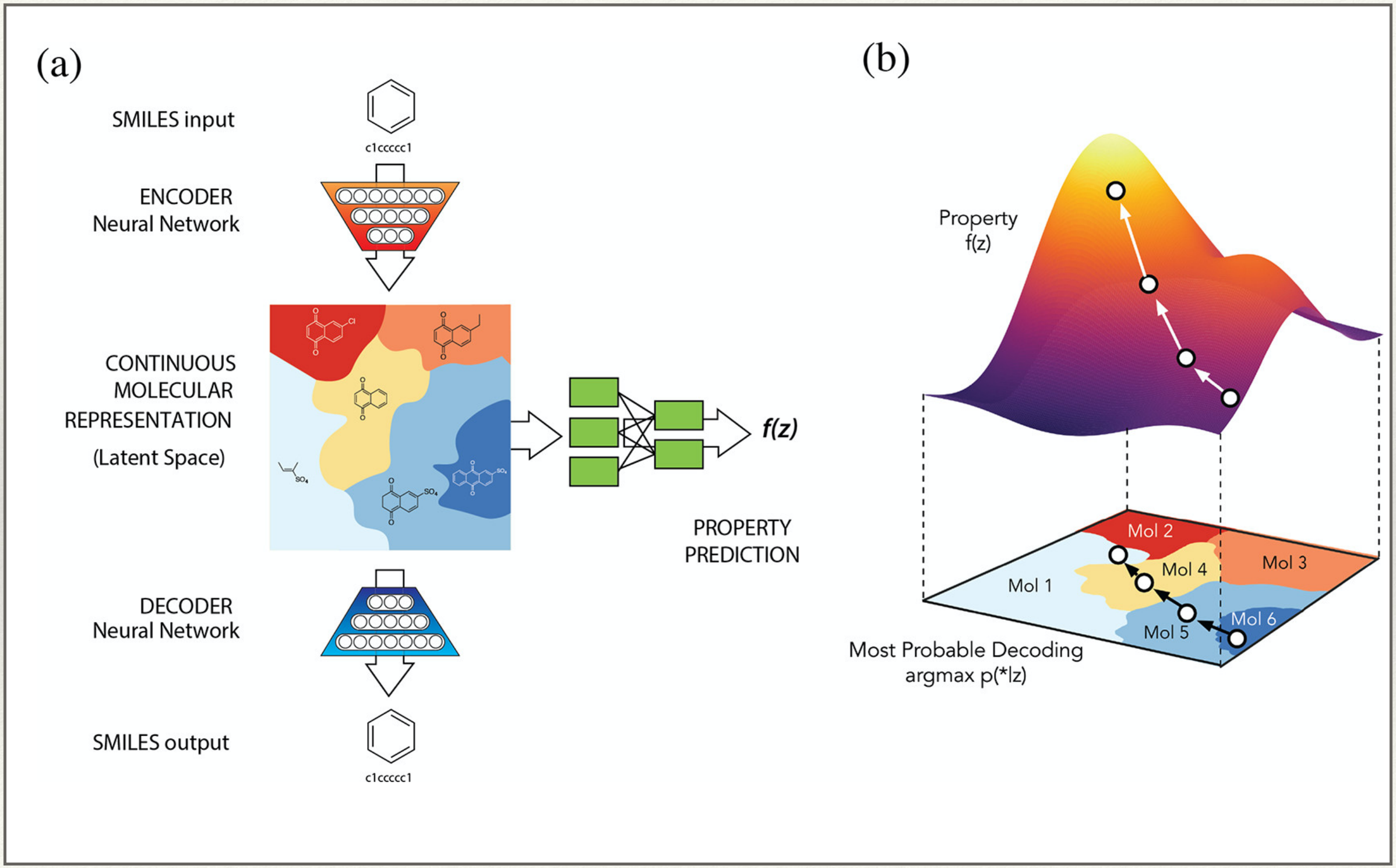
DDR1 (Target for fibrosis)

Model that takes synthetic feasibility, novelty, and biological activity into account



Zhavoronkov, A.; et al. *Nat Biotechnol* **2019**, *37* (9), 1038–1040.

21 Days to Drug Candidate: Using Reinforcement Learning to Expedite Drug Discovery



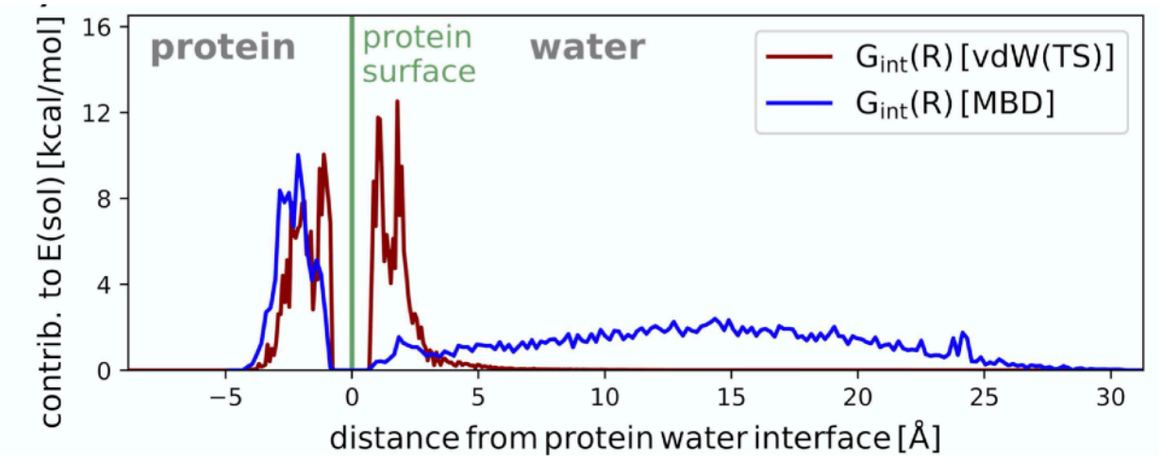
Gómez-Bombarelli, R.; *et al*.. *ACS Cent. Sci*. **2018**, *4* (2), 268–276.

Challenges for the Field

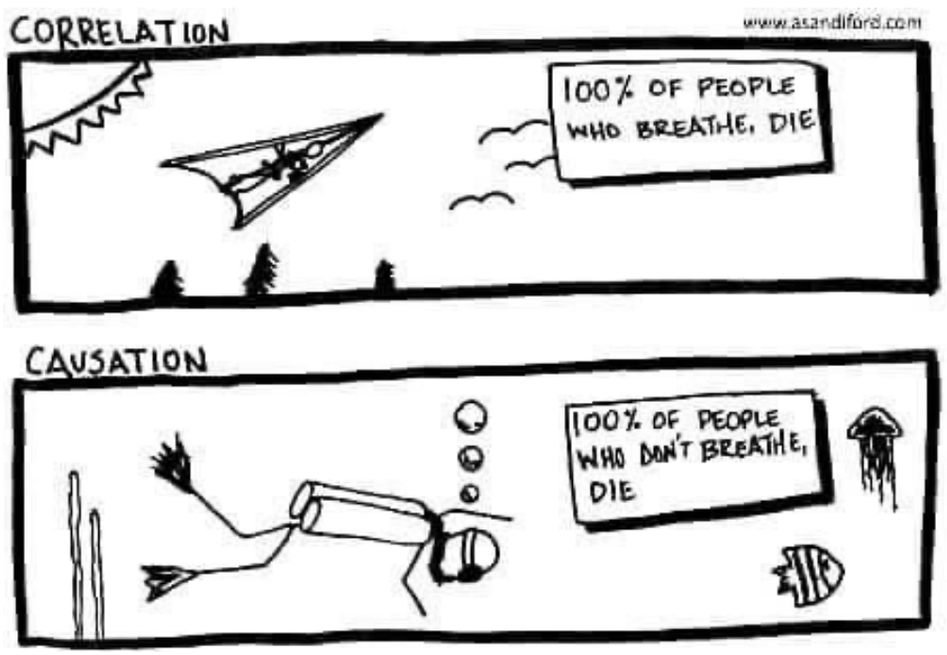
Benchmarks, Reproducibility and Comparability

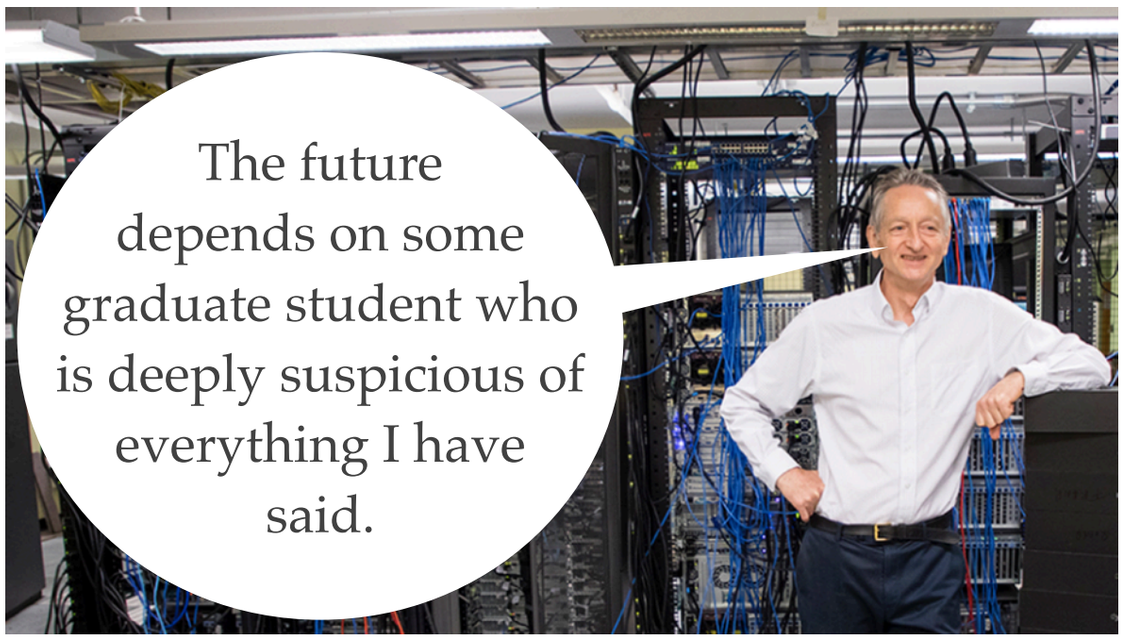
	MNIST	CIFAR10	ImageNet
FGS	0.996	0.911	0.881
JSMA	0.995	0.966	-
Deepfool	0.996	0.960	0.908
CarliniL2	0.989	0.929	0.907
BIM	0.994	0.907	0.820

Incorporation of long-range interactions



Causal Inference





Pearl, J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *arXiv:1801.04016 [cs, stat]* **2018**.

Stöhr, M.; Tkatchenko, A. *Sci. Adv.* **2019**, *5* (12), eaax0024.