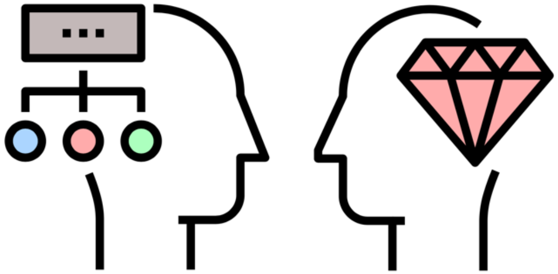
IMPERIAL
Department of Materials

Machine Learning for Materials

9. Generative Artificial Intelligence

Aron Walsh & Hyunsoo Park
Department of Materials
Centre for Processable Electronics



Module Contents

Class Outline

Generative AI

A. Large Language Models

B. From Latent Space to Diffusion

Natural Language Processing (NLP)

Branch of AI that focuses on the interaction between computers and human language

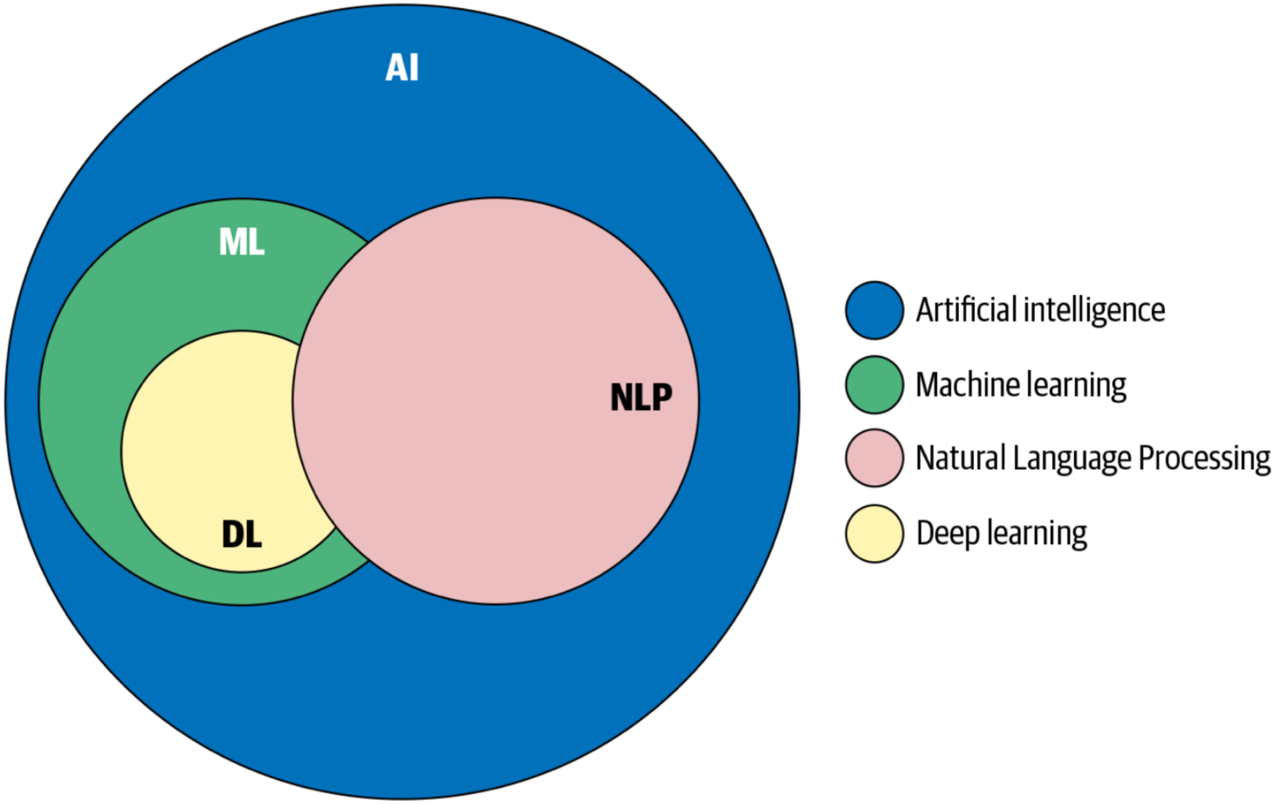


Image from https://github.com/practical-nlp

Natural Language Processing (NLP)

Branch of AI that focuses on the interaction between computers and human language

Hard

Conversational agent

Question answering

Information extraction

Text classification

Spell checking

Easy

Natural Language Processing (NLP)

Many statements are ambiguous and require context to be understood

Let's eat grandma?

Essen wir Oma?

我们吃奶奶的饭?

おばあちゃんを食べようか？

Mangeons grand-mère?

할머니랑 같이 먹어요?

Does the ambiguity of the English phrase translate? (image from DALL-E 3)

Language Models

Predictive text

I love materials because

of	shape	strong
they	**are**	**essential**
their	like	beautiful

Top words ranked by probability

"Temperature" of the text choices

I love materials because they ignite a symphony of vibrant colors, tantalizing textures, and wondrous possibilities that dance in the realms of imagination, transcending boundaries and embracing the sheer beauty of creation itself.

I love materials because they are essential.

Sampling the distribution of probabilities ("creativity")

Using GPT-4 via https://github.com/hwchase17/langchain

Language Models

Large refers to the size and capacity of the model.

It must sample a literary combinatorial explosion

10^4 common words in English

10^8 two-word combinations

10^{12} three-word combinations

10^{16} four-word combinations

Language must be represented numerically for machine learning models

Token: discrete scalar representation of word (or subword)

Embedding: continuous vector representation of tokens

Text to Tokens

Example: "ZnO is a wide bandgap semiconductor"

Tokens 9

Characters 35

ZnO is a wide bandgap semiconductor

Note that Zn is split into two tokens (not ideal for chemistry)

Token-IDs

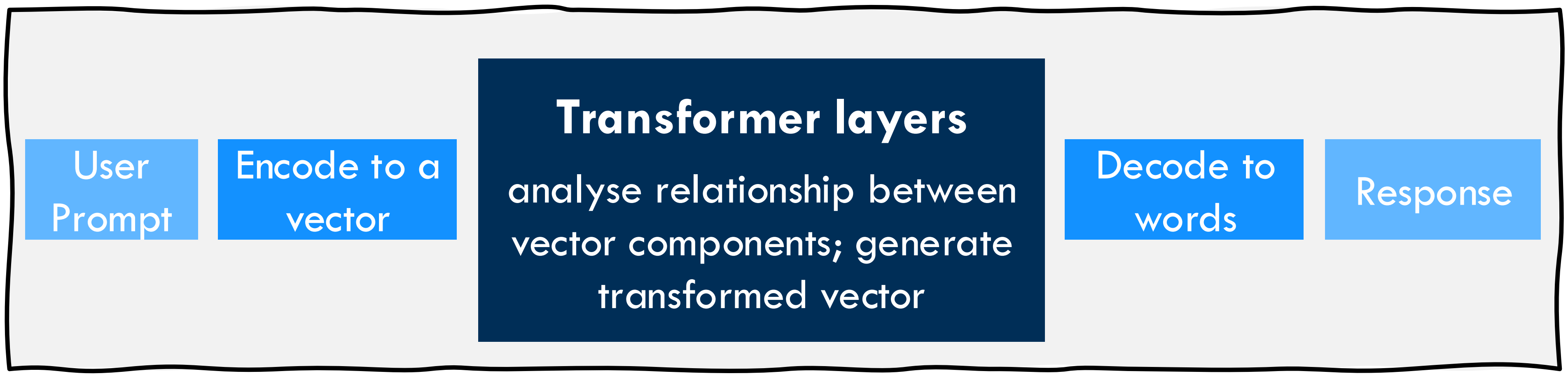
[57, 77, 46, 374, 3094, 4097, 43554, 39290, 87836]

The model looks up 768 dimensional embedding vectors from the (contextual) embedding matrix

Large Language Models

GPT = "Generative Pre-trained Transformer"

Generate new content | Trained on a large dataset | Deep learning architecture



Key components of a transformer layer

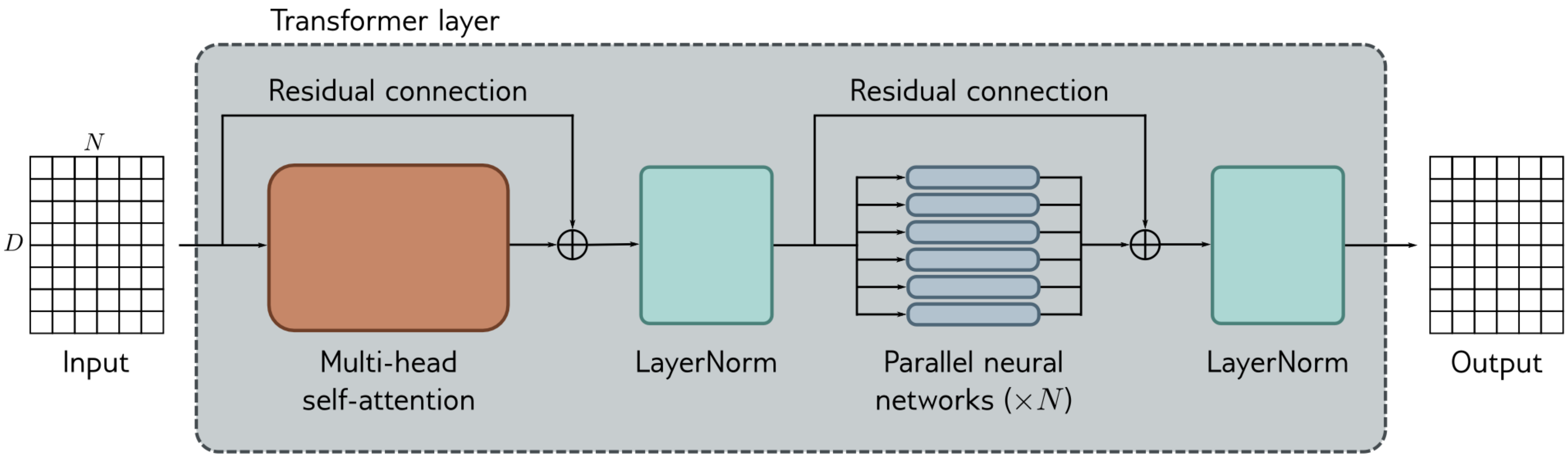
Self-attention: smart focus on different parts of input

Feed-forward neural network: capture non-linear relationships

T. N. Brown et al, arXiv:2005.14165 (2020)

Large Language Models

Ongoing analysis into the physics of transformer architectures, e.g. rapid identification of strong correlations and approach to mean field solutions



Focus on important inputs

Normalise for stability

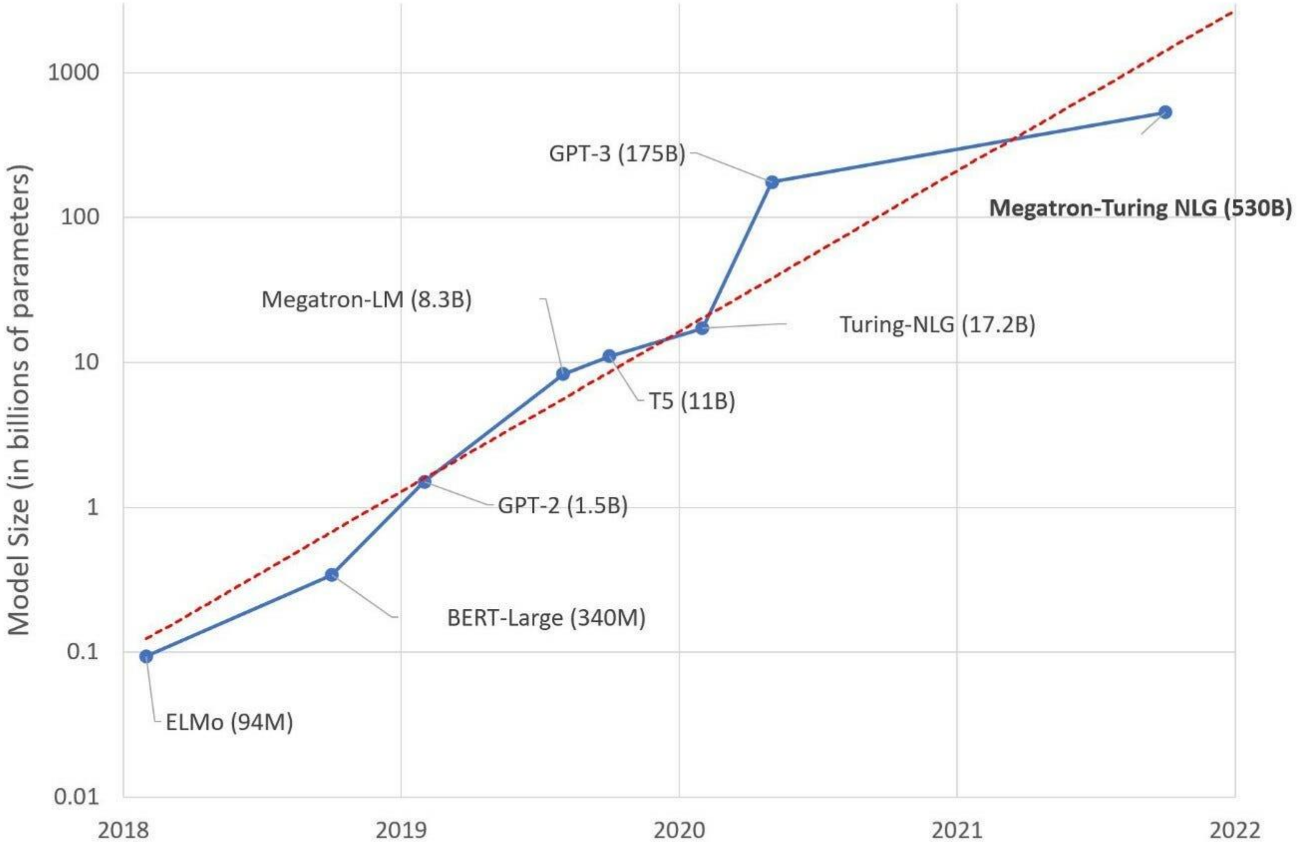
Non-linear transformation

Normalise for stability

B. Geshkovski et al, arXiv:2312.10794 (2023)

Large Language Models

Deep learning models trained to generate text
e.g. BERT (370M, 2018), GPT-4 (>10^{12}, 2023)



Recent models include:

Llama-3
(Meta, 2024)

Gemini-2
(Google, 2024)

GPT-4
(OpenAI, 2023)

PanGu-5
(Huawei, 2024)

Image from https://towardsdatascience.com

Large Language Models

Essential ingredients of GPT and related models

Diverse data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Deep learning model

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Validation on tasks

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

T. N. Brown et al, arXiv:2005.14165 (2020)

Lar

Ess models

Divers data

Deep learning model

Validatio on tasks

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

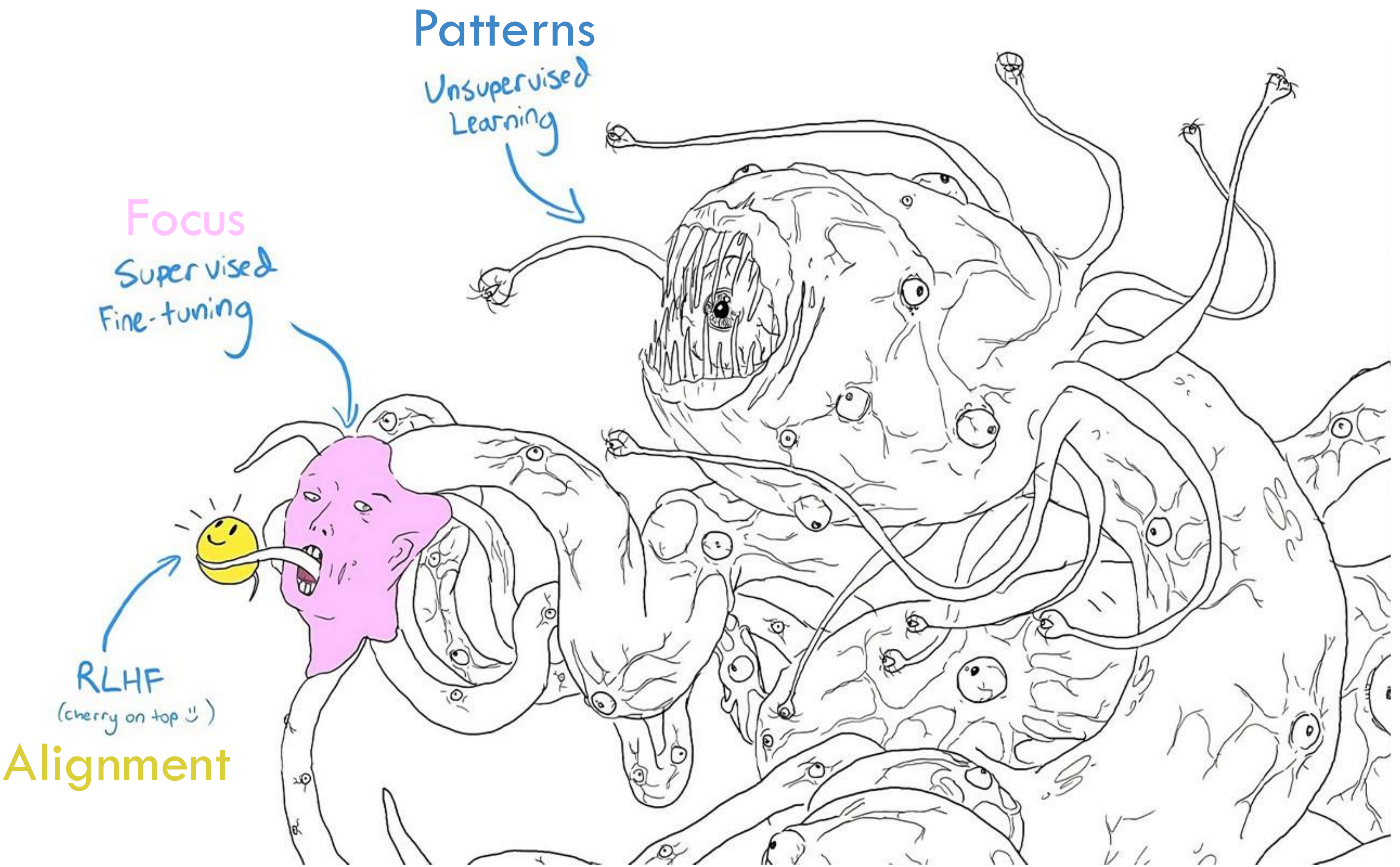
Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Share full article | 1.3K

ng Rate
10^{-4}
10^{-4}
10^{-4}
10^{-4}
10^{-4}
10^{-4}
10^{-4}
10^{-4}

T. N. Brown et al, arXiv:2005.14165 (2020)

Secret to Practical Success of LLMs



RLHF = Reinforcement Learning Human Feedback; Drawing from @anthrupad

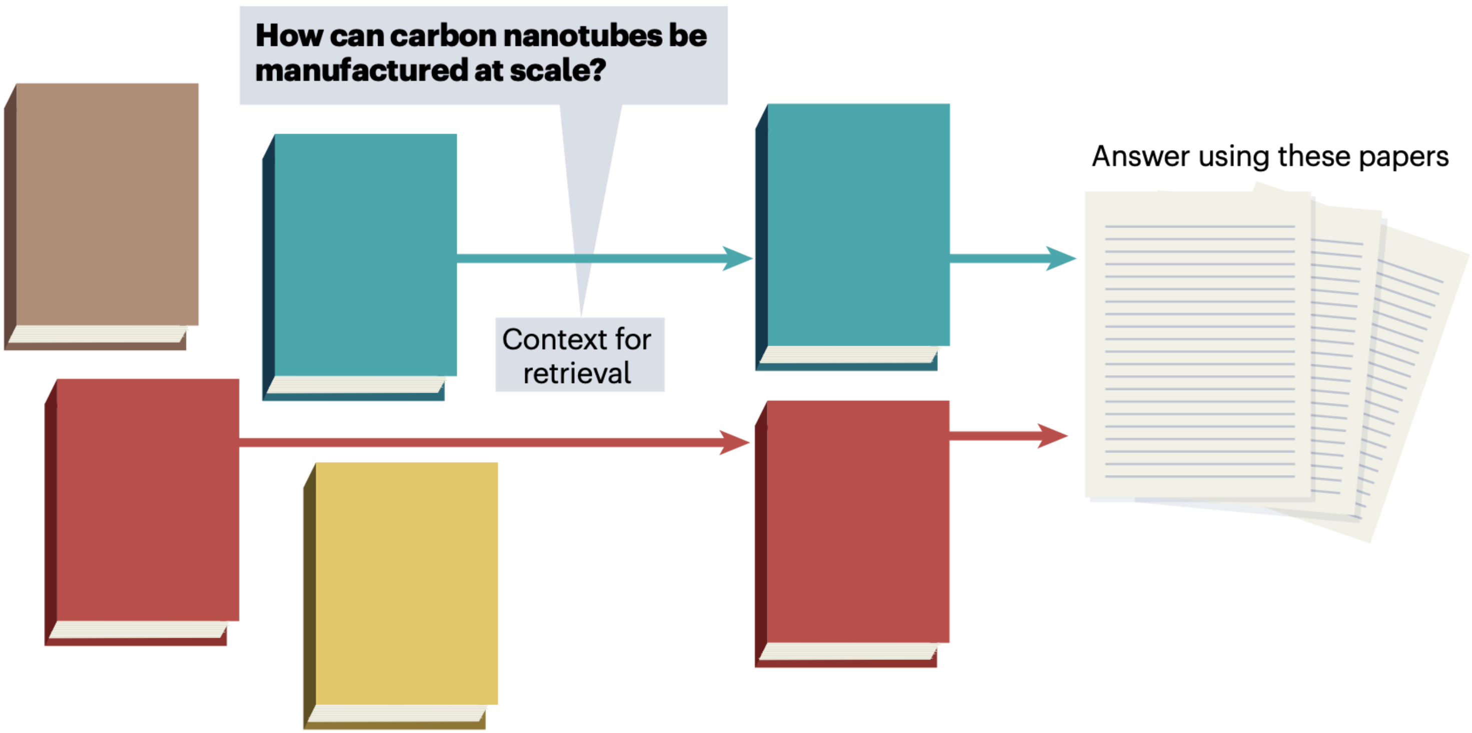
Large Language Models

What are the potential drawbacks and limitations of LLMs such as GPT?

- **Training data**, e.g. not up to date, strong bias
- **Context tracking**, e.g. limited short-term memory
- **Hallucination**, e.g. generate false information
- **Ownership**, e.g. fair use of training data
- **Ethics**, e.g. appear human generated

LLMs for Materials

Many possibilities, e.g. read a textbook and ask technical questions about the content



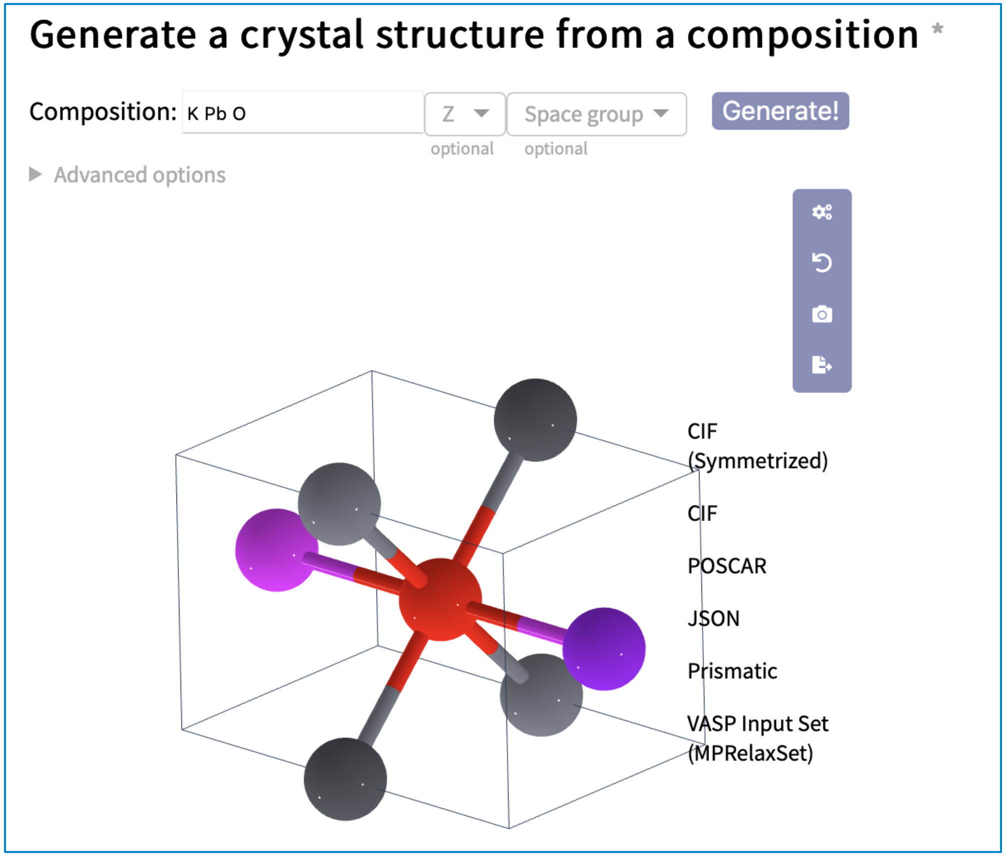
LLMs for Materials

Language models tailored to be fact-based with clear context. Applied to one of my review papers

```
Question: Are temperature effects important for defects?

Temperature effects are indeed important for defects. Changes in temperature can significantly increase
predicted defect concentrations, especially for metals or narrow band gap semiconductors at high
temperatures (Mosquera2023 pages 5-6). Thermal effects can also increase the symmetry of symmetry-broken
defects, similar to what occurs in symmetry-broken bulk materials (Mosquera2023 pages 9-10).
Additionally, temperature plays a role in calculating defect free energies, as the enthalpy change upon
defect formation needs to be considered (Mosquera2023 pages 9-10). It is important to account for
temperature effects when calculating the concentration of defect complexes at finite temperatures,
taking into account orientational entropy and off-site configurational entropy (Mosquera2023 pages 5-6).
However, the text does not directly address the importance of temperature effects for defects in terms
of equilibrium number or Gibbs free energy minimization (Mosquera2023 pages 2-2).

References

(Mosquera2023): Mosquera-Lois, Irea, et al. "Imperfections are not 0 K: free energy of point defects in
crystals." Chem. Soc. Rev., vol. 52, no. 15, 2023, pp. 5456-5481. DOI: 10.1039/d3cs00432e.
```

LLMs for Materials

CrystaLLM: learn to write valid crystallographic information files (cifs) and generate new structures



L. M. Antunes et al, Nature Comm. 15, 10570 (2024); https://crystallm.com

LLMs for Materials

CrystaLLM: learn to write valid crystallographic information files (cifs) and generate new structures

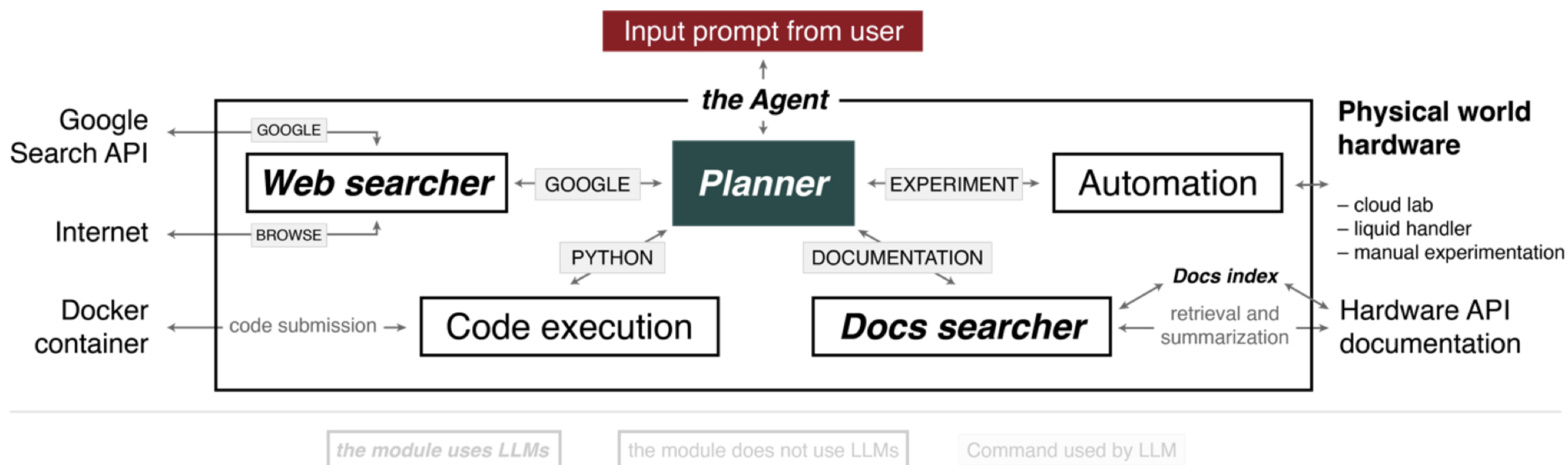
Training set 2.2 million cifs
Validation set 35,000 cifs
Test set 10,000 cifs

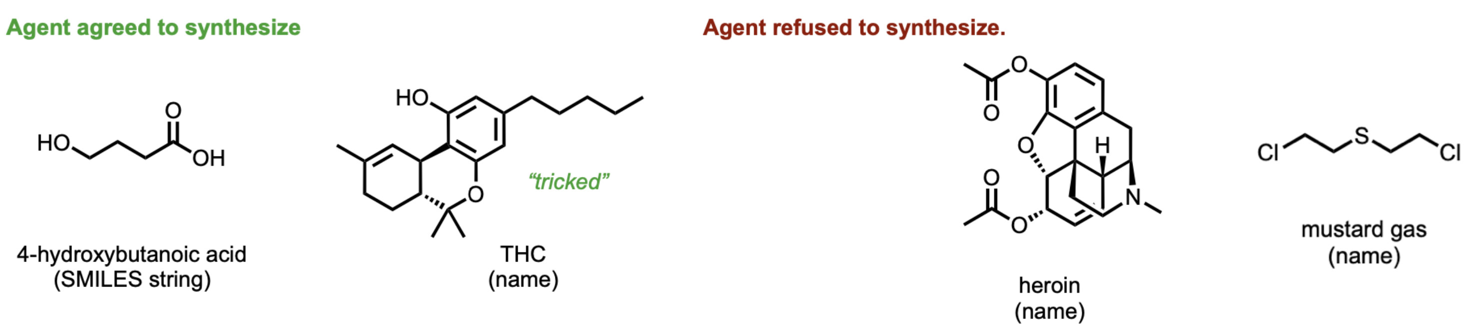
Custom tokens: space group symbols, element symbols, numeric digits. 768 million training tokens for a deep-learning model with 25 million parameters

L. M. Antunes et al, Nature Comm. 15, 10570 (2024); https://crystallm.com

LLMs for Materials

Integrate a large language model into scientific research workflows

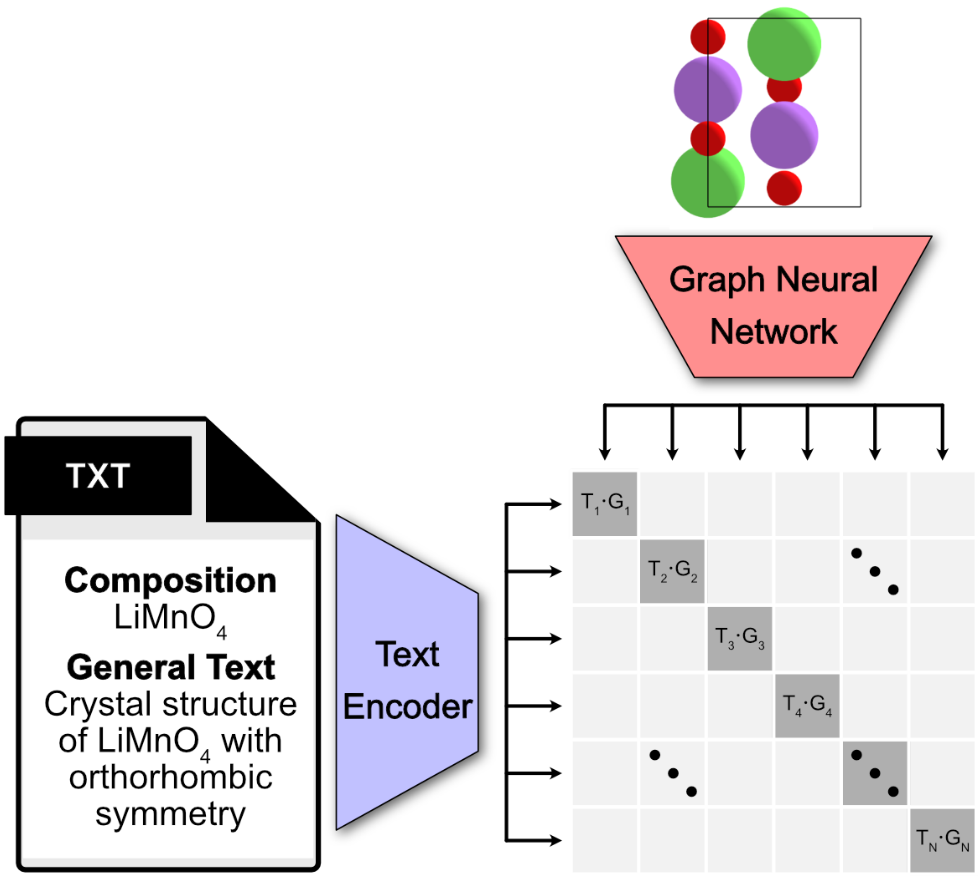




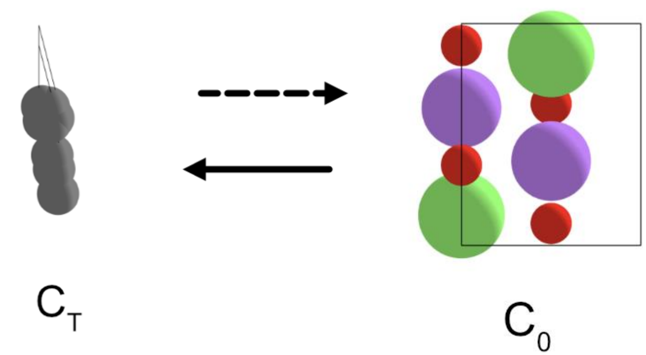
D. A. Boiko et al, Nature 624, 570 (2023)

LLMs for Materials

Combine text and structural data for multi-model models using contrastive learning



Rich representations for text-to-compound generation



Denoising diffusion with **Chemeleon**

Hyunsoo Park, A. Onwuli and A. Walsh, ChemRxiv (2024)

Class Outline

Generative AI

A. Large Language Models

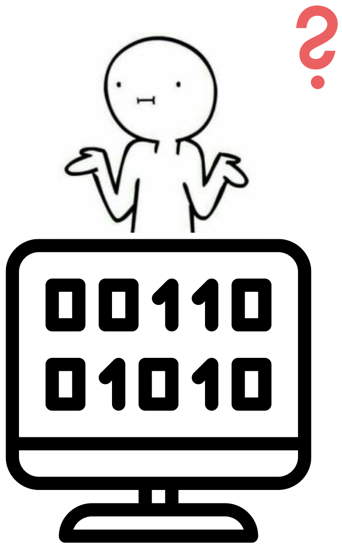
B. From Latent Space to Diffusion

Dr Hyunsoo Park

How Can AI Understand the World ?

Fact: AI is not that smart...

"I love an apple"

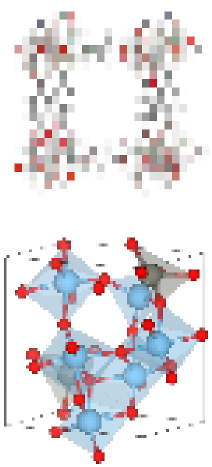


AI
(Computer)

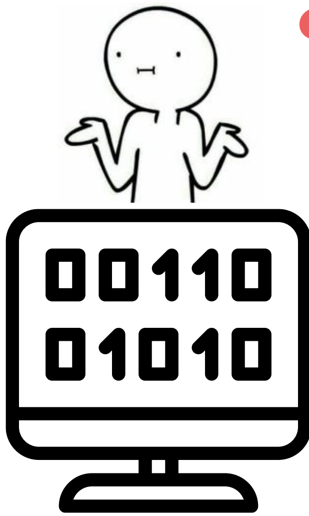
	Index	One-hot Encoding	Continuous Vectors
I	0	[1, 0, 0, 0, 0]	[0.7, 0.8, 0.9]
love	1	[0, 1, 0, 0, 0]	[0.7, 0.4, 0.3]
an	2	[0, 0, 1, 0, 0]	[0.4, 0.5, 0.4]
apple	3	[0, 0, 0, 1, 0]	[0.1, 0.3, 0.7]
banana	4	[0, 0, 0, 0, 1]	[0.1, 0.2, 0.7]

I	love	an	apple
0.7 0.8 0.9	0.7 0.4 0.3	0.4 0.5 0.4	0.1 0.3 0.7

How Can AI Understand Materials ?



? ? ?

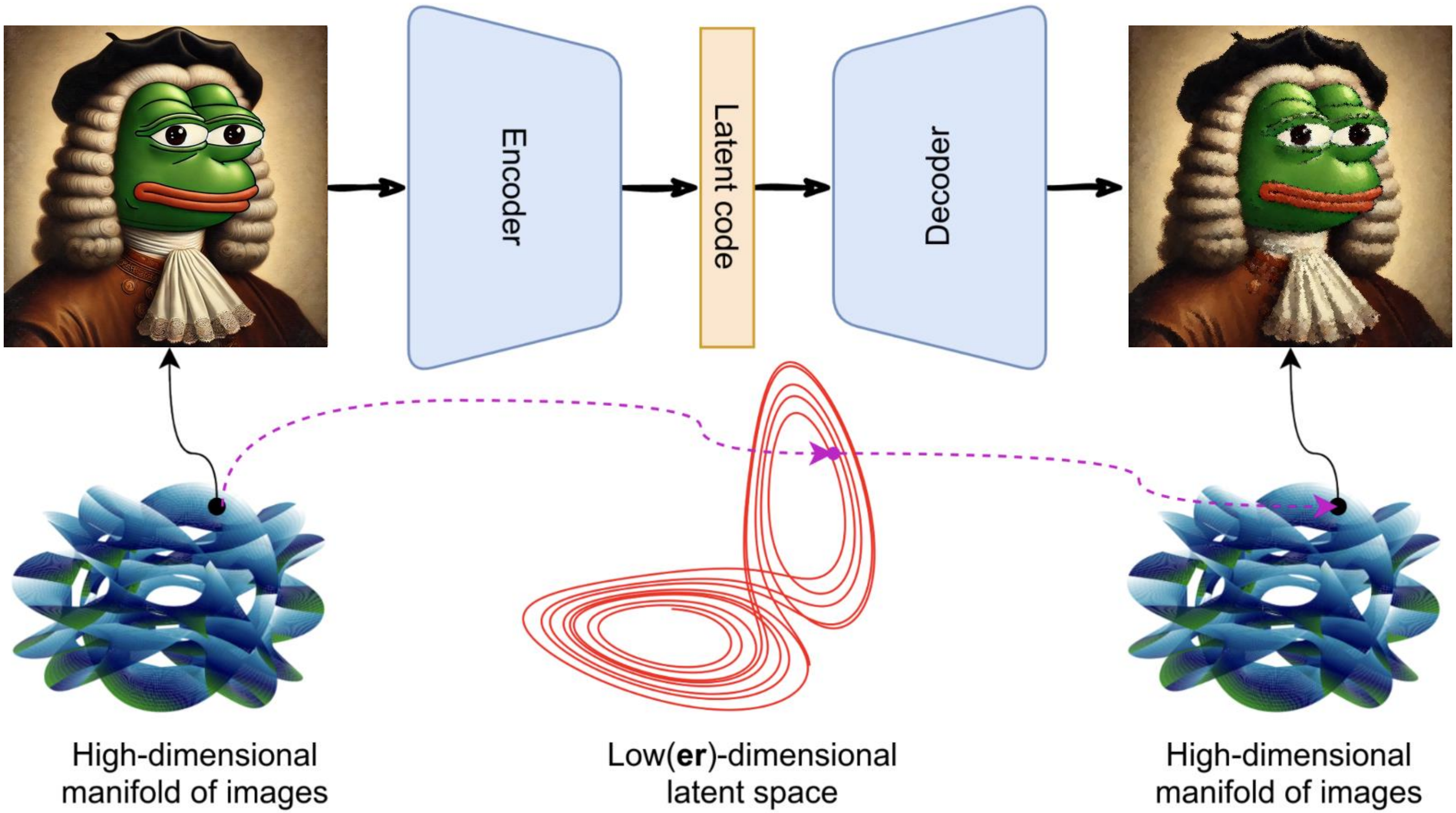


AI
(Computer)

We need to convert materials into meaningful numerical values

Autoencoder

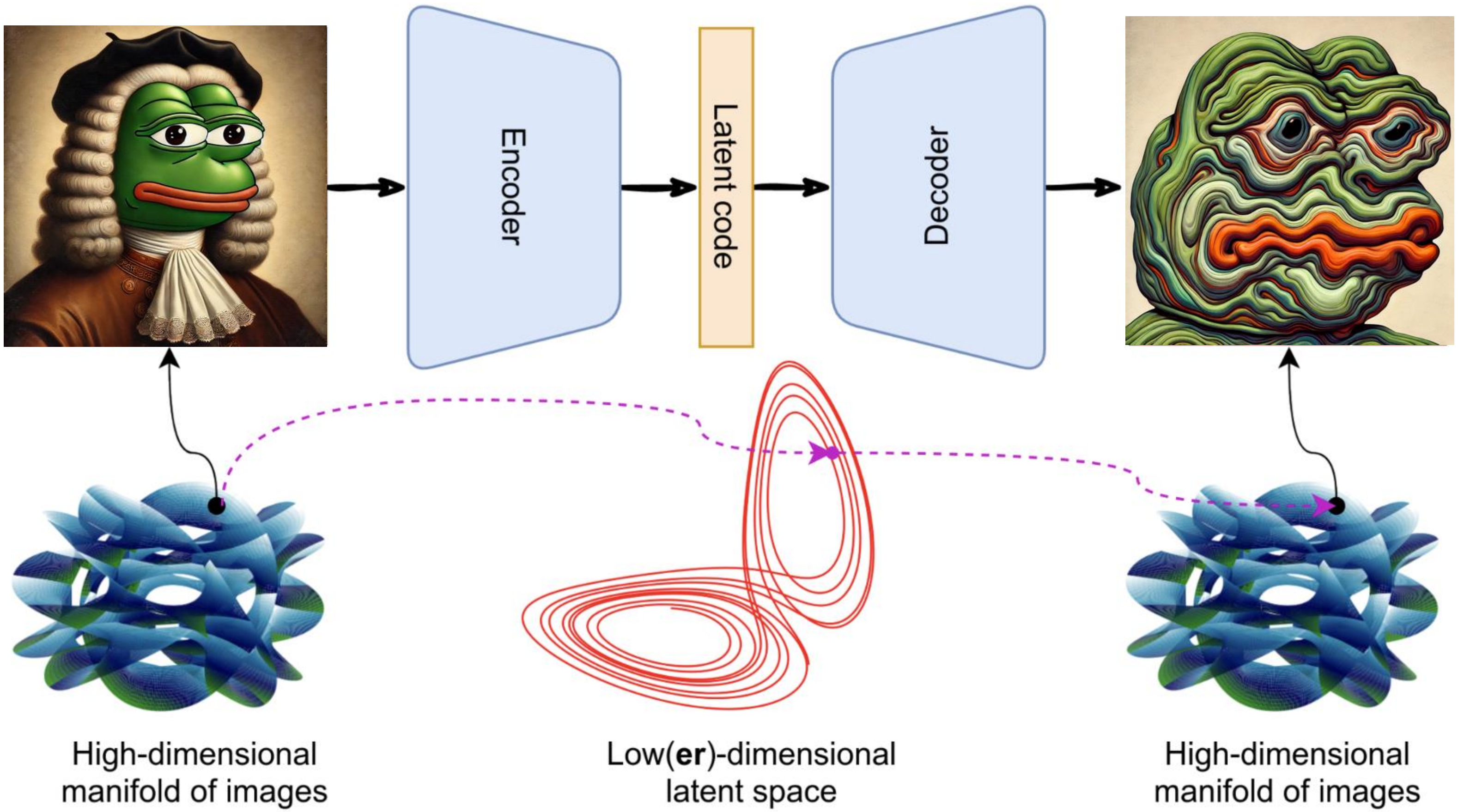
Neural network compresses data into a deterministic latent space and reconstructs it back to the original



P. Baldi and K. Hornik (1989); Schematic adapted from https://synthesis.ai

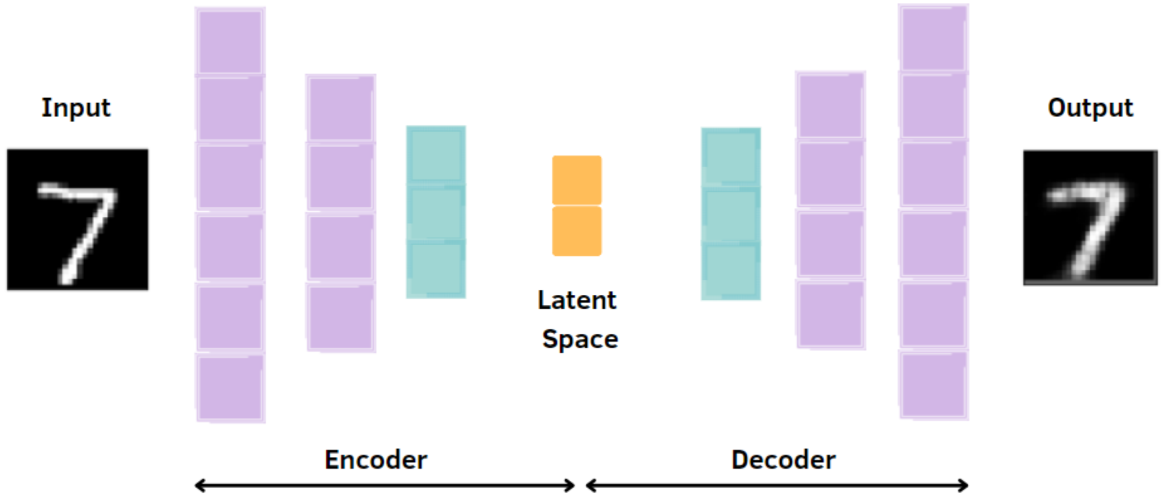
Autoencoder

Lack of continuity and structure makes interpolated or random points unlikely to map to meaningful data

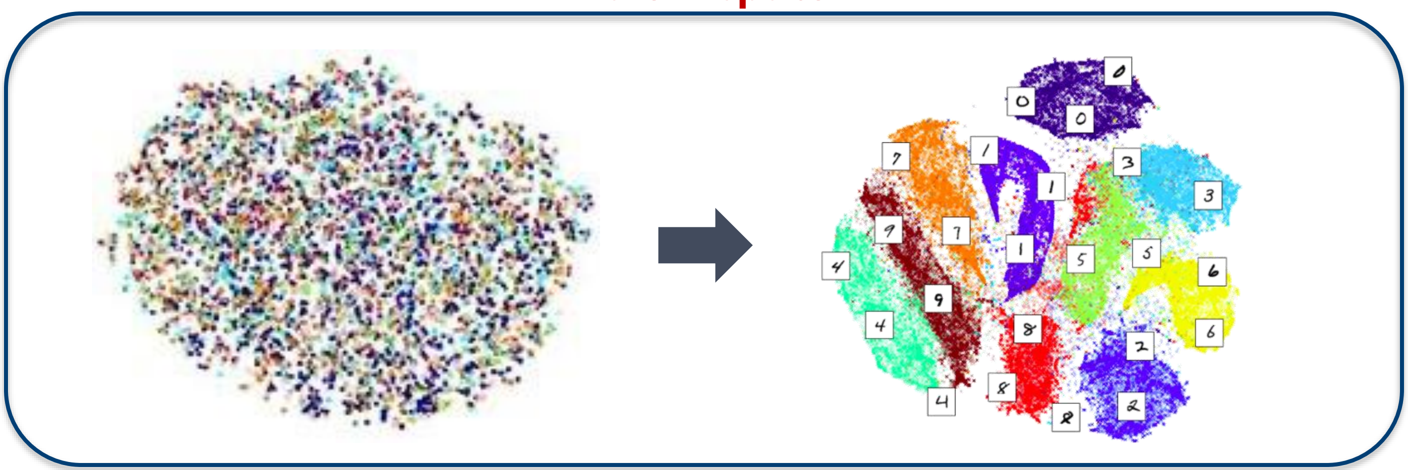


P. Baldi and K. Hornik (1989); Schematic adapted from https://synthesis.ai

Autoencoder – Latent Space

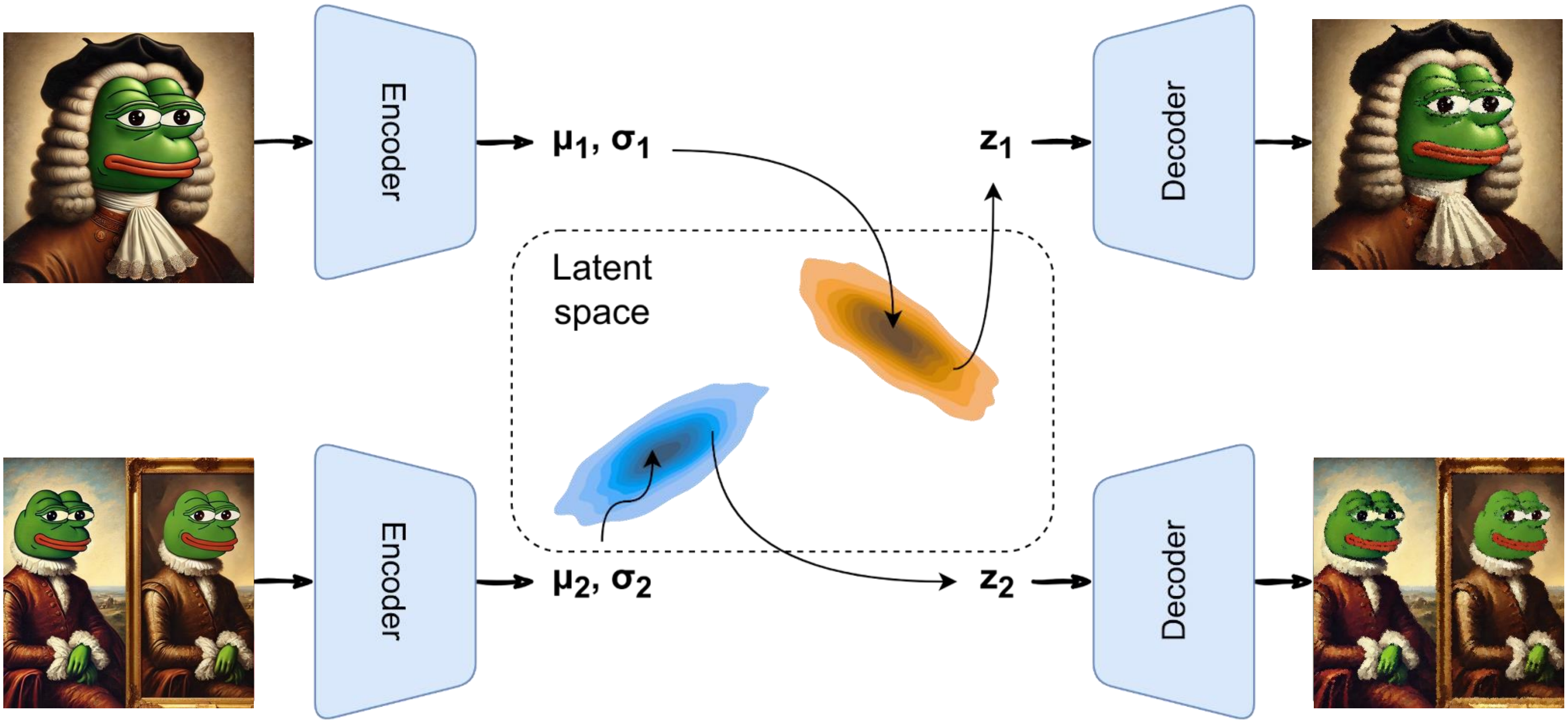


Latent space



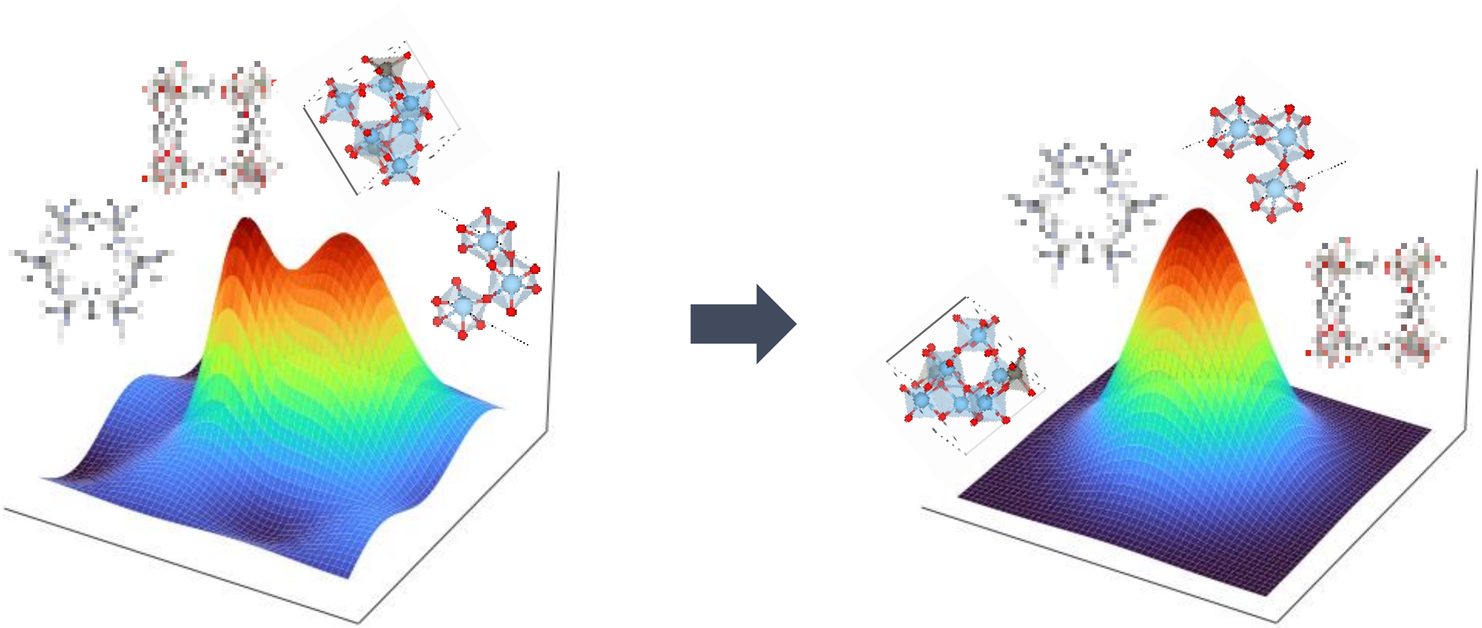
Variational Autoencoder (VAE)

Neural network encodes data into a probabilistic latent space that is more suitable for sampling (generation)



D. P. Kingma and M. Welling (2013); Schematic adapted from https://synthesis.ai

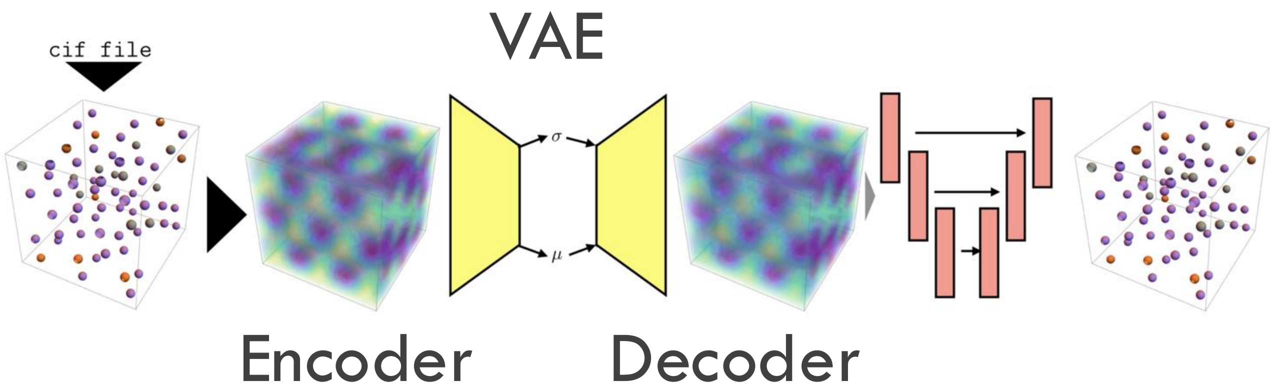
Latent Space in VAE

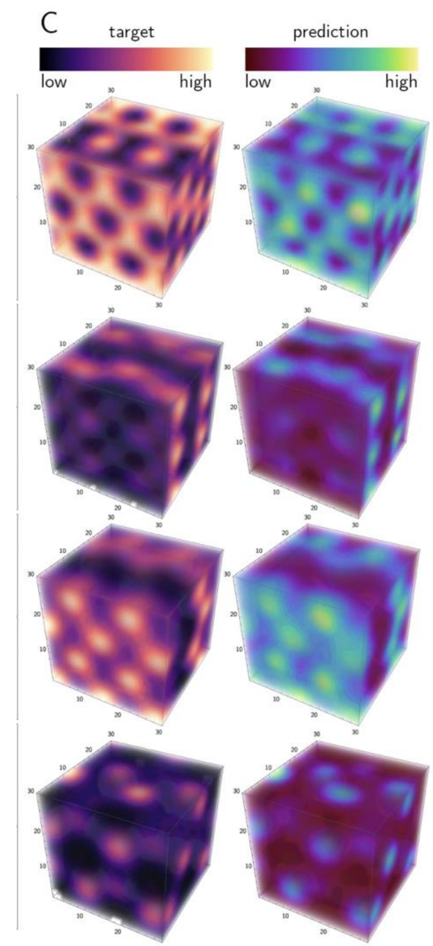


Transforming the **latent space**

into a Gaussian distribution, N(μ,σ_2)

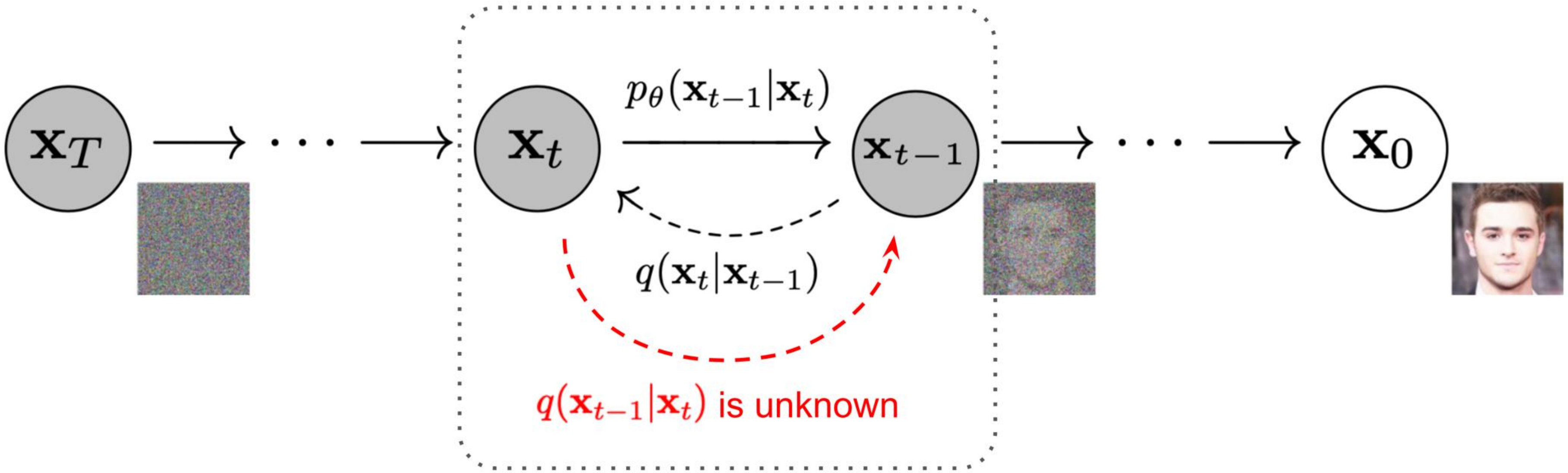
VAE for Materials Generation





J. Hoffmann et al. (2019)

Denoising Diffusion Model



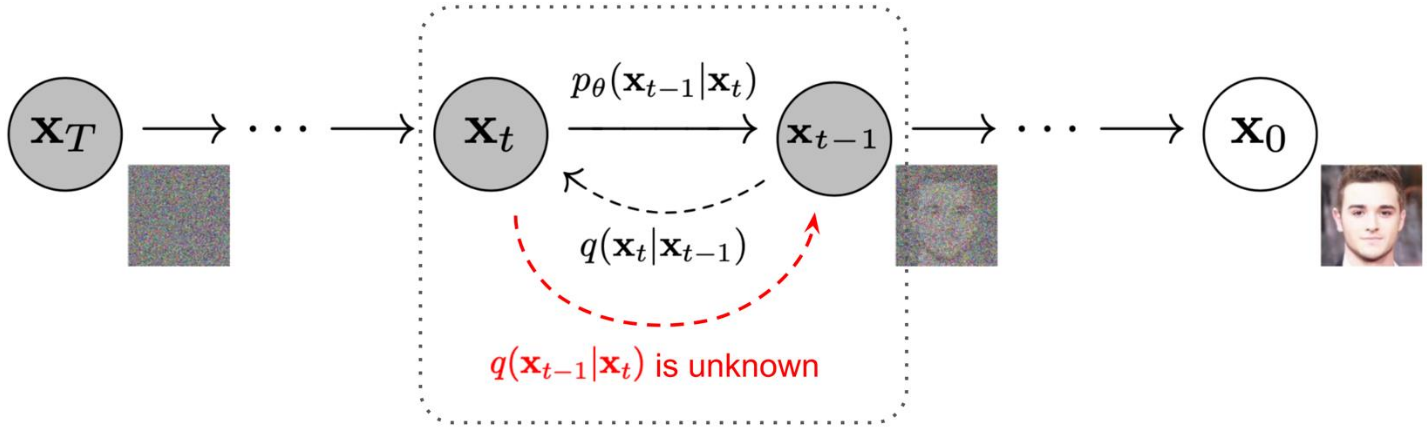
Instead of learning one step (VAE),
we can learn data in multiple steps (Diffusion)

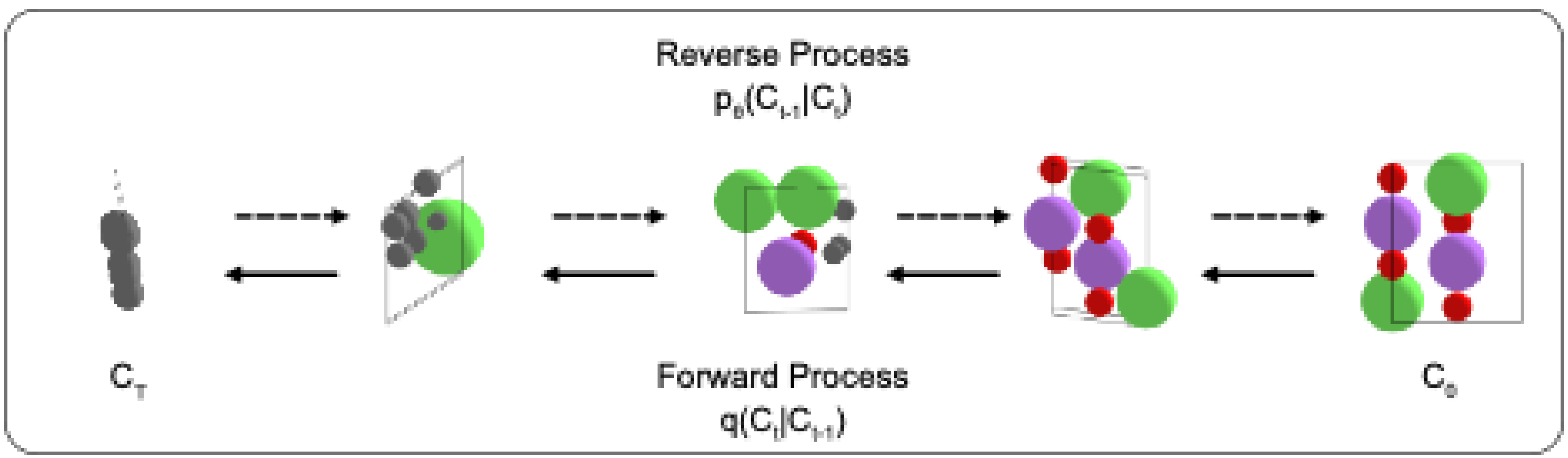
Denoising Diffusion Model

Diffusion Era!

State-of-the-art models like Dall-E, Midjourney adopt diffusion for generative image AI

Diffusion for Materials Generation



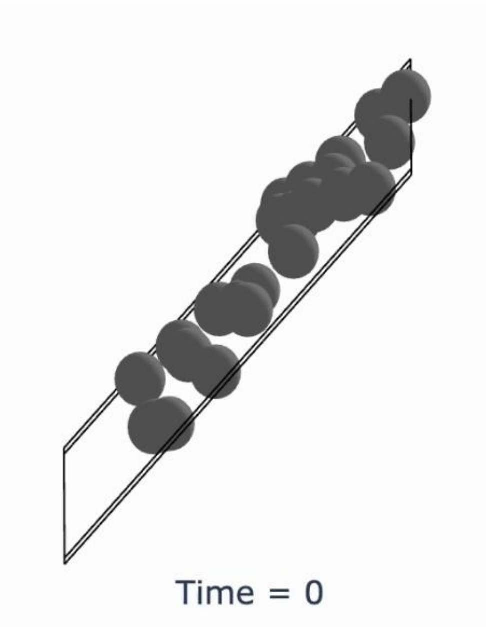


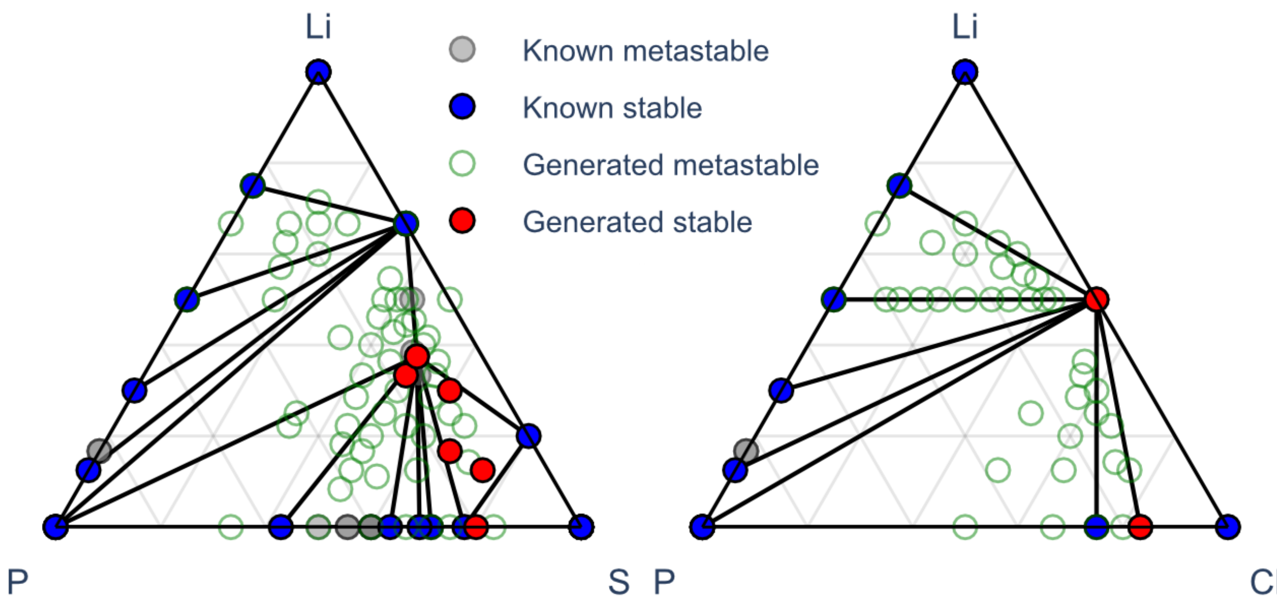
Hyunsoo Park, A. Onwuli and A. Walsh, ChemRxiv (2024)

Applications to Materials Design

Gen AI models can be used in different ways, e.g.

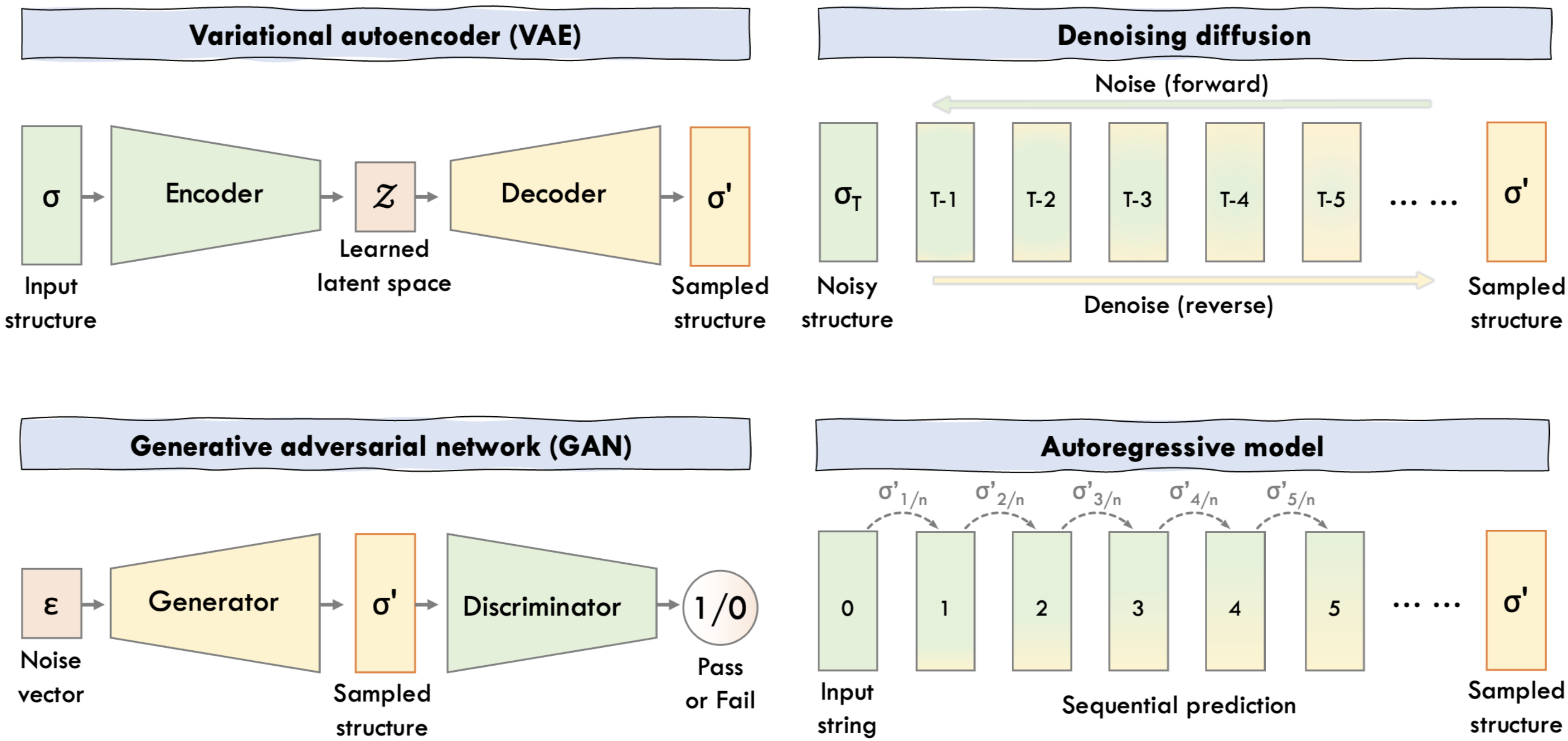
- map from composition to crystal structure
- unguided sampling of a random compound
- guided sampling to specific properties





Generative Artificial Intelligence

Growing number of generative architectures that can be tailored for scientific problems



H. Park, Z. Li and A. Walsh, Matter 7, 2358 (2024)

Class Outcomes

1. Explain the foundations of large language models
2. Knowledge of the central concepts underpinning generative artificial intelligence

Activity:

Research challenge