Autoregressive model: alphabets, actions, and atoms

A unified perspective to LLM, RL, and atomistic modeling

Lei Wang (王磊)
Institute of Physics, CAS
https://wangleiphy.github.io

Plan

1. Generative AI

2. Autoregressive models
   - -principle
   - -architecture
   - -training
   - -inference

3. Applications

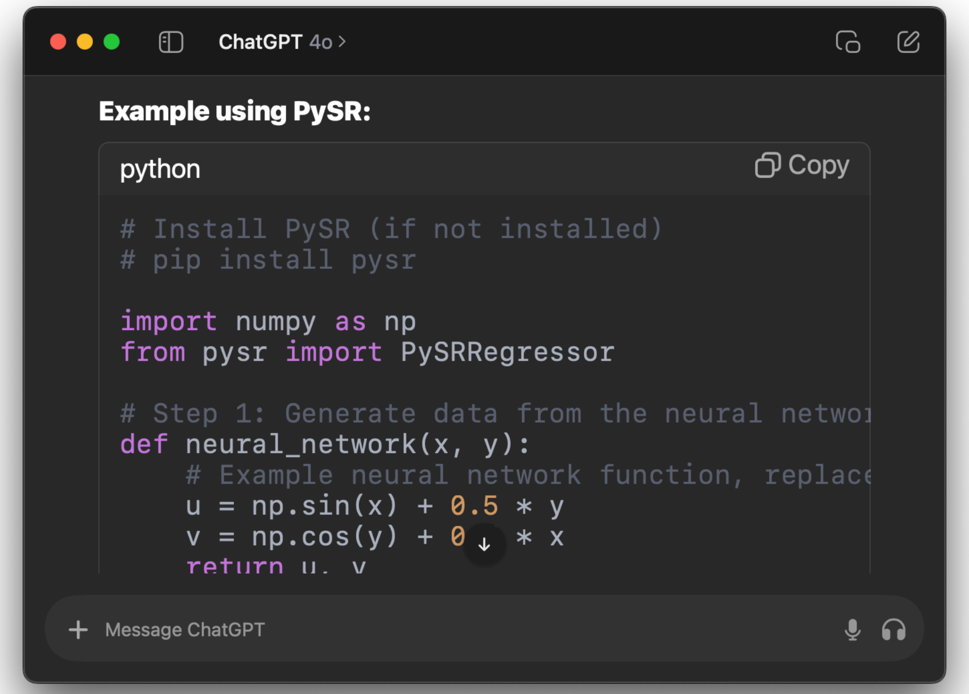
Probabilistic modeling with generative AI

$$p(\boldsymbol{X})$$

pixels, words, atoms, ...

How to express, learn, and sample from a high-dimensional probability distribution ?

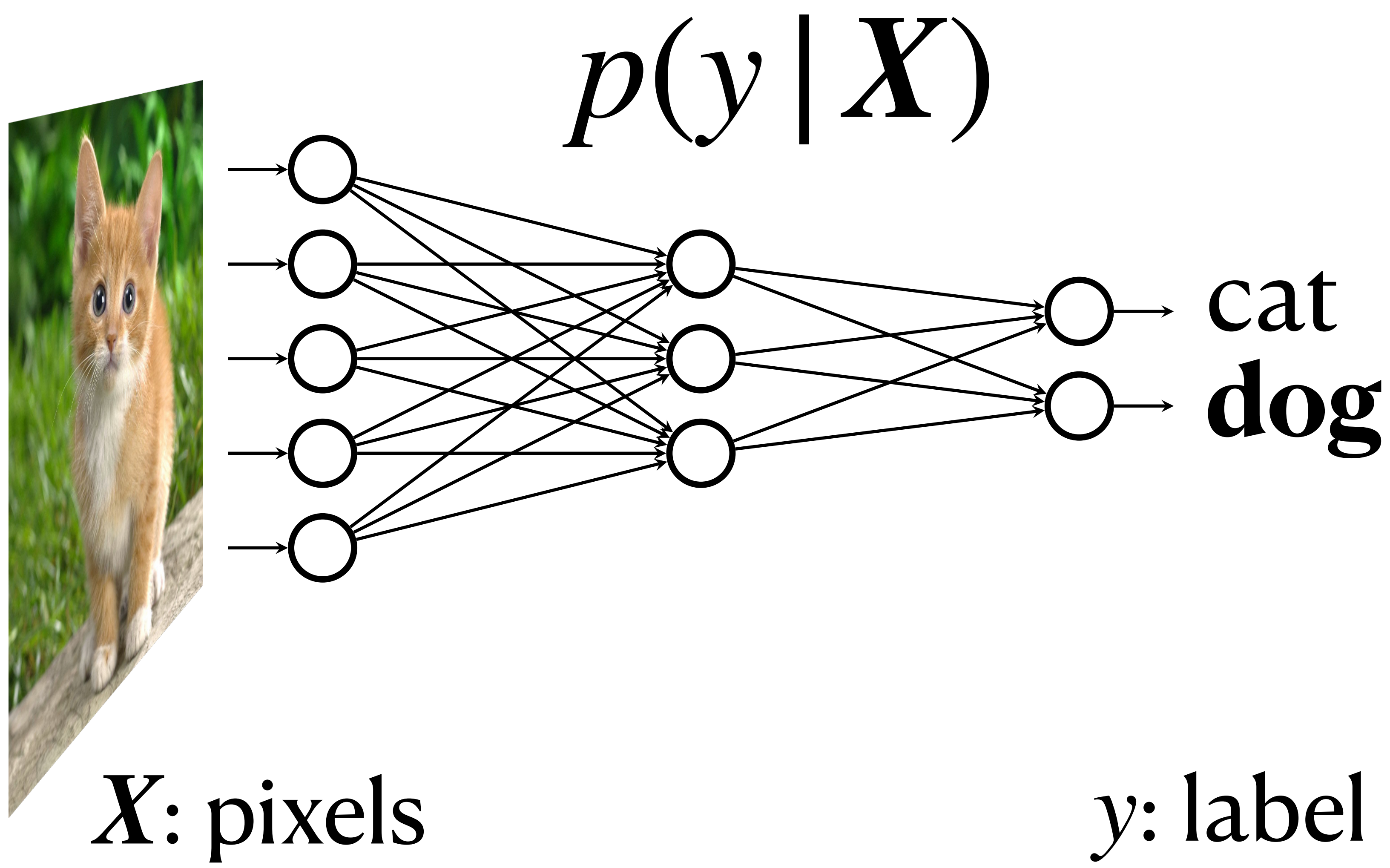
DaLL-E



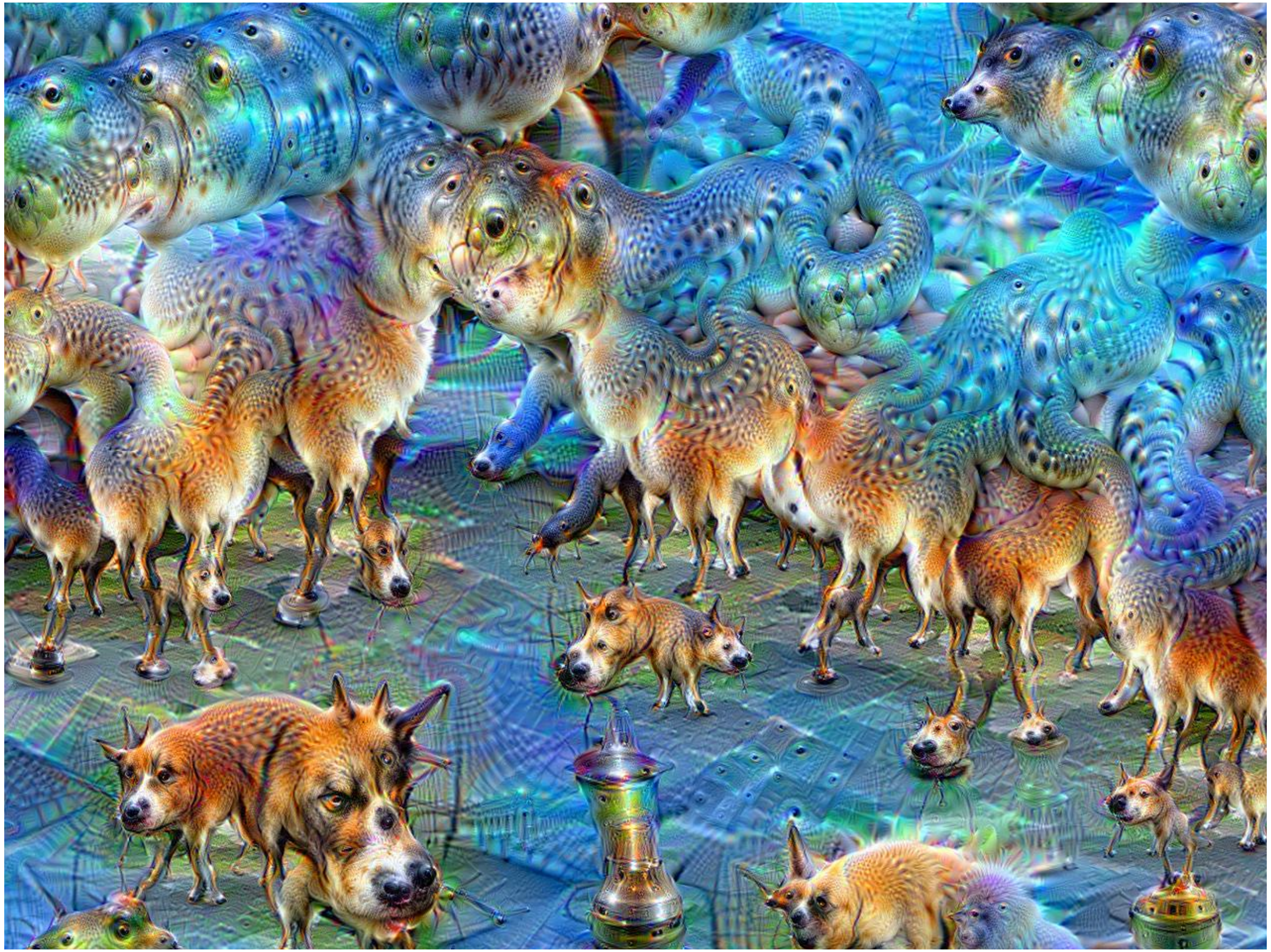
ChatGPT

AlphaFold3

Discriminative AI is not enough



$$\nabla_{\text{pixels}} p(\text{dog} \mid \text{pixels})$$



Deepdream: http://googleresearch.blogspot.ch/2015/06/inceptionism-going-deeper-into-neural.html

Bayes rule

$$\underset{\text{Inverse design}}{\overset{\text{posterior}}{p(\boldsymbol{X} \mid y)}} \propto \overset{\text{prior}}{\color{red}{p(\boldsymbol{X})}} \underset{\text{Forward prediction}}{\overset{\text{likelihood}}{p(y \mid \boldsymbol{X})}}$$

Probability theory 101

Conditional probability $p(y \mid \boldsymbol{X})$

Joint probability $p(\boldsymbol{X}, y)$

Product rule $p(\boldsymbol{X}, y) = p(y \mid \boldsymbol{X})p(\boldsymbol{X})$

Sum rule $$p(\boldsymbol{X}) = \sum_y p(\boldsymbol{X}, y)$$


generative model $p(\boldsymbol{X})$
data $\boldsymbol{X}$

Two sides of the same coin

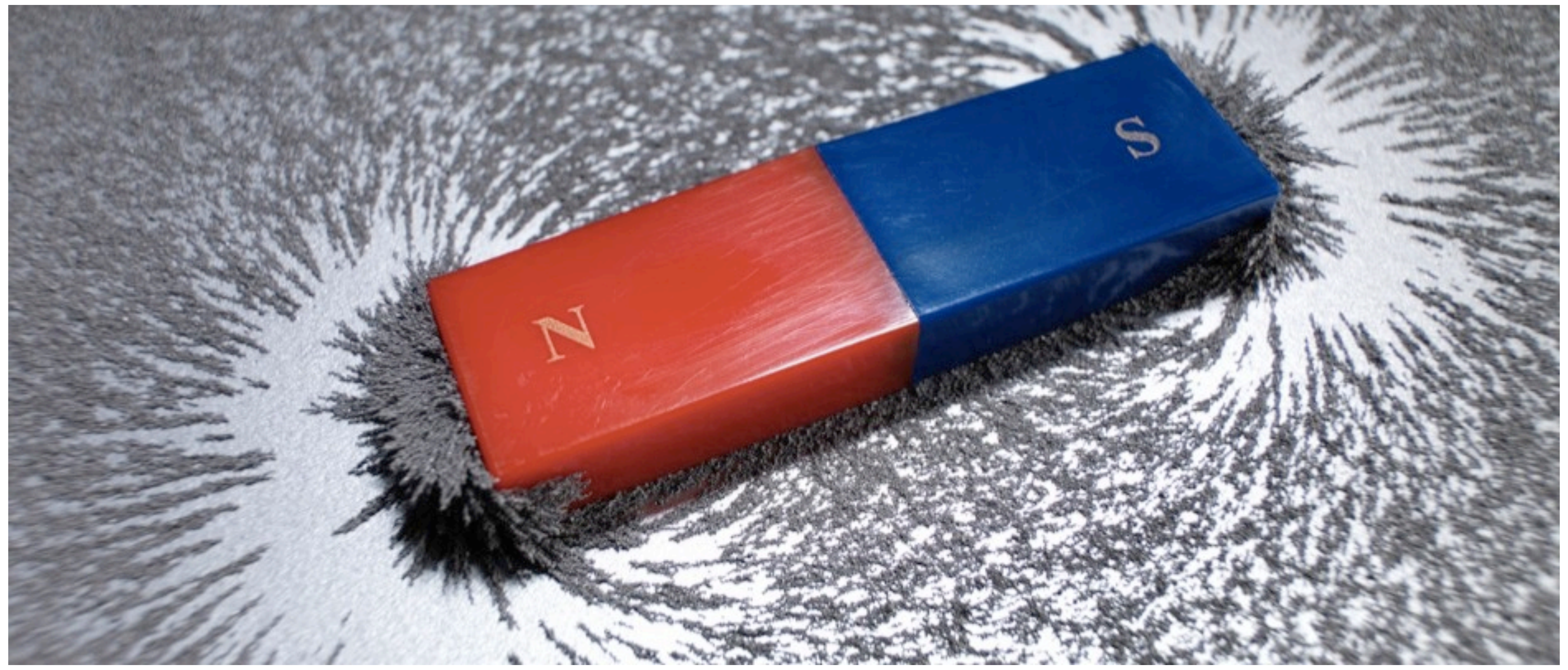
Generative modeling

"learn from data"

Mamixmum likelihood estimation

$$\mathscr{L} = -\,\mathbb{E}_{\boldsymbol{X}\sim\text{data}}\left[\ln p(\boldsymbol{X})\right]$$

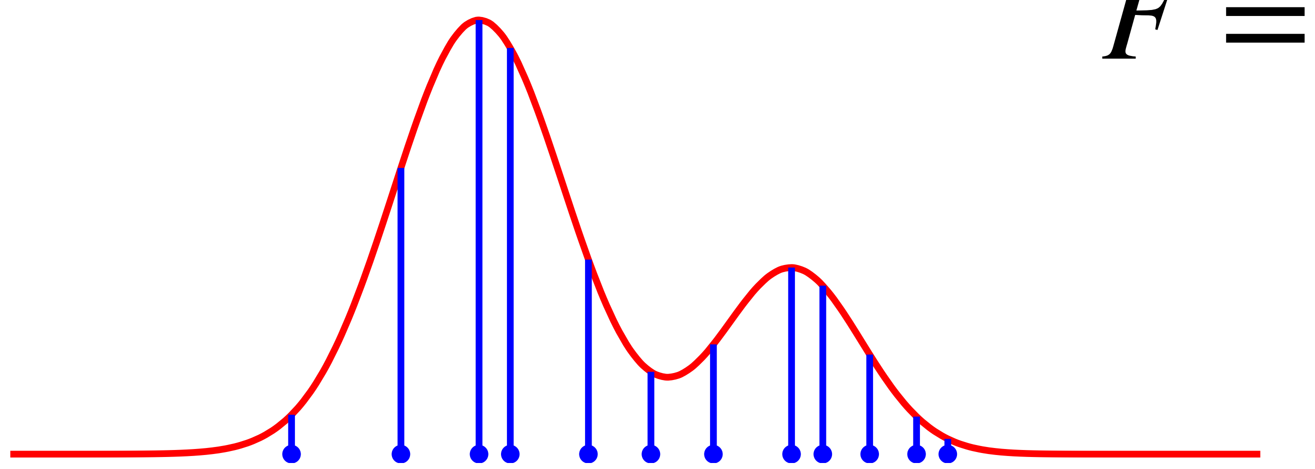
Statistical physics



"learn from energy"

Variational free energy

$$F = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim p(\boldsymbol{X})}\left[E(\boldsymbol{X}) + k_B T \ln p(\boldsymbol{X})\right]$$



$$\mathbb{KL}(\text{data} \,\|\, p) \quad \textbf{vs} \quad \mathbb{KL}(p \,\|\, e^{-E/k_BT})$$

Kullback–Leibler divergence

$$\mathbb{KL}(\pi \parallel p) \equiv \sum_{X} \pi(\boldsymbol{X}) \left[\ln \pi(\boldsymbol{X}) - \ln p(\boldsymbol{X}) \right]$$

$$\mathbb{KL}(\pi \parallel p) \geq 0$$

$$\mathbb{KL}(\pi \parallel p) = 0 \iff \pi(\boldsymbol{X}) = p(\boldsymbol{X})$$

$$\mathbb{KL}(\pi \parallel p) \neq \mathbb{KL}(p \parallel \pi)$$

Learn from data

$$\pi(\boldsymbol{X}) \propto \sum_{\boldsymbol{d} \in \text{dataset}} \delta(\boldsymbol{X} - \boldsymbol{d})$$

$$\min_{\theta} \mathbb{KL}(\pi \parallel p_{\theta}) \iff \min_{\theta} \left\{ \mathbb{E}_{\boldsymbol{X} \sim \text{dataset}} \left[-\ln p_{\theta}(\boldsymbol{X}) \right] \right\}$$

target model

Maximum likelihood estimation

The lower bound is the entropy of the dataset: complete memorization

Learn from Energy

$$\pi(\boldsymbol{X}) \propto e^{-E/k_B T}$$

$$\min_{\theta} \mathbb{KL}(p_\theta \,\|\, \pi) \iff \min_{\theta} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{X} \sim p_\theta(\boldsymbol{X})} \left[E(\boldsymbol{X}) + k_B T \ln p_\theta(\boldsymbol{X}) \right] \right\}$$

model target

Variational free energy

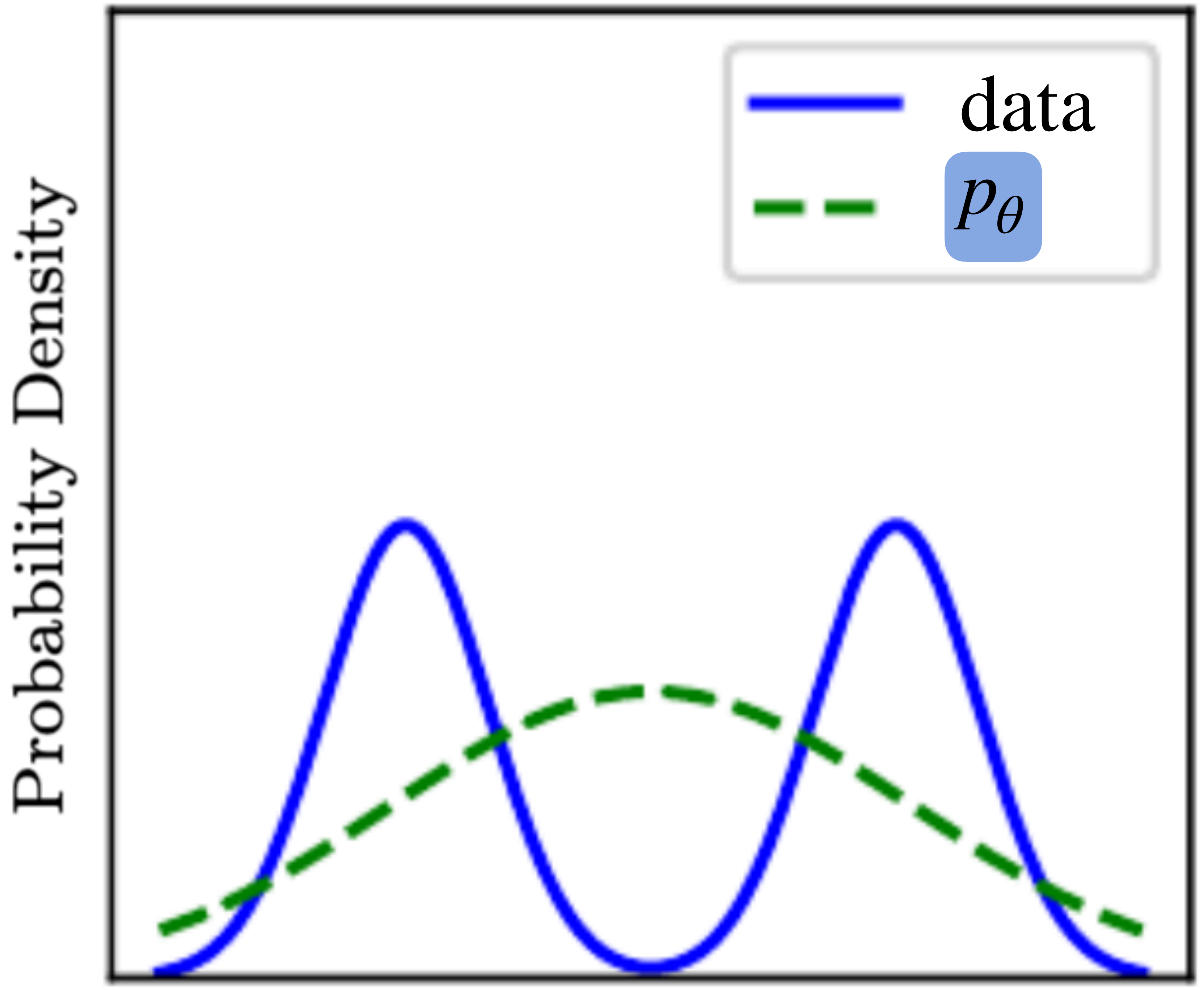
The lower bound is the true free energy: exact solution

Forward KL or Reverse KL ?

Maximum likelihood estimation

$$\min_{\theta} \mathbb{KL}(\text{data} \parallel p_\theta)$$

Mode covering

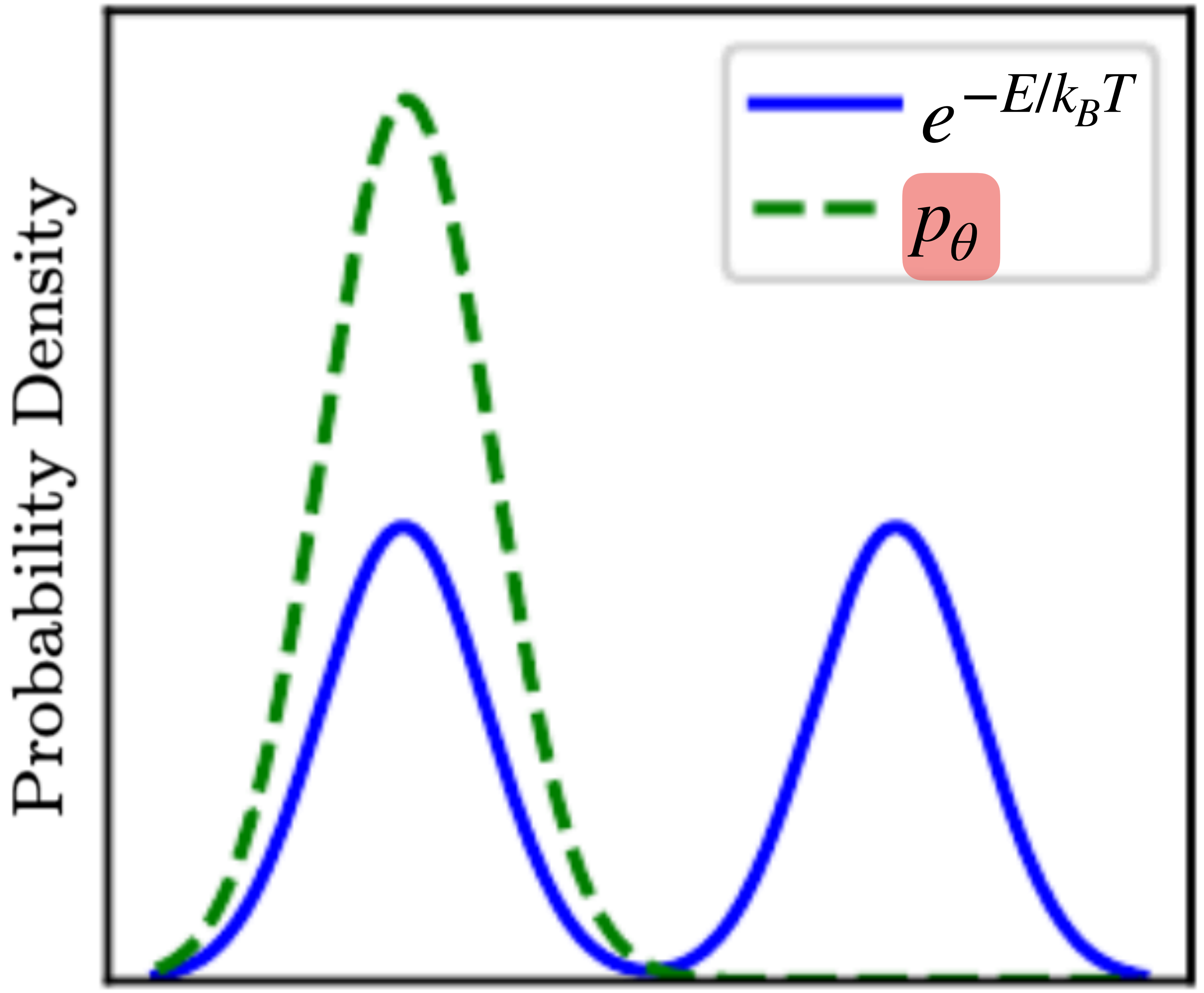


Failure mode: hallucination

Variational free energy

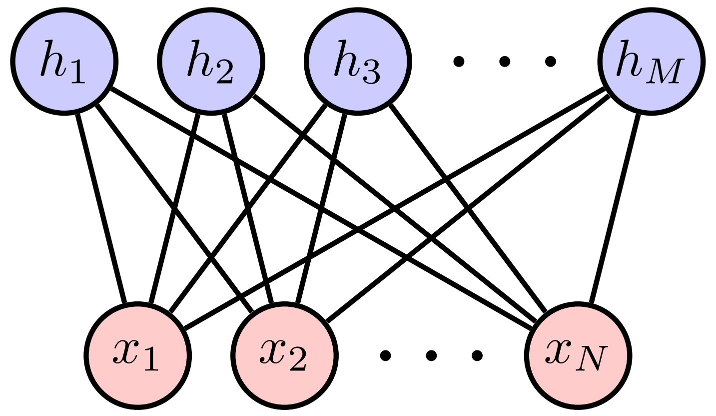
$$\min_{\theta} \mathbb{KL}(p_\theta \parallel e^{-E/k_BT})$$

Mode seeking



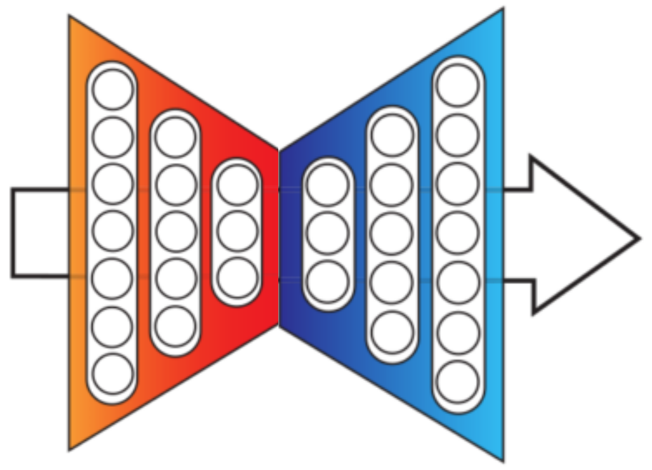
Failure mode: local minima

Goodfellow et al, Deep Learning



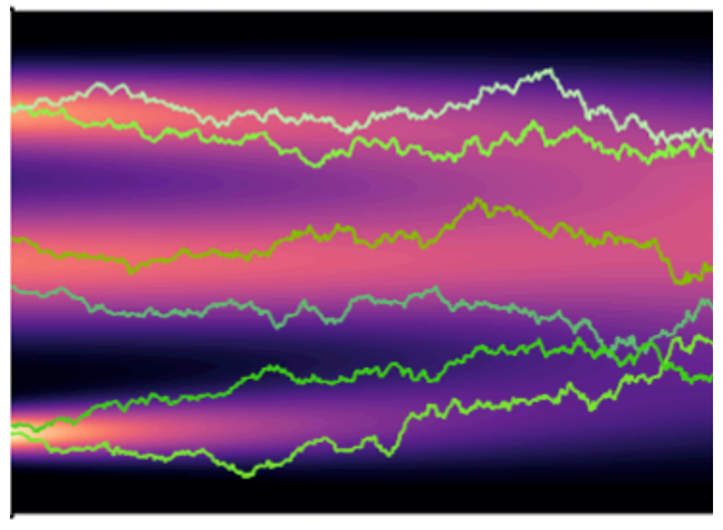
Boltzmann Machine

1985



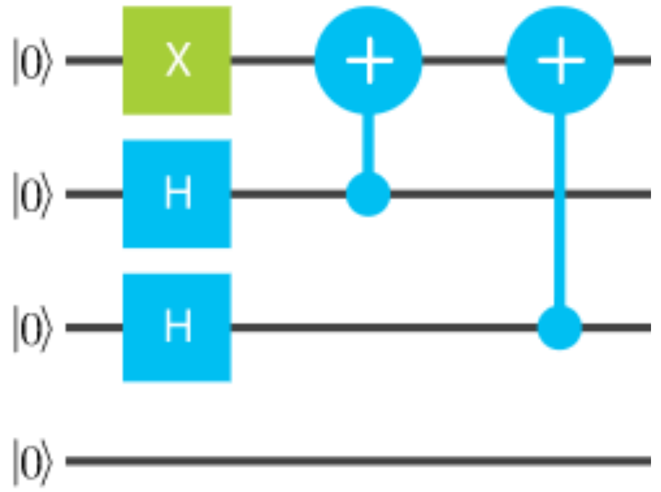
Variational Autoencoder

2013



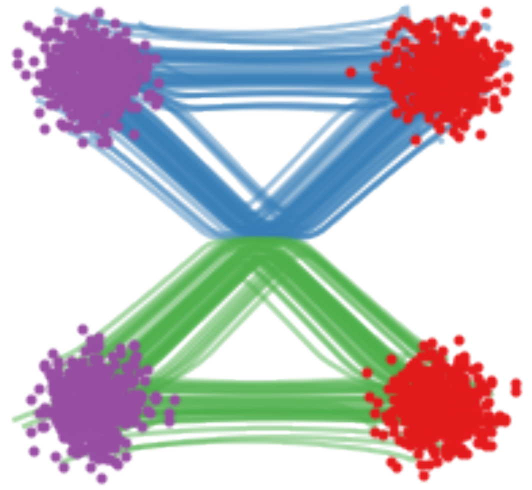
Diffusion Model

2015



Born Machine

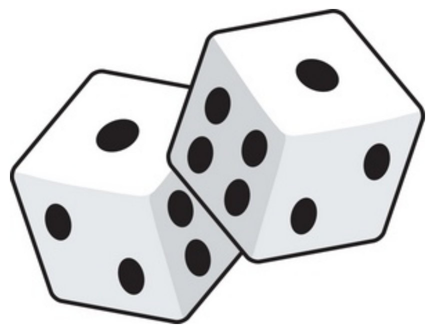
2017



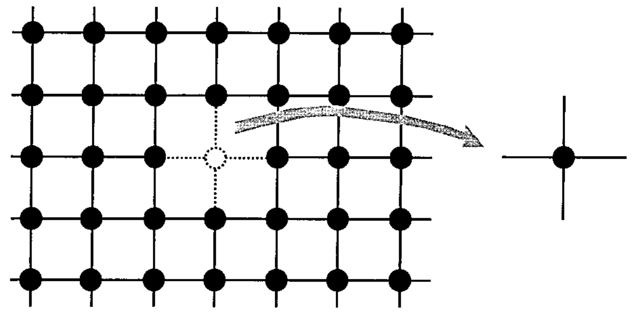
Flow Matching

2022

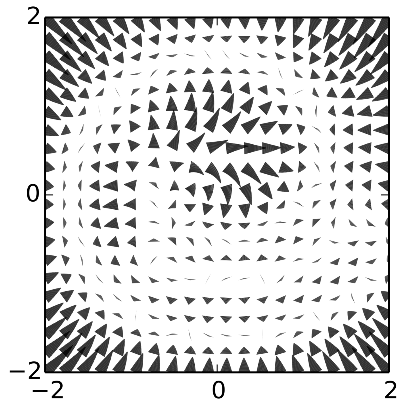
Monte Carlo
Ising model



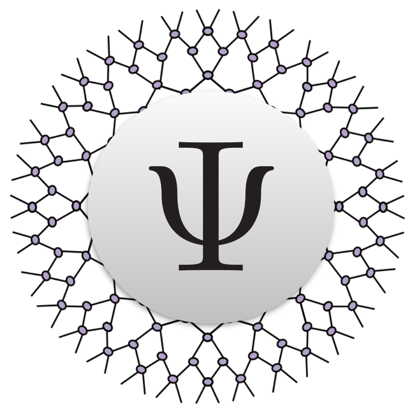
Variational
mean field



Nonequilibrium
thermodynamics



Tensor networks
Quantum circuits



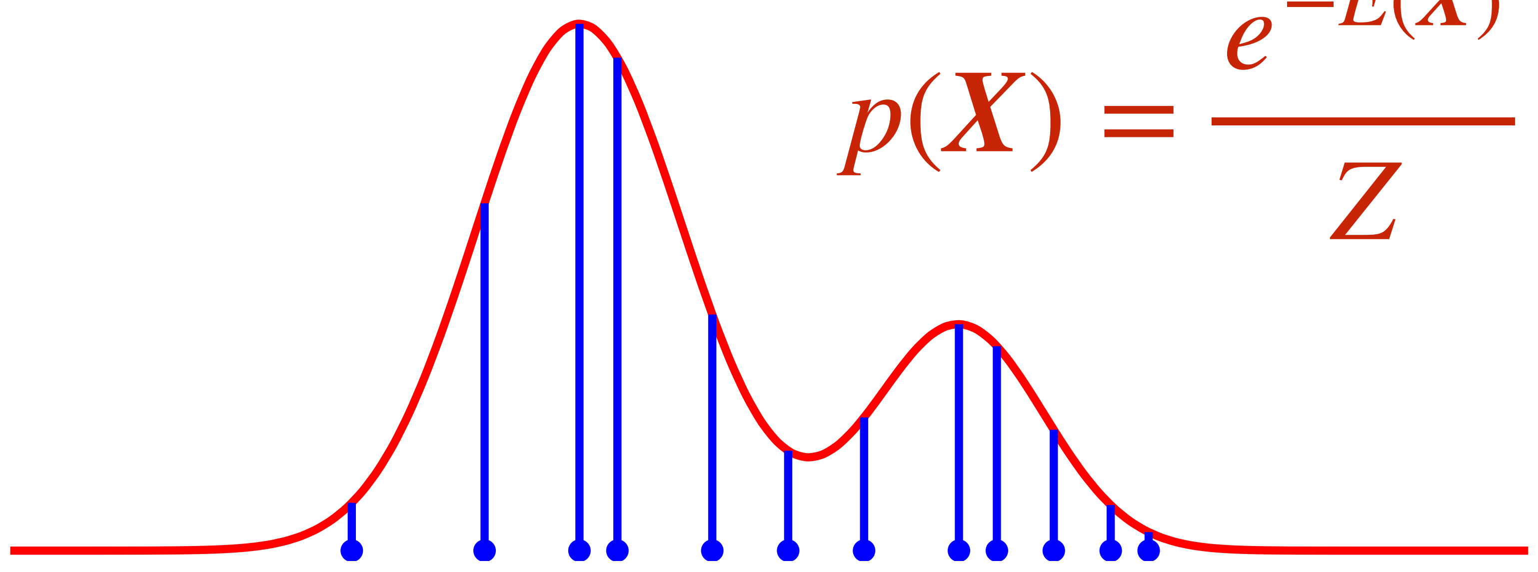
Fluid optimal
transportation

$$\frac{\partial p(\boldsymbol{X},t)}{\partial t} + \nabla \cdot \left[p(\boldsymbol{X},t)\boldsymbol{v}\right] = 0$$

Statistical, quantum, fluid, … physics insights into generative models

Leverage the power of modern generative models for science

Boltzmann machines 

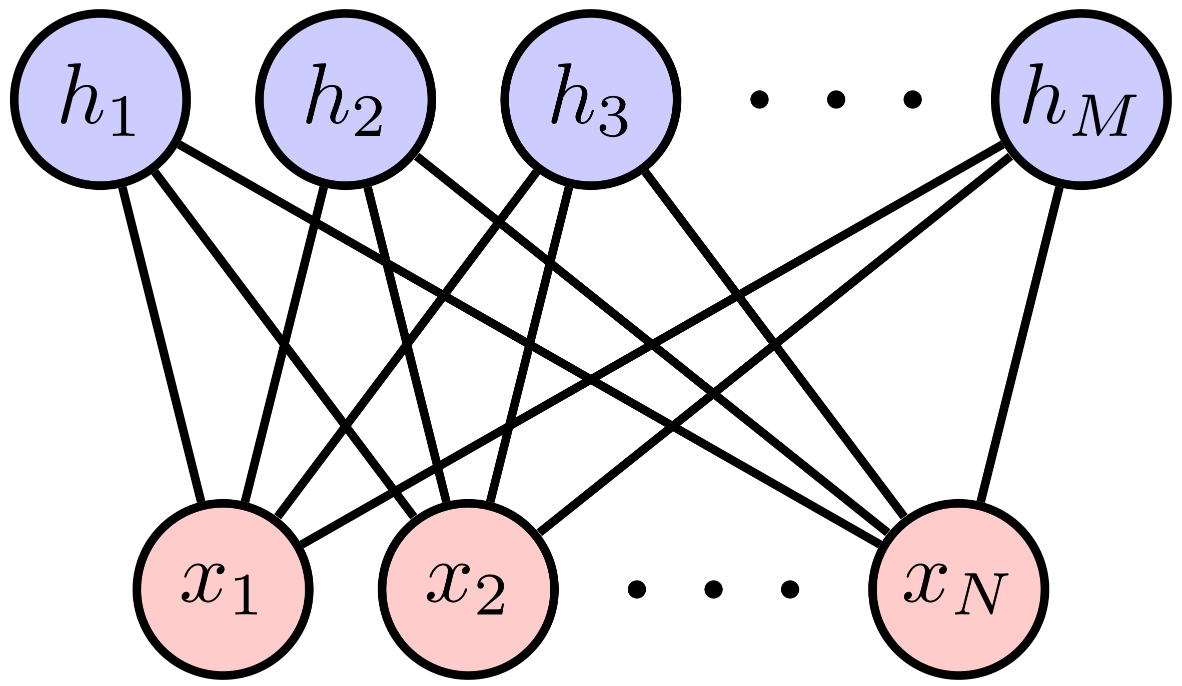


$$p(\boldsymbol{X}) = \frac{e^{-E(\boldsymbol{X})}}{Z}$$

Ackley, Hinton, Sejnowski, 1985

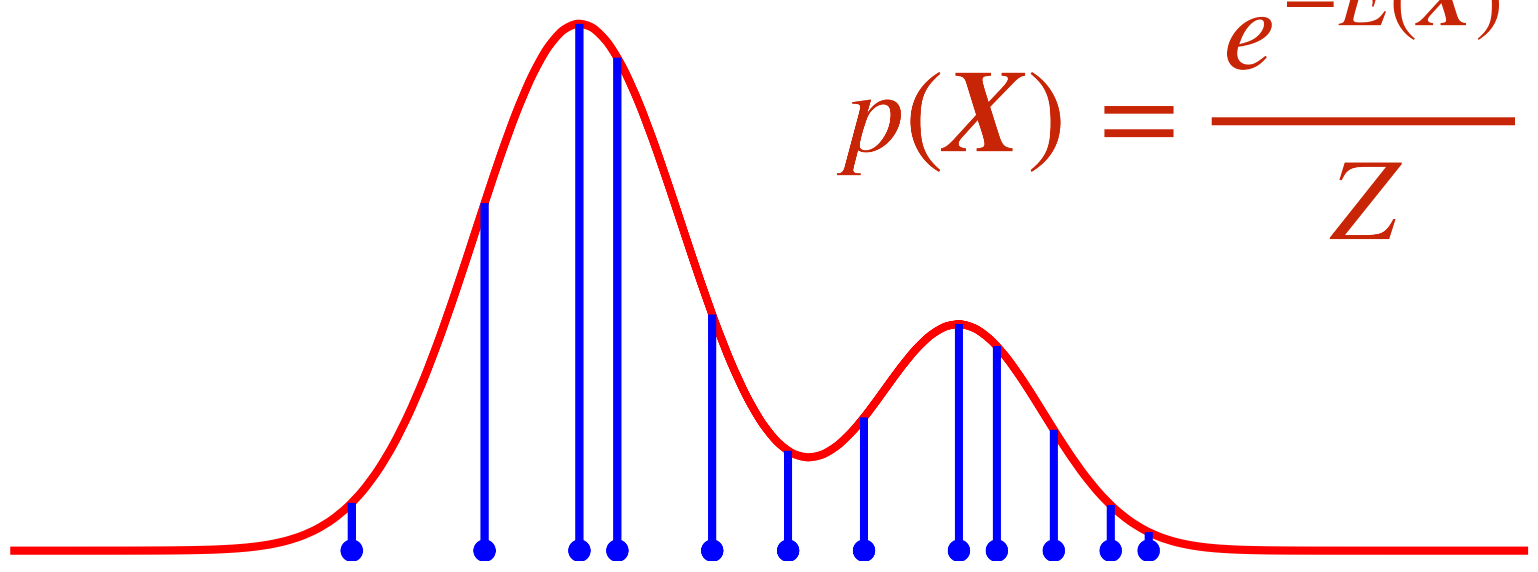
Learn

$$\mathscr{L} = \mathbb{E}_{\boldsymbol{X}\sim\text{data}}\left[-\ln p(\boldsymbol{X})\right]$$



$$\nabla_\theta \mathscr{L} = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\text{dataset}}\left[\nabla_\theta E\right] - \mathop{\mathbb{E}}_{\boldsymbol{X}\sim p(\boldsymbol{X})}\left[\nabla_\theta E\right]$$

Boltzmann machines 

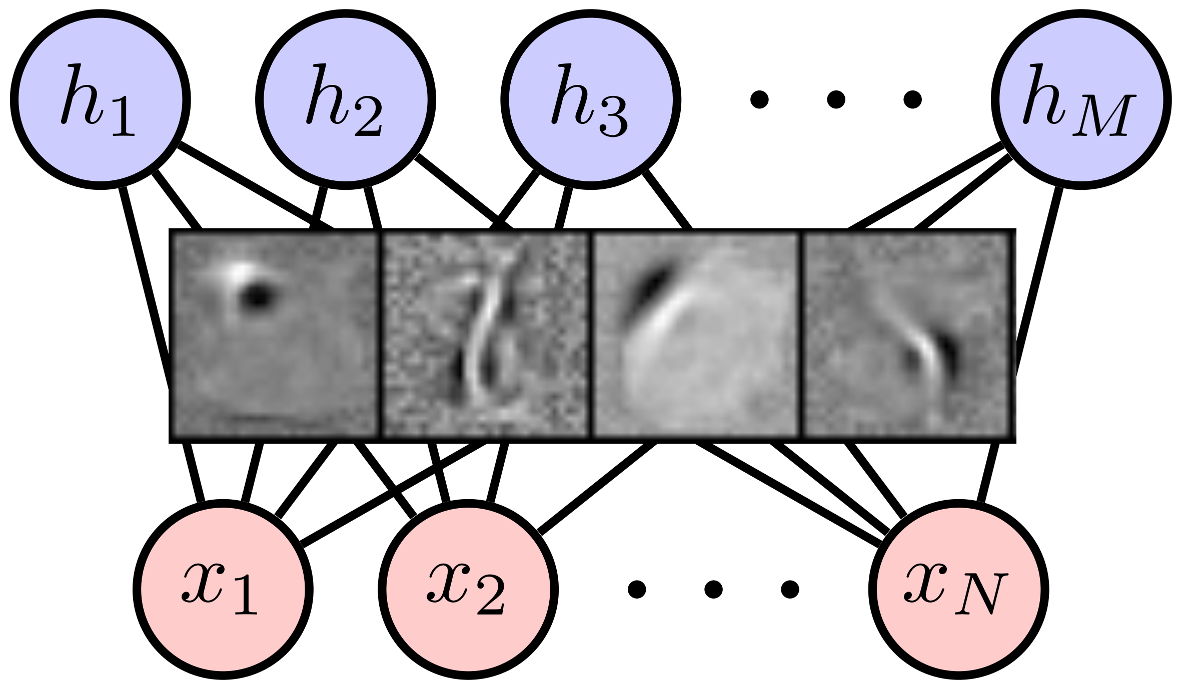


$$p(\boldsymbol{X}) = \frac{e^{-E(\boldsymbol{X})}}{Z}$$

Ackley, Hinton, Sejnowski, 1985

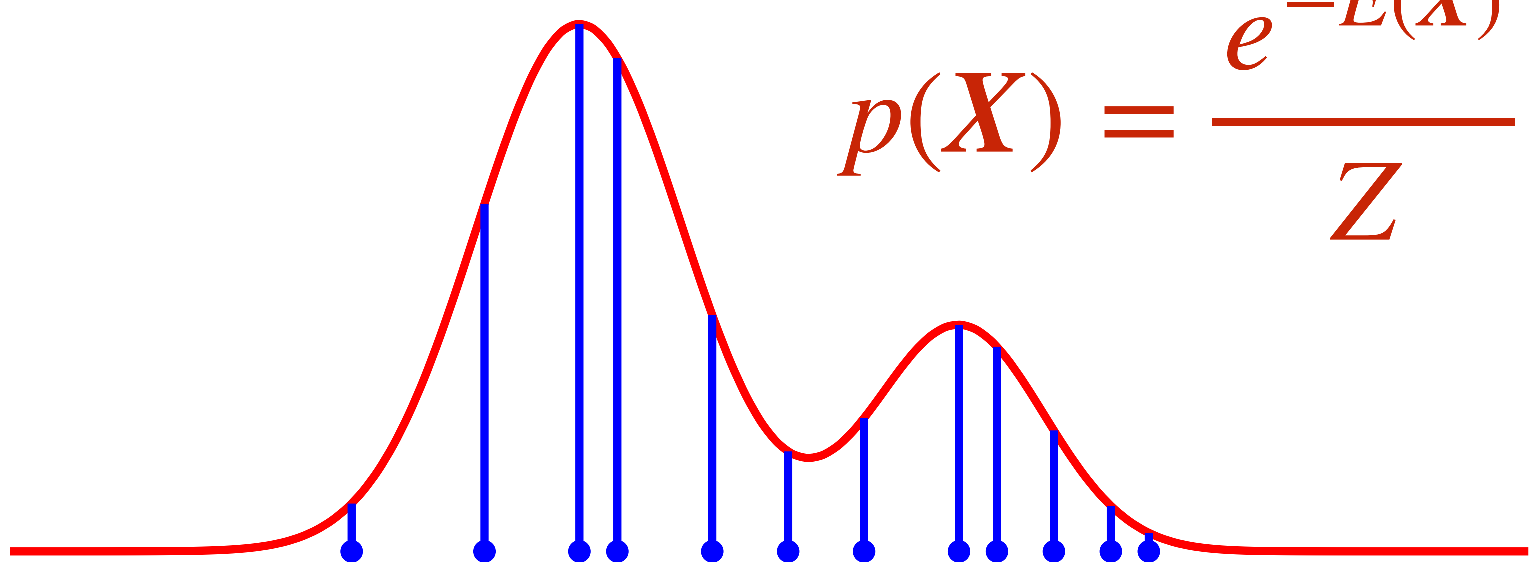
Learn

$$\mathscr{L} = \mathbb{E}_{\boldsymbol{X}\sim\text{data}}\left[-\ln p(\boldsymbol{X})\right]$$



$$\nabla_\theta \mathscr{L} = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\text{dataset}}\left[\nabla_\theta E\right] - \mathop{\mathbb{E}}_{\boldsymbol{X}\sim p(\boldsymbol{X})}\left[\nabla_\theta E\right]$$

Boltzmann machines 

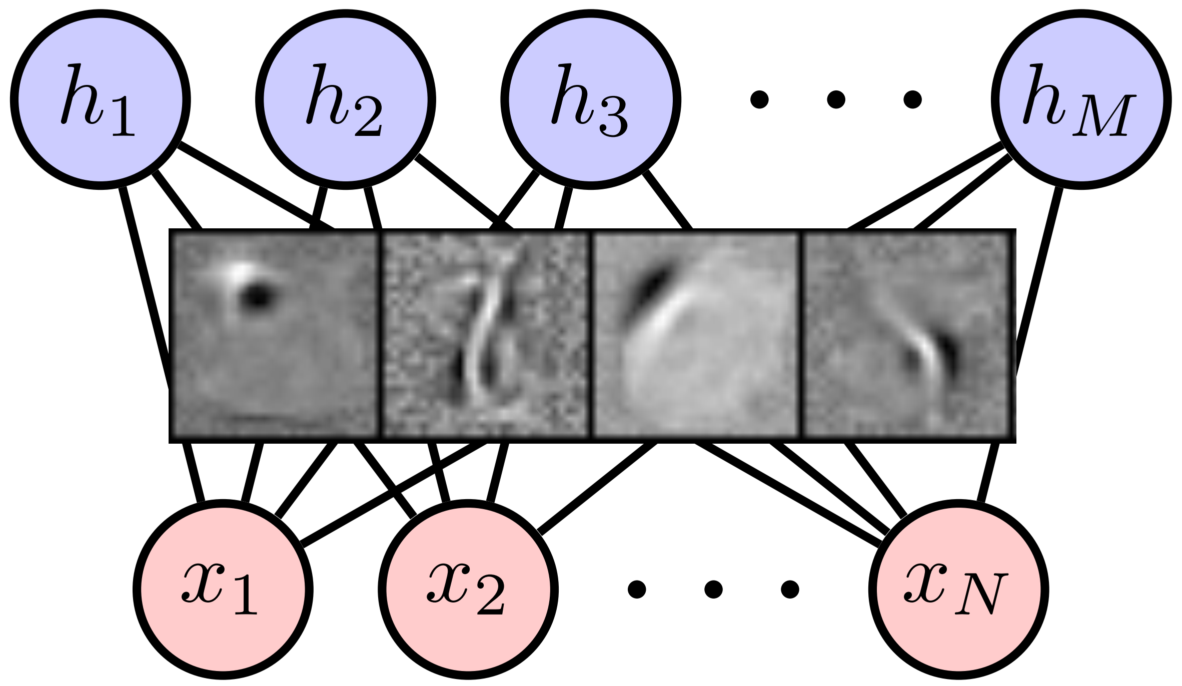


$$p(\boldsymbol{X}) = \frac{e^{-E(\boldsymbol{X})}}{Z}$$

Ackley, Hinton, Sejnowski, 1985

Learn

$$\mathscr{L} = \mathbb{E}_{X \sim \text{data}}\left[-\ln p(\boldsymbol{X})\right]$$

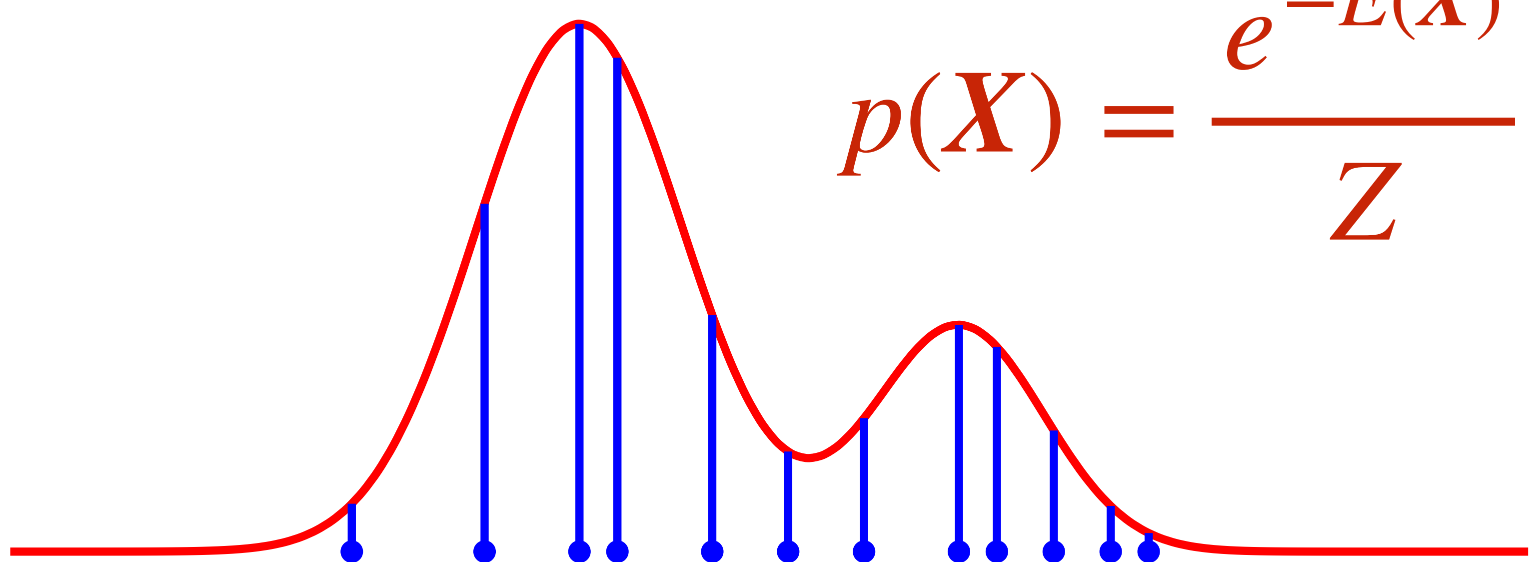


Generate

$$\boldsymbol{X} \sim p(\boldsymbol{X})$$

$$\nabla_\theta \mathscr{L} = \mathop{\mathbb{E}}_{X \sim \text{dataset}}\left[\nabla_\theta E\right] - \mathop{\mathbb{E}}_{X \sim p(X)}\left[\nabla_\theta E\right]$$

Boltzmann machines 

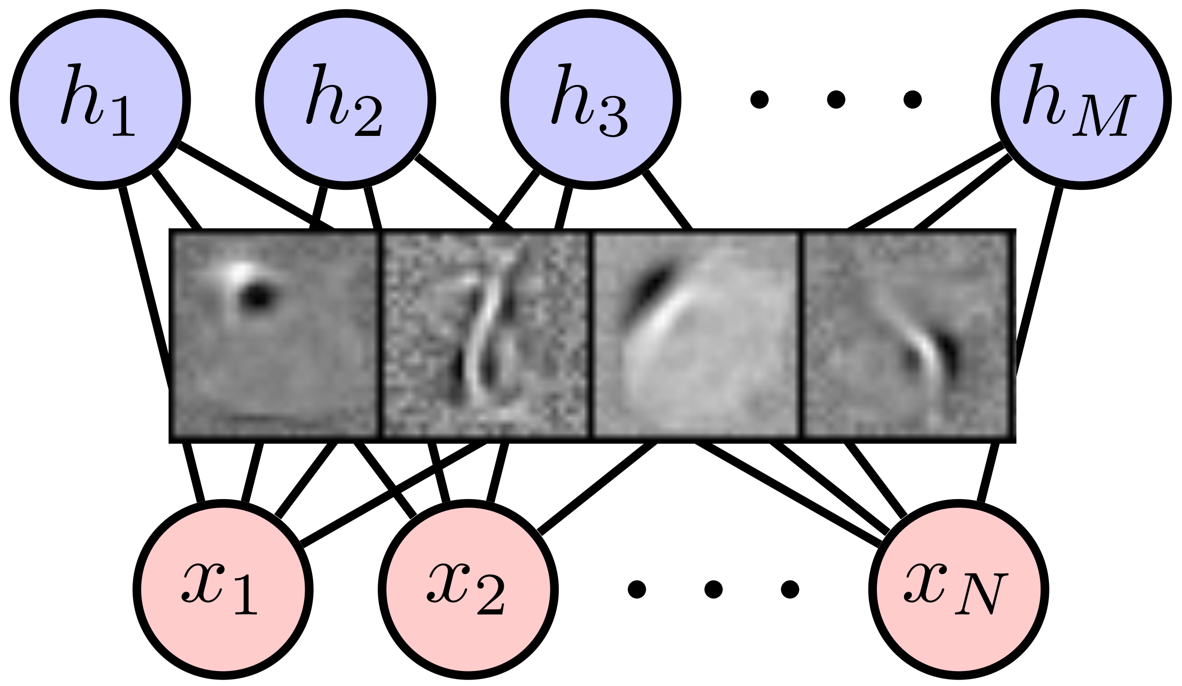


$$p(\boldsymbol{X}) = \frac{e^{-E(\boldsymbol{X})}}{Z}$$

Ackley, Hinton, Sejnowski, 1985

Learn

$$\mathscr{L} = \mathbb{E}_{X\sim\text{data}}\left[-\ln p(\boldsymbol{X})\right]$$

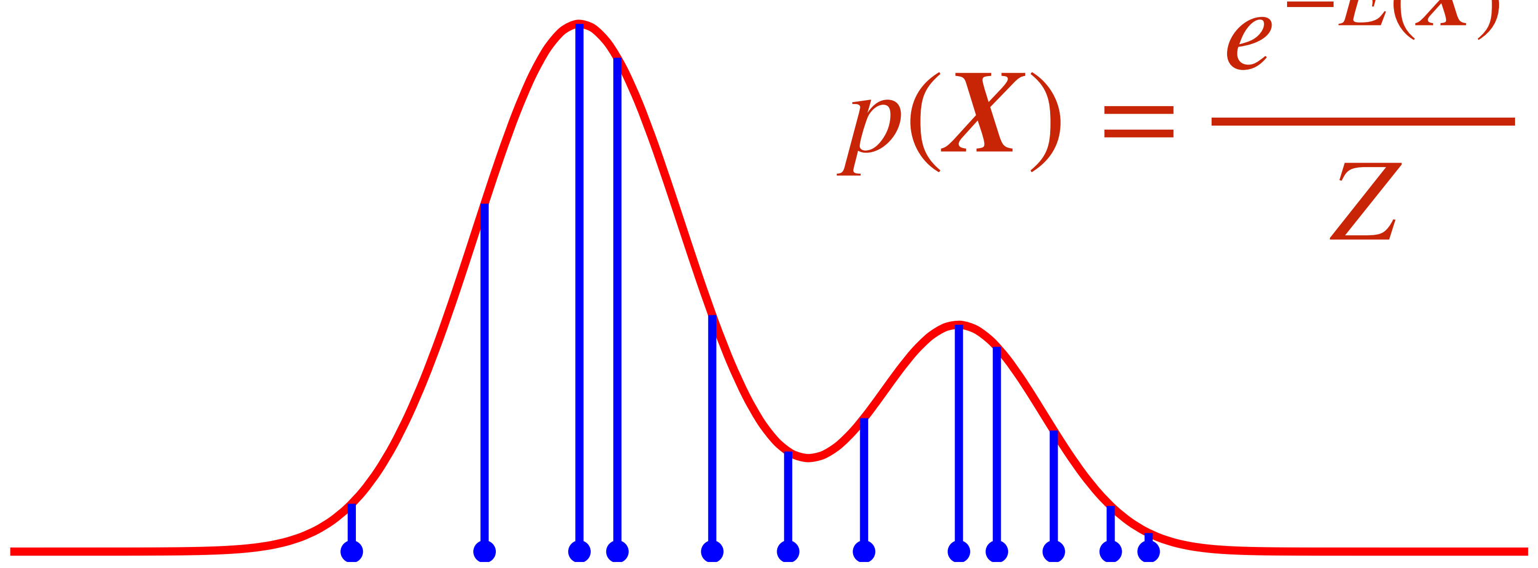


Generate

$$\boldsymbol{X} \sim p(\boldsymbol{X})$$

$$\nabla_\theta \mathscr{L} = \mathop{\mathbb{E}}_{X\sim\text{dataset}}\left[\nabla_\theta E\right] - \mathop{\mathbb{E}}_{X\sim p(X)}\left[\nabla_\theta E\right]$$

Boltzmann machines 



$$p(\boldsymbol{X}) = \frac{e^{-E(\boldsymbol{X})}}{Z}$$

Ackley, Hinton, Sejnowski, 1985

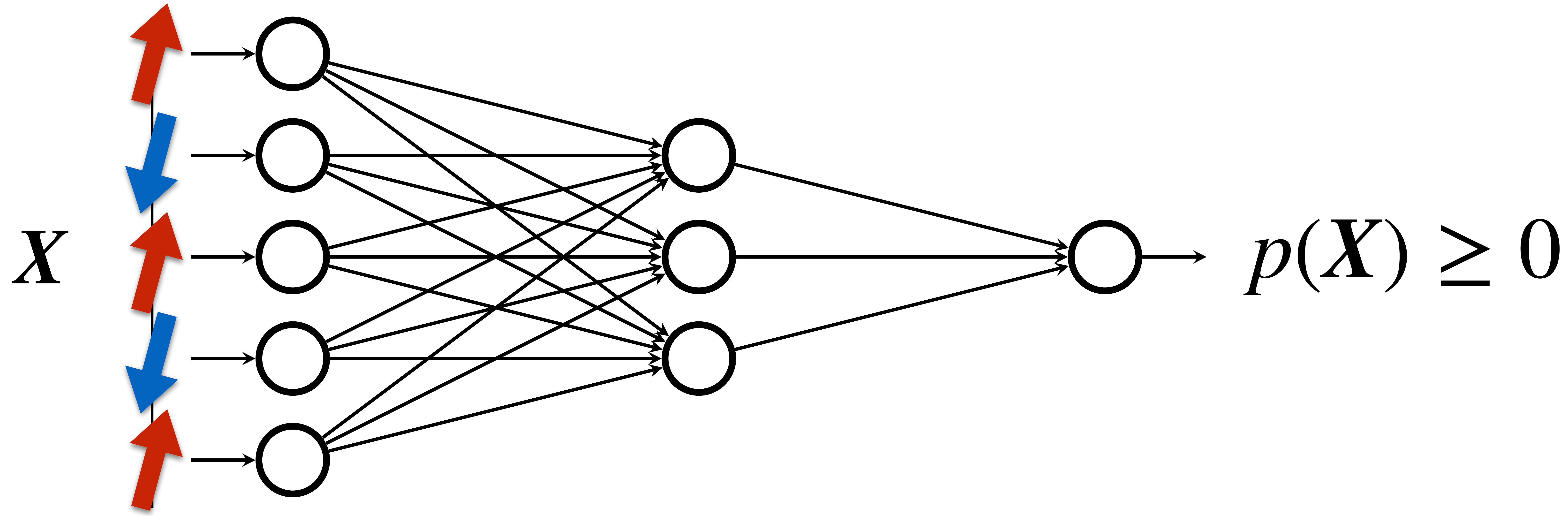
2210.10318

GAUSSIAN-BERNOULLI RBMS WITHOUT TEARS 😂

Renjie Liao*[1], Simon Kornblith[2], Mengye Ren[3], David J. Fleet[2,4,5], Geoffrey Hinton[2,4,5]

$$\nabla_\theta \mathscr{L} = \mathop{\mathbb{E}}_{X \sim \text{dataset}} \left[\nabla_\theta E\right] - \mathop{\mathbb{E}}_{X \sim p(X)} \left[\nabla_\theta E\right]$$

So, why bother ?



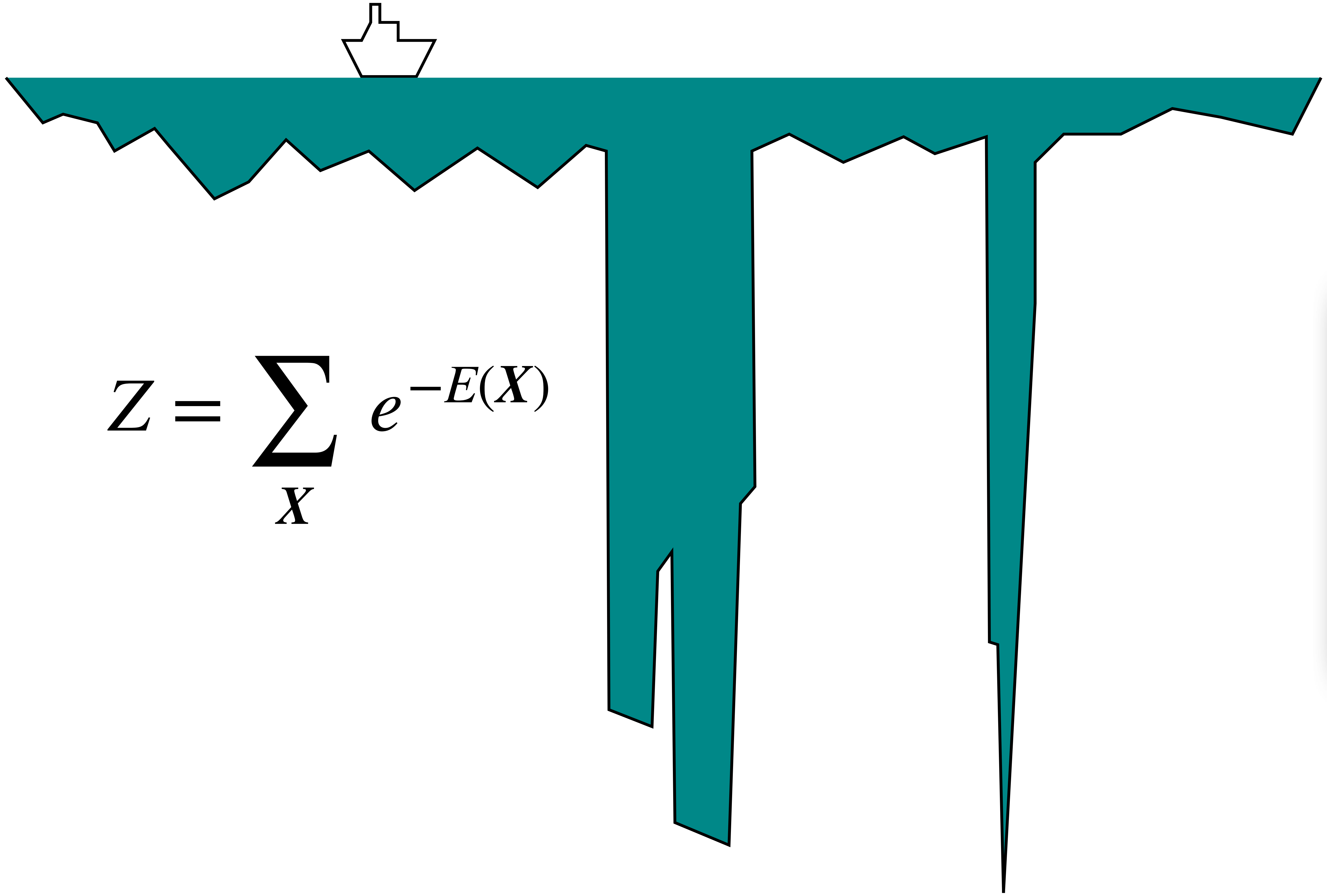
Normalization ?

$$\sum_{X} p(X)$$

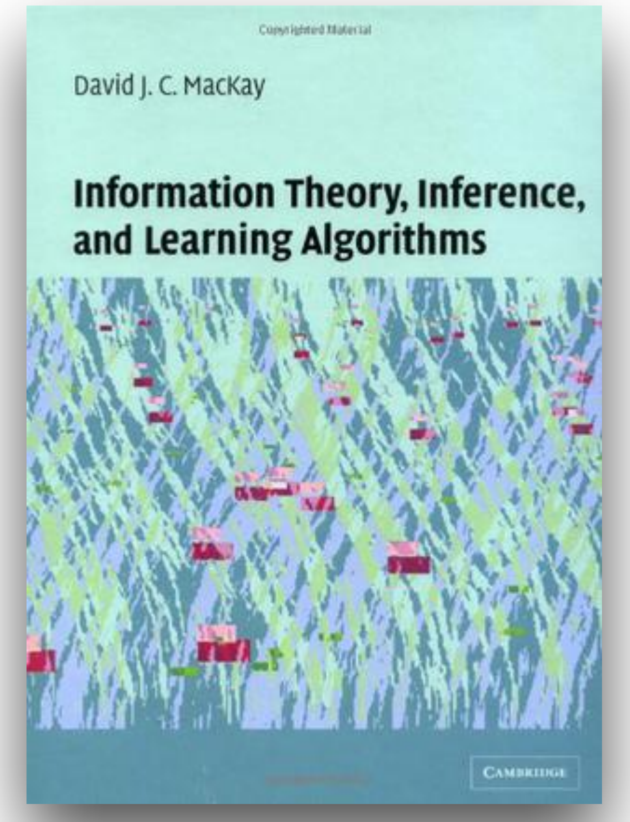
Sampling ?

$$\boldsymbol{X} \sim p(\boldsymbol{X})$$

The difficulty of normalization



$$Z = \sum_{X} e^{-E(X)}$$

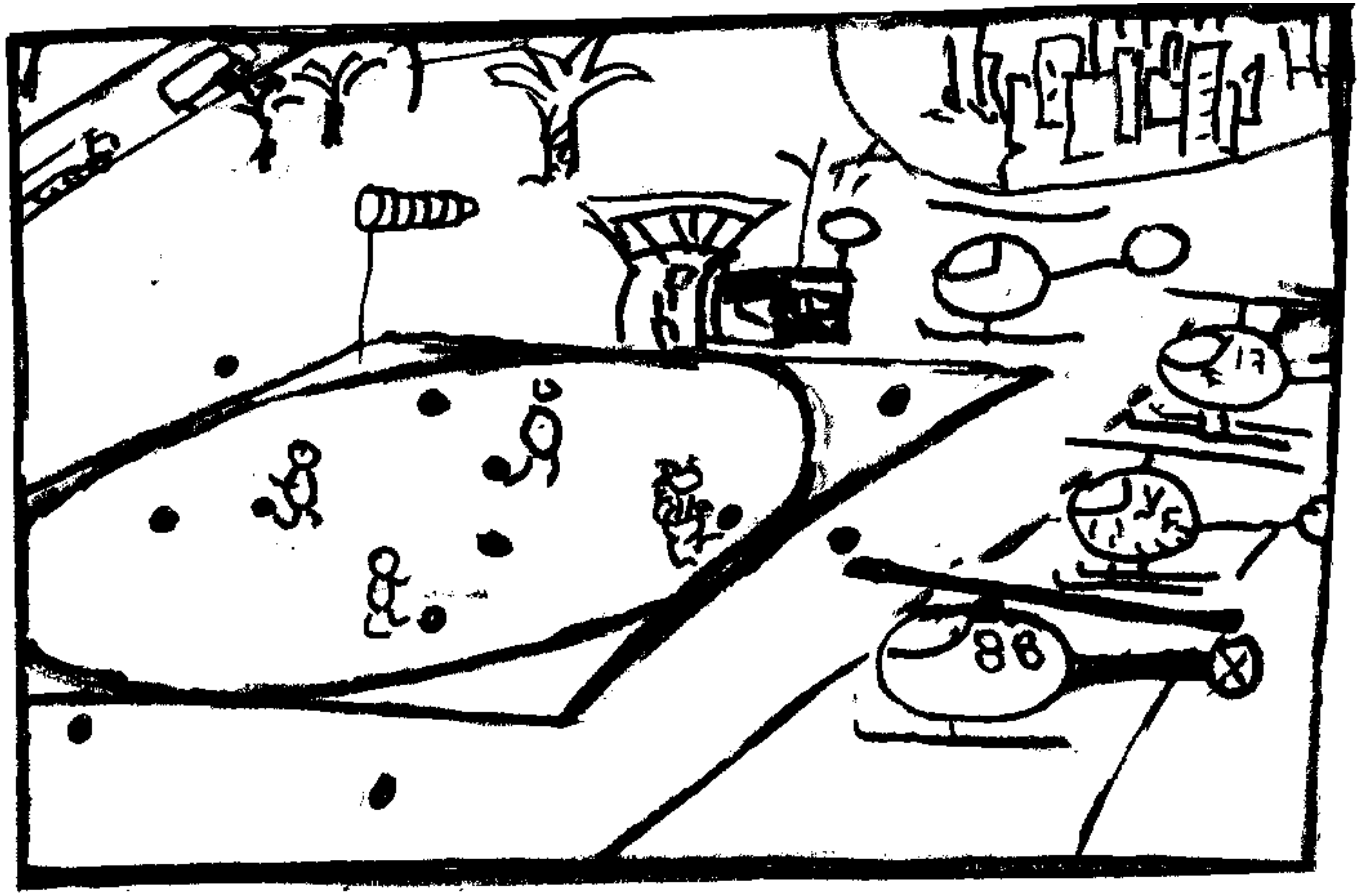


“Intractable” partition function Z

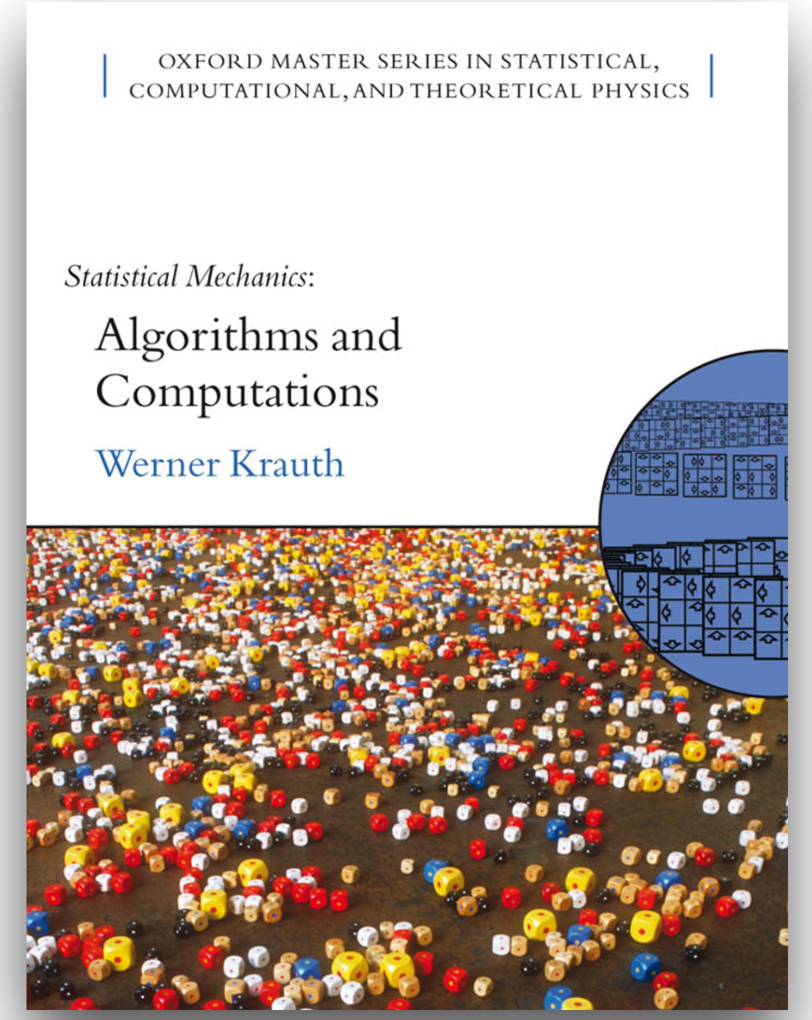
appears widely in machine learning and statistical physics (entropy and free energy calculation)

The difficulty of sampling

$$\boldsymbol{X} \sim p(\boldsymbol{X})$$

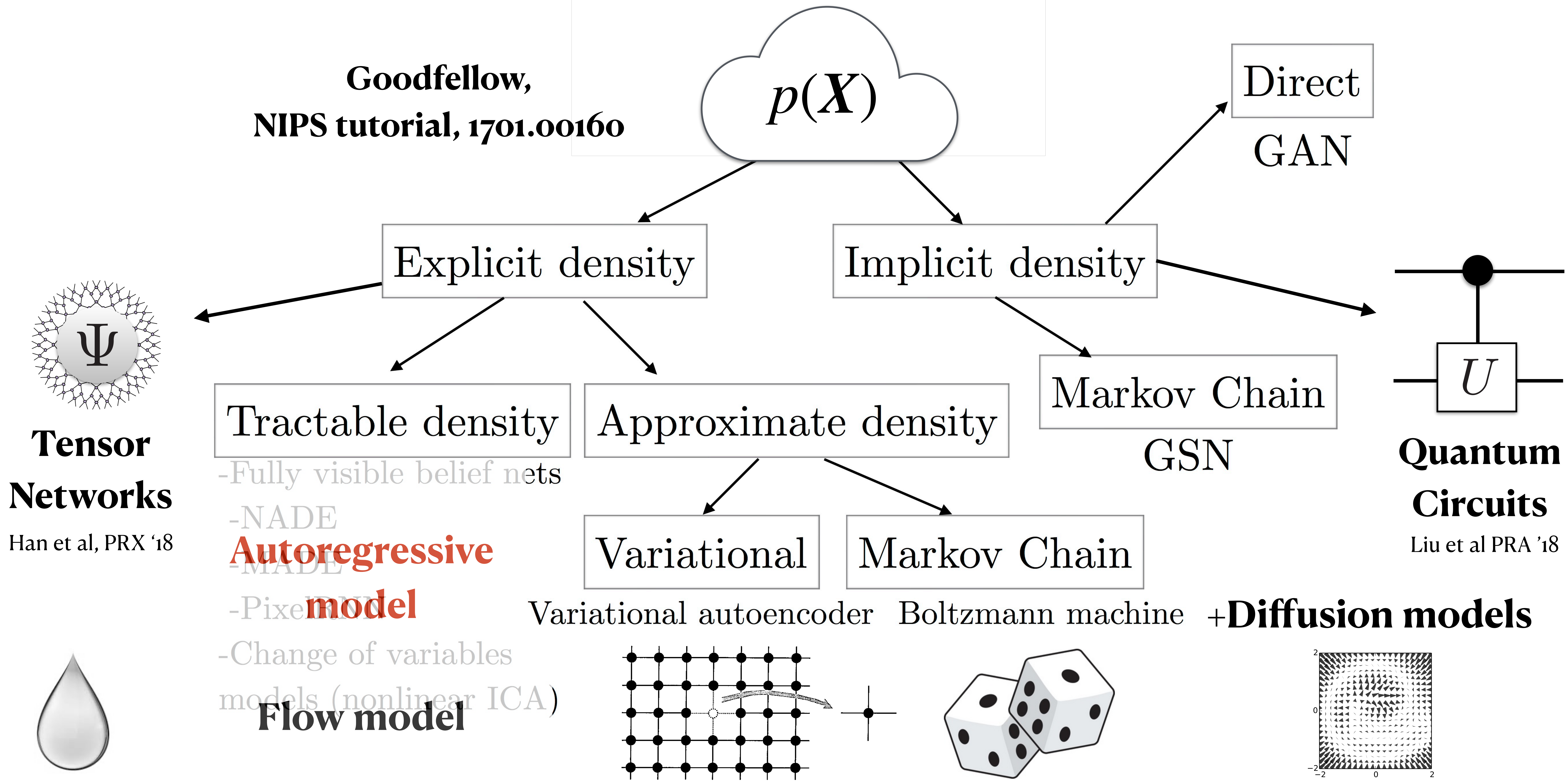


Adults computing the number π at the Monte Carlo heliport.



Direct sampling is generally difficult in high-dimensional space

Generative models and their physics genes



Autoregressive model

$$p(\boldsymbol{X}) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1, x_2)\cdots$$

“... the murderer is ___”

$p(\_ \,|\, \ldots)$

Normalization

$$\sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \,|\, x_1) \sum_{x_3} p(x_3 \,|\, x_1, x_2)\cdots$$

Sampling

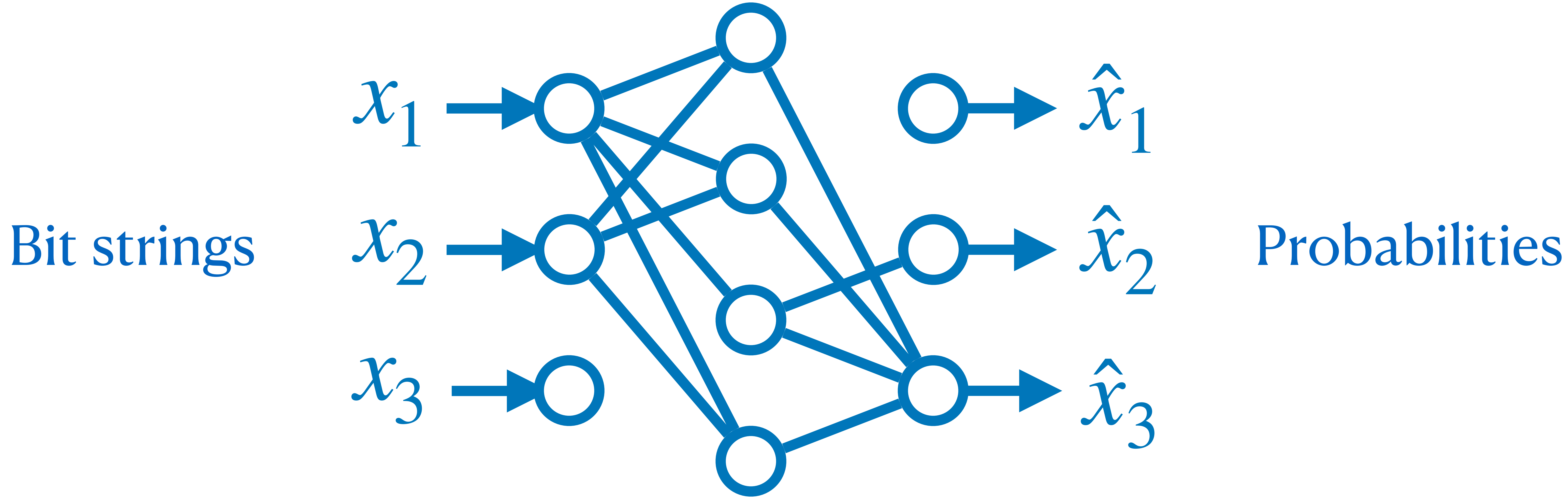
$$x_1 \sim p(x_1)$$

$$x_2 \sim p(x_2 \,|\, x_1)$$

$$\vdots$$

Implementation: autoregressive masks

Masked Autoencoder Germain et al, 1502.03509



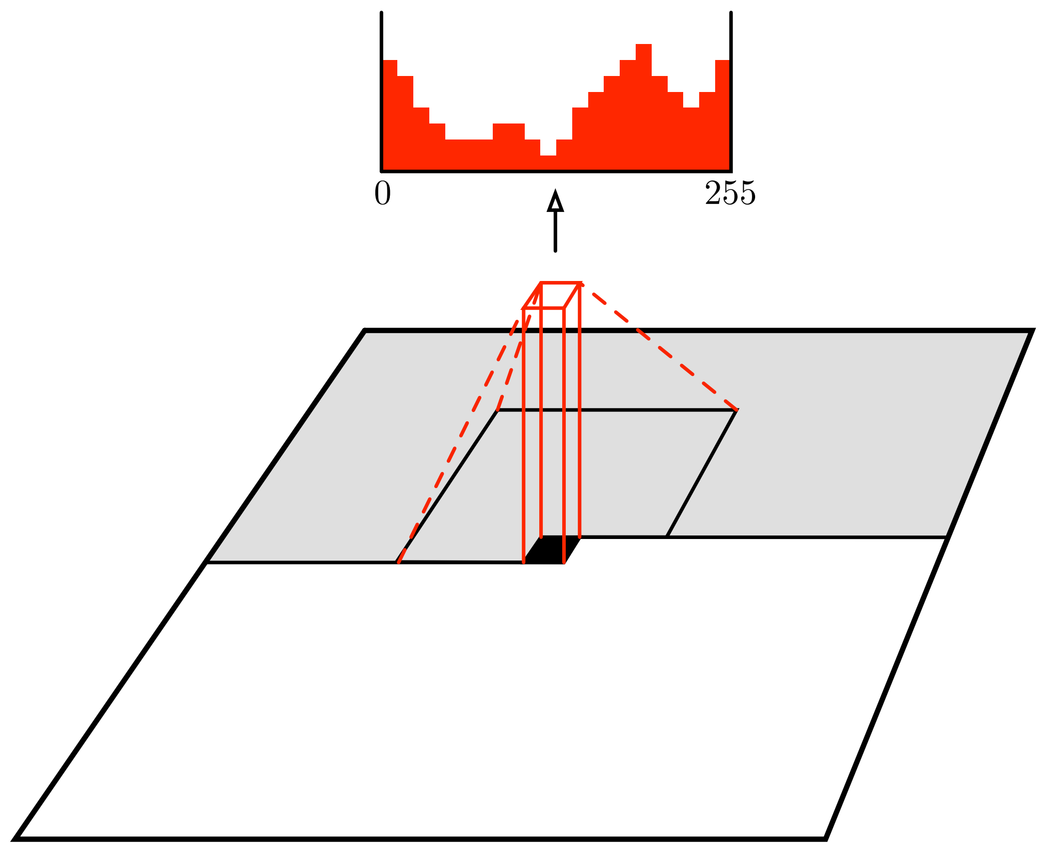
$p(x_1) = \text{Bernoulli}(\hat{x}_1)$ $\quad p(x_2 \mid x_1) = \text{Bernoulli}(\hat{x}_2)$ $\quad p(x_3 \mid x_1, x_2) = \text{Bernoulli}(\hat{x}_3)$

https://github.com/karpathy/pytorch-made

Implementation: autoregressive masks

Mask convolutional kernel

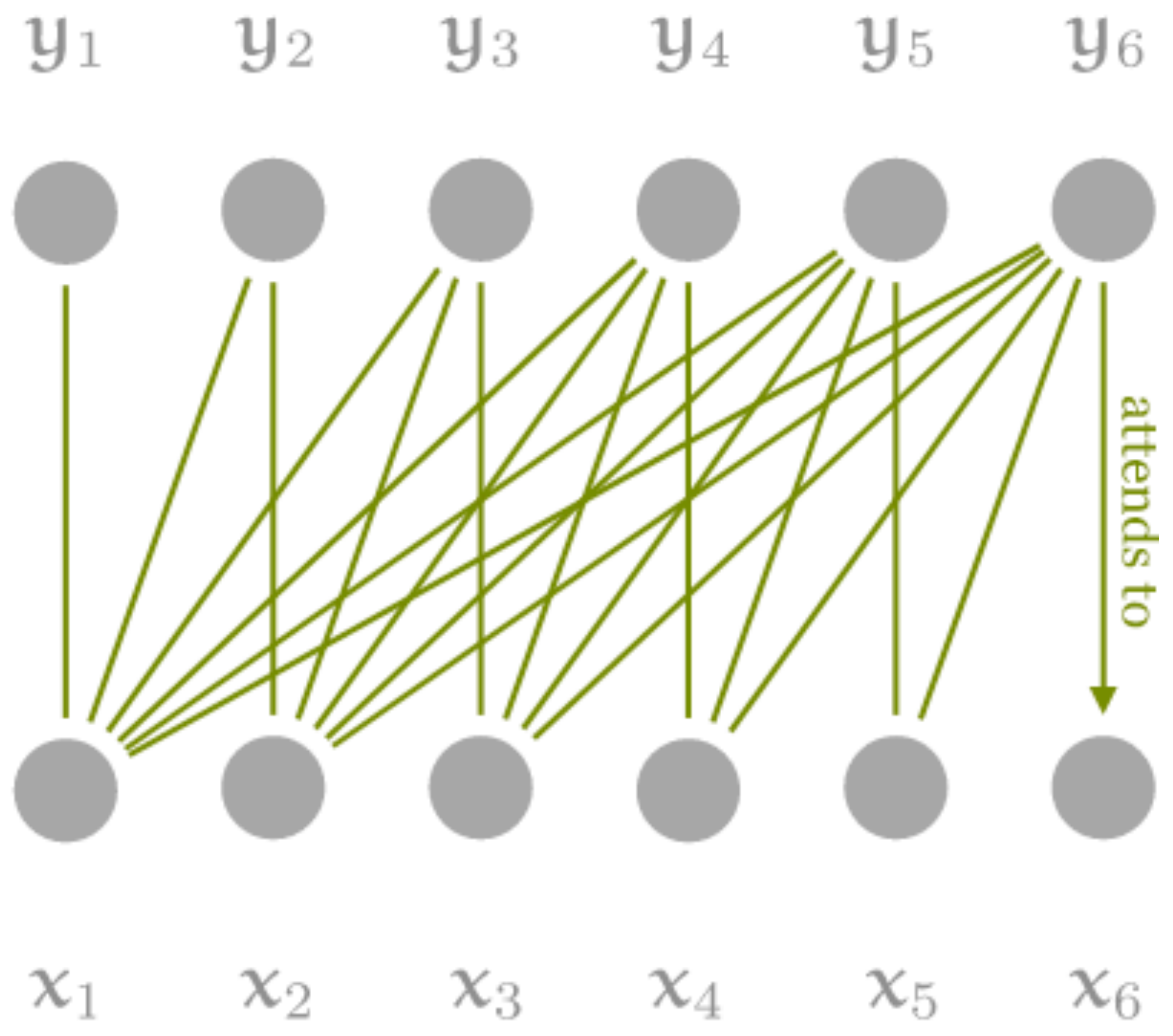
PixelCNN, van den Oord et al, 1601.06759



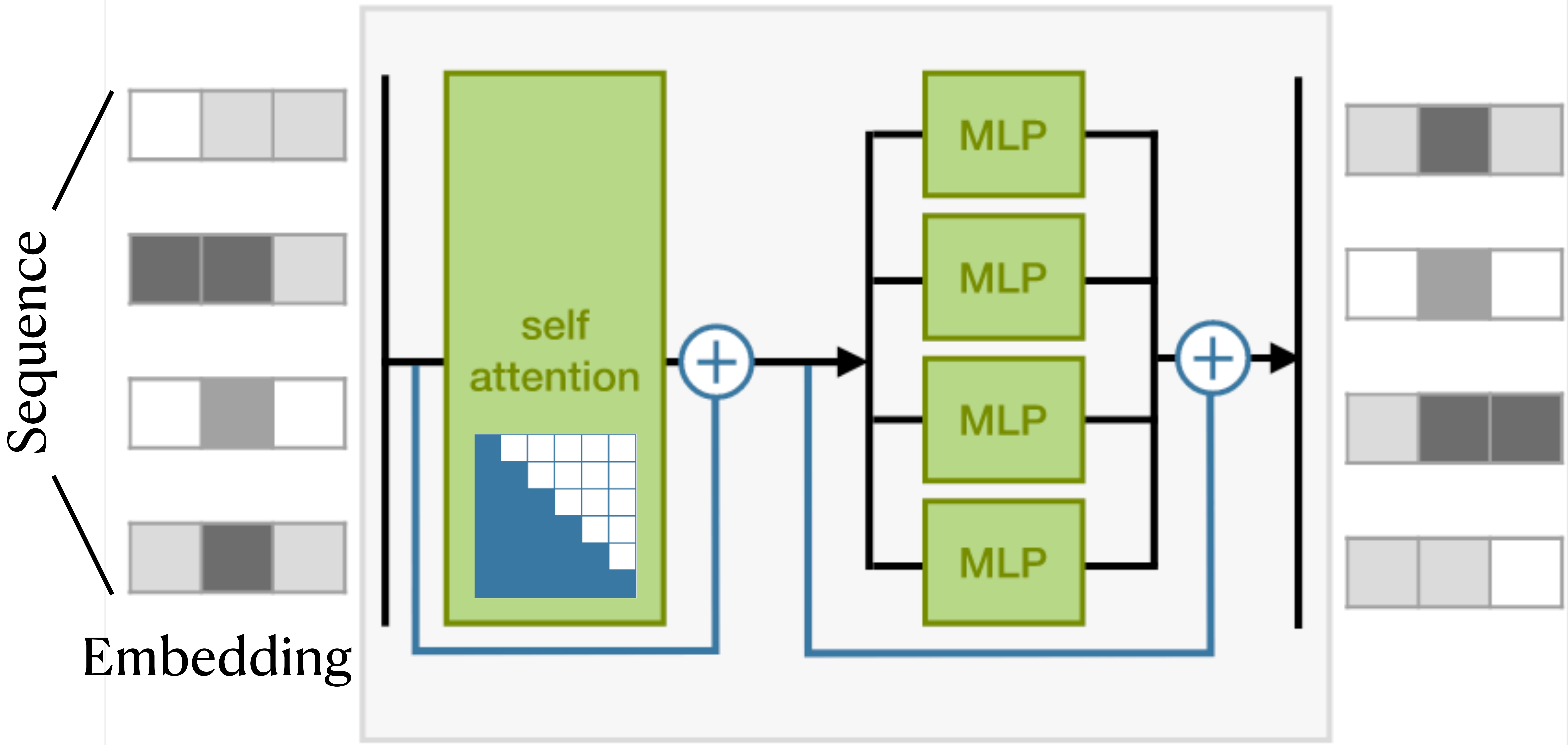
1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Mask self-attention matrix

Causal transformer, Vaswani et al 1706.03762



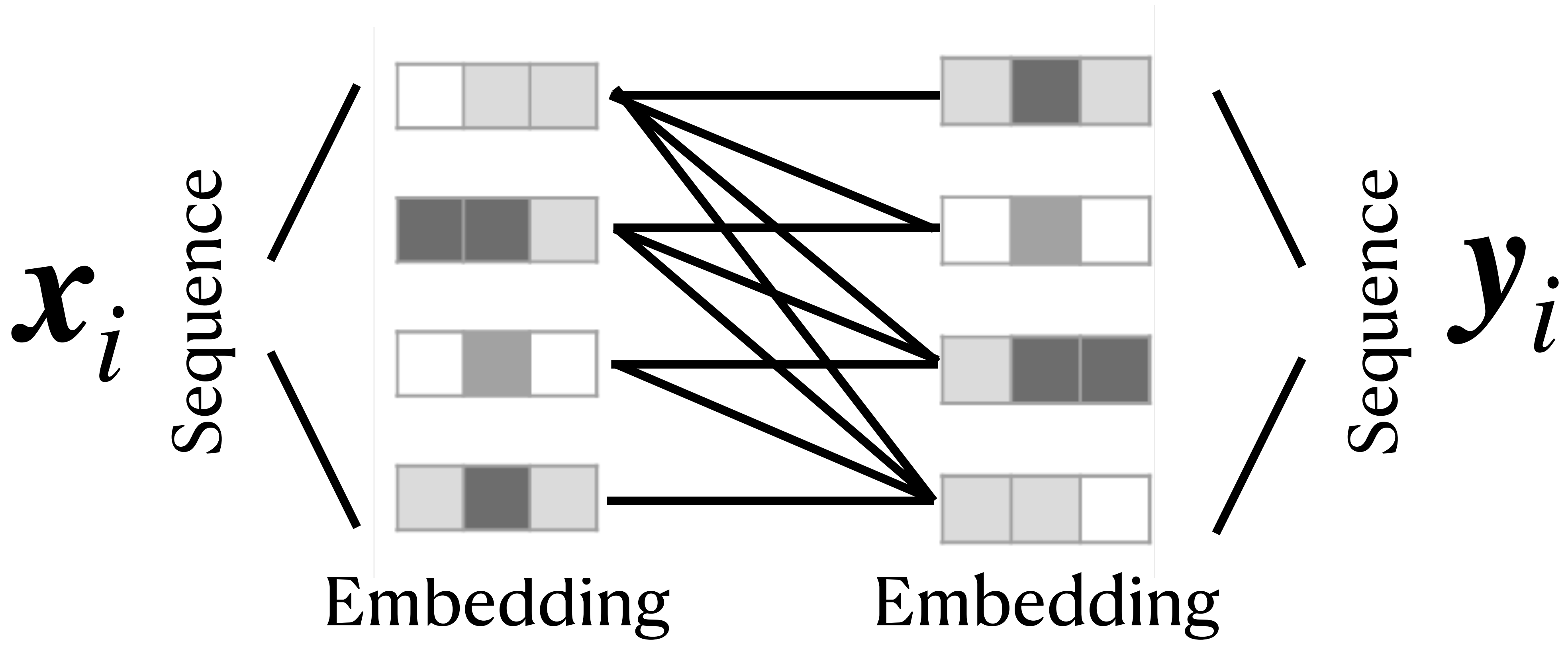
The autoregressive transformer

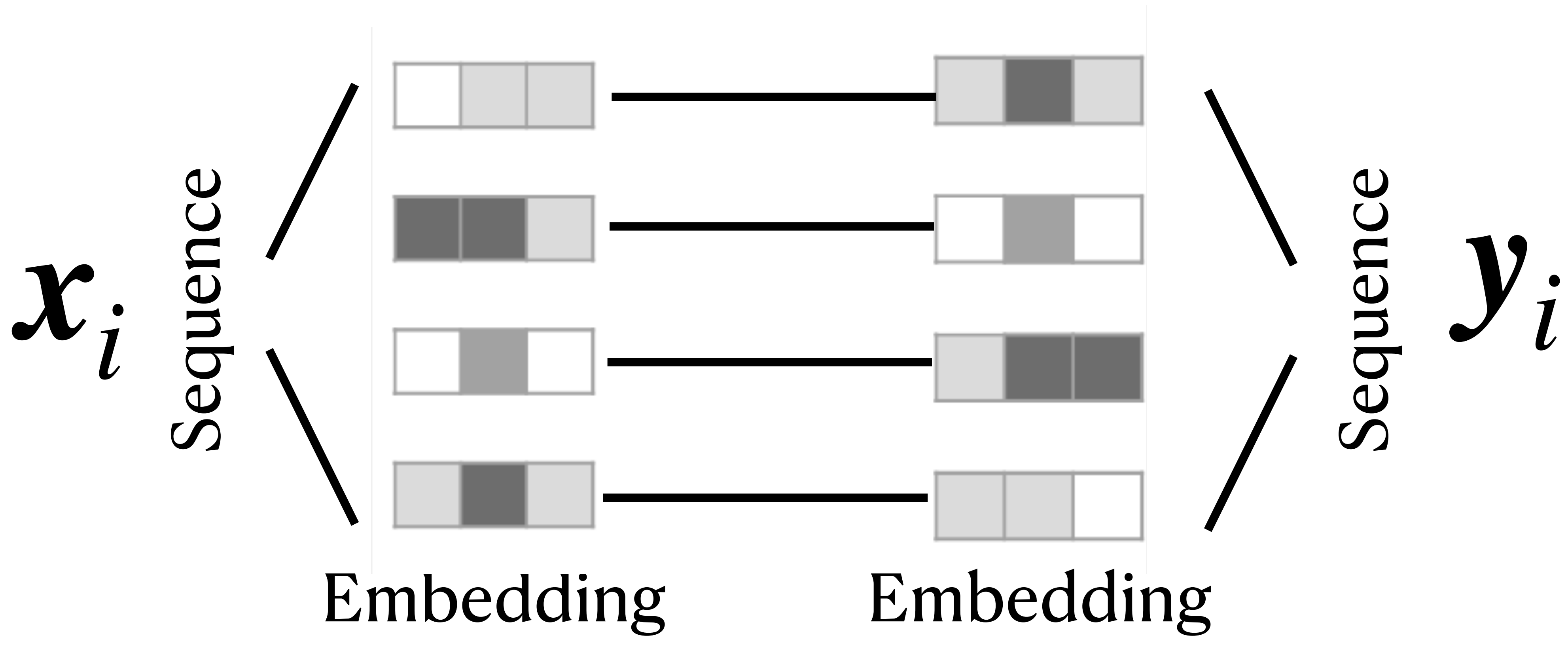


Masked attention matrix => lower triangular Jacobian matrix => autoregressive model

Great at capturing long-range dependence; friendly to backpropagation and GPUs

Transformers from scratch, https://peterbloem.nl/blog/transformers





1 Self-attention

$$y_i = \sum_j \alpha(x_i, x_j) x_j$$

attention weights

Mixing signal along sequence direction

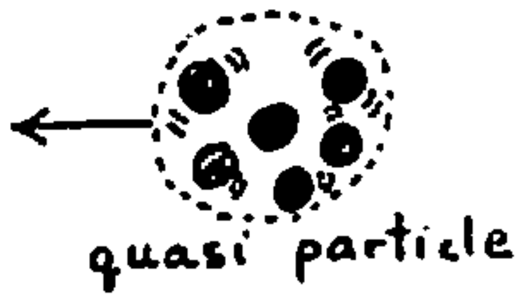
2 Multi-layer perceptrons

$$y_i = \sigma \circ \cdots \sigma(x_i W + b)$$

Nonlinear activation

Transforming signals locally

Feynman's backflow as an attention layer



$$z_i = \boldsymbol{x}_i + \sum_{j\neq i} \eta(|\boldsymbol{x}_i - \boldsymbol{x}_j|)(\boldsymbol{x}_j - \boldsymbol{x}_i)$$

Quasi-particle coordinates

Feynman & Cohen 1956
wavefunction for liquid Helium

Electron coordinates

Each particle attends to its surrounding and dresses up as a quasi-particle

c.f. Lu et al, 1906.02762 transformer as convection-diffusion multi-particle dynamics picoGPT — wanglei@bright90:~ — vi gpt2_pico.py — 116×63

```python
import numpy as np

def gelu(x):
    return 0.5 * x * (1 + np.tanh(np.sqrt(2 / np.pi) * (x + 0.044715 * x**3)))

def softmax(x):
    exp_x = np.exp(x - np.max(x, axis=-1, keepdims=True))
    return exp_x / np.sum(exp_x, axis=-1, keepdims=True)

def layer_norm(x, g, b, eps: float = 1e-5):
    mean = np.mean(x, axis=-1, keepdims=True)
    variance = np.var(x, axis=-1, keepdims=True)
    return g * (x - mean) / np.sqrt(variance + eps) + b

def linear(x, w, b):
    return x @ w + b

def ffn(x, c_fc, c_proj):
    return linear(gelu(linear(x, **c_fc)), **c_proj)

def attention(q, k, v, mask):
    return softmax(q @ k.T / np.sqrt(q.shape[-1]) + mask) @ v

def mha(x, c_attn, c_proj, n_head):
    x = linear(x, **c_attn)
    qkv_heads = list(map(lambda x: np.split(x, n_head, axis=-1), np.split(x, 3, axis=-1)))
    causal_mask = (1 - np.tri(x.shape[0], dtype=x.dtype)) * -1e10
    out_heads = [attention(q, k, v, causal_mask) for q, k, v in zip(*qkv_heads)]
    x = linear(np.hstack(out_heads), **c_proj)
    return x

def transformer_block(x, mlp, attn, ln_1, ln_2, n_head):
    x = x + mha(layer_norm(x, **ln_1), **attn, n_head=n_head)
    x = x + ffn(layer_norm(x, **ln_2), **mlp)
    return x

def gpt2(inputs, wte, wpe, blocks, ln_f, n_head):
    x = wte[inputs] + wpe[range(len(inputs))]
    for block in blocks:
        x = transformer_block(x, **block, n_head=n_head)
    return layer_norm(x, **ln_f) @ wte.T

def generate(inputs, params, n_head, n_tokens_to_generate):
    from tqdm import tqdm
    for _ in tqdm(range(n_tokens_to_generate), "generating"):
        logits = gpt2(inputs, **params, n_head=n_head)
        next_id = np.argmax(logits[-1])
        inputs.append(int(next_id))
    return inputs[len(inputs) - n_tokens_to_generate :]

def main(prompt: str, n_tokens_to_generate: int = 40, model_size: str = "124M", models_dir: str = "models"):
    from utils import load_encoder_hparams_and_params
    encoder, hparams, params = load_encoder_hparams_and_params(model_size, models_dir)
    input_ids = encoder.encode(prompt)
    assert len(input_ids) + n_tokens_to_generate < hparams["n_ctx"]
    output_ids = generate(input_ids, params, hparams["n_head"], n_tokens_to_generate)
    output_text = encoder.decode(output_ids)
    return output_text

if __name__ == "__main__":
    import fire
    fire.Fire(main)
"gpt2_pico.py" 62L, 2330B
```

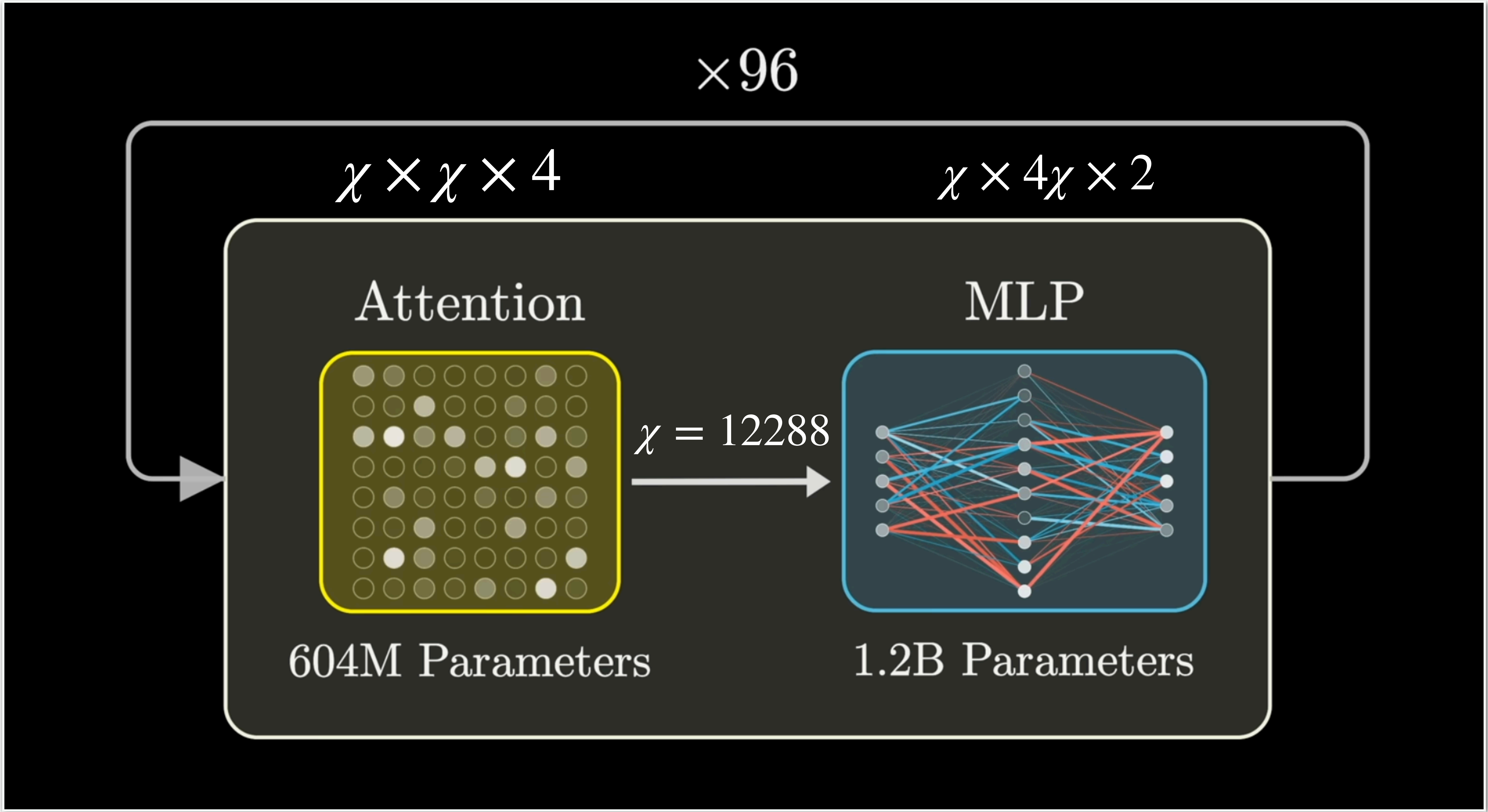
GPT2 in 60 lines of numpy

https://jaykmody.com/blog/gpt-from-scratch

Params count in GPT3

Brown et al, 2005.14165

3Blue1Brown, https://youtu.be/9-JlodxWQs8?t=940



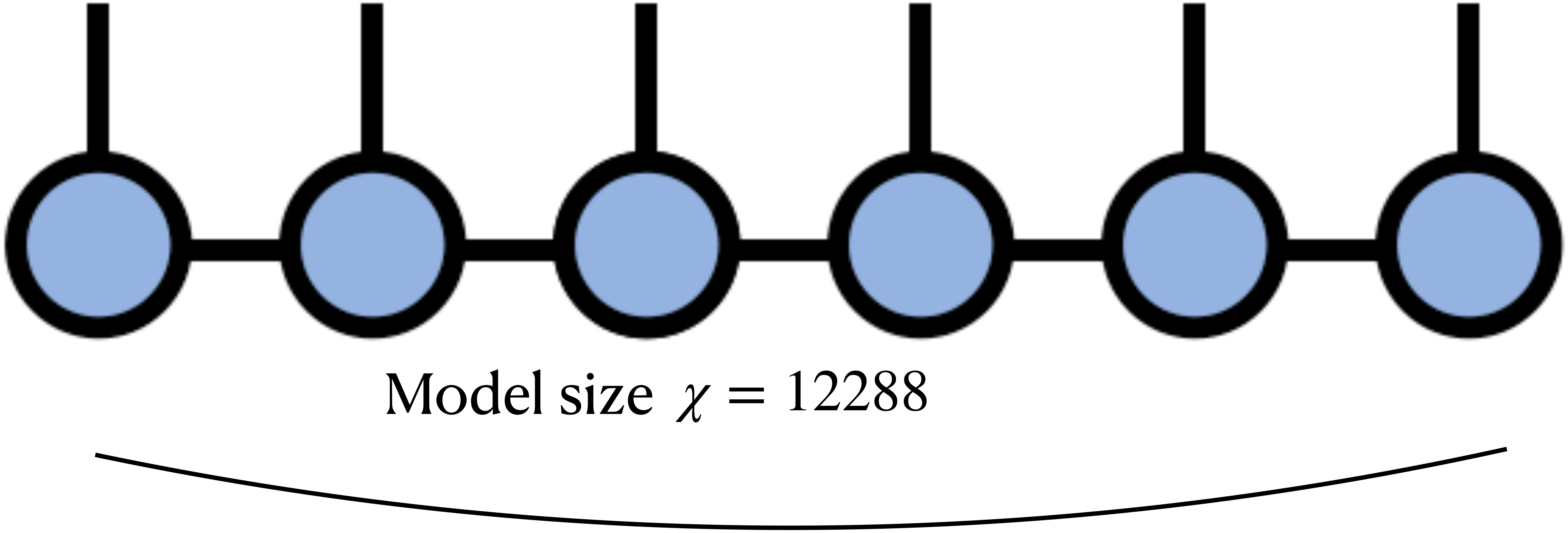
175B in total

Independent of the context length (4096) and vocabulary size (50257, almost)

An MPS analog

Vocabulary

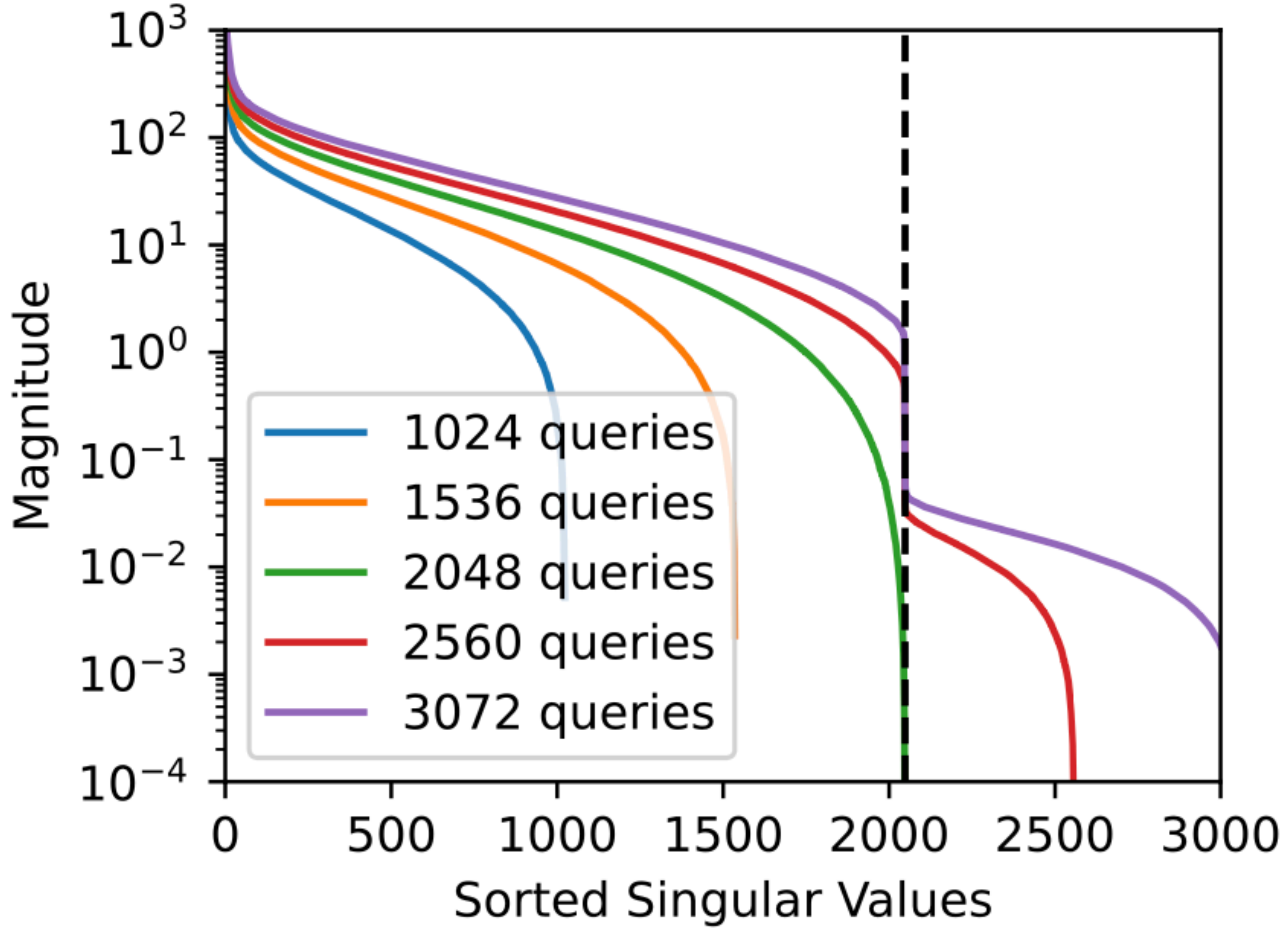
$d = 50257$



Model size $\chi = 12288$

Contex length $L = 4096$

Aside: SVD attack



What is χ?

If you only have access to d-dim logits via the LLM API

$$\text{svd}\begin{bmatrix} \boxed{\cdot \quad \cdot \quad \cdot \quad \cdot} \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \end{bmatrix}_{\vdots \times d}$$

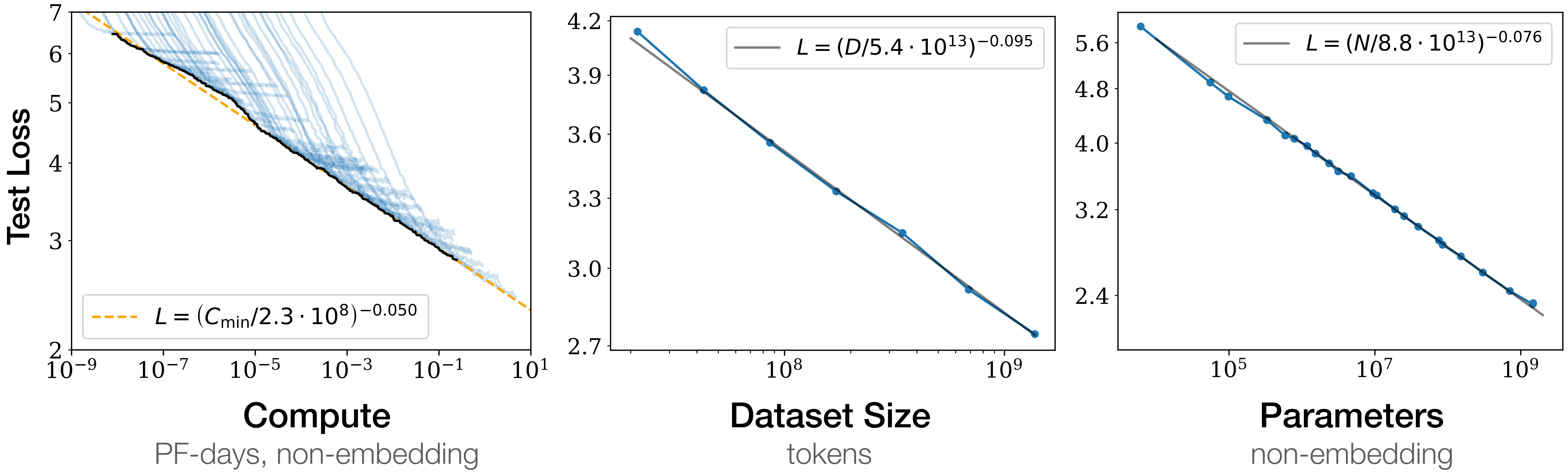
$\chi < d$

Carlini et al, Stealing Part of a Production Language Model, 2403.06634
Finlayson et al, Logits of API-Protected LLMs Leak Proprietary Information, 2403.09539

Scaling law of the loss function

$$\mathscr{L} = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\text{dataset}} \left[-\ln p(\boldsymbol{X})\right]$$

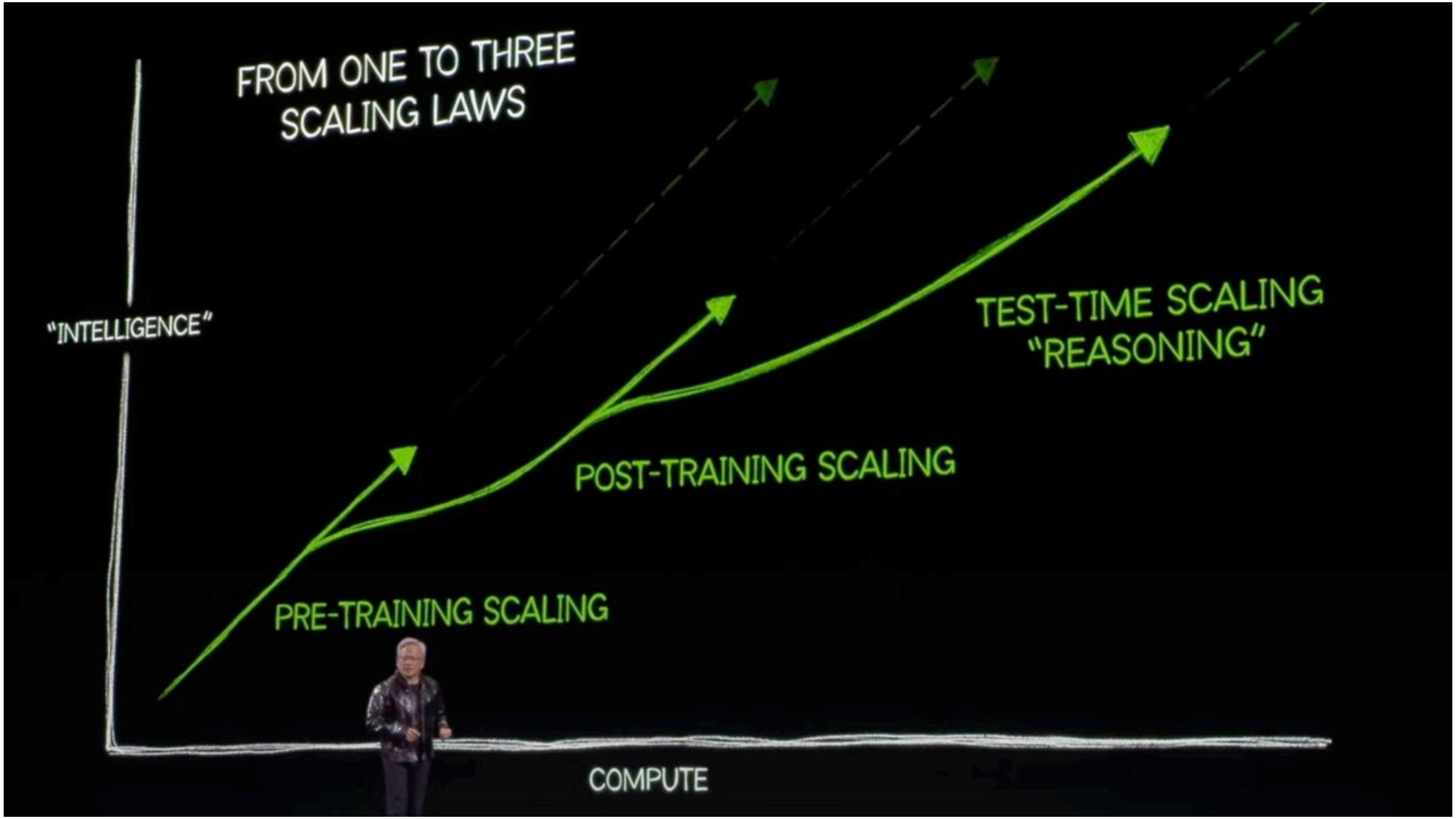
Kaplan et al, 2001.08361



"Predict resouces needed to sovle increasingly difficult tasks" — Sam McCandlish, Aspen talk '19

https://sites.google.com/view/phys4ml/home

A trillion $ plot

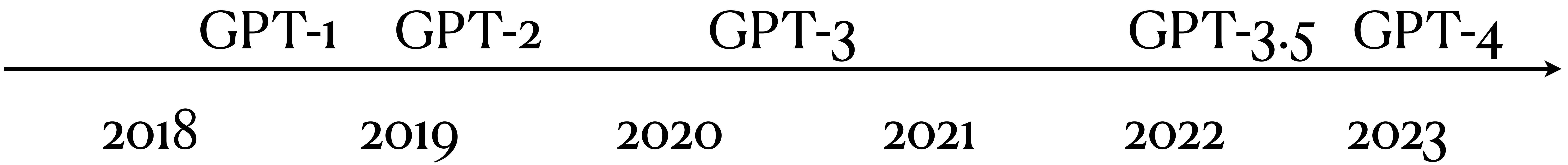


A visionary discussion section

Scaling Laws for Neural Language Models, Kaplan et al, 2001.08361
Scaling Laws for Autoregressive Generative Modeling, Henighan et al, 2010.14701

It is natural to conjecture that the scaling relations will apply to other generative modeling tasks with a maximum likelihood loss, and perhaps in other settings as well. To this purpose, it will be interesting to test these relations on other domains, such as images, audio, and video models, and perhaps also for random network distillation. At this point we do not know which of our results depend on the structure of natural language data, and which are universal. It would also be exciting to find a theoretical framework from which the scaling relations can be derived: a 'statistical mechanics' underlying the 'thermodynamics' we have observed. Such a theory might make it possible to derive other more precise predictions, and provide a systematic understanding of the limitations of the scaling laws.

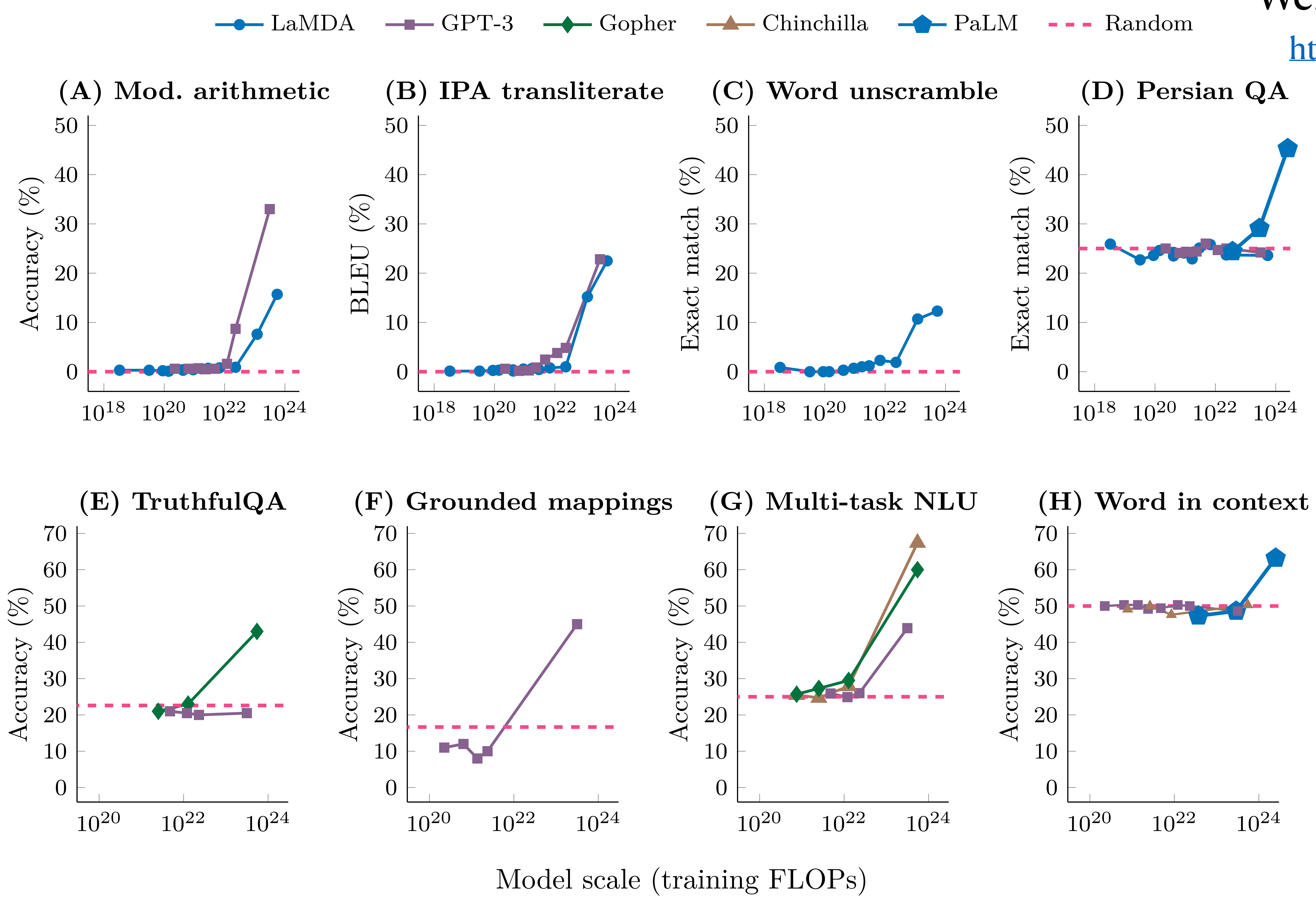
In the domain of natural language, it will be important to investigate whether continued improvement on the loss translates into improvement on relevant language tasks. Smooth quantitative change can mask major qualitative improvements: "more is different". For example, the smooth aggregate growth of the economy provides no indication of the specific technological developments that underwrite it. Similarly, the smooth improvements in language model loss may hide seemingly qualitative changes in capability.



Emergent abilities: more is different

Wei et al, 2206.07682

https://www.jasonwei.net/blog/emergence



Avogadro constant number of FLOPs

Autoregressive model is more than language modeling

“Language” => token sequence => bitstream => **ANYTHING**

Speech: WaveNet 1609.03499

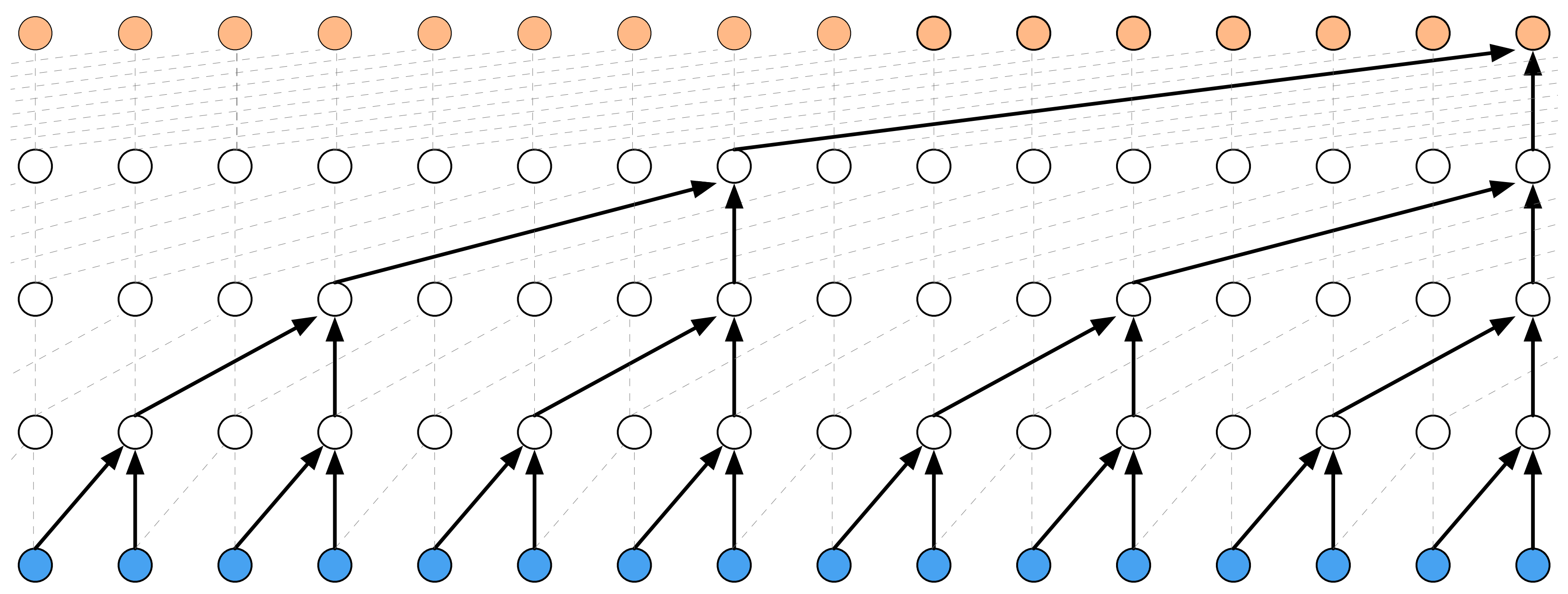
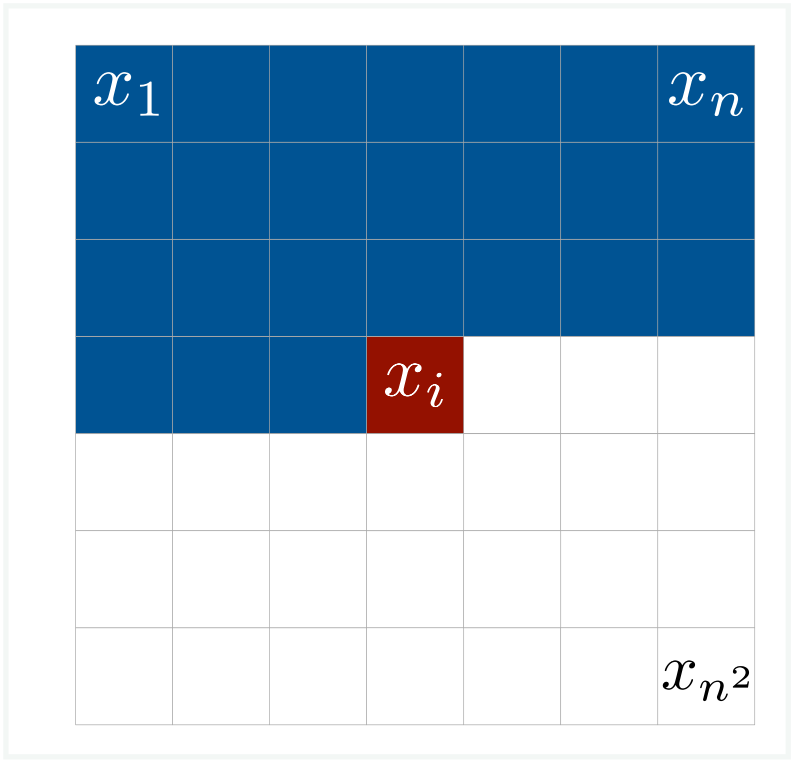
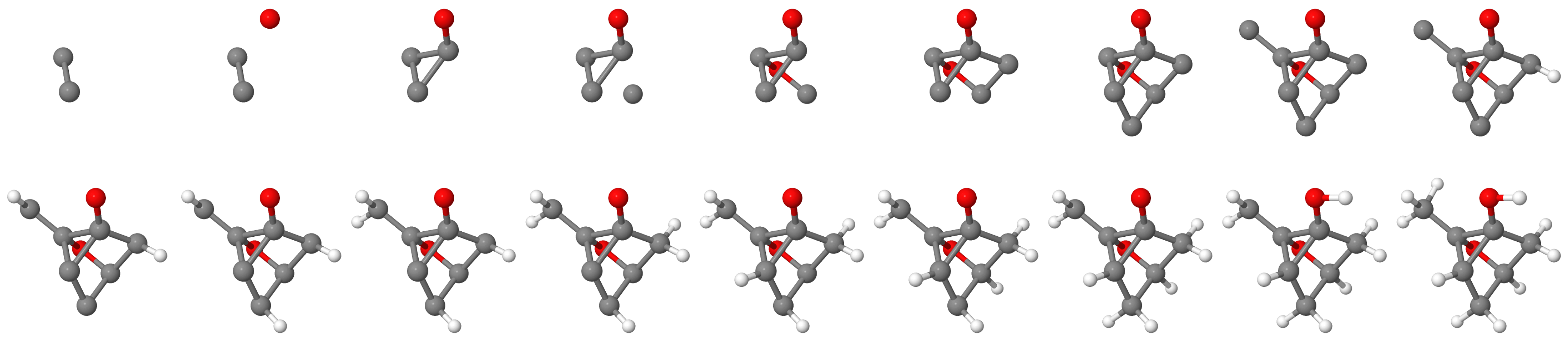


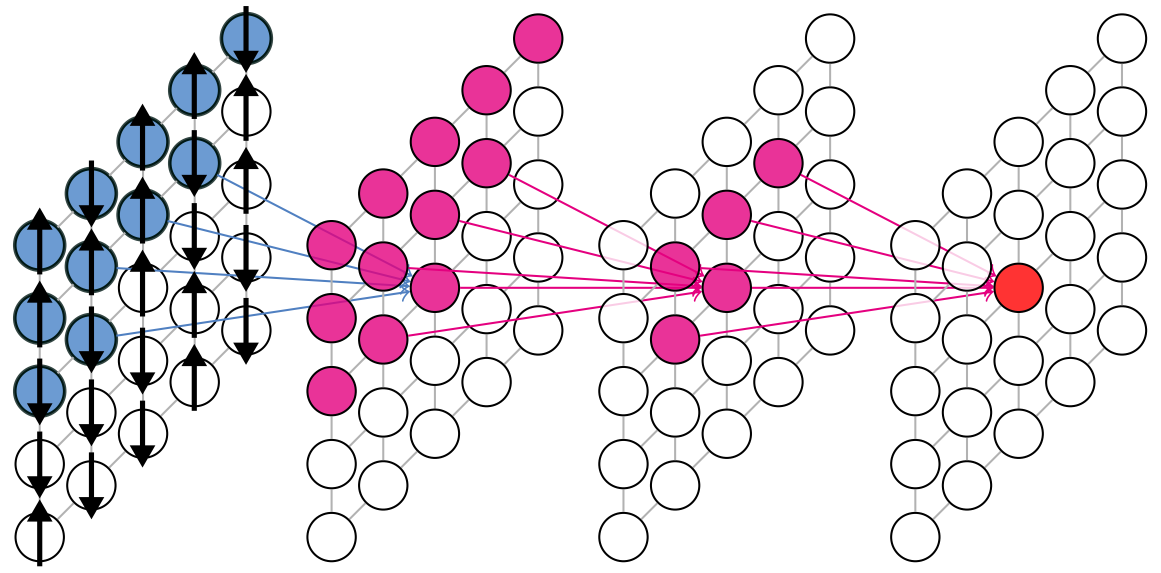
Image: PixelCNN 1601.06759



Molecules: 1810.11347

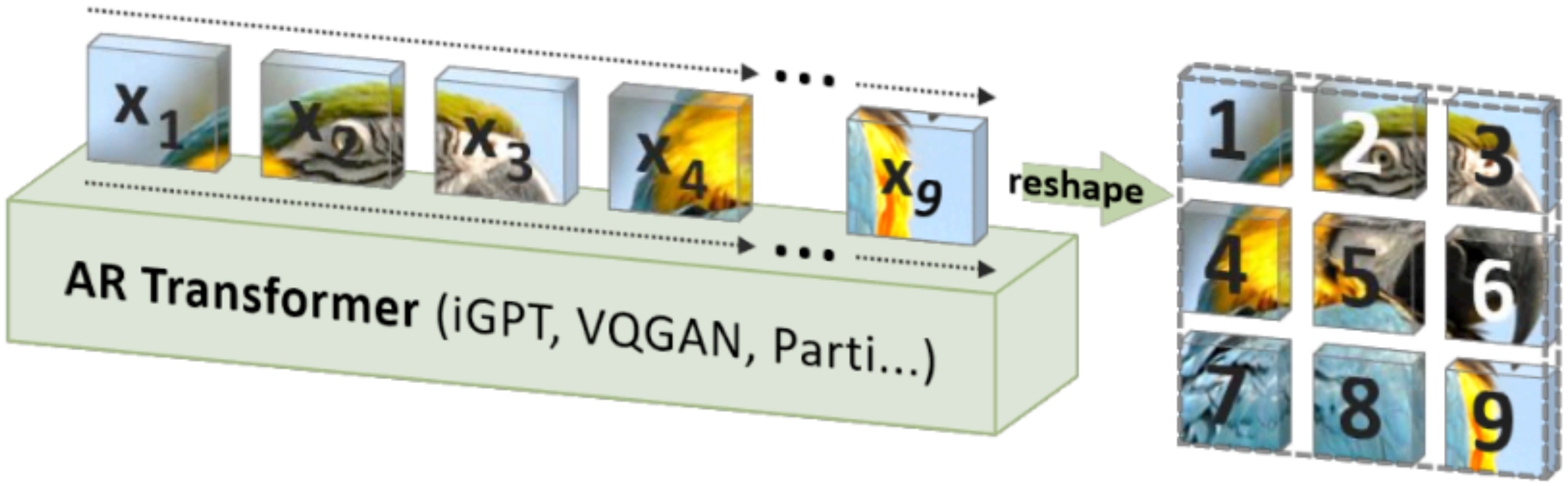


Ising spins: 1809.10606



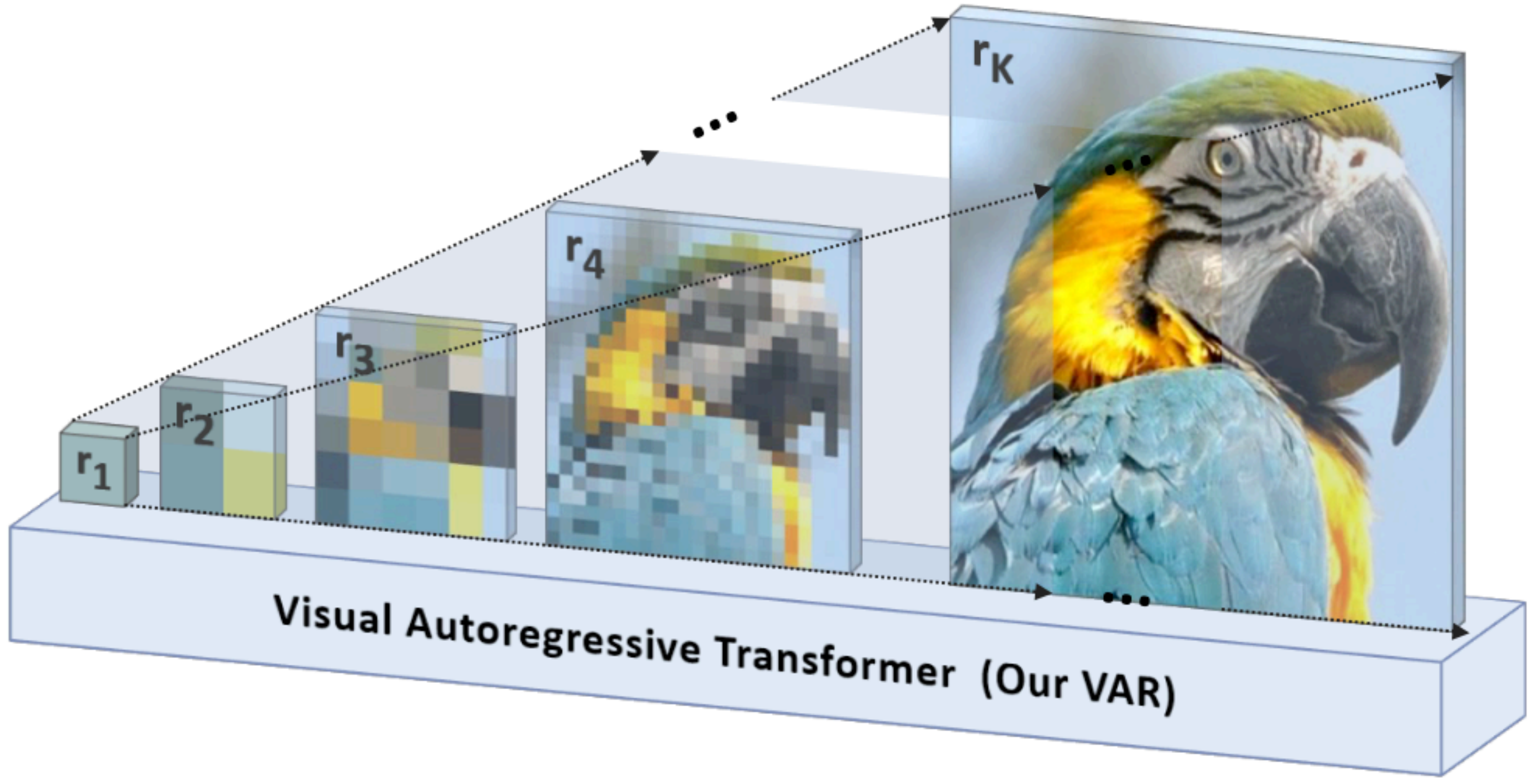
Autoregressive models for images

Chen et al, PMLR '20, Esser et al, 2012.09841



Next pixel (patch) prediction

Reed et al, 1703.03664
Tian et al, 2404.02905, Li et al, 2502.17437



Next **scale** prediction

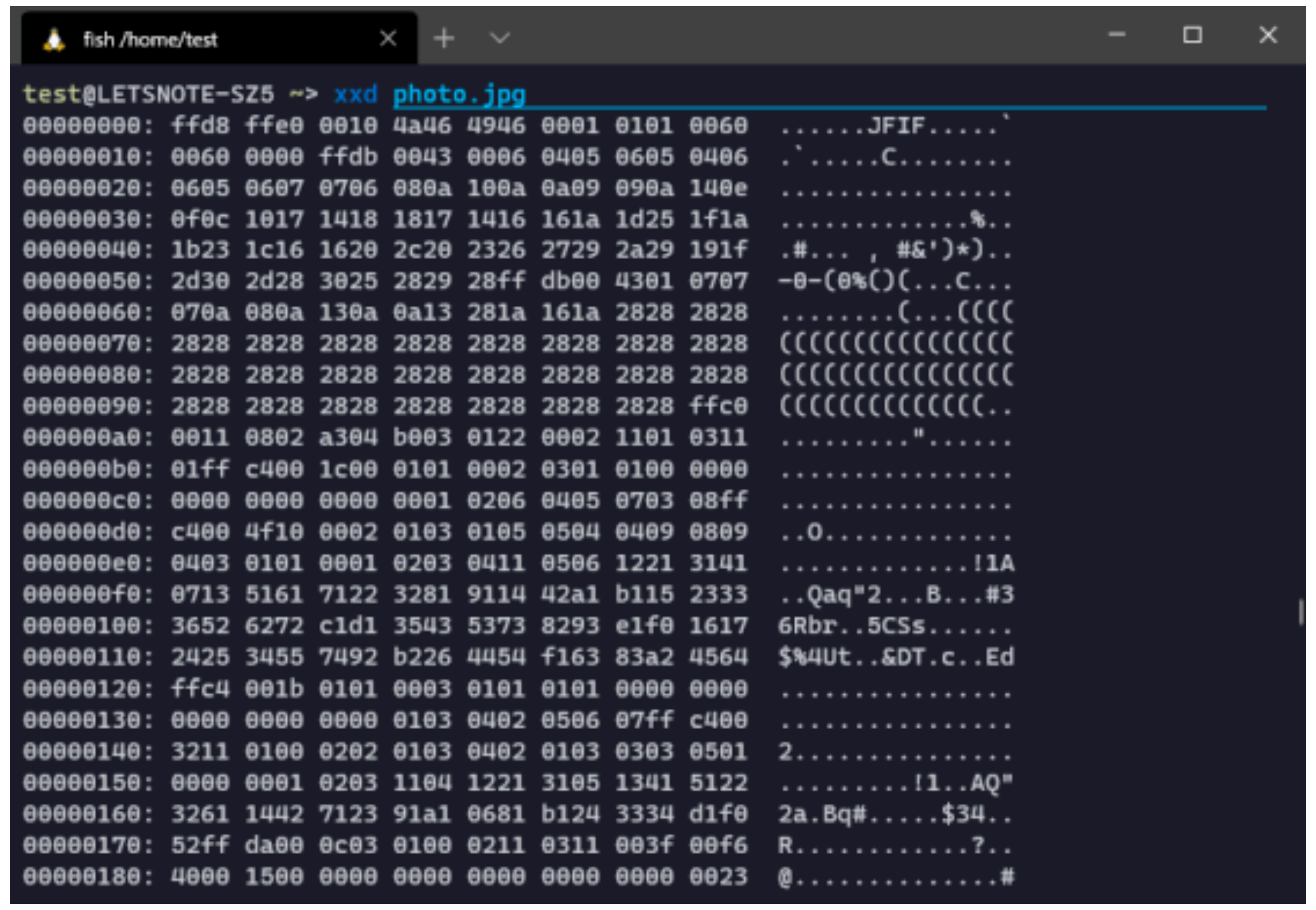
What is the suitable 1D ordering of 2D images ?

Autoregressive model for images

Han et al, 2408.08459

"Language" => token sequence => bitstream => **ANYTHING**

Compress



westlake.jpeg

Further compress

JPEG-LM

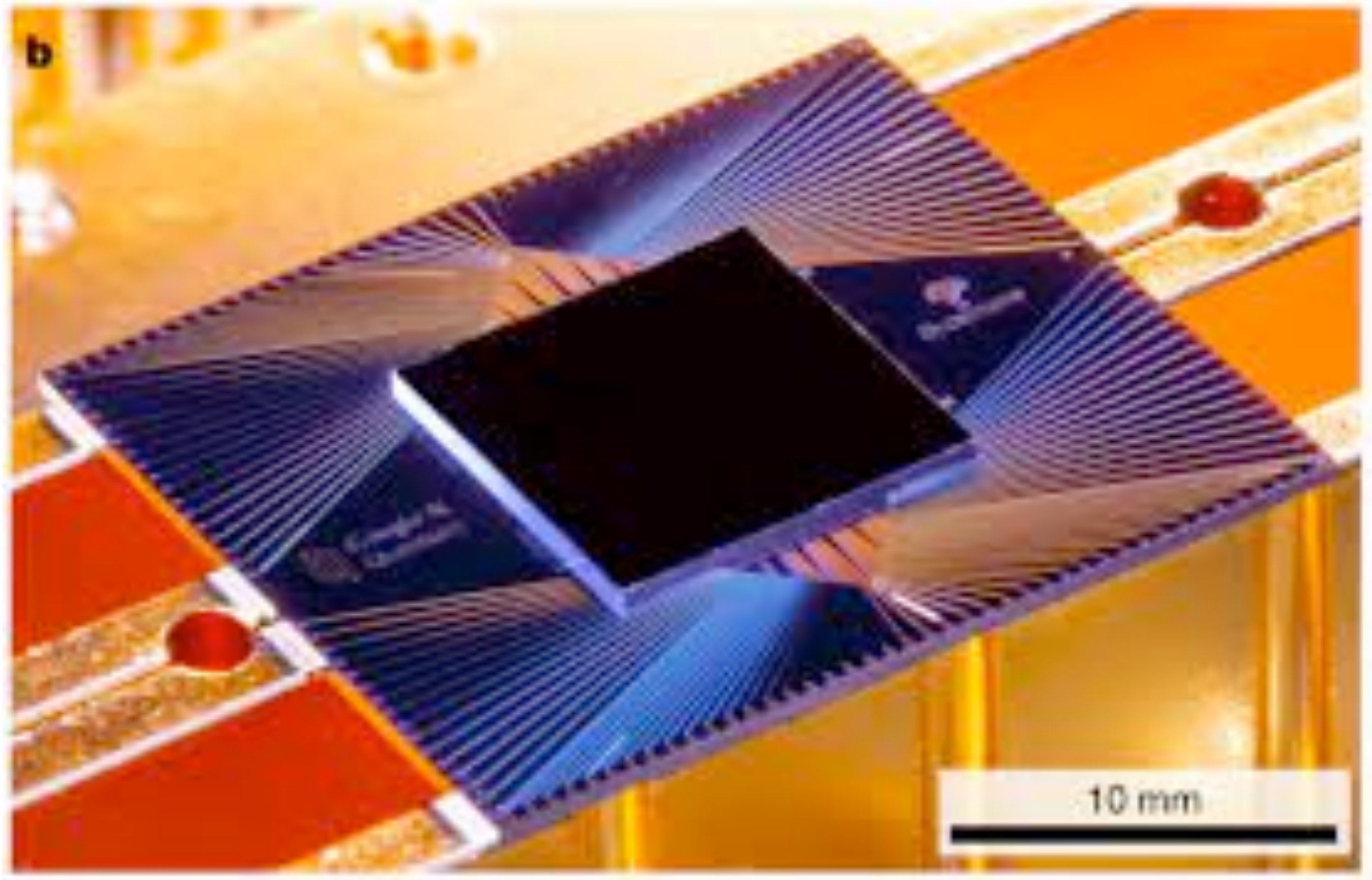
jpeg is a common lossy compression format for digital images

1) compute weights on predefined high-and-low frequency patches

2) throw away high-frequency weights; lossless compress low-frequency weights

Demo: Generative model of Sycamore data

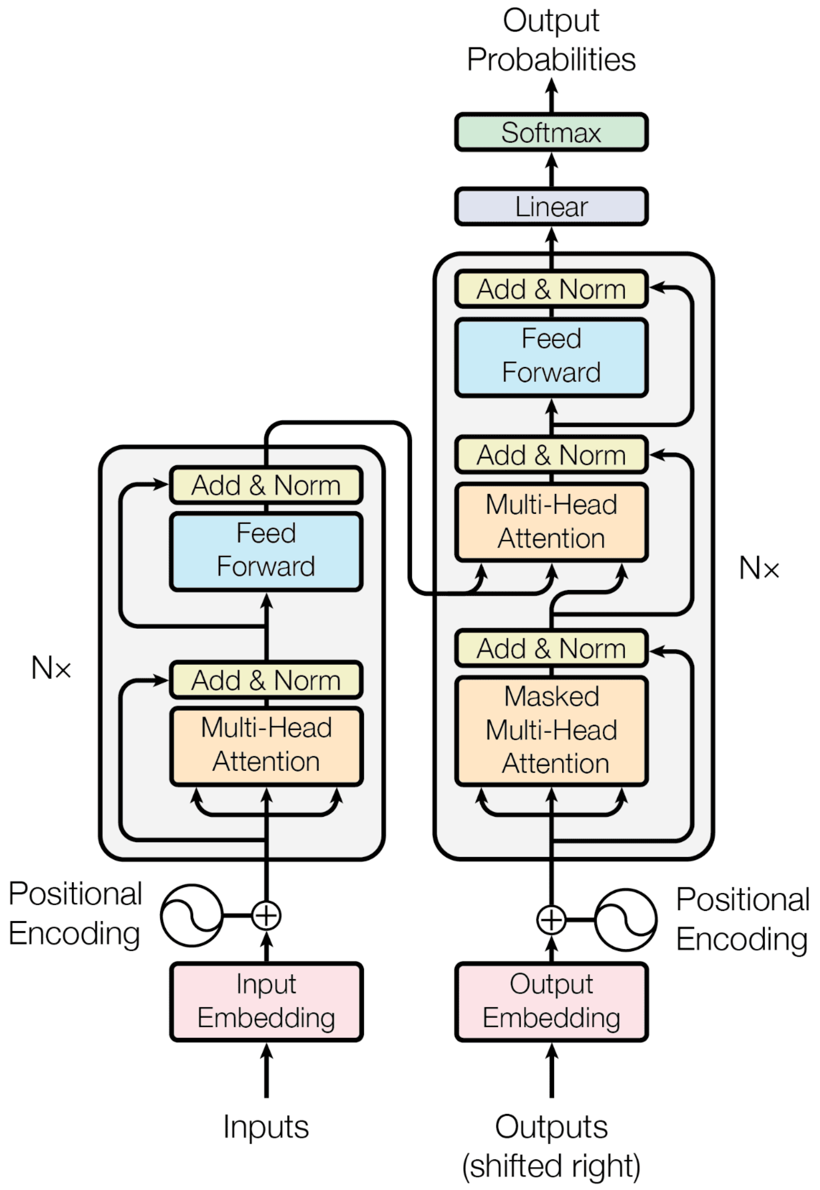
Quantum chip



bitstrings $\sim |\Psi(X)|^2$

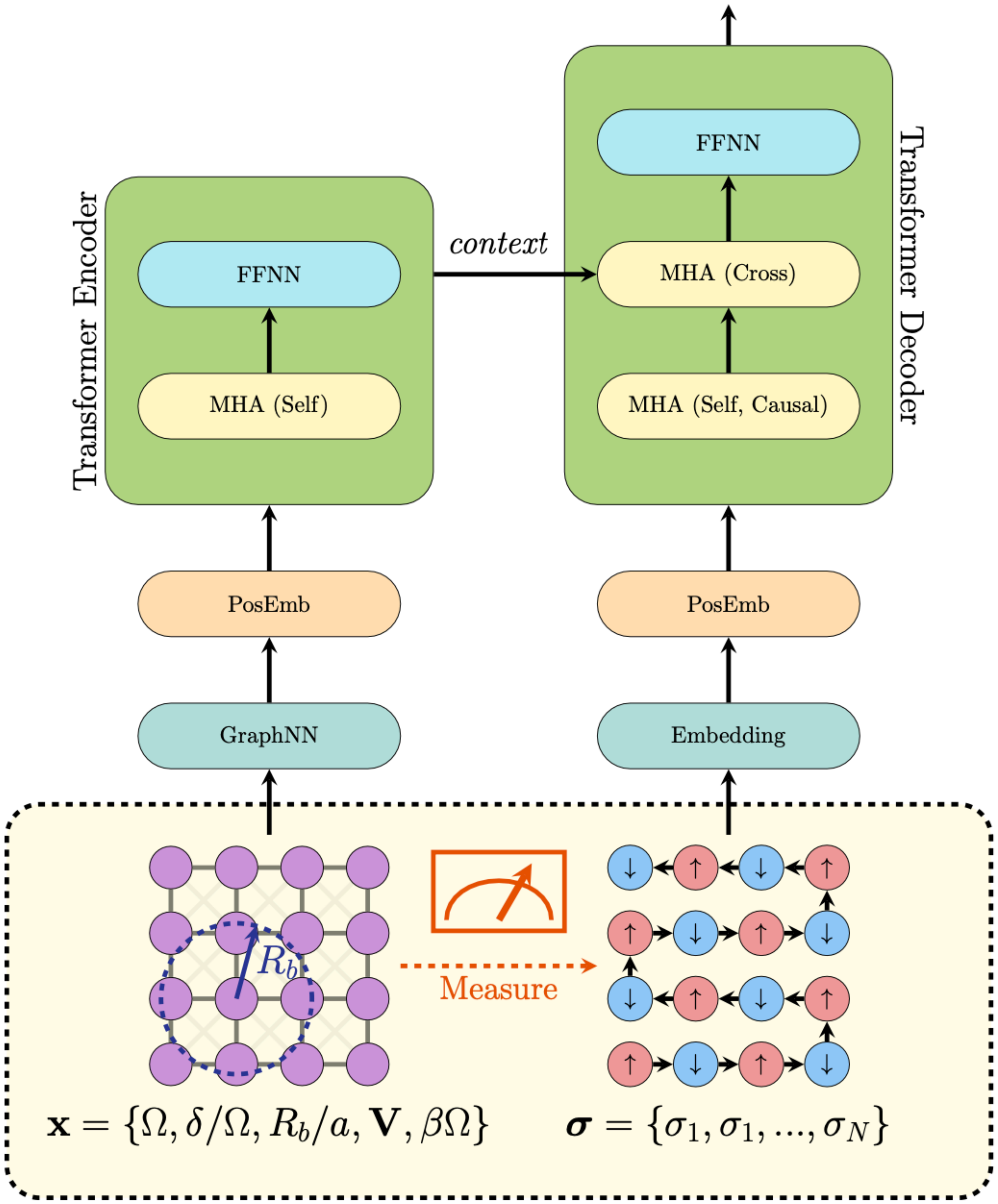
```
011110110100
100001111011
100110110111
100110100010
010100011000
010001000000
010101101100
100001111000
100101001001
001000001010
```

Transformer



Can we fake the measurement of the sycamore quantum circuit by training a transformer?

https://colab.research.google.com/drive/11WaroqULkudKT3h2i5J6r_EmA4wFKkoZ?usp=sharing



Rydberg GPT

Fitzek et al, 2405.21052

Prompts

$\boldsymbol{x}$: Hamiltonian parameters

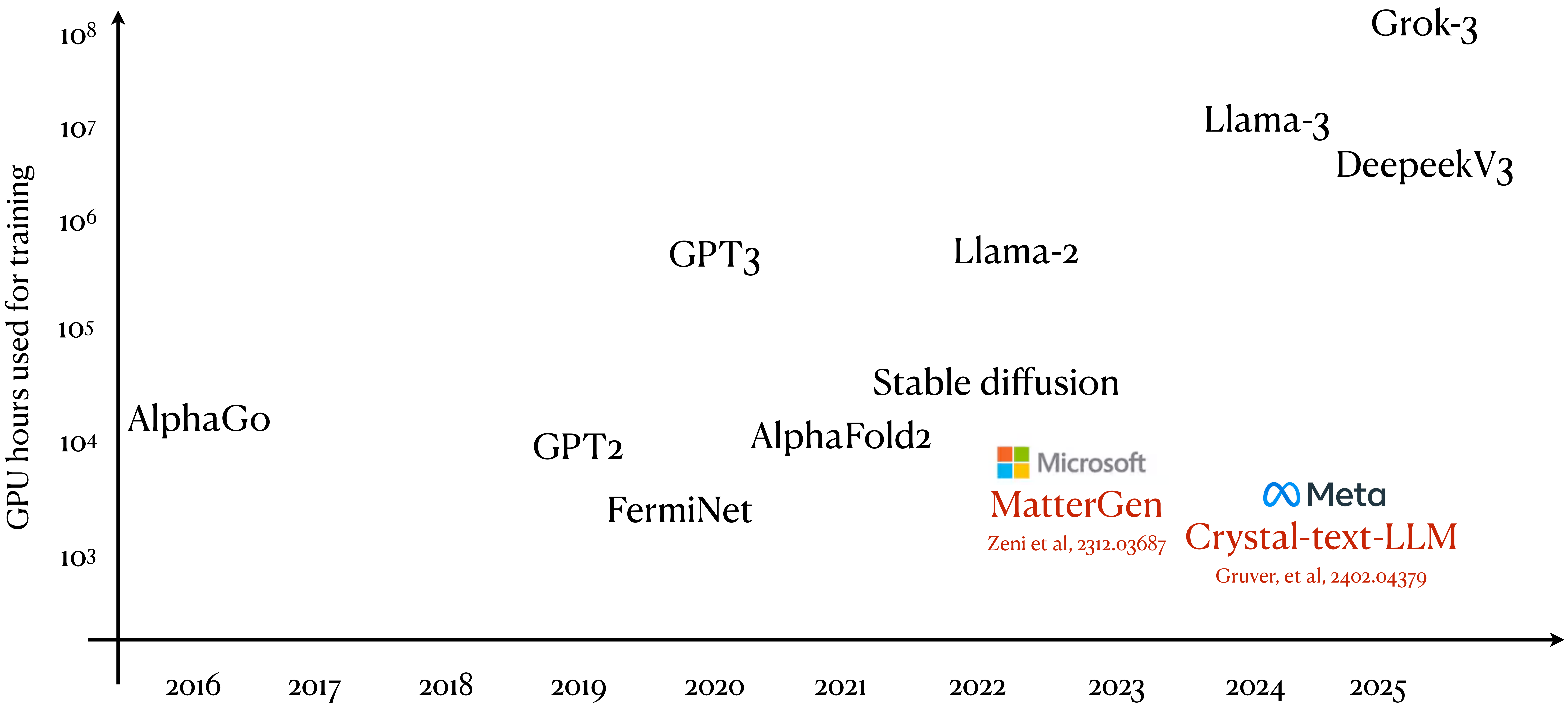
Answers

$\boldsymbol{\sigma}$: Projective measurements

$$p(\boldsymbol{\sigma} \,|\, \boldsymbol{x}) = p(\sigma_1 \,|\, \boldsymbol{x}) p(\sigma_2 \,|\, \sigma_1, \boldsymbol{x}) \ldots$$

"an image of beautiful crystals in 16:9"

$$\text{pixels} \sim p(\text{pixels} \mid \text{texts})$$



Is there a bitter lesson ?

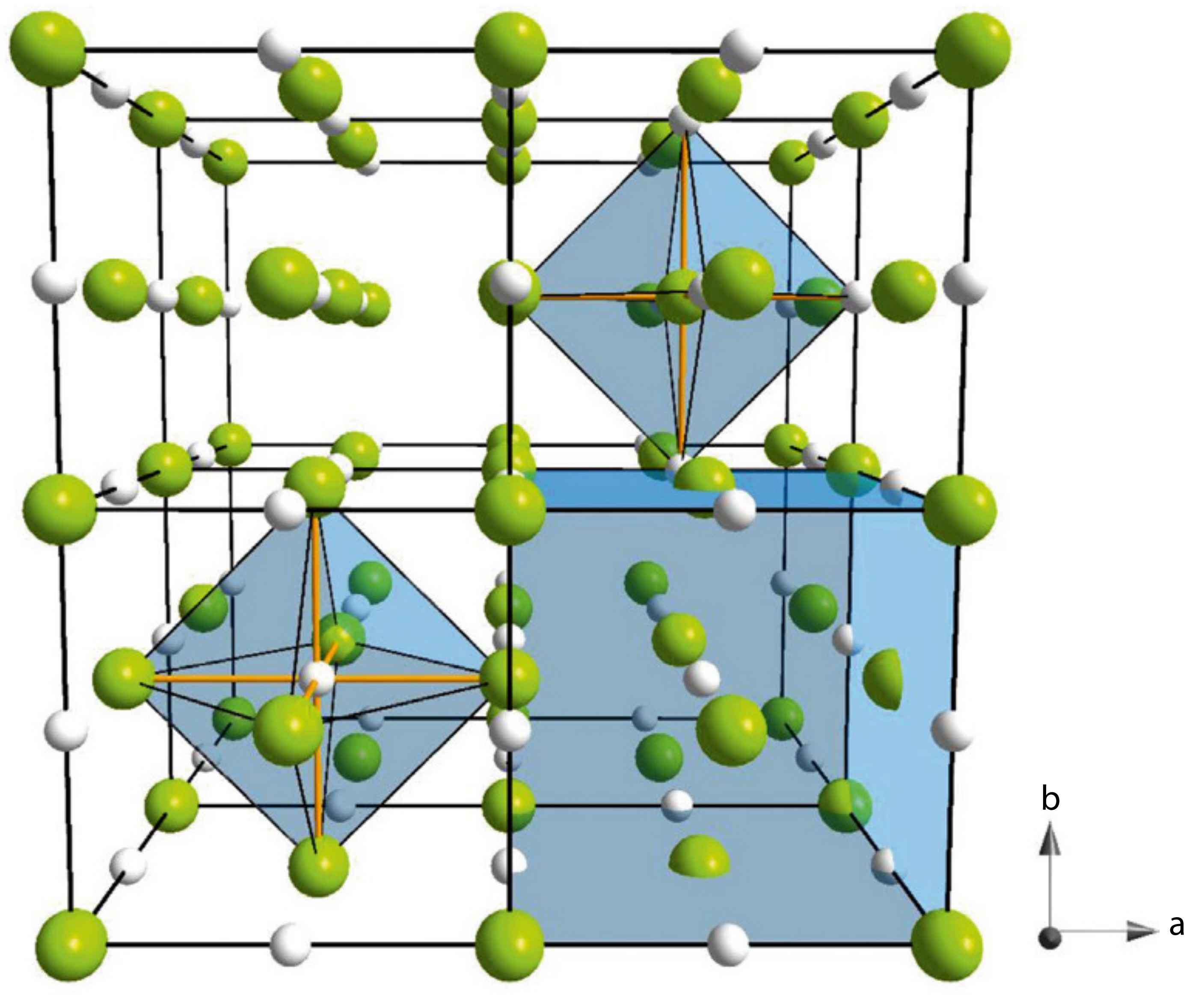
"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective"

—Rich Sutton 2019

```
data_Na1Cl1
_symmetry_space_group_name_H-M   'P1'
_cell_length_a   3.9893
_cell_length_b   3.9893
_cell_length_c   3.9893
_cell_angle_alpha   60.0000
_cell_angle_beta   60.0000
_cell_angle_gamma   60.0000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   NaCl
_chemical_formula_sum   'Na1 Cl1'
_cell_volume   44.8931
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Cl  Cl0  1  0.0000  0.0000  0.0000  1
  Na Na1 1  0.5000  0.5000  0.5000  1
```

Flam-Shepherd et al, 2305.05708
Antunes et al, 2307.04340
Gruver, et al, 2402.04379...

CALYPSO
USPEX
AIRSS
GNoME, ...



Large language model

Energy-based structure prediction

more data and compute

more physics and symmetries

We have much less crystal data

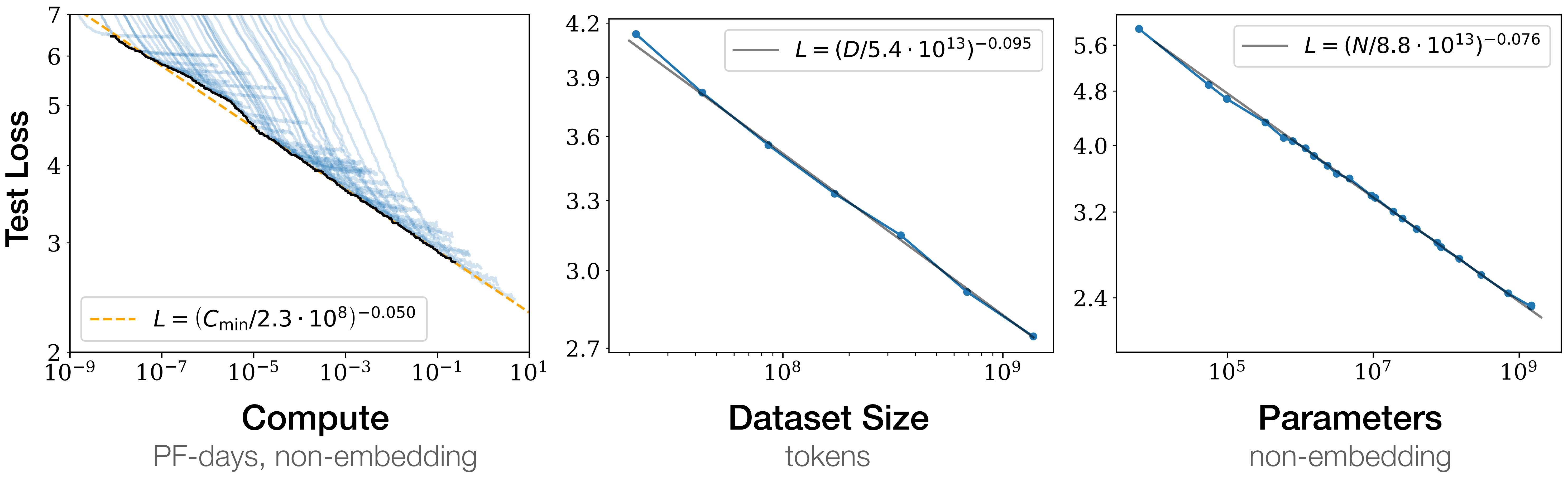
COMMON CRAWL

Over 250 billion pages

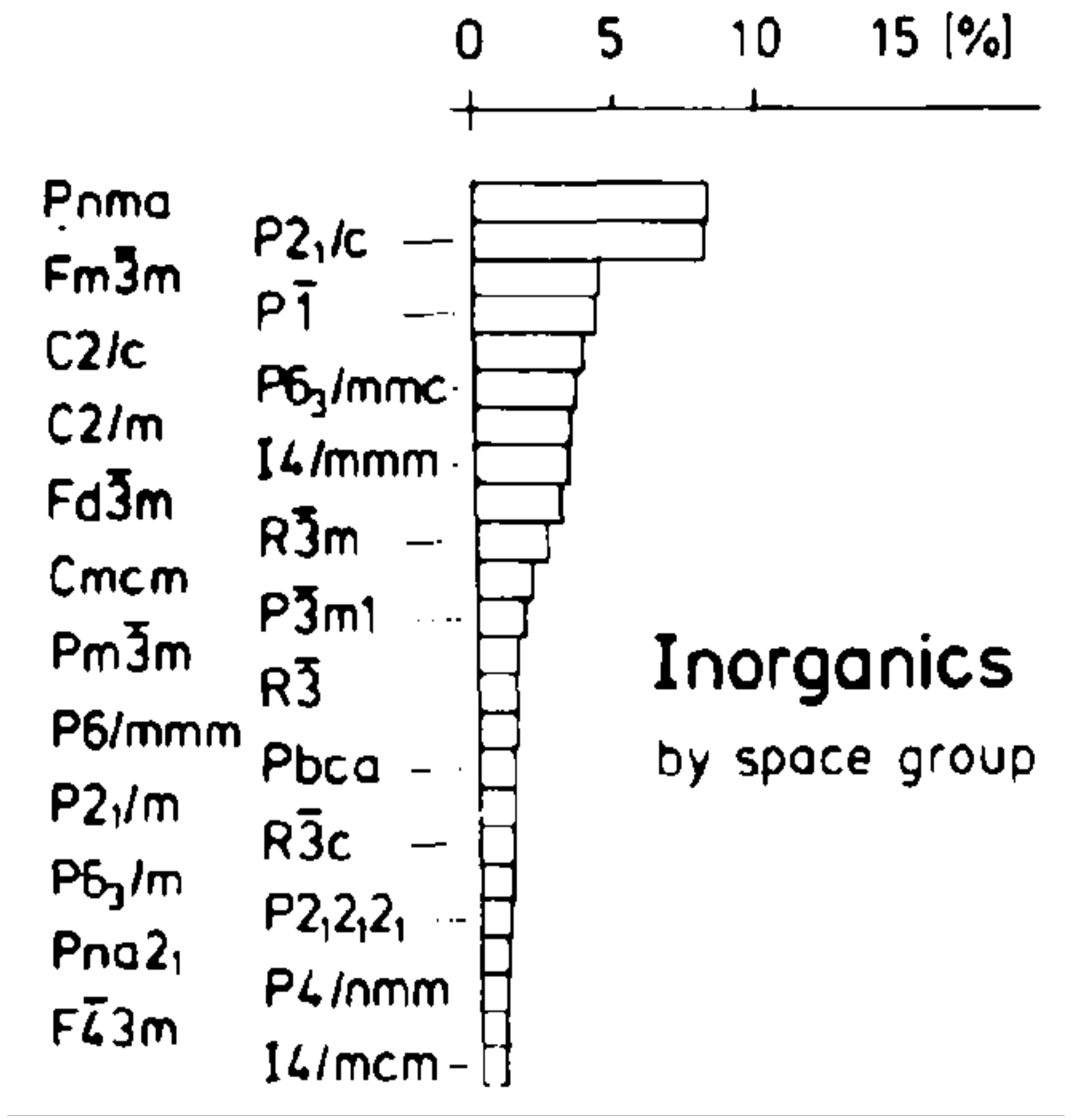
ICSD

> 291,000 crystal structures

Data, compute, and parameters need to scale simultaneously Kaplan et al, 2001.08361



Space groups quantify Nature's preference over symmetry



Wyckoff Positions of Group *P*1 (No. 1)

P1 is rare!

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
1	a	1	(x,y,z)

...

Wyckoff Positions of Group *Ia-3d* (No. 230)

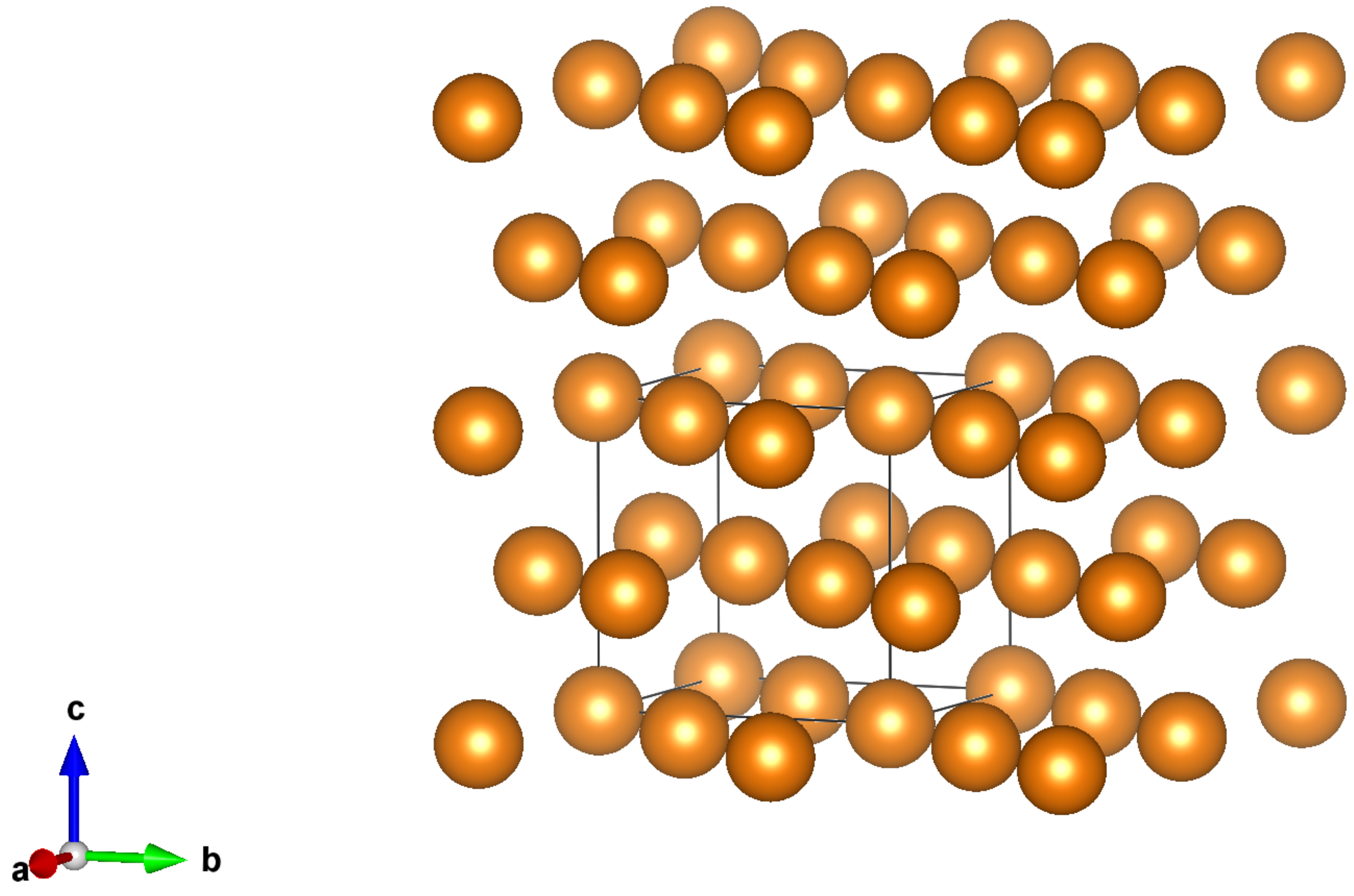
Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			(0,0,0) + (1/2,1/2,1/2) +
96	h	1	(x,y,z) (-x+1/2,-y,z+1/2) (-x,y+1/2,-z+1/2) (x+1/2,-y+1/2,-z) (z,x,y) (z+1/2,-x+1/2,-y) (-z+1/2,-x,y+1/2) (-z,x+1/2,-y+1/2) (y,z,x) (-y,z+1/2,-x+1/2) (y+1/2,-z+1/2,-x) (-y+1/2,-z,x+1/2) (y+3/4,x+1/4,-z+1/4) (-y+3/4,-x+3/4,-z+3/4) (y+1/4,-x+1/4,z+3/4) (-y+1/4,x+3/4,z+1/4) (x+3/4,z+1/4,-y+1/4) (-x+1/4,z+3/4,y+1/4) (-x+3/4,-z+3/4,-y+3/4) (x+1/4,-z+1/4,y+3/4) (z+3/4,y+1/4,-x+1/4) (z+1/4,-y+1/4,x+3/4) (-z+1/4,y+3/4,x+1/4) (-z+3/4,-y+3/4,-x+3/4) (-x,-y,-z) (x+1/2,y,-z+1/2) (x,-y+1/2,z+1/2) (-x+1/2,y+1/2,z) (-z,-x,-y) (-z+1/2,x+1/2,y) (z+1/2,x,-y+1/2) (z,-x+1/2,y+1/2) (-y,-z,-x) (y,-z+1/2,x+1/2) (-y+1/2,z+1/2,x) (y+1/2,z,-x+1/2) (-y+1/4,-x+3/4,z+3/4) (y+1/4,x+1/4,z+1/4) (-y+3/4,x+3/4,-z+1/4) (y+3/4,-x+1/4,-z+3/4) (-x+1/4,-z+3/4,y+3/4) (x+3/4,-z+1/4,-y+3/4) (x+1/4,z+1/4,y+1/4) (-x+3/4,z+3/4,-y+1/4) (-z+1/4,-y+3/4,x+3/4) (-z+3/4,y+3/4,-x+1/4) (z+3/4,-y+1/4,-x+3/4) (z+1/4,y+1/4,x+1/4)
48	g	..2	(1/8,y,-y+1/4) (3/8,-y,-y+3/4) (7/8,y+1/2,y+1/4) (5/8,-y+1/2,y+3/4) (-y+1/4,1/8,y) (-y+3/4,3/8,-y) (y+1/4,7/8,y+1/2) (y+3/4,5/8,-y+1/2) (y,-y+1/4,1/8) (-y,-y+3/4,3/8) (y+1/2,y+1/4,7/8) (-y+1/2,y+3/4,5/8) (7/8,-y,y+3/4) (5/8,y,y+1/4) (1/8,-y+1/2,-y+3/4) (3/8,y+1/2,-y+1/4) (y+3/4,7/8,-y) (y+1/4,5/8,y) (-y+3/4,1/8,-y+1/2) (-y+1/4,3/8,y+1/2) (-y,y+3/4,7/8) (y,y+1/4,5/8) (-y+1/2,-y+3/4,1/8) (y+1/2,-y+1/4,3/8)
48	f	2..	(x,0,1/4) (-x+1/2,0,3/4) (1/4,x,0) (3/4,-x+1/2,0) (0,1/4,x) (0,3/4,-x+1/2) (3/4,x+1/4,0) (3/4,-x+3/4,1/2) (x+3/4,1/2,1/4) (-x+1/4,0,1/4) (0,1/4,-x+1/4) (1/2,1/4,x+3/4) (-x,0,3/4) (x+1/2,0,1/4) (3/4,-x,0) (1/4,x+1/2,0) (0,3/4,-x) (0,1/4,x+1/2) (1/4,-x+3/4,0) (1/4,x+1/4,1/2) (-x+1/4,1/2,3/4) (x+3/4,0,3/4) (0,3/4,x+3/4) (1/2,3/4,-x+1/4)
32	e	.3.	(x,x,x) (-x+1/2,-x,x+1/2) (-x,x+1/2,-x+1/2) (x+1/2,-x+1/2,-x) (x+3/4,x+1/4,-x+1/4) (-x+3/4,-x+3/4,-x+3/4) (x+1/4,-x+1/4,x+3/4) (-x+1/4,x+3/4,x+1/4) (-x,-x,-x) (x+1/2,x,-x+1/2) (x,-x+1/2,x+1/2) (-x+1/2,x+1/2,x) (-x+1/4,-x+3/4,x+3/4) (x+1/4,x+1/4,x+1/4) (-x+3/4,x+3/4,-x+1/4) (x+3/4,-x+1/4,-x+3/4)
24	d	-4..	(3/8,0,1/4) (1/8,0,3/4) (1/4,3/8,0) (3/4,1/8,0) (0,1/4,3/8) (0,3/4,1/8) (3/4,5/8,0) (3/4,3/8,1/2) (1/8,1/2,1/4) (7/8,0,1/4) (0,1/4,7/8) (1/2,1/4,1/8)
24	c	2.2 2	(1/8,0,1/4) (3/8,0,3/4) (1/4,1/8,0) (3/4,3/8,0) (0,1/4,1/8) (0,3/4,3/8) (7/8,0,3/4) (5/8,0,1/4) (3/4,7/8,0) (1/4,5/8,0) (0,3/4,7/8) (0,1/4,5/8)
16	b	.32	(1/8,1/8,1/8) (3/8,7/8,5/8) (7/8,5/8,3/8) (5/8,3/8,7/8) (7/8,7/8,7/8) (5/8,1/8,3/8) (1/8,3/8,5/8) (3/8,5/8,1/8)
16	a	.-3.	(0,0,0) (1/2,0,1/2) (0,1/2,1/2) (1/2,1/2,0) (3/4,1/4,1/4) (3/4,3/4,3/4) (1/4,1/4,3/4) (1/4,3/4,1/4)

https://www.cryst.ehu.es/cryst/get_wp.html

Wyckoff Positions of Group *Fm*-3*m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates (0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +
192	l	1	(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z) (z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y) (y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x) (y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z) (x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y) (z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x) (-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z) (-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y) (-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x) (-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z) (-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y) (-z,-y,x) (-z,y,-x) (z,-y,-x) (z,y,x)
96	k	..m	(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z) (z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x) (x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x) (x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z) (x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x) (z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)
96	j	m..	(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z) (z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y) (y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0) (y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z) (0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y) (z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)
48	i	m.m 2	(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y) (y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y) (y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)
48	h	m.m 2	(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y) (y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y) (y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)
48	g	2.m m	(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4) (1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4) (x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)
32	f	.3m	(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x) (x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)
24	e	4m. m	(x,0,0) (-x,0,0) (0,x,0) (0,-x,0) (0,0,x) (0,0,-x)
24	d	m.m m	(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4) (1/4,1/4,0) (3/4,1/4,0)
8	c	-43m	(1/4,1/4,1/4) (1/4,1/4,3/4)
4	b	m-3m	(1/2,1/2,1/2)
4	a	m-3m	(0,0,0)

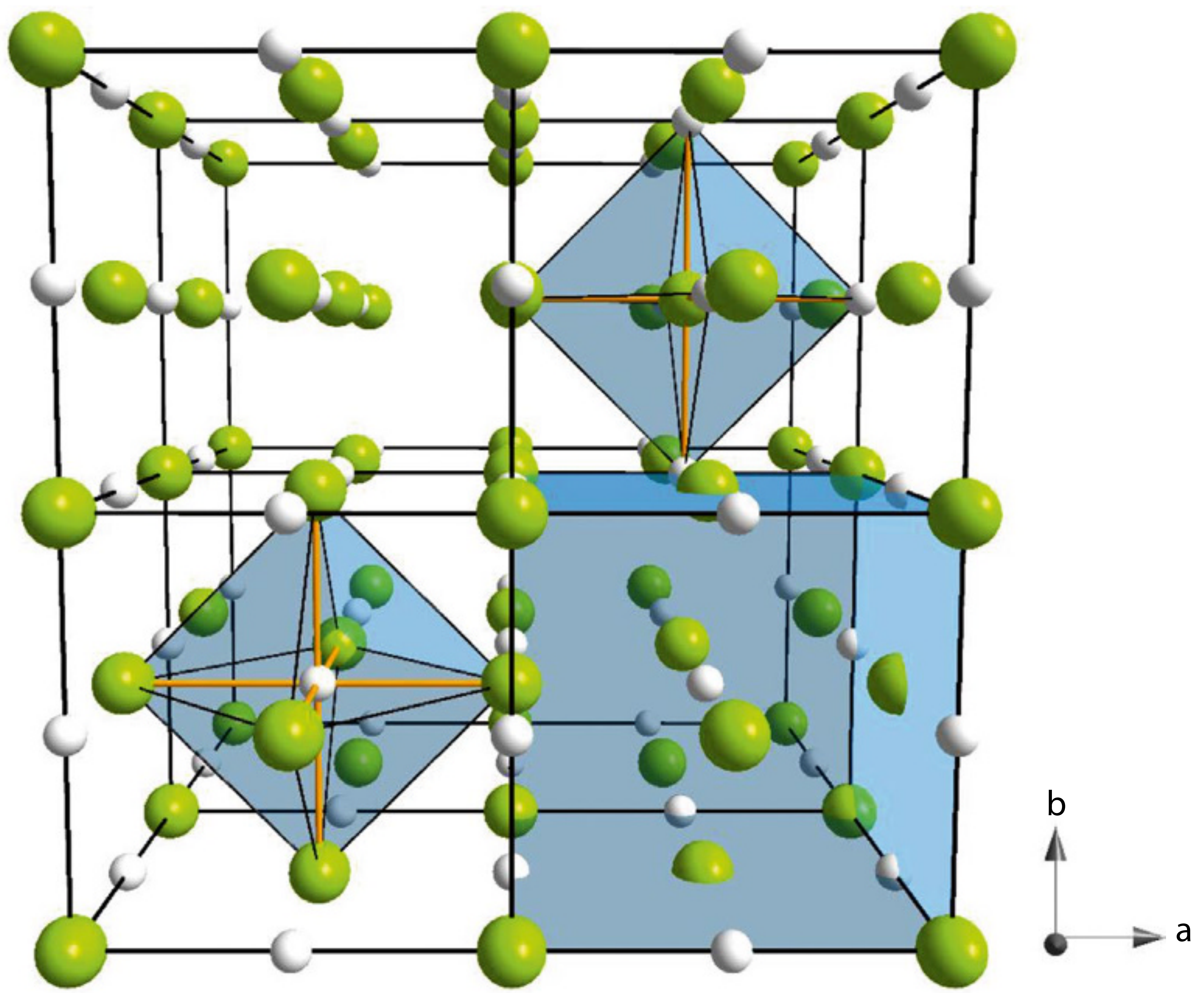
Copper



Wyckoff Positions of Group *Fm*-3*m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates (0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +
192	l	1	(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z) (z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y) (y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x) (y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z) (x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y) (z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x) (-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z) (-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y) (-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x) (-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z) (-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y) (-z,-y,x) (-z,y,-x) (z,-y,-x) (z,y,x)
96	k	..m	(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z) (z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x) (x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x) (x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z) (x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x) (z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)
96	j	m..	(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z) (z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y) (y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0) (y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z) (0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y) (z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)
48	i	m.m 2	(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y) (y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y) (y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)
48	h	m.m 2	(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y) (y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y) (y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)
48	g	2.m m	(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4) (1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4) (x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)
32	f	.3m	(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x) (x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)
24	e	4m. m	(x,0,0) (-x,0,0) (0,x,0) (0,-x,0) (0,0,x) (0,0,-x)
24	d	m.m m	(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4) (1/4,1/4,0) (3/4,1/4,0)
8	c	-43m	(1/4,1/4,1/4) (1/4,1/4,3/4)
4	b	m-3m	(1/2,1/2,1/2)
4	a	m-3m	(0,0,0)

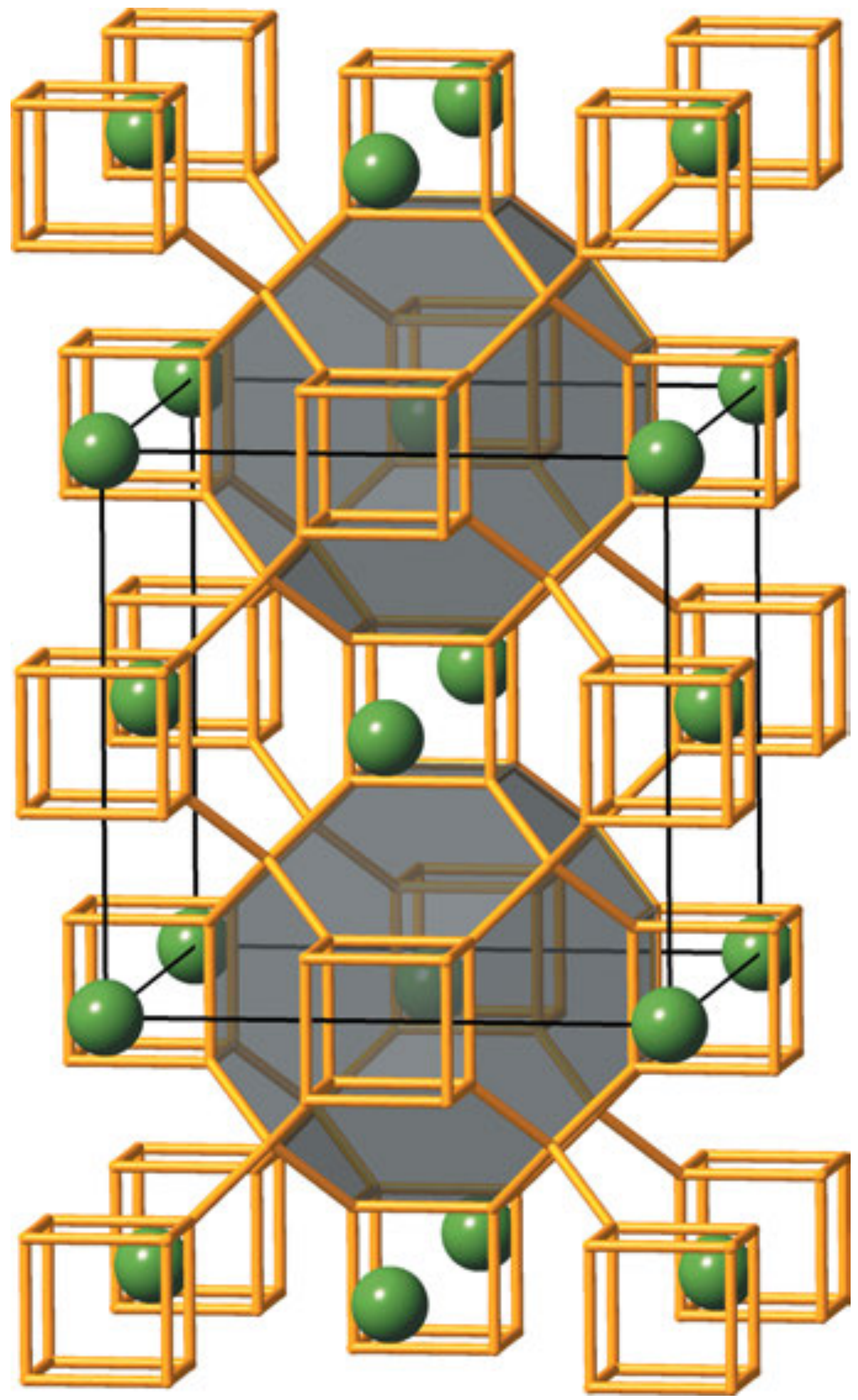
NaCl



Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates (0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +
192	l	1	(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z) (z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y) (y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x) (y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z) (x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y) (z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x) (-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z) (-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y) (-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x) (-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z) (-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y) (-z,-y,x) (-z,y,-x) (z,-y,-x) (z,y,x)
96	k	..m	(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z) (z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x) (x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x) (x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z) (x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x) (z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)
96	j	m..	(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z) (z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y) (y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0) (y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z) (0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y) (z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)
48	i	m.m 2	(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y) (y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y) (y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)
48	h	m.m 2	(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y) (y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y) (y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)
48	g	2.m m	(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4) (1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4) (x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)
32	f	.3m	(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x) (x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)
24	e	4m. m	(x,0,0) (-x,0,0) (0,x,0) (0,-x,0) (0,0,x) (0,0,-x)
24	d	m.m m	(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4) (1/4,1/4,0) (3/4,1/4,0)
8	c	-43m	(1/4,1/4,1/4) (1/4,1/4,3/4)
4	b	m-3m	(1/2,1/2,1/2)
4	a	m-3m	(0,0,0)

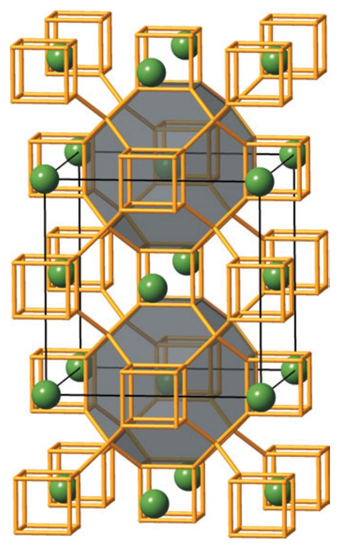
LaH_{10}



CrystalFormer

Zhendong Cao, Xiaoshan Luo,
Jian Lv, and LW, 2403.15734
deepmodeling/CrystalFormer

Space Group Informed Transformer for Crystals



225-a-La-0-0-0-c-H-1/4-1/4-1/4-f-H-0.375-0.375-0.375-X-5.1-5.1-5.1-90-90-90

"Grammar" ~ Solid state chemistry

"Synonyms" ~ Elememt substitution

"Idioms" ~ Coordination polyhedra

"Rhythm" ~ Wyckoff positions

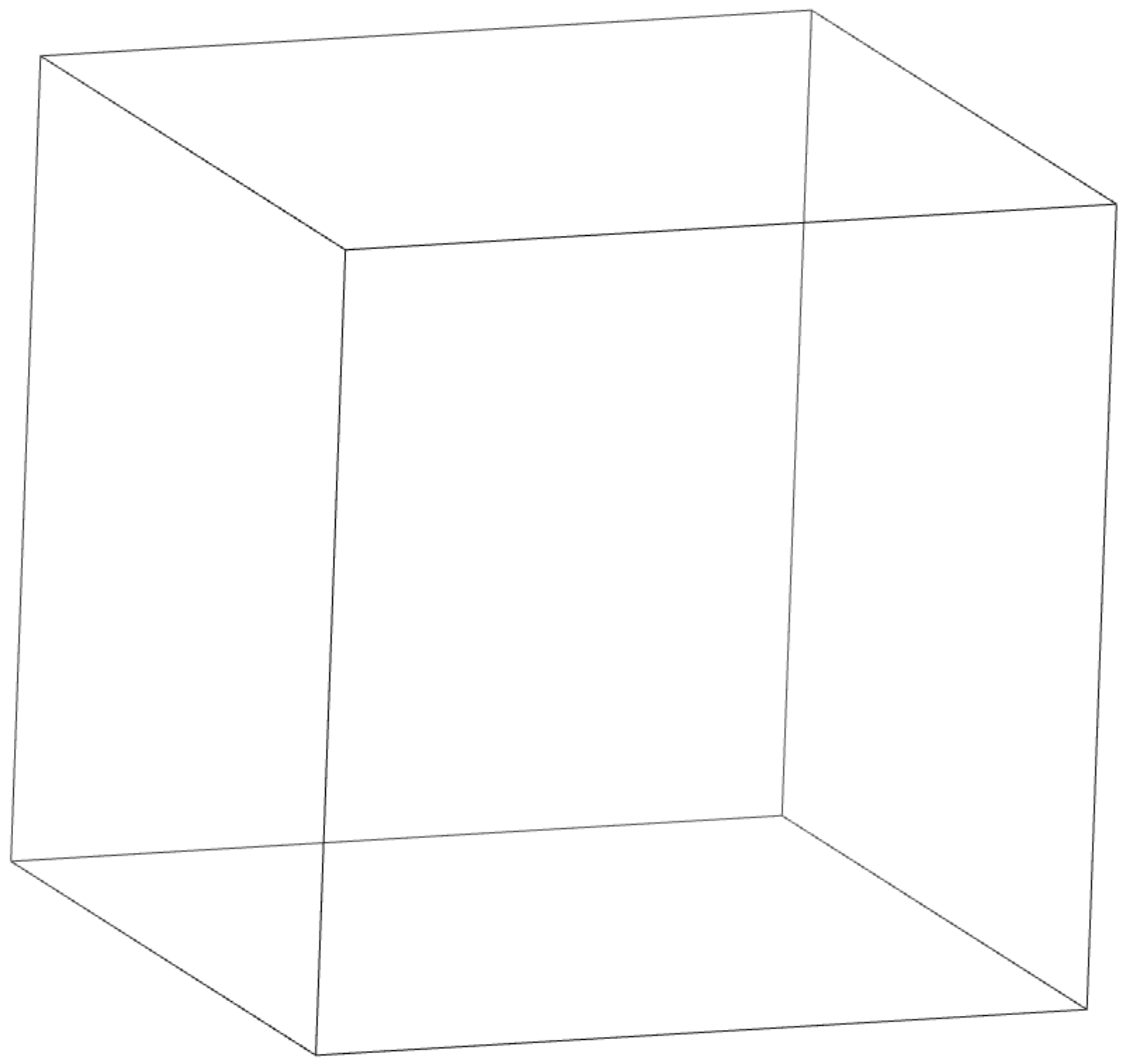
Nature's codebook for tokenization
discrete, pre-compression

more data and compute ⟵ CRYSTALFORMER ⟶ more physics and symmetries

Not a large language model, **nor** a potential energy surface

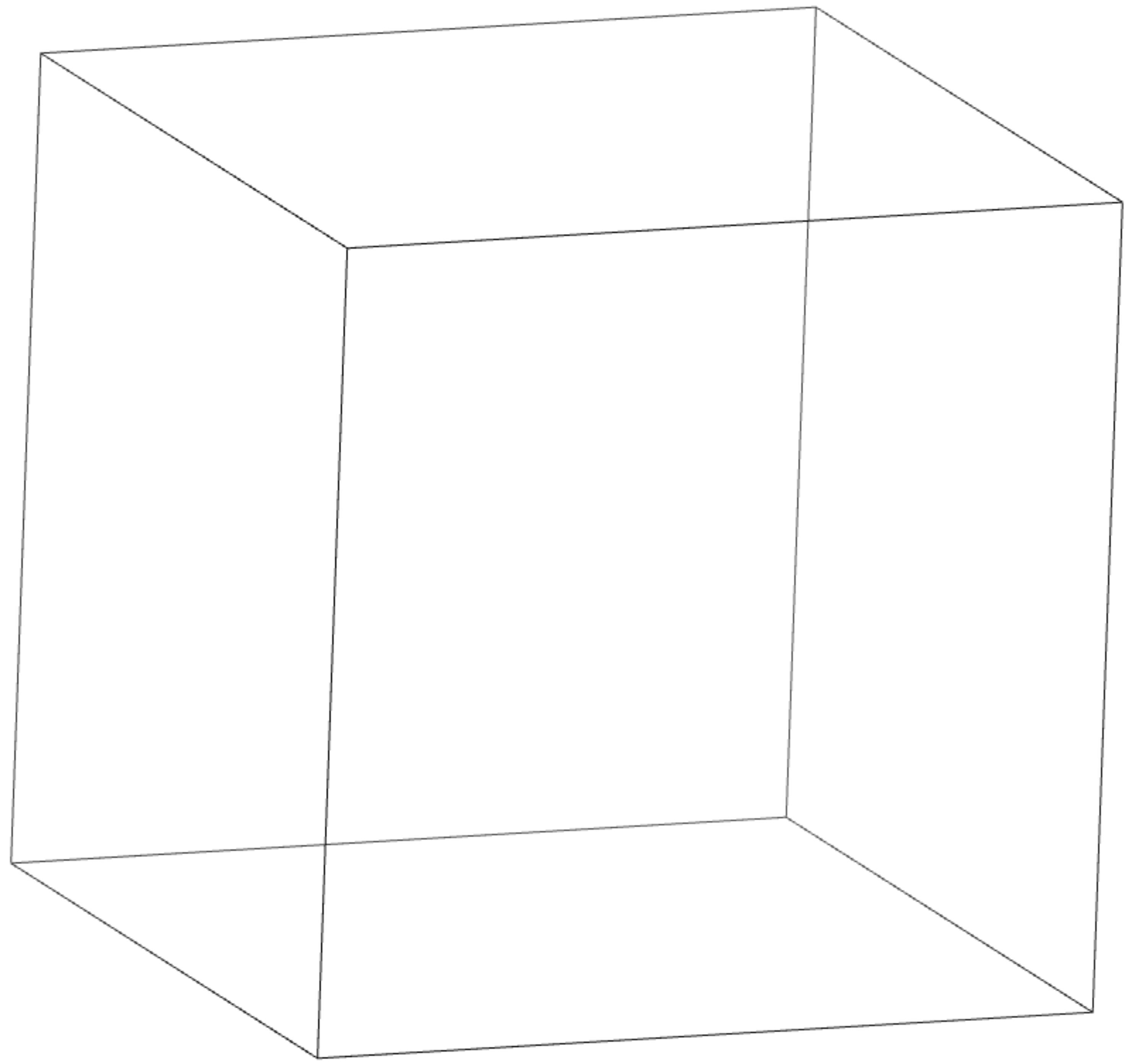
Autoregressive sampling of a crystal

$Cs_2ZnFe(CN)_6$



Autoregressive sampling of a crystal

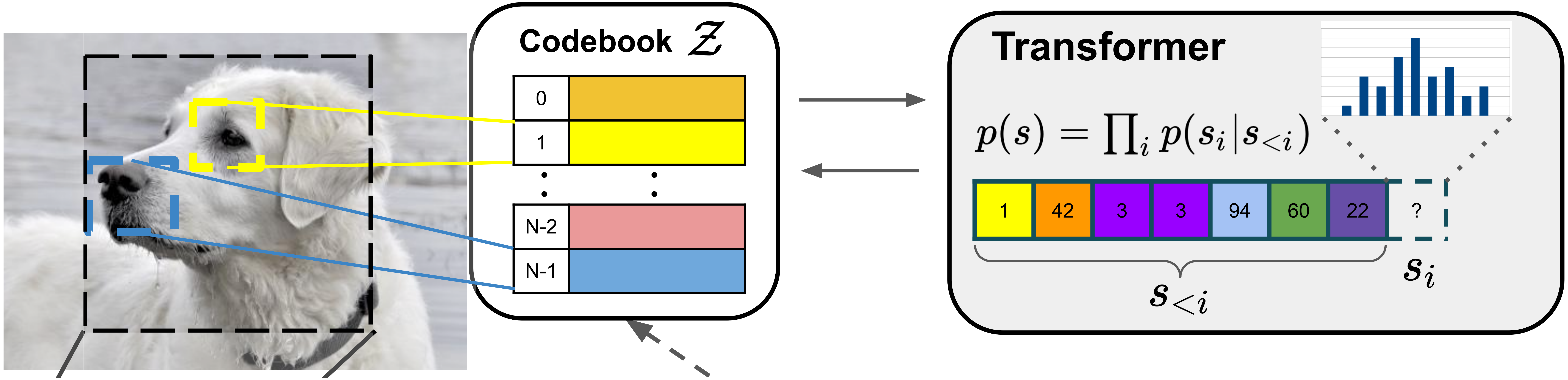
$Cs_2ZnFe(CN)_6$



225-a-Fe-0-0-0-b-Zn-1/2-1/2-1/2-c-Cs-1/4-1/4-1/4-e-C-0.18-0-0-e-N-0.29-0-0-X-10.45-10.45-10.45-90-90-90

Aside: autoregressive transformer for images

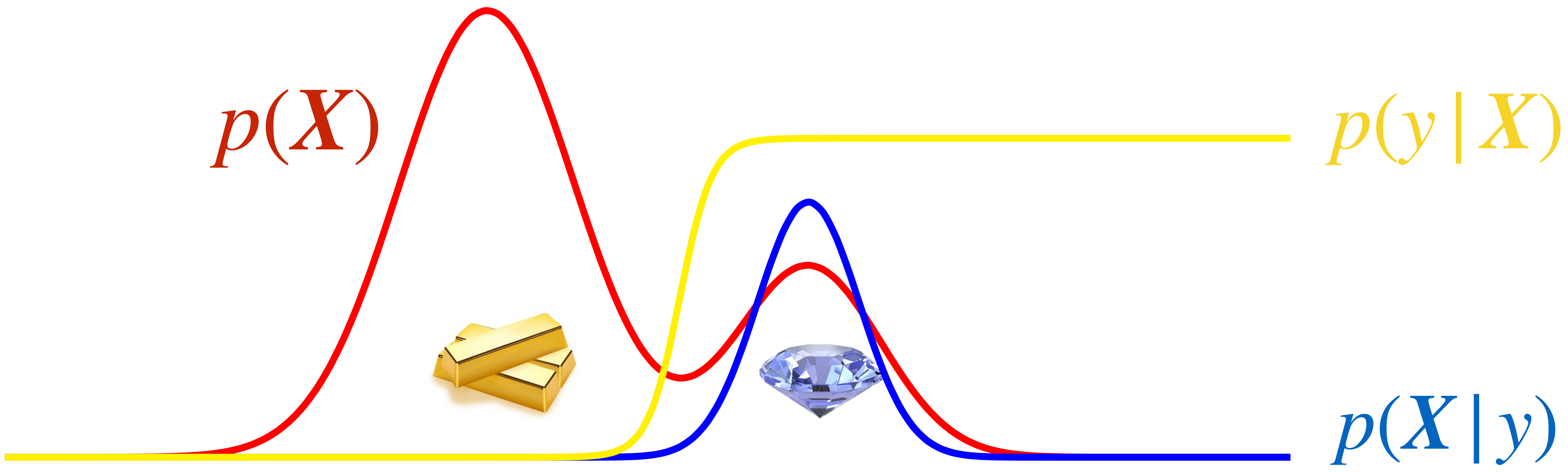
Esser et al, Taming Transformers for High-Resolution Image Synthesis (VQGAN), 2012.09841



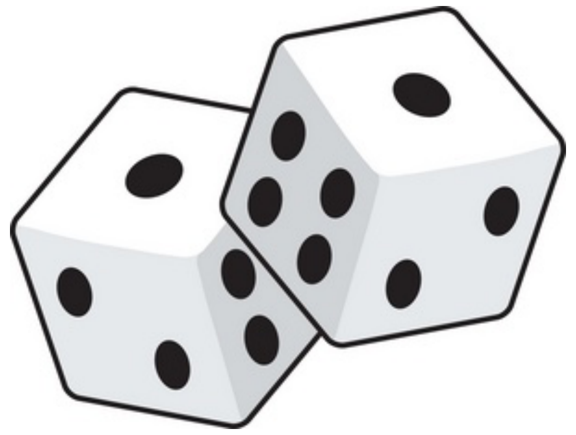
learned codebook, see also Tian et al, 2404.02905

CrystalFormer leverages Nature's codebook: the Wyckoff position table

Bayes rule for materials inverse design



How to sample from $p(\boldsymbol{X}\,|\,y)$? Two approaches originated in physics



Markov chain Monte Carlo

Metropolis et al, 1953, Hastings 1970

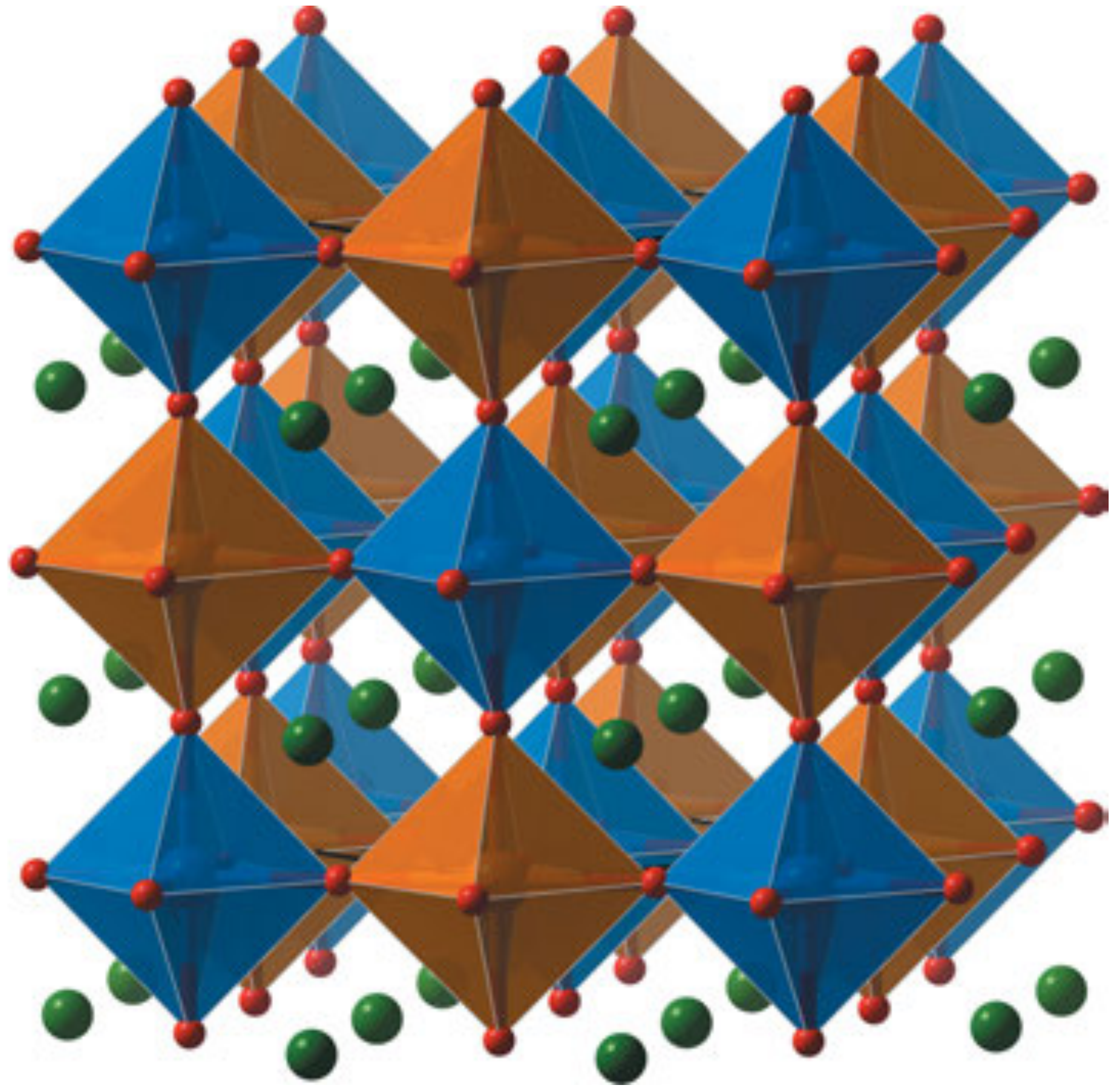
$$\nabla F$$

Variational inference

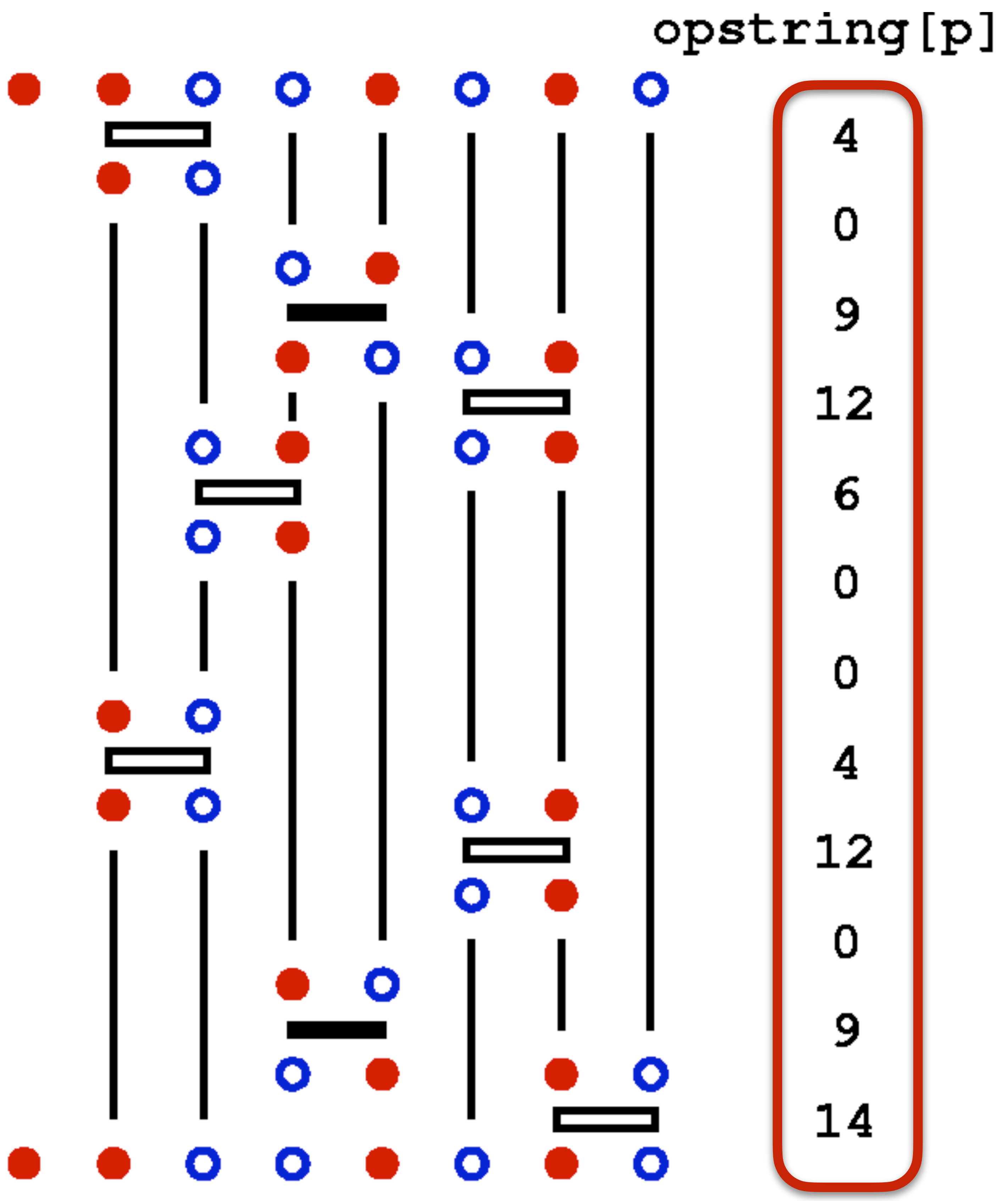
Gibbs, Feynman, Bogoliubov,..., Jordan et al 1999

MCMC sampling from the posterior

Generate more double perovskites $A_2BB'O_6$



225-a-[?]-0-0-0-b-[?]-1/2-1/2-1/2-c-[?]-1/4-1/4-1/4-e-O-[?]-0-0



Solve crystal cloze test via MCMC sweep through the "crystal string"

$$A(X \to X') = \min\left[1, \frac{p(X')}{p(X)}\right]$$

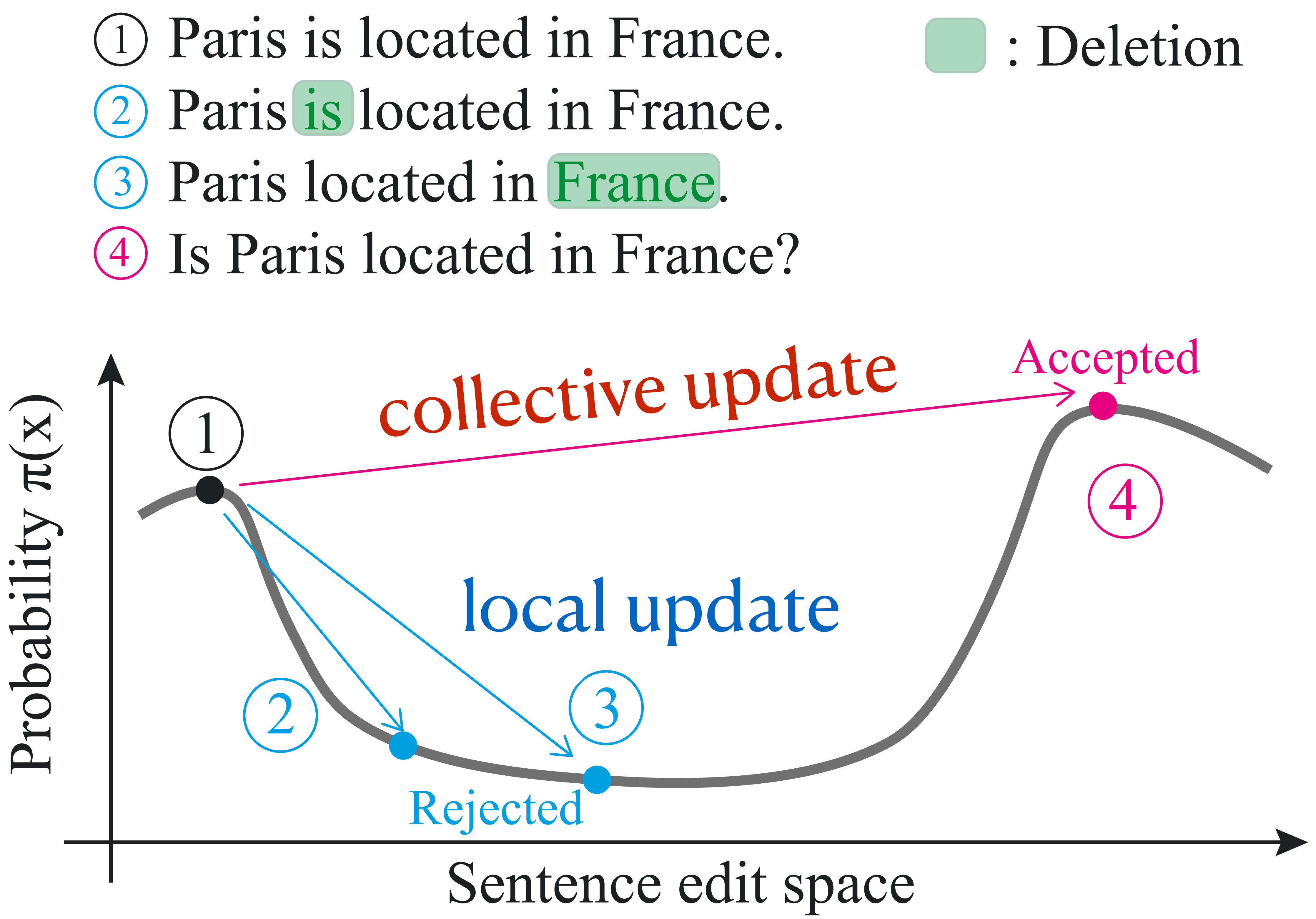
MCMC sweep through the "operator string" in Sandvik's SSE algorithm

Aside: Constrained sentence generation in languate modeling

$$\pi(x) \propto P_{\mathrm{LM}}(x) \cdot \mathrm{Constraint}(x)$$

"traverses the probabilistic space of high-quality sentences more effectively"

Miao et al, 1811.10996, Zhang et al, 2011.12334



The sequence length for inorganic crystals is ~100 with vocabulary size ~100

So, even naive Metropolis-Hastings with annealing works fine

Variational inference the posterior

$$\mathbb{KL}\left(q(\boldsymbol{X}) \parallel p(\boldsymbol{X} \mid y)\right) = \underset{\boldsymbol{X}\sim q(\boldsymbol{X})}{\mathbb{E}}\left[-\ln p(y \mid \boldsymbol{X})\right] + \mathbb{KL}\left(q(\boldsymbol{X}) \| p(\boldsymbol{X})\right)$$

Variational probability

Likelihood function

Prior

$q(\boldsymbol{X})$ is easy to sample, e.g. another autoregressive model

Variational inference turns a sampling problem into a stochastic optimization problem

Also known as: reinforcement fine-tuning

$$\mathbb{KL}\left(q(\boldsymbol{X}) \parallel p(\boldsymbol{X} \mid y)\right) = \mathop{\mathbb{E}}_{\boldsymbol{X} \sim q(\boldsymbol{X})} \left[-r(\boldsymbol{X})\right] + \mathbb{KL}\left(q(\boldsymbol{X}) \| p(\boldsymbol{X})\right)$$

$\uparrow$ Fine-tuned model

$\uparrow$ Reward function

$\uparrow$ Remain close to the pretrained model

"RL with KL penalties is better viewed as Bayesian inference" Korbar et al, 2205.11275

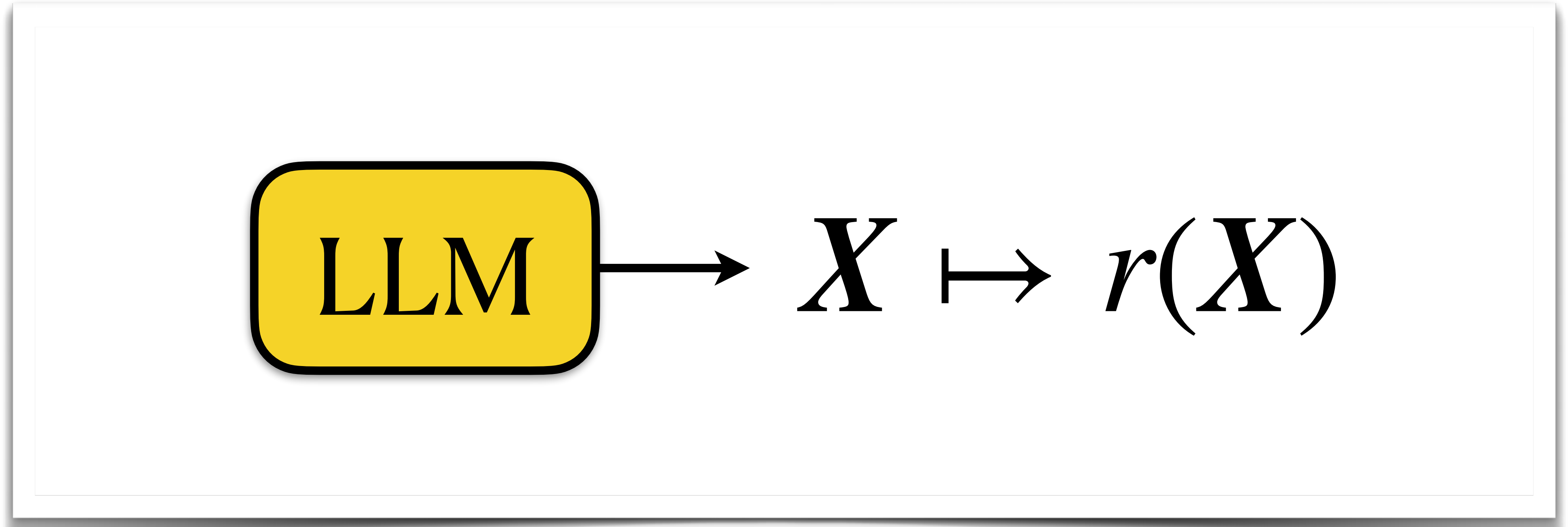
Two sides of the same coin

❶ Pre-training

learn from data
to be a generalist

$$\mathscr{L} = -\,\mathbb{E}_{\boldsymbol{X}\sim\text{data}}\left[\ln p(\boldsymbol{X})\right]$$

❷ Post-training



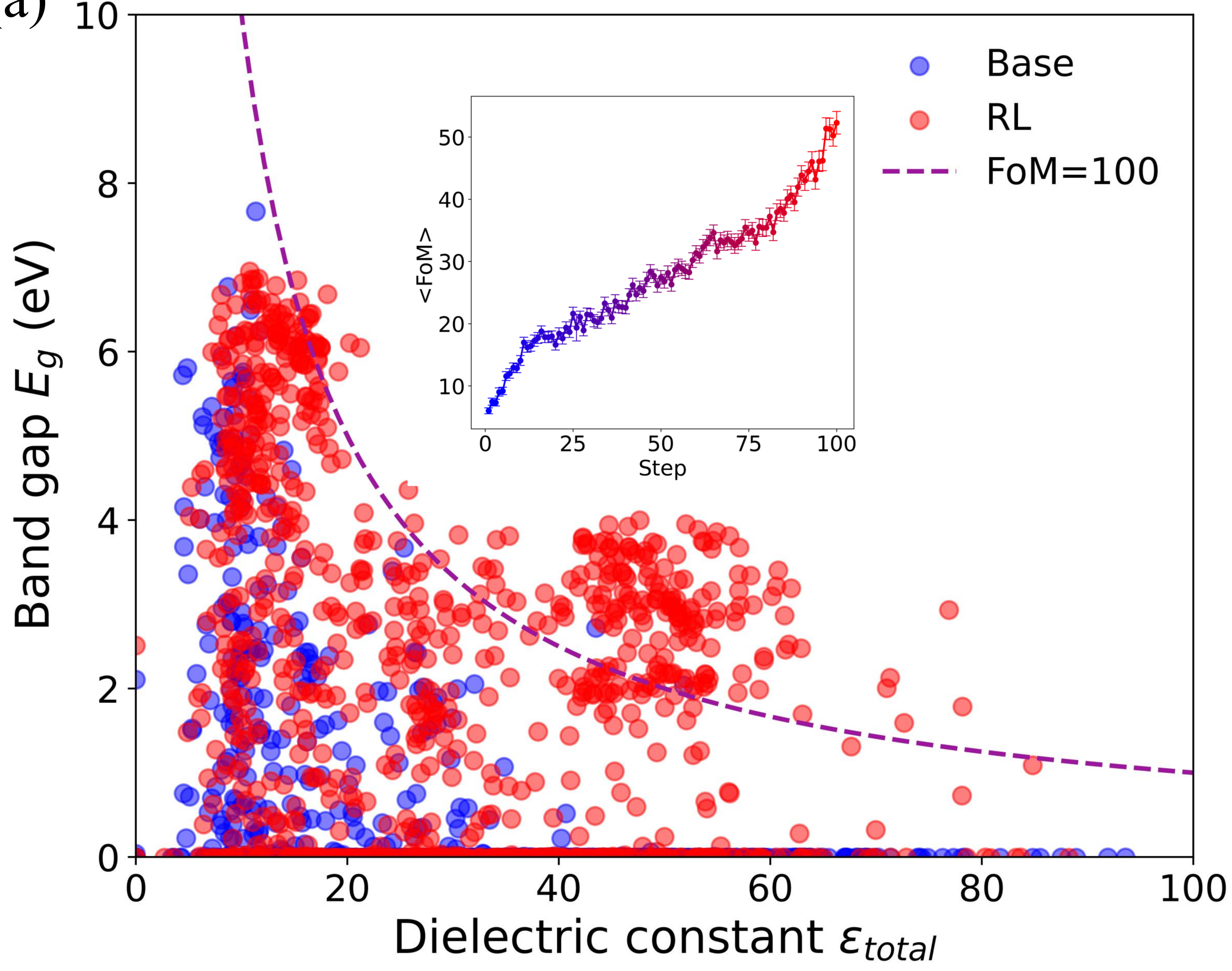
learn from reward
to be a specialized generalist

$$\mathscr{L} = \mathop{\mathbb{E}}_{X\sim q(X)}\left[-r(\boldsymbol{X})\right] + \mathbb{KL}\left(q(\boldsymbol{X})\|p(\boldsymbol{X})\right)$$

$\mathbb{KL}(\text{data} \,\|\, p)$ vs $\mathbb{KL}(q \,\|\, pe^{r})$

Reinforcement fine-tuning for materials design

(a)



CrystalFormer-RL, Cao et al, 2504.02367

$$\mathop{\mathbb{E}}_{\boldsymbol{X}\sim q(\boldsymbol{X})}\left[r(\boldsymbol{X})\right] + \mathbb{KL}\left(q(\boldsymbol{X})\|p(\boldsymbol{X})\right)$$

Reward

Pretrained Crystalformer

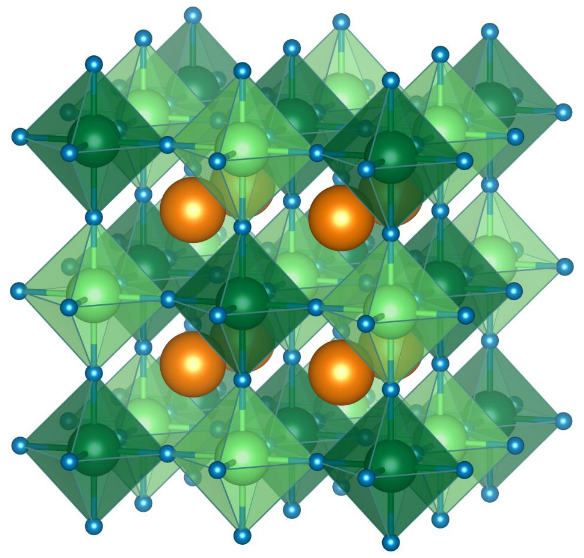
Reward = Band gap x dielectric constant

(Two usually anti-correlated properties)

$\varepsilon_{total} = 21.32$

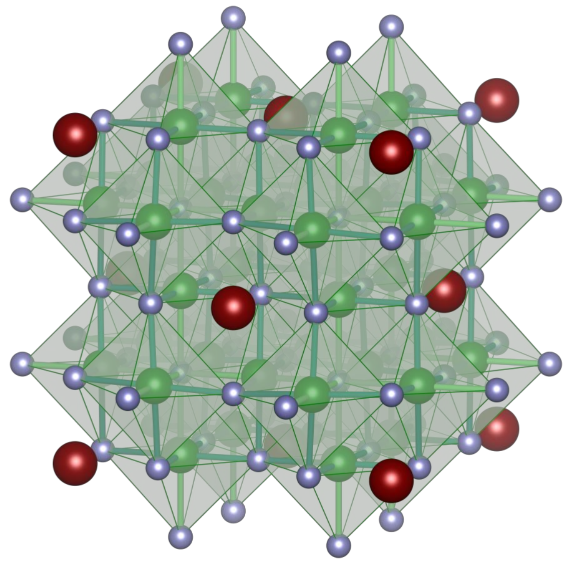
$E_g = 5.38$eV

$E_{hull} = 0.04$eV/atom



225-a-Sr-b-Ba-c-Cs-e-F

Cs_2BaSrF_6



225-a-Pb-c-Cl-d-Li-e-Cl

Li_6PbCl_8

$\varepsilon_{total} = 25.19$

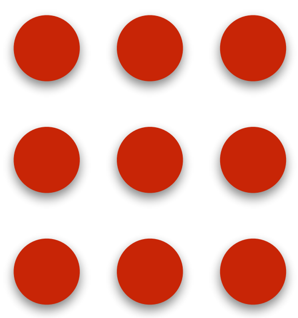
$E_g = 4.64$ eV

$E_{hull} = 0.08$eV/atom

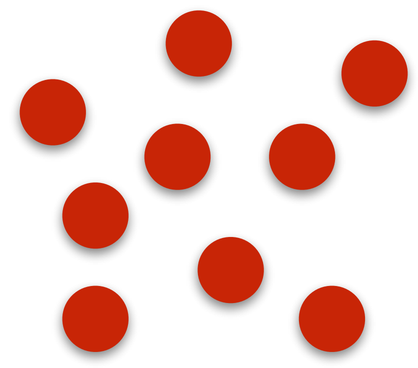
Nature tries to minimize free energy

$$F = E - TS$$

energy



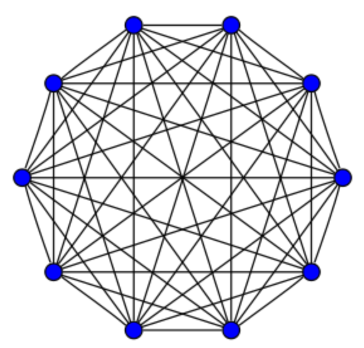
entropy



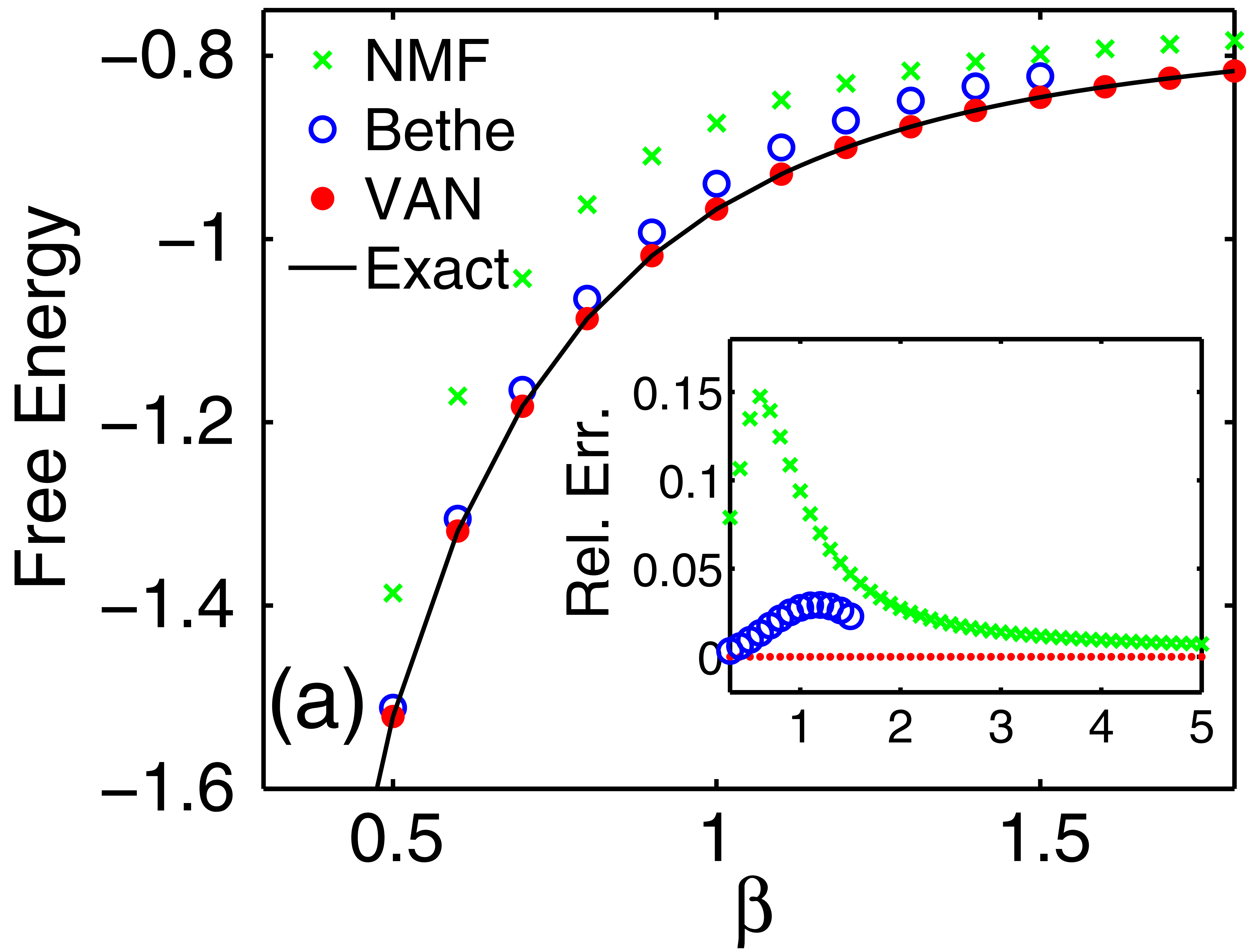
F is a cost function given by Nature

The *same* cost function for training deep generative models

Variational autoregressive network for statistical mechanics



Sherrington-Kirkpatrick spin glass



Objective function: variational free-energy

$$F = \mathop{\mathbb{E}}_{\boldsymbol{X} \sim p(\boldsymbol{X})} \left[E(\boldsymbol{X}) + k_B T \ln p(\boldsymbol{X}) \right]$$

Naive mean-field
factorized probability

$$p(\boldsymbol{X}) = \prod_i p(x_i)$$

Bethe approximation
pairwise interaction

$$p(\boldsymbol{X}) = \prod_i p(x_i) \prod_{(i,j)\in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$$

Variational autoregressive network

$$p(\boldsymbol{X}) = \prod_i p(x_i | \boldsymbol{x}_{<i})$$

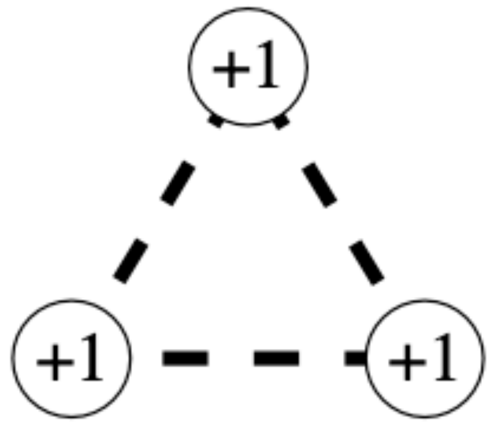
Wu, LW, Zhang, PRL '19

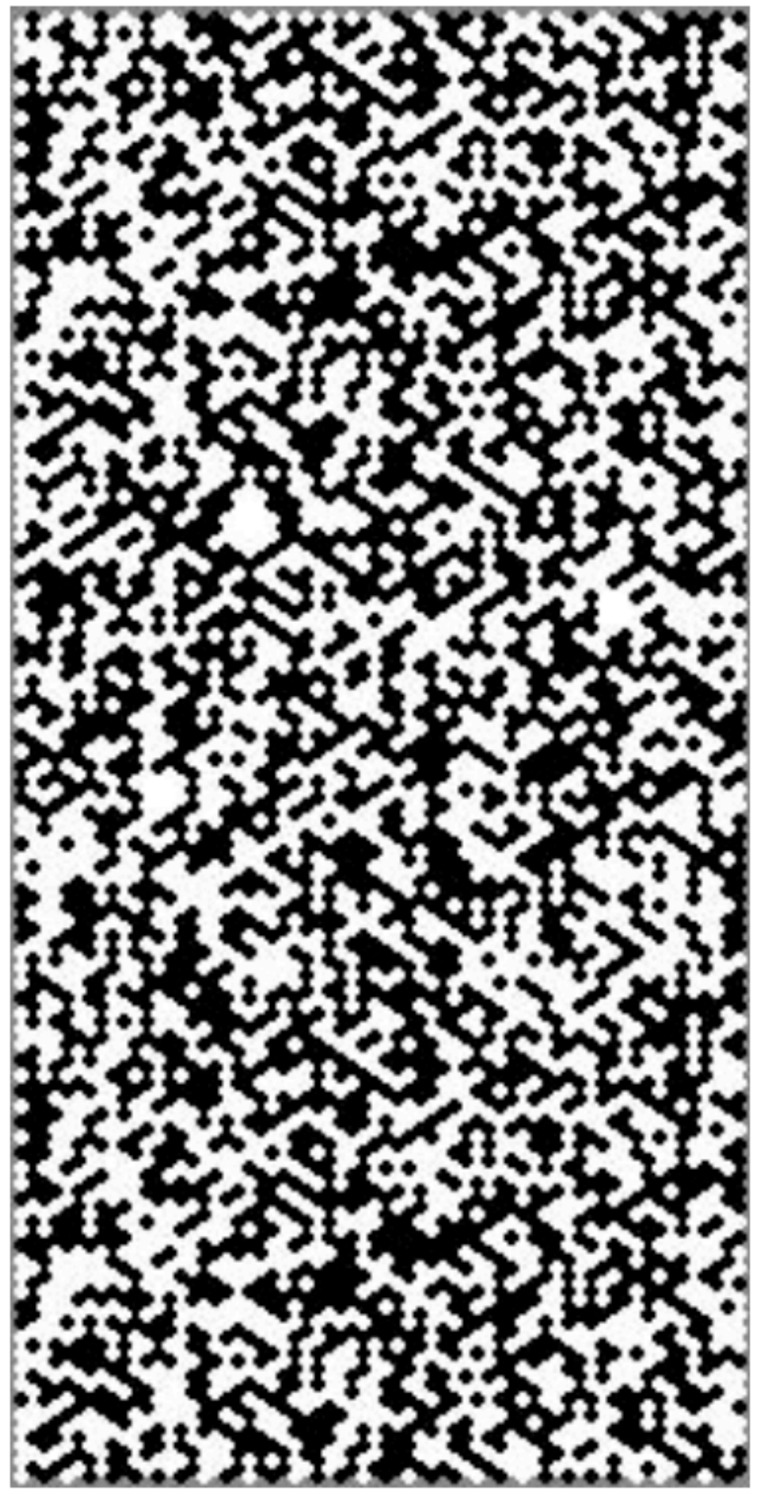
github.com/wdphy16/stat-mech-van

VAN for triangular Ising

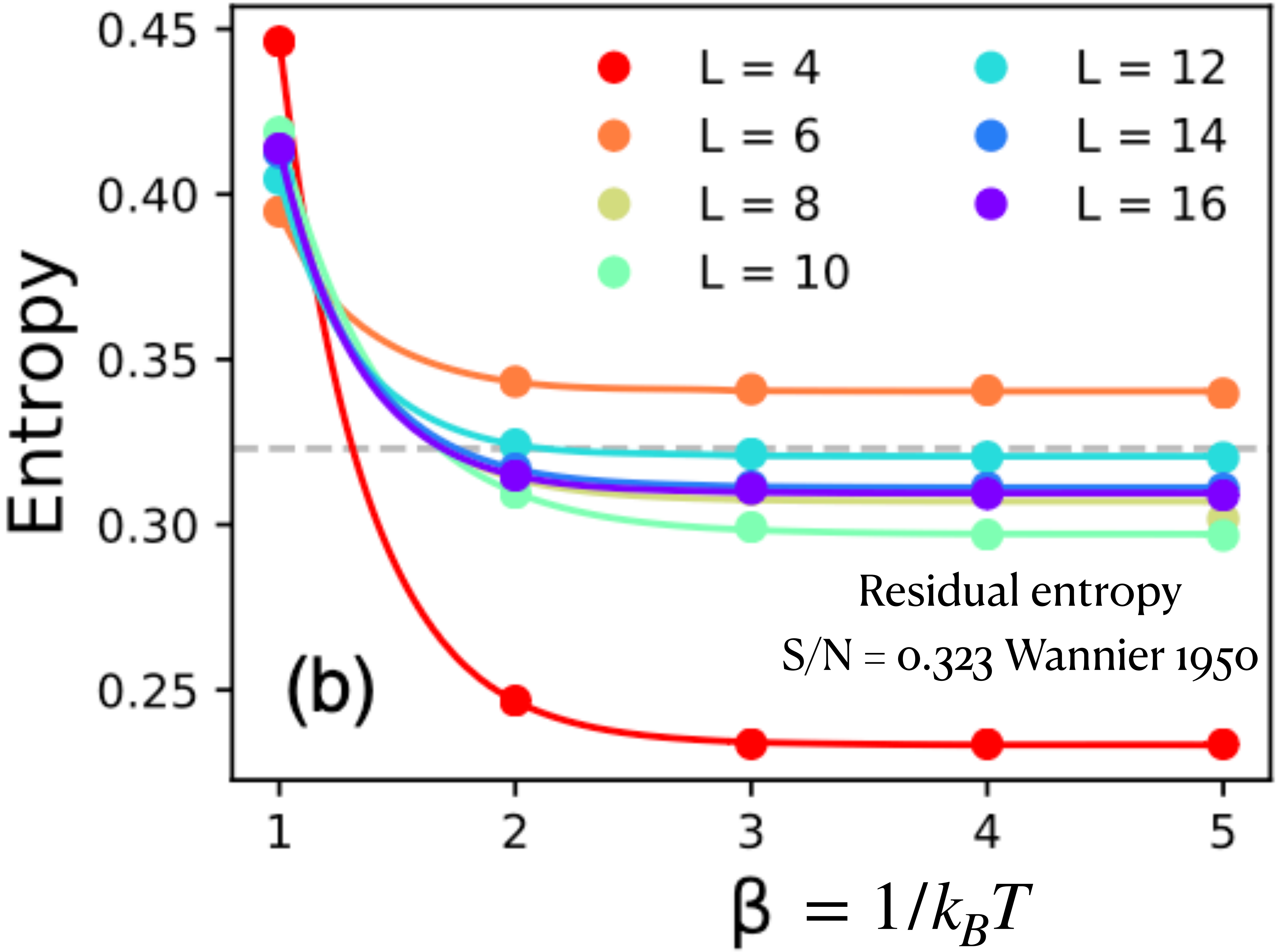
Wu, LW, Zhang, PRL '19

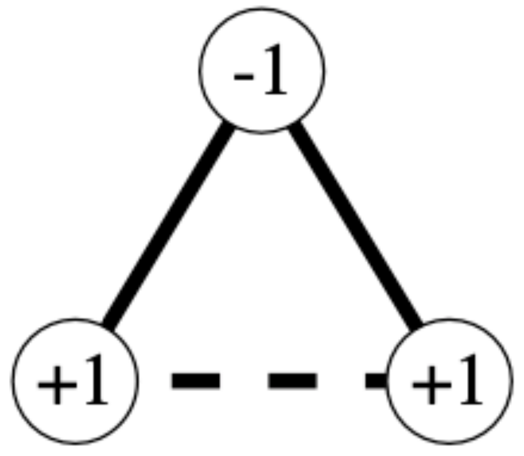
$$F = \mathop{\mathbb{E}}_{\boldsymbol{X} \sim p(\boldsymbol{X})} \left[E(\boldsymbol{X}) + k_B T \ln p(\boldsymbol{X}) \right]$$

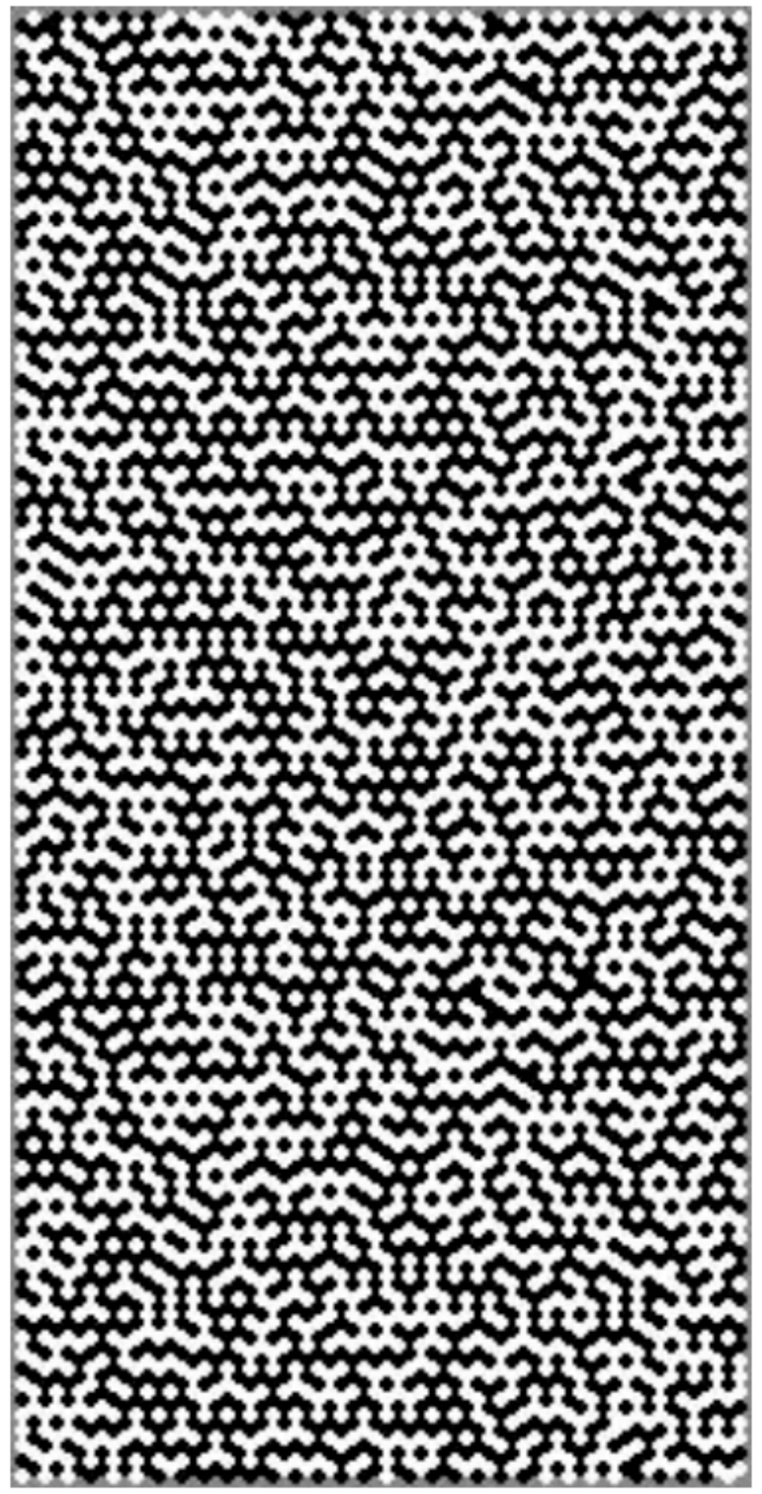




Hot configuration







Cold configuration
MacKay, 2006

VAN (aka RL) for 8-queens problem

$$\mathscr{L} = \mathop{\mathbb{E}}_{\boldsymbol{X} \sim q(\boldsymbol{X})} \left[-r(\boldsymbol{X}) + \ln q(\boldsymbol{X}) \right]$$

Energy exploitation

Entropy exploration

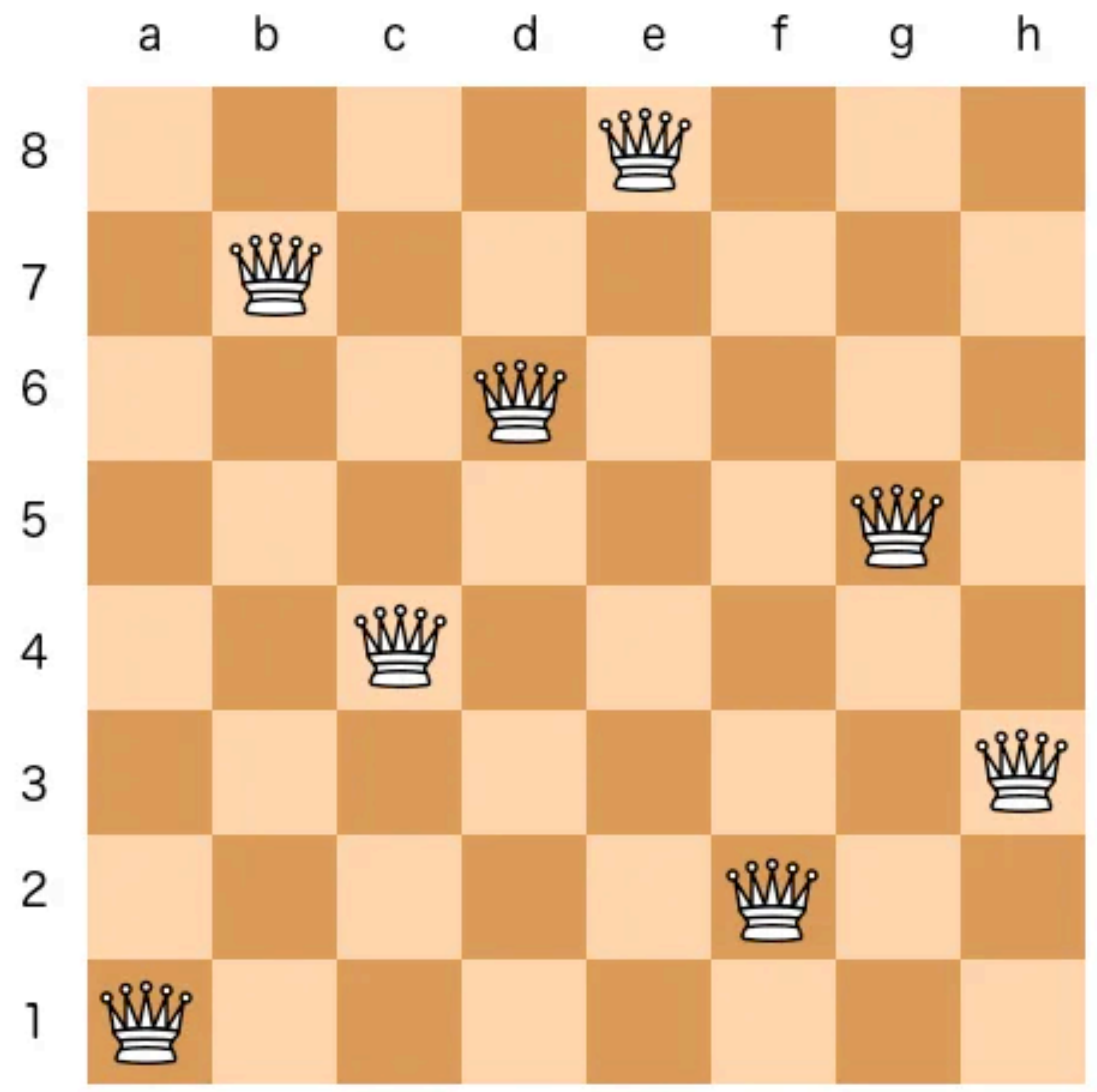
Reward $r(\boldsymbol{X}) = \begin{cases} 1 & \text{if no attack} \\ 0 & \text{otherwise} \end{cases}$

Policy network

$q(\boldsymbol{X}) = q(x_1)q(x_2 \,|\, x_1)\ldots$

$\boldsymbol{X}$: a sequence of actions

a1—b7—c4—d6—e8—f2—g5—h3



VAN (aka RL) for 8-queens problem

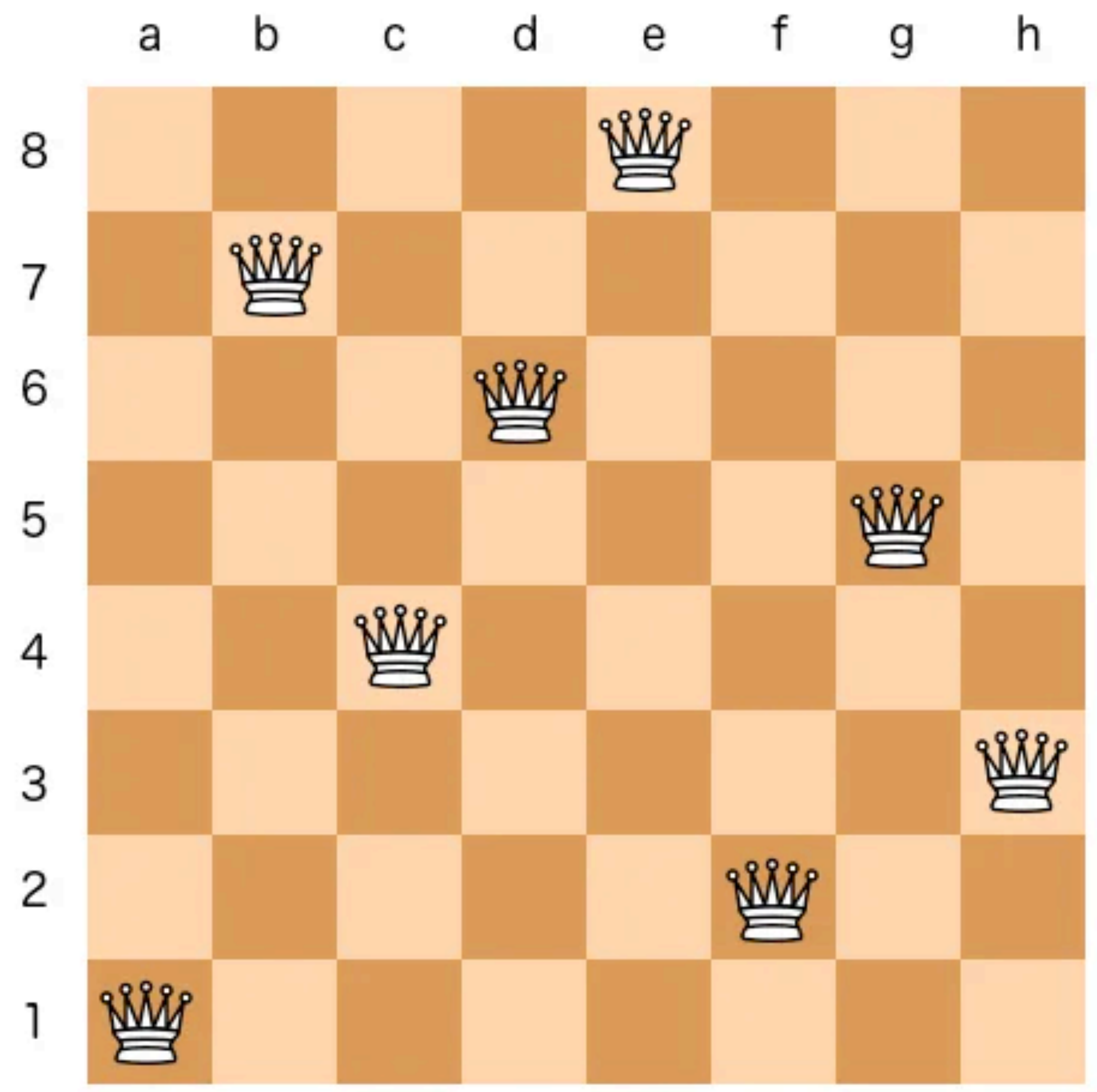
$$\mathscr{L} = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim q(\boldsymbol{X})} \left[-r(\boldsymbol{X}) + \ln q(\boldsymbol{X}) \right]$$

Energy exploitation

Entropy exploration

Board size	Solutions
8	92
12	14,200
16	14,772,512
20	39,029,188,884
24	227,514,171,973,736
28	???

Can you solve it ?



Variational autoregressive quantum states

$$\Psi(\boldsymbol{\sigma}) = \Psi(\sigma_1)\Psi(\sigma_2\,|\,\sigma_1)\Psi(\sigma_3\,|\,\sigma_1,\sigma_2)\cdots$$

Objective function: ground state energy

McMillan 1965, Carleo & Troyer Science 2017

$$\frac{\langle\Psi\,|\,\hat{H}\,|\,\Psi\rangle}{\langle\Psi\,|\,\Psi\rangle} = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim|\Psi(\boldsymbol{\sigma})|^2}\left[\frac{\hat{H}\Psi(\boldsymbol{\sigma})}{\Psi(\boldsymbol{\sigma})}\right]$$

Heisenberg and Hubbard models

Sharir et al, PRL '20, Hibat-Allah et al, PRResarch '20

Humeniuk et al, SciPost '23

Ibarra-García-Padilla et al, 2411.07144 Moss et al, 2502.17144

Quantum chemistry problems

Barrett et al, Nat. Mach. Intell. '22

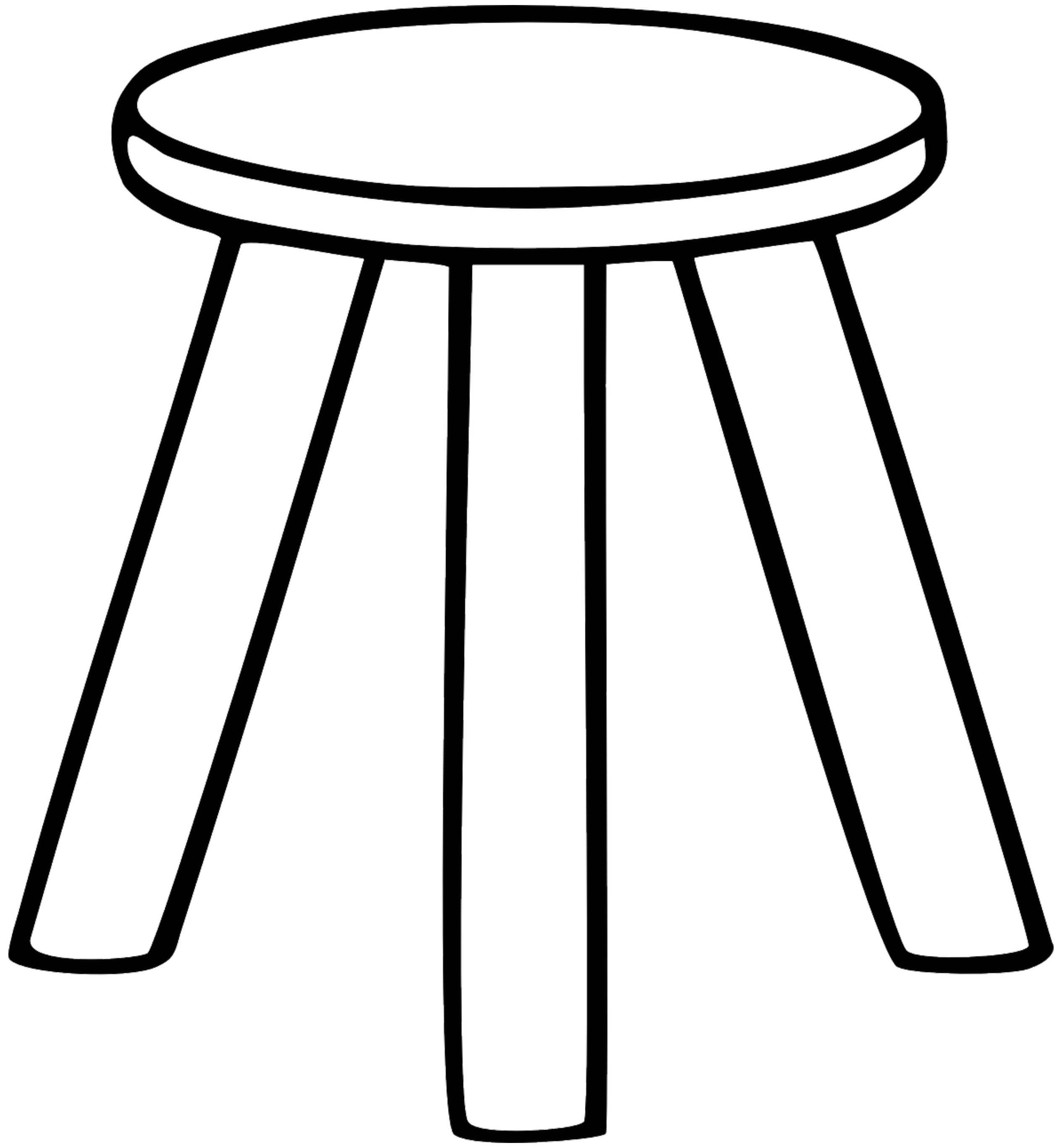
Zhao et al, MLST. '23 Shang et al, 2307.09343

Malyshev et al, 2310.04166 Malyshev et al, 2408.07625

Deep learning for variational calculations

Turning physics problems into stochastic optimization

Leverages the deep learning engine 



RBM, FermiNet,...

Representation

Carleo and Troyer, Science '17

Pfau et al, PR Research '20

Automatic differentiation,

Wasserstein gradient, KFAC, ...

Optimization

Liao et al, PRX '19

Neklyudov, et al, 2307.07050

Chen et al, Nat.Phys. '24

Sampling

Wu et al, PRL '19

Humeniuk et al, SciPost '23

Malyshev et al, 2408.07625

Autoregressive sampling,

Gumbel-top-k sampling,...

Two kinds of variational Monte Carlo

Variational ground state energy $T = 0$

McMillan 1965, Carleo & Troyer Science 2017, Pfau et al, FermiNet, ...

$$E[\Psi] = \mathop{\mathbb{E}}_{X \sim |\psi(X)|^2} \left[\frac{\hat{H}\Psi(X)}{\Psi(X)} \right]$$

Ψ: ANY neural network that respects physical symmetries

Variational free energy $T > 0$

Gibbs–Bogolyubov-Feynman, Li and LW, PRL '18, Wu, LW, Zhang, PRL '19, ...

$$F[p] = \mathop{\mathbb{E}}_{X \sim p(X)} \left[E(\boldsymbol{X}) + k_B T \ln p(\boldsymbol{X}) \right]$$

p: probabilistic models with tractable normalization

Three kinds of variational Monte Carlo

Quantum Ground state

McMillan 1965
Carleo & Troyer Science 2017, Pfau et al, FermiNet, ...

$$E[\Psi] = \mathop{\mathbb{E}}_{X\sim|\psi(X)|^2}\left[\frac{\hat{H}\Psi(X)}{\Psi(X)}\right]$$

Ψ: ANY neural network that respects physical symmetries

Quantum Stat-Mech

$$\rho$$

Classical Stat-Mech

Gibbs–Bogolyubov-Feynman
Li and LW, PRL '18, Wu, LW, Zhang, PRL '19, ...

$$F[p] = \mathop{\mathbb{E}}_{X\sim p(X)}\left[E(X) + k_B T \ln p(X)\right]$$

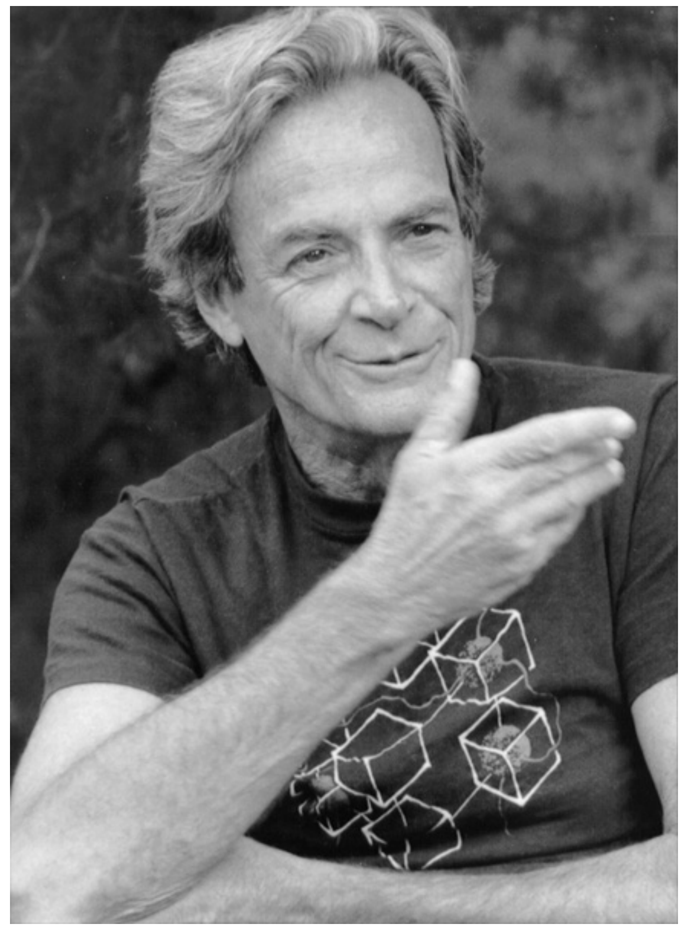
p: probabilistic models with tractable normalization

The variational free energy principle

Gibbs–Bogolyubov-Feynman-Delbrück–Molière

$$\min \quad F[\rho] = \mathrm{Tr}(H\rho) + k_B T\, \mathrm{Tr}(\rho \ln \rho) \quad \geq F$$

$\downarrow$ variational density matrix $\quad \downarrow$ energy $\quad \downarrow$ entropy 😱



Difficulties in Applying the Variational Principle to Quantum Field Theories[1]

Richard P. Feynman

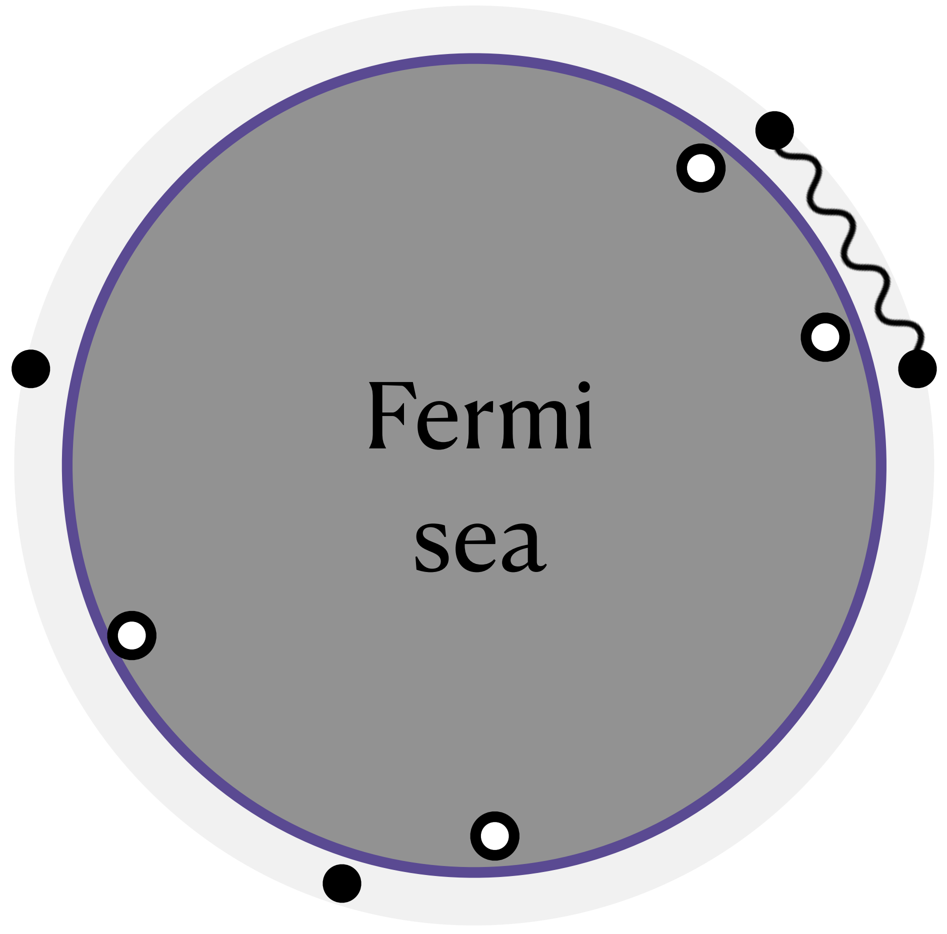
[1]transcript of Professor Feynman's talk in 1987

ρ ?

Generative models !

Example: the variational density matrix of electron gas

Xie, Zhang, LW, SciPost Physics '23



$$\rho = \sum_{\boldsymbol{K}} p(\boldsymbol{K}) \left| \Psi_{\boldsymbol{K}} \rangle \langle \Psi_{\boldsymbol{K}} \right|$$

Normalized probability distribution

Orthonormal many-electron basis

Low-energy excited states are labeled in the same way as the ideal Fermi gas

$\boldsymbol{K} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_N\}$

1 $\sum_{\boldsymbol{K}} p(\boldsymbol{K}) = 1$

autoregressive model

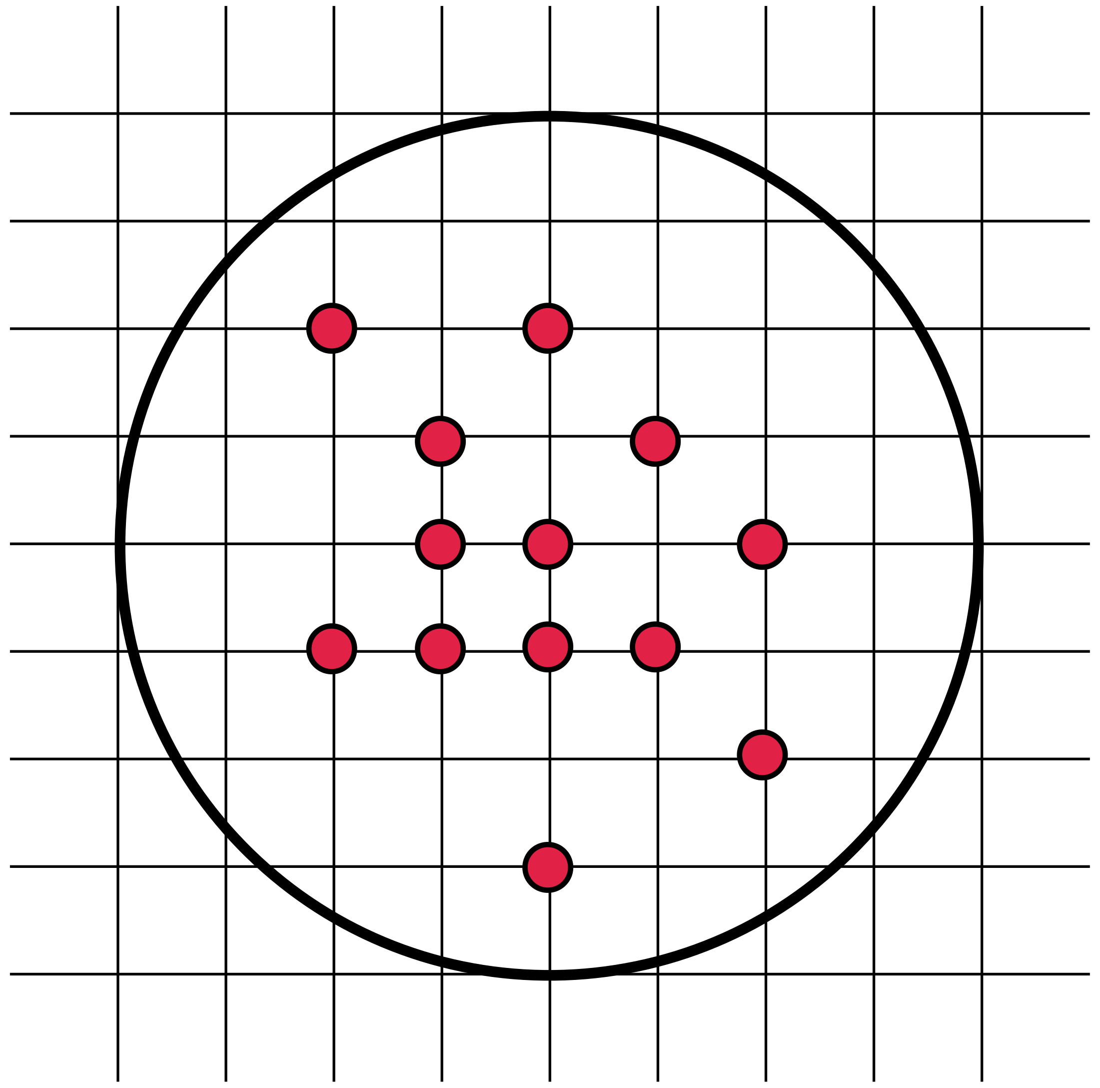
2 $\langle \Psi_{\boldsymbol{K}} | \Psi_{\boldsymbol{K}'} \rangle = \delta_{\boldsymbol{K},\boldsymbol{K}'}$

$\sqrt{\text{flow}}$

There will also be interesting twists for physics considerations

❶ Variational autoregressive network for $p(\boldsymbol{K})$

Fermionic occupation in k-space



$$p(\boldsymbol{K}) = p(\boldsymbol{k}_1)p(\boldsymbol{k}_2|\boldsymbol{k}_1)p(\boldsymbol{k}_3|\boldsymbol{k}_1,\boldsymbol{k}_2)\cdots$$

$\binom{M}{N}$ probability space

N	# of fermions	# of words
M	Momentum cutoff	Vocabulary

Pauli exclusion: we are modeling a *set of words* with no repetitions and no order

We use masked casual self-attention Vaswani et al 1706.03762; Alternative solution: Hibat-Allah et al, 2002.02793, Barrett et al, 2109.12606

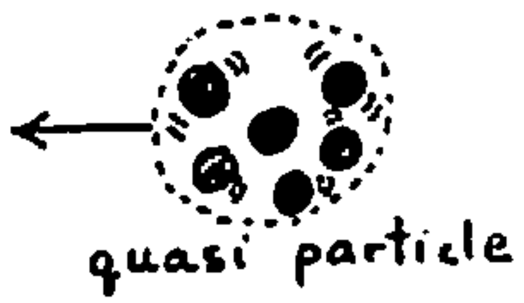
② $\sqrt{\text{flow}}$ for $|\Psi_K\rangle$

$$\Psi_K(X) = \frac{\det(e^{ik_i \cdot z_j})}{\sqrt{N!}} \cdot \left| \det\left(\frac{\partial Z}{\partial X}\right) \right|^{\frac{1}{2}}$$

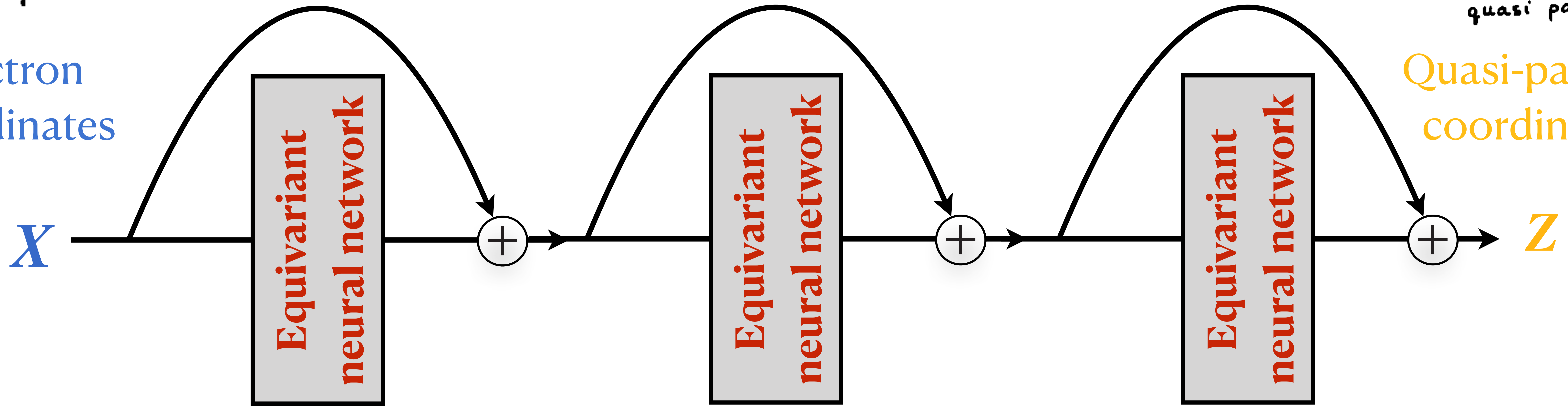
Xie, Zhang, LW, SciPost '23

Orthonormal many-body states

Electron coordinates



Quasi-particle coordinates



$X \leftrightarrow Z$: unitary backflow between particle and quasi-particle coordinates

Fermion statistics: permutation equivariant flow We use FermiNet layer Pfau et al, 1909.02487

The objective function of variational density matrix

$$\boxed{\rho = \sum_{\boldsymbol{K}} p(\boldsymbol{K}) \left| \Psi_{\boldsymbol{K}} \rangle \langle \Psi_{\boldsymbol{K}} \right|}$$

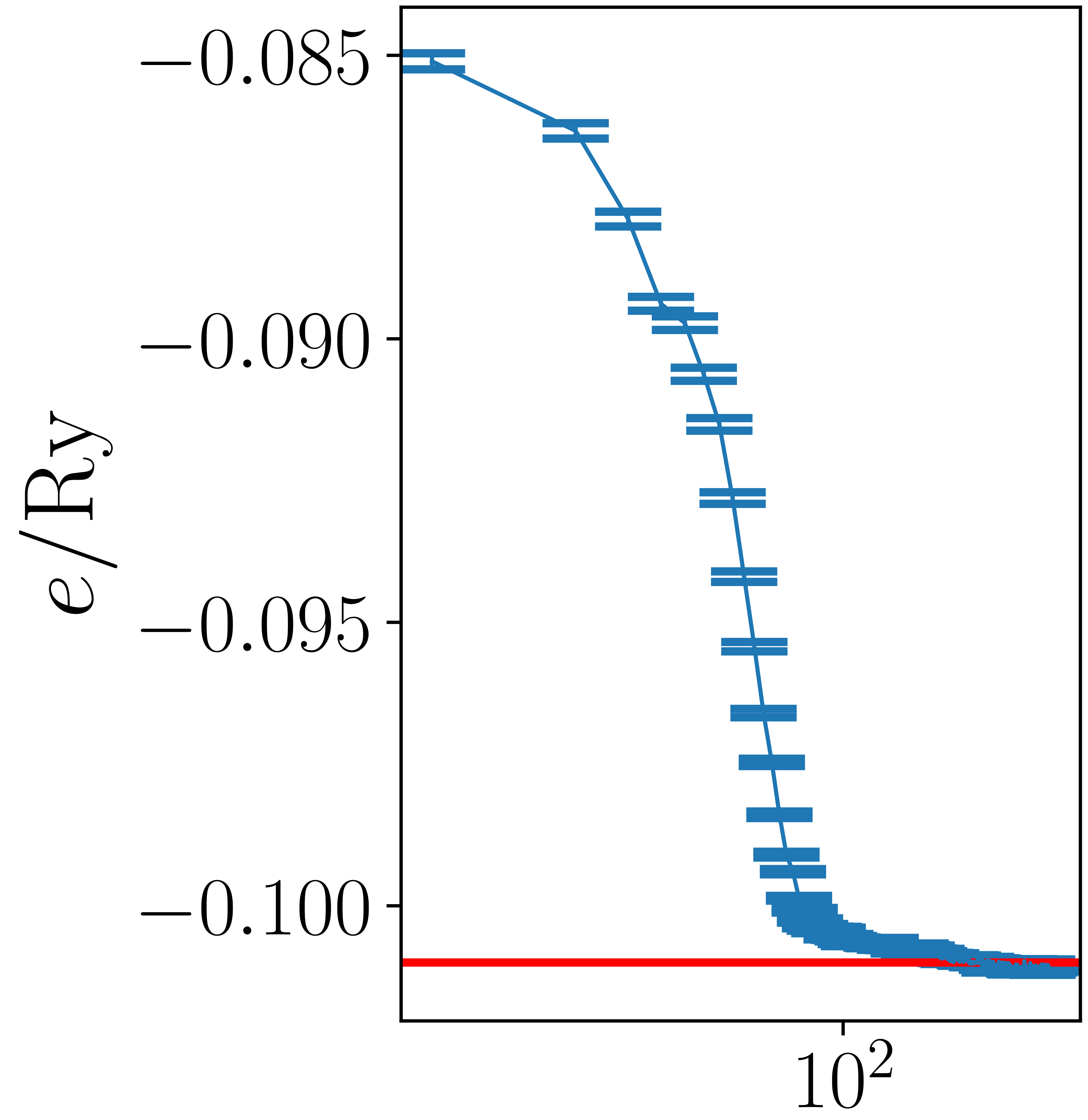
$$F = \mathop{\mathbb{E}}_{\boldsymbol{K}\sim p(\boldsymbol{K})} \left[\underbrace{k_B T \ln p(\boldsymbol{K})}_{\text{Boltzmann distribution}} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\left|\Psi_{\boldsymbol{K}}(\boldsymbol{X})\right|^2}}_{\text{Born probability}} \left[\frac{H\Psi_{\boldsymbol{K}}(\boldsymbol{X})}{\Psi_{\boldsymbol{K}}(\boldsymbol{X})} \right] \right]$$

Jointly optimize $p(\boldsymbol{K})$ and $\Psi_{\boldsymbol{K}}(\boldsymbol{X})$ to minimize the variational free energy

Benchmarks on uniform electron gas

Xie, Zhang, LW, SciPost Physics '23

$r_s = 10$, $T/T_F=0.0625$, $N = 33$



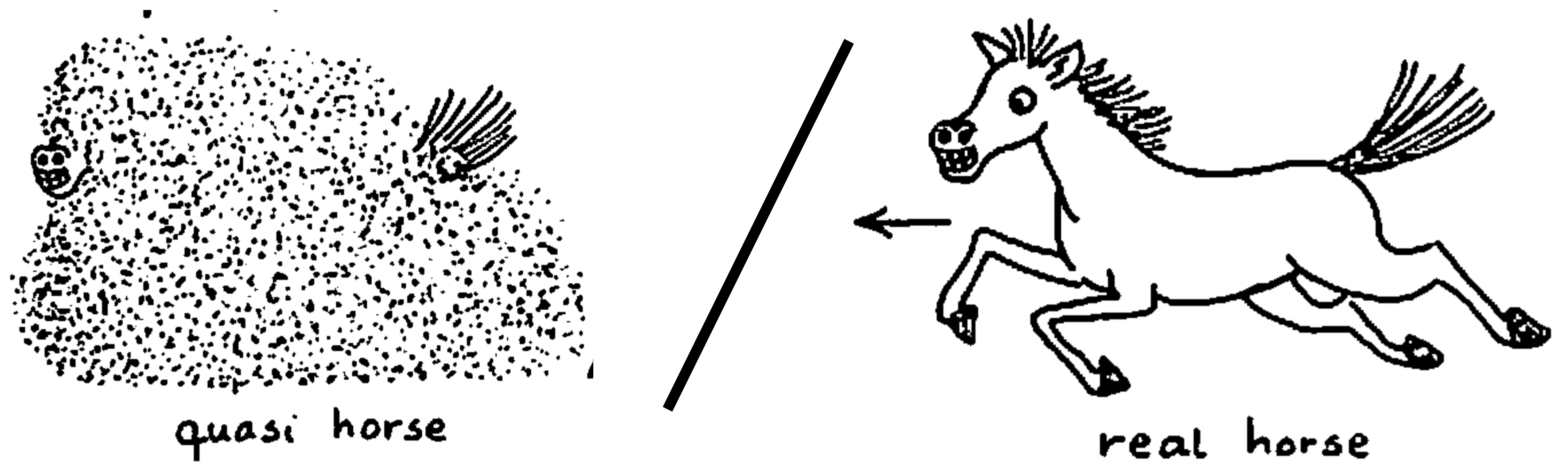
metals: $2< r_s < 6$

r_s	Θ	$\langle sign \rangle$	E_{tot}^{exact}	E_{tot}
4.0	0.0625	−0.00055(62)	−0.5(1)	−0.1023(7)
10.0	0.0625	−0.002(1)	−0.16(2)	−0.1010(1)

Brown et al, PRL '13 Restricted PIMC

see also Schoof et al PRL '15, Malone et al PRL '16

Application to 2DEG effective mass



Quansi-particle effective mass
contradicting experiments

$m*/m > 1$

VOLUME 91, NUMBER 4 PHYSICAL REVIEW LETTERS week ending 25 JULY 2003

Spin-Independent Origin of the Strongly Enhanced Effective Mass in a Dilute 2D Electron System

$m*/m < 1$

PRL **101,** 026402 (2008) PHYSICAL REVIEW LETTERS week ending 11 JULY 2008

Effective Mass Suppression in Dilute, Spin-Polarized Two-Dimensional Electron Systems

Medini Padmanabhan, T. Gokmen, N. C. Bishop, and M. Shayegan

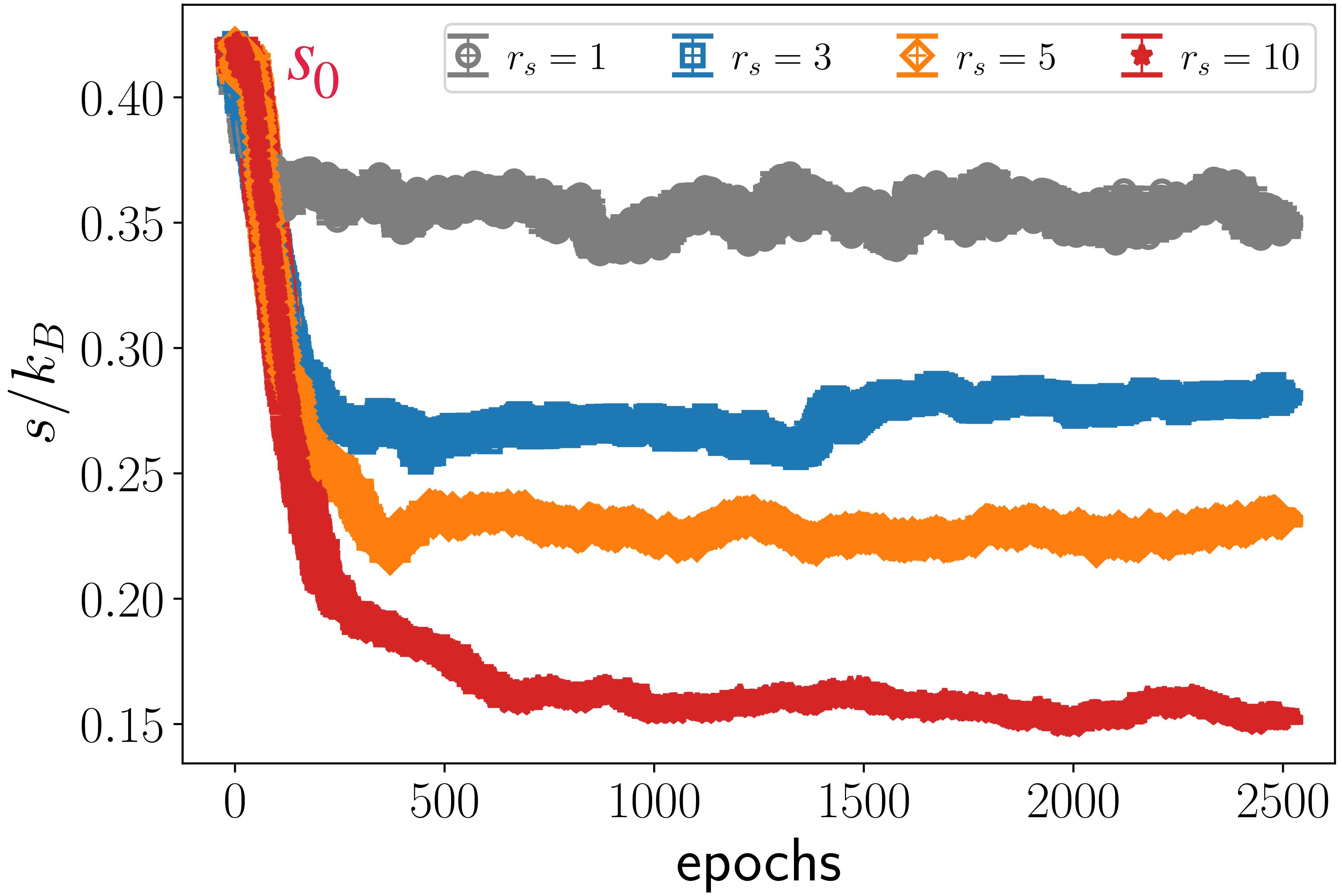
Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA

(Received 19 September 2007; published 7 July 2008)

Hao Xie et al, SciPost Physics '23

Thermal entropy of 2D electron gas

$$\frac{m^*}{m} = \frac{s}{s_0} < 1$$

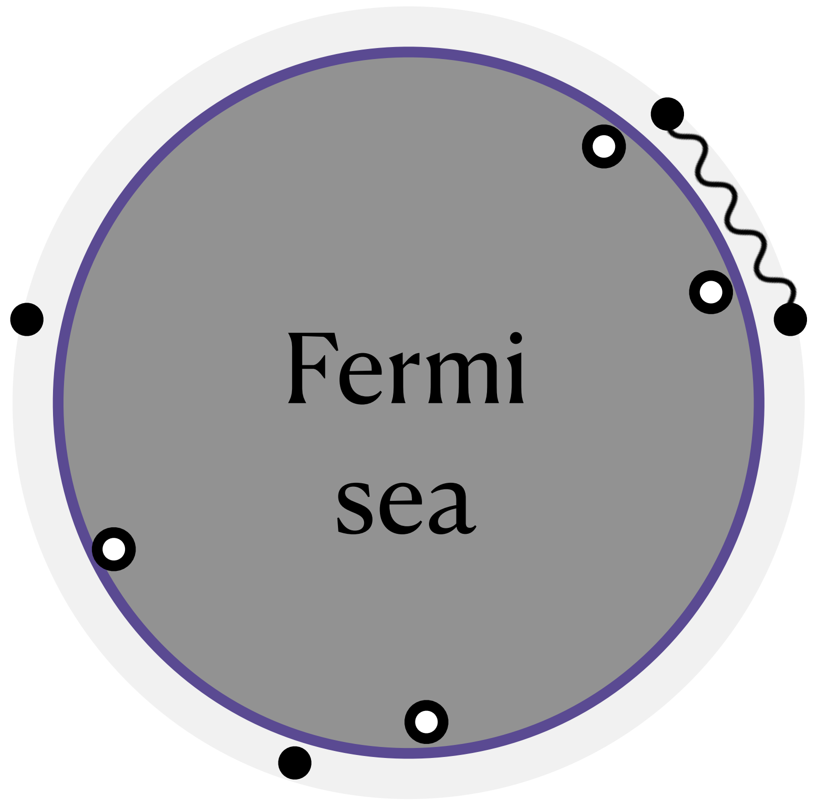


Deep variational free-energy for electrons and atoms

Ideal Fermi gas

↕

Fermi liquid

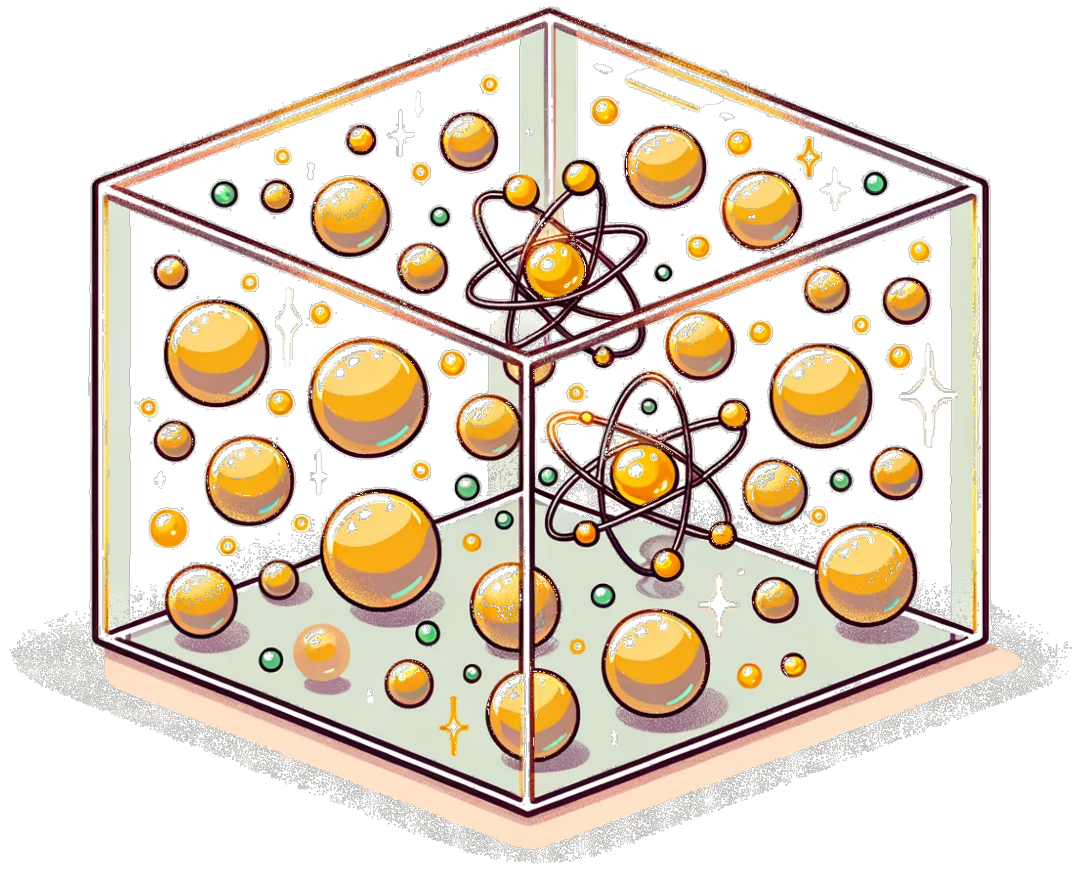


Low-temperature properties
of Coulomb gas
JML ’22 and SciPost Physics ’23
(~50 electrons)

Hartree-Fock states

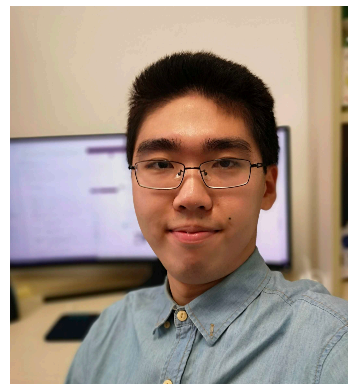
↕

Interacting electrons



Equation of states of
dense hydrogen
PRL ’23 and ongoing
(~50 e-p pairs)

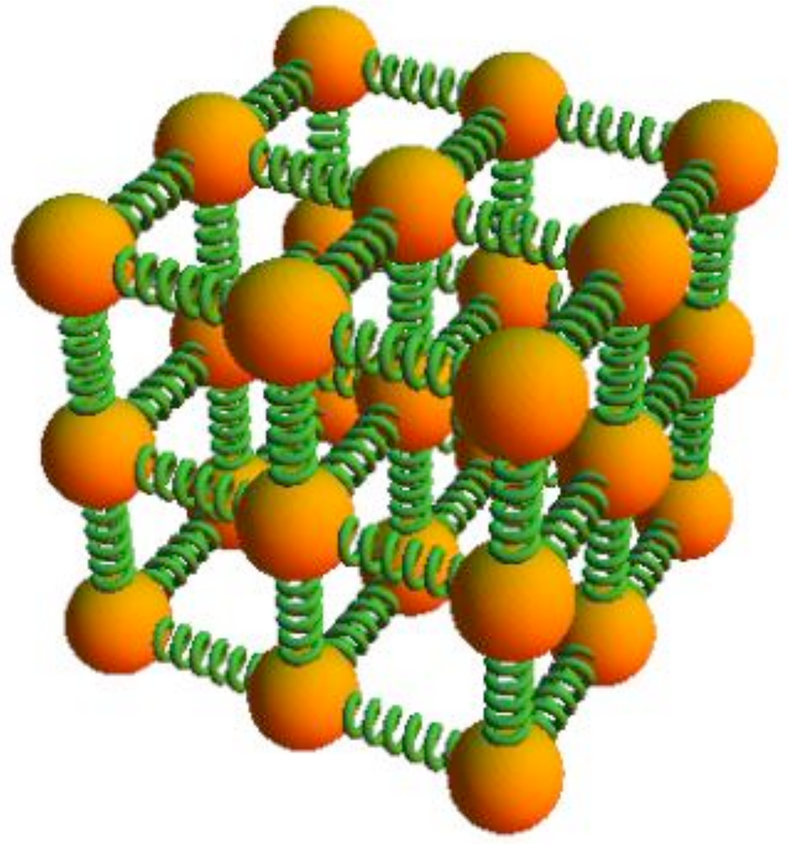
Poster by Zihang Li



Harmonic oscillators

↕

Anharmonic crystal



Vibrational spectra of
molecules and quantum solids
JCP ’24 and 2412.12451
(~500 atoms)

Poster by Qi Zhang

Generative AI for It

① $$p(\boldsymbol{X}\,|\,y) \propto p(\boldsymbol{X})p(y\,|\,\boldsymbol{X})$$

Matter inverse design

Exploiting intuitions in data

② $$F[\rho] = E - TS$$

Nature's cost function

Variational free energy is finally practical

Autoregressive modeling

1 Ordering

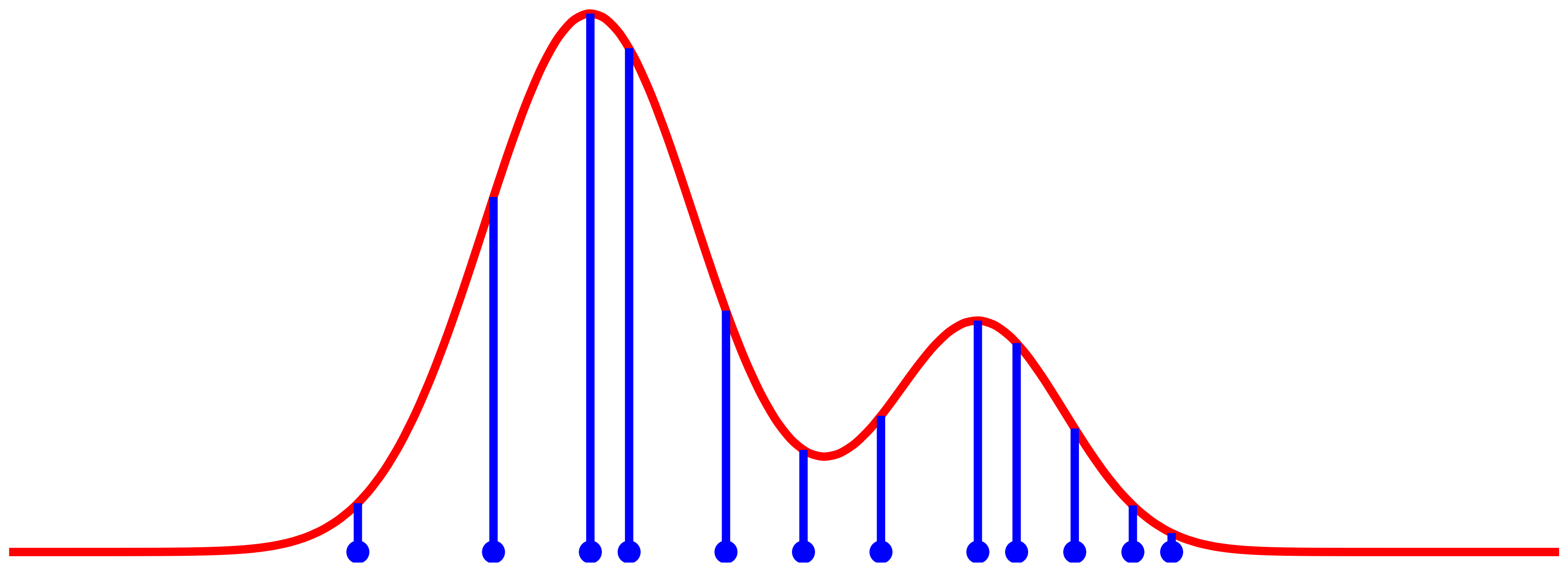
2 Tokenization

3 Objective function

4 Inference

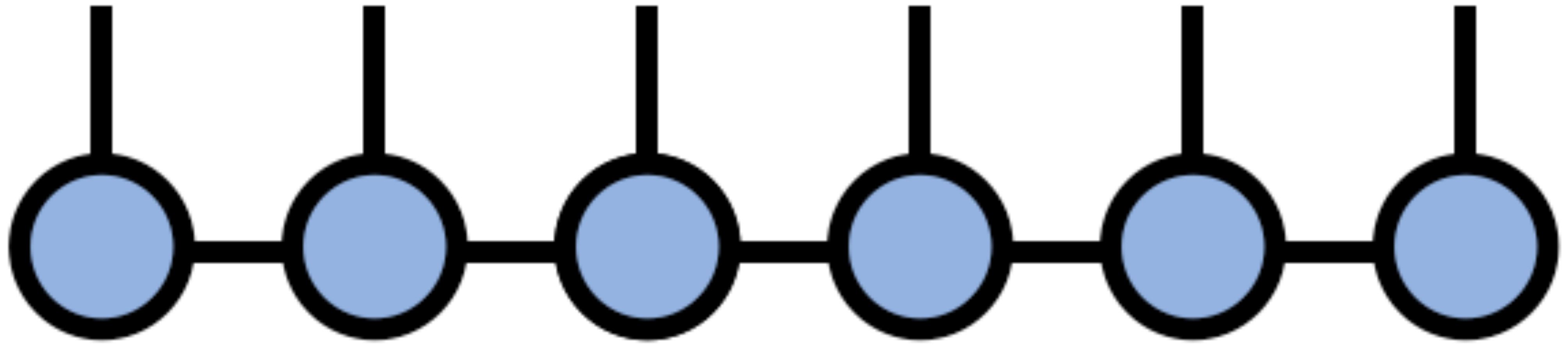
Wyckoff Positions of Group *Fm-3m* (No. 225)

Multiplicity	Wyckoff letter	Site symmetry	Coordinates
			(0,0,0) + (0,1/2,1/2) + (1/2,0,1/2) + (1/2,1/2,0) +
192	l	1	(x,y,z) (-x,-y,z) (-x,y,-z) (x,-y,-z) (z,x,y) (z,-x,-y) (-z,-x,y) (-z,x,-y) (y,z,x) (-y,z,-x) (y,-z,-x) (-y,-z,x) (y,x,-z) (-y,-x,-z) (y,-x,z) (-y,x,z) (x,z,-y) (-x,z,y) (-x,-z,-y) (x,-z,y) (z,y,-x) (z,-y,x) (-z,y,x) (-z,-y,-x) (-x,-y,-z) (x,y,-z) (x,-y,z) (-x,y,z) (-z,-x,-y) (-z,x,y) (z,x,-y) (z,-x,y) (-y,-z,-x) (y,-z,x) (-y,z,x) (y,z,-x) (-y,-x,z) (y,x,z) (-y,x,-z) (y,-x,-z) (-x,-z,y) (x,-z,-y) (x,z,y) (-x,z,-y) (-z,-y,x) (-z,y,-x) (z,-y,-x) (z,y,x)
96	k	..m	(x,x,z) (-x,-x,z) (-x,x,-z) (x,-x,-z) (z,x,x) (z,-x,-x) (-z,-x,x) (-z,x,-x) (x,z,x) (-x,z,-x) (x,-z,-x) (-x,-z,x) (x,x,-z) (-x,-x,-z) (x,-x,z) (-x,x,z) (x,z,-x) (-x,z,x) (-x,-z,-x) (x,-z,x) (z,x,-x) (z,-x,x) (-z,x,x) (-z,-x,-x)
96	j	m..	(0,y,z) (0,-y,z) (0,y,-z) (0,-y,-z) (z,0,y) (z,0,-y) (-z,0,y) (-z,0,-y) (y,z,0) (-y,z,0) (y,-z,0) (-y,-z,0) (y,0,-z) (-y,0,-z) (y,0,z) (-y,0,z) (0,z,-y) (0,z,y) (0,-z,-y) (0,-z,y) (z,y,0) (z,-y,0) (-z,y,0) (-z,-y,0)
48	i	m.m 2	(1/2,y,y) (1/2,-y,y) (1/2,y,-y) (1/2,-y,-y) (y,1/2,y) (y,1/2,-y) (-y,1/2,y) (-y,1/2,-y) (y,y,1/2) (-y,y,1/2) (y,-y,1/2) (-y,-y,1/2)
48	h	m.m 2	(0,y,y) (0,-y,y) (0,y,-y) (0,-y,-y) (y,0,y) (y,0,-y) (-y,0,y) (-y,0,-y) (y,y,0) (-y,y,0) (y,-y,0) (-y,-y,0)
48	g	2.m m	(x,1/4,1/4) (-x,3/4,1/4) (1/4,x,1/4) (1/4,-x,3/4) (1/4,1/4,x) (3/4,1/4,-x) (1/4,x,3/4) (3/4,-x,3/4) (x,1/4,3/4) (-x,1/4,1/4) (1/4,1/4,-x) (1/4,3/4,x)
32	f	.3m	(x,x,x) (-x,-x,x) (-x,x,-x) (x,-x,-x) (x,x,-x) (-x,-x,-x) (x,-x,x) (-x,x,x)
24	e	4m. m	(x,0,0) (-x,0,0) (0,x,0) (0,-x,0) (0,0,x) (0,0,-x)
24	d	m.m m	(0,1/4,1/4) (0,3/4,1/4) (1/4,0,1/4) (1/4,0,3/4) (1/4,1/4,0) (3/4,1/4,0)
8	c	-43m	(1/4,1/4,1/4) (1/4,1/4,3/4)
4	b	m-3m	(1/2,1/2,1/2)
4	a	m-3m	(0,0,0)



$$\mathbb{KL}(\text{data} \parallel p) \quad \text{vs} \quad \mathbb{KL}(p \parallel e^{-E/k_B T})$$

Comparision/Connection to MPS



Han et al, PRX '18

Both require 1D ordering

Both support direct sampling and tractable normaliztion

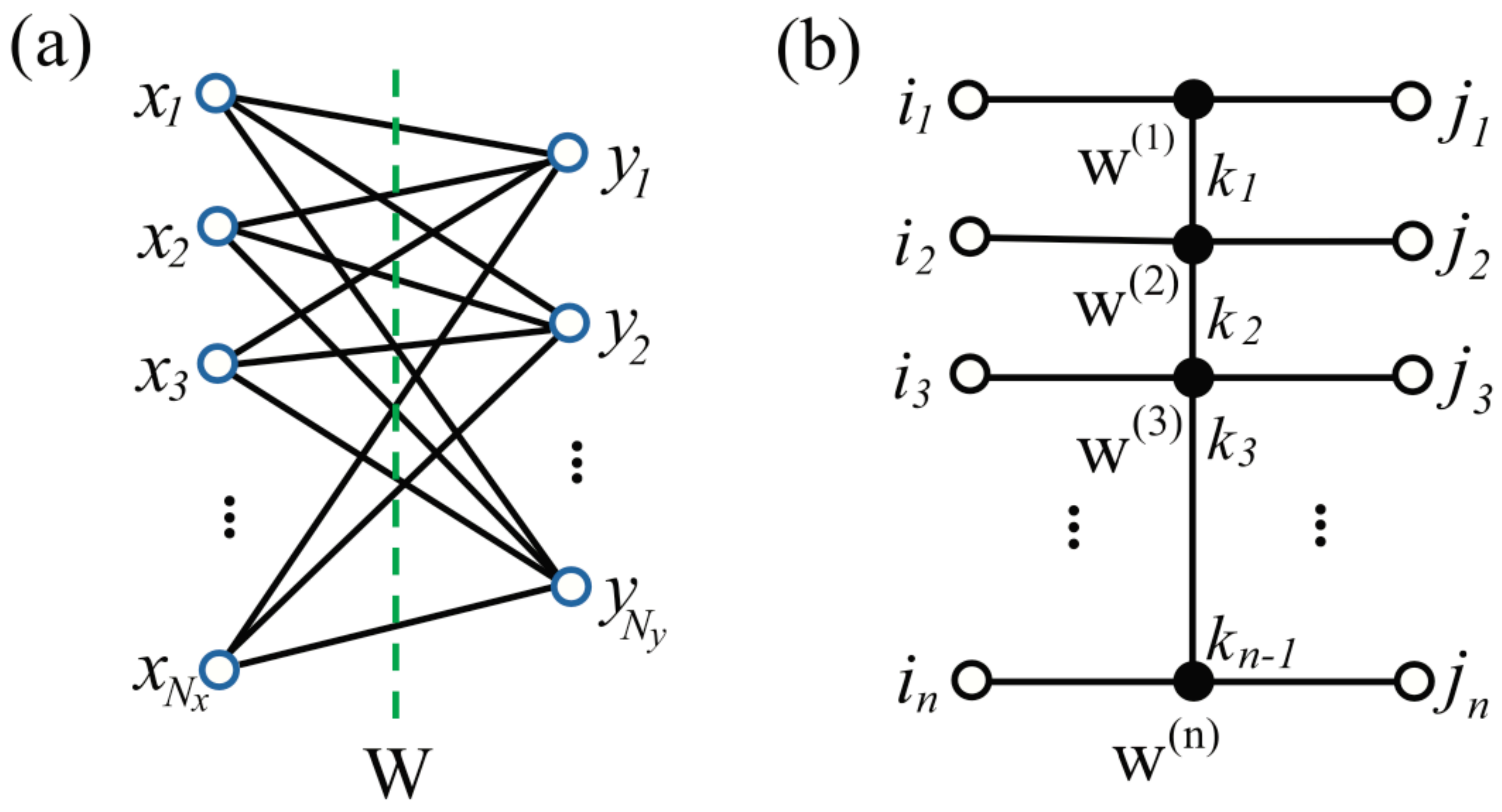
MPS is bidirectional

MPS mediates long range correlation via virtual bonds

Similar to recurrent neural net

Can we tensorize GPT ? What does it good for ?

Tensor network-based compression and finetuning



Gao et al, PRResearch '20

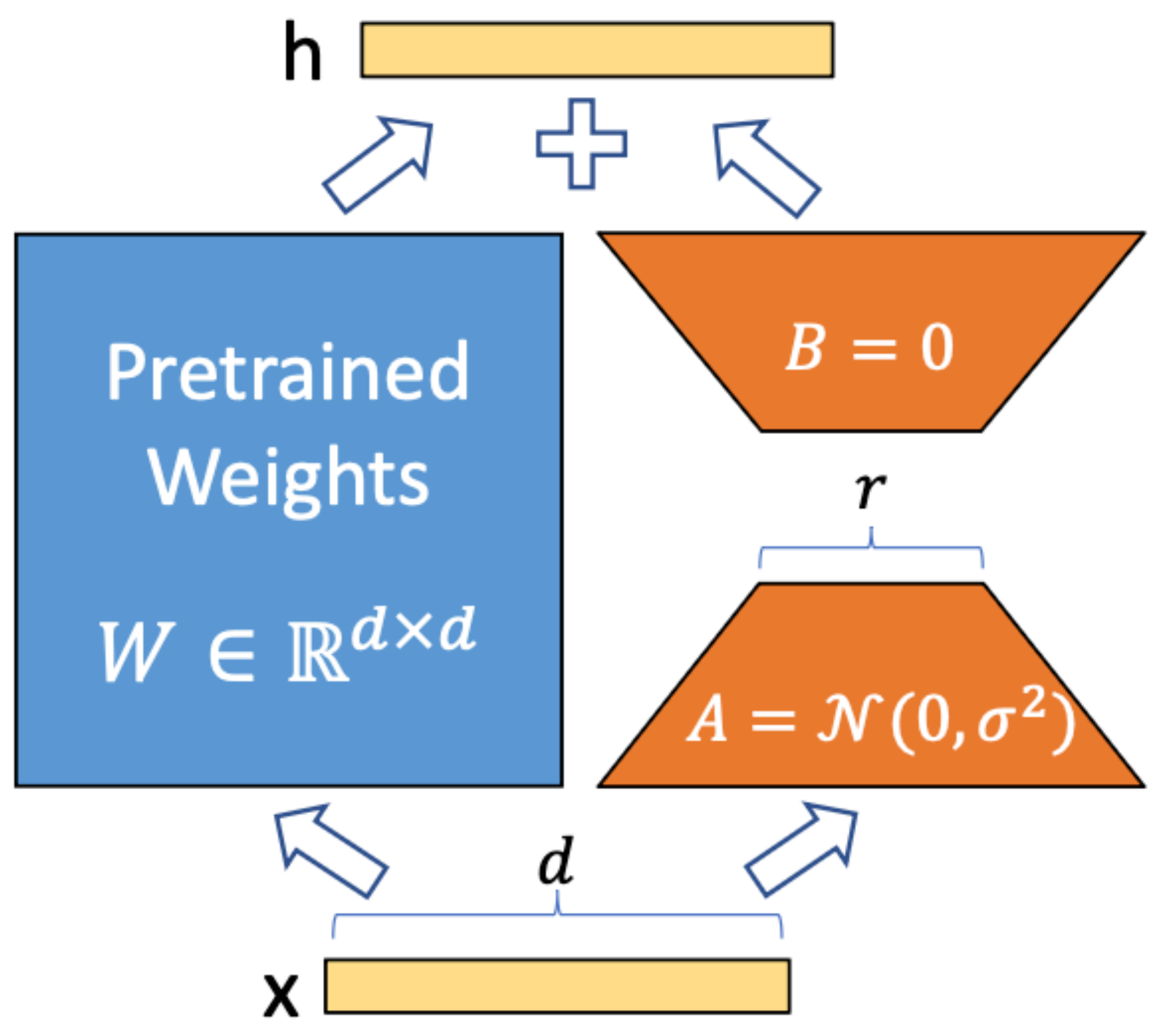
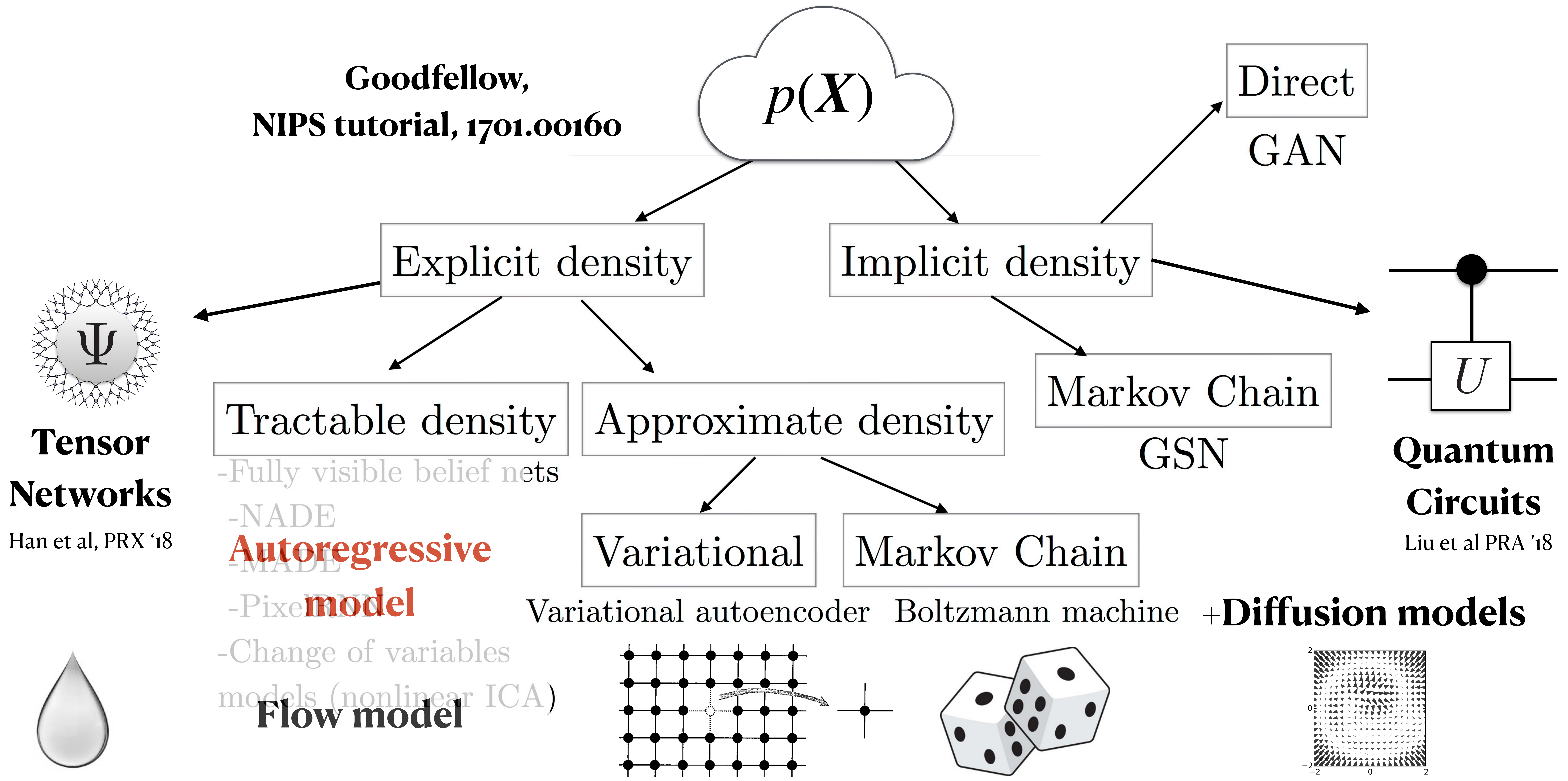


Figure 1: Our reparametrization. We only train A and B.

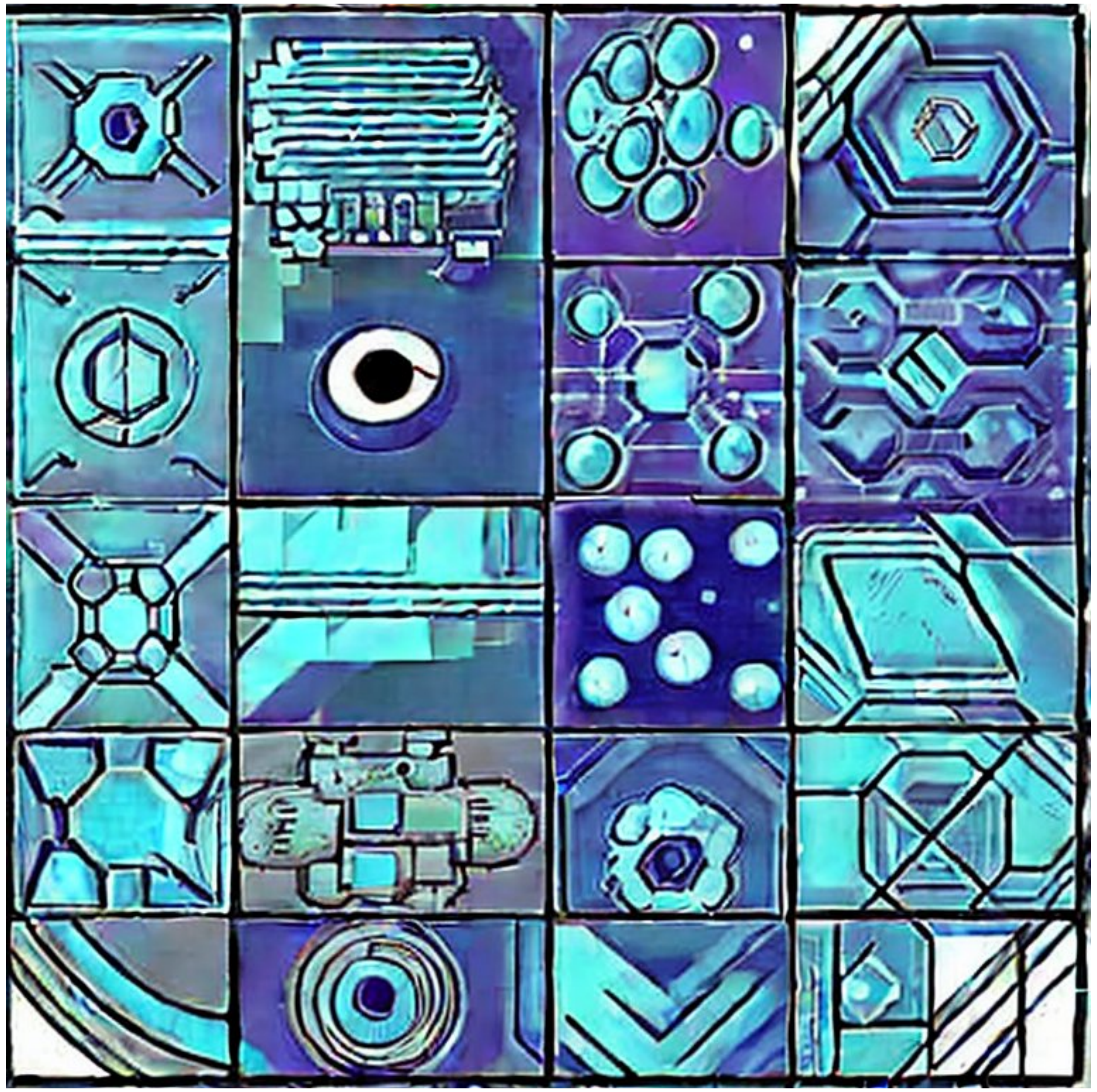
Low-Rank Adaptation (LoRA)
Hu et al, 2106.09685

Generative models and their physics genes



A crash course offerred at IOP 2023 spring

2.23	Overview
3.2	Machine learning practices
3.9	A hitchhiker's guide to deep learning
3.16	Research projects hands-on
3.23	Symmetries in machine learning
3.30	Differentiable programming
4.6	Generative models-I
4.13	Generative models-II
4.20	Research projects presentation
4.27	AI for science: why now ?



Machine learning for physicists

https://github.com/wangleiphy/ml4p