

# GENERATIVE AI FOR DRUG DISCOVERY: DREAMING UP NEW MOLECULES

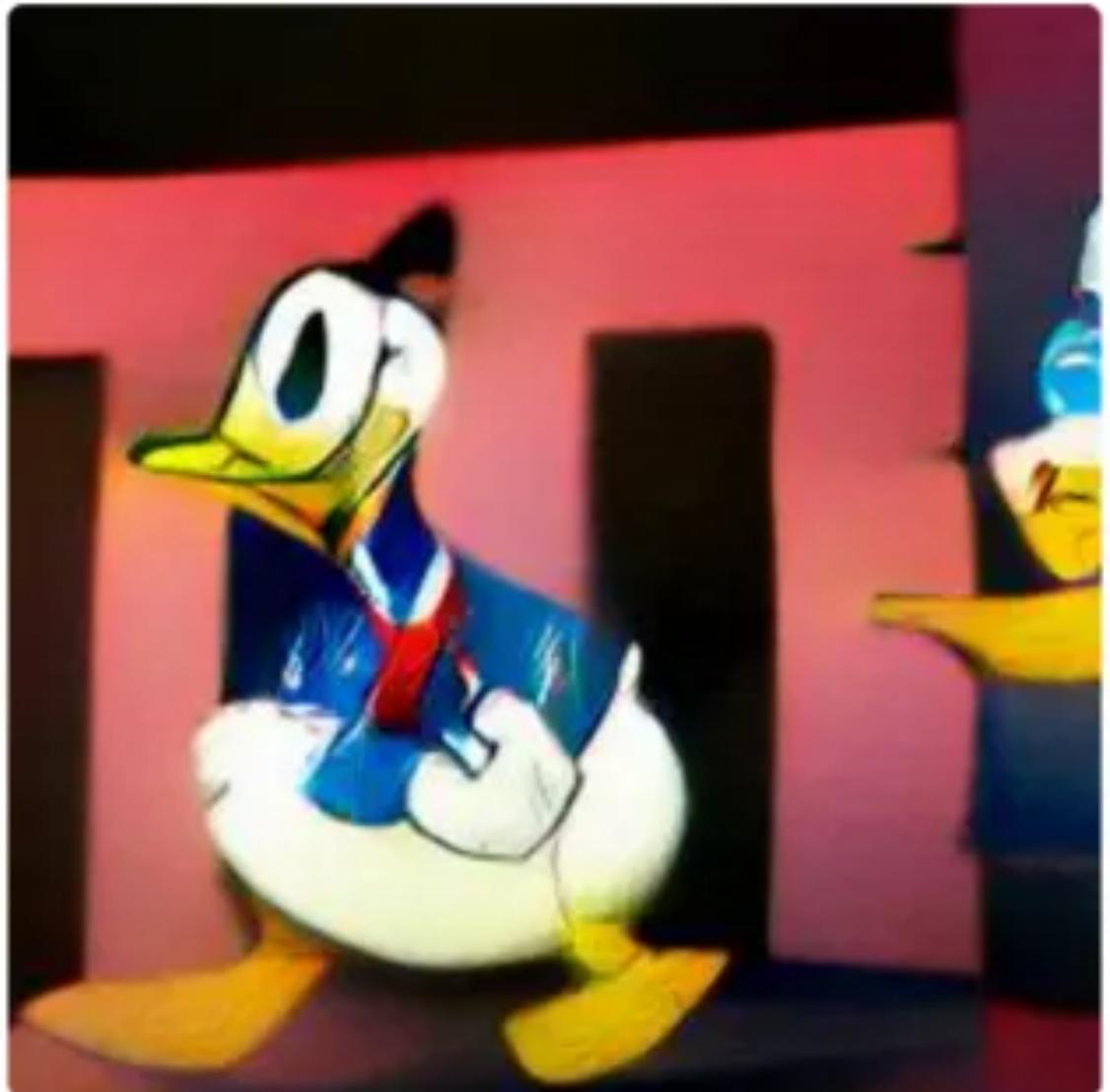
Gerrit Großmann

09.11.2024

[gerritgrossmann.de/#cerfa](http://gerritgrossmann.de/#cerfa)

# Motivation

2022



<https://www.sciencefocus.com/news/dall-e-mini-creator-explains-blurred-faces-going-viral-and-the-future-of-the-project>

2023



Jeroen Pixel  
@pixelprotest\_

I used Midjourney v6 to rebuild famous artworks with lego blocks. The results are chunky af.

"The Girl with a Pearl Earring" by Johannes Vermeer

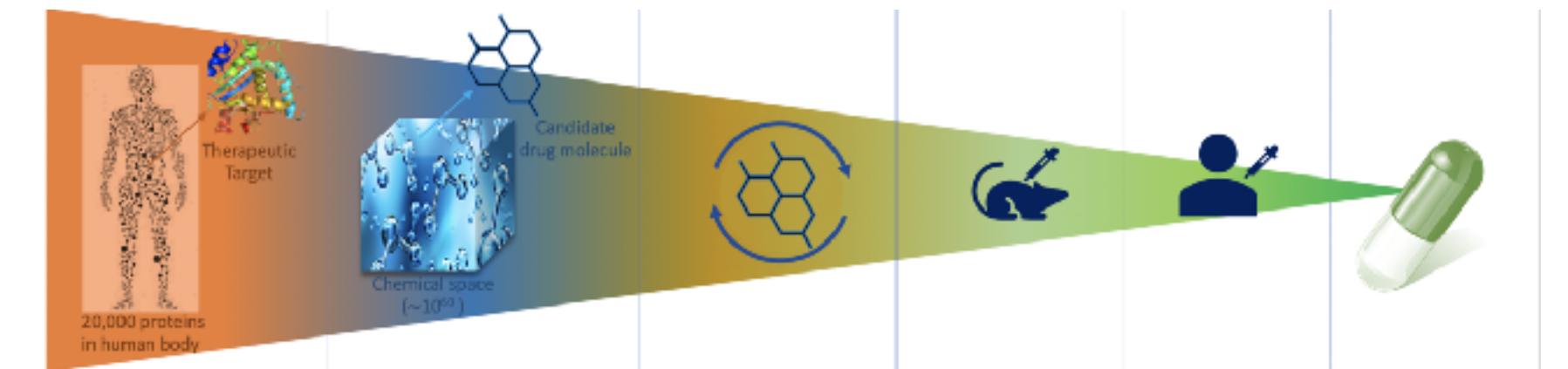


[youtube.com/watch?v=TU1gMl0k&ab\\_channel=MagnaAI](https://youtube.com/watch?v=TU1gMl0k&ab_channel=MagnaAI)

2024

# Motivation

“What works for images should also work for medicine!”



[linkedin.com/pulse/ai-drug-discovery-vijay-morampudi](https://www.linkedin.com/pulse/ai-drug-discovery-vijay-morampudi)

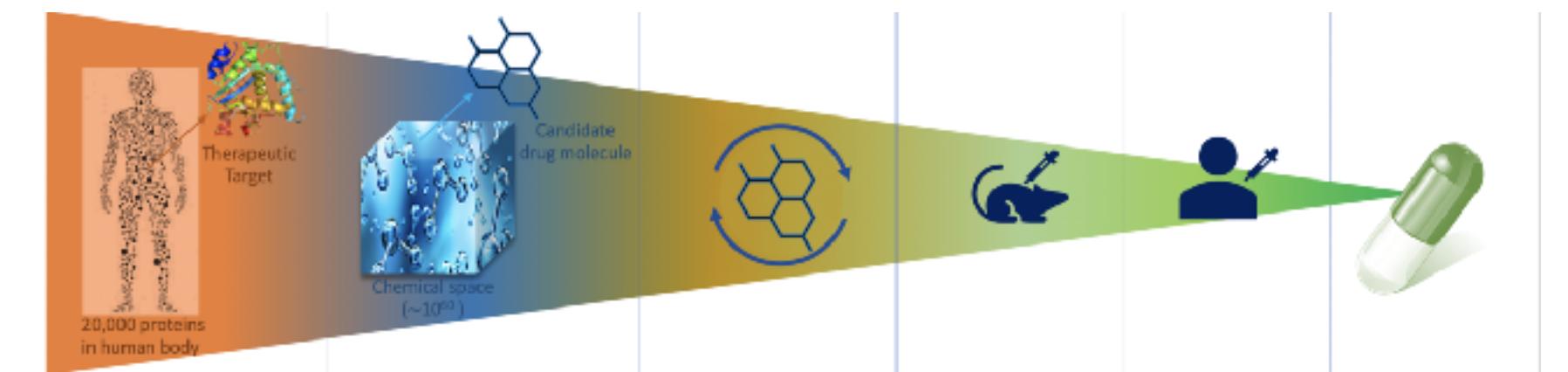
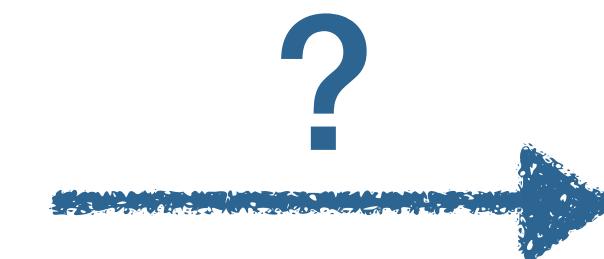
Jeroen Pixel [twitter.com/pixelprotest\\_/status/1743227361888186773](https://twitter.com/pixelprotest_/status/1743227361888186773)

# Motivation

“What works for images should also work for medicine!”



“Painting images and making drugs is not the same thing!”



[linkedin.com/pulse/ai-drug-discovery-vijay-morampudi](https://www.linkedin.com/pulse/ai-drug-discovery-vijay-morampudi)

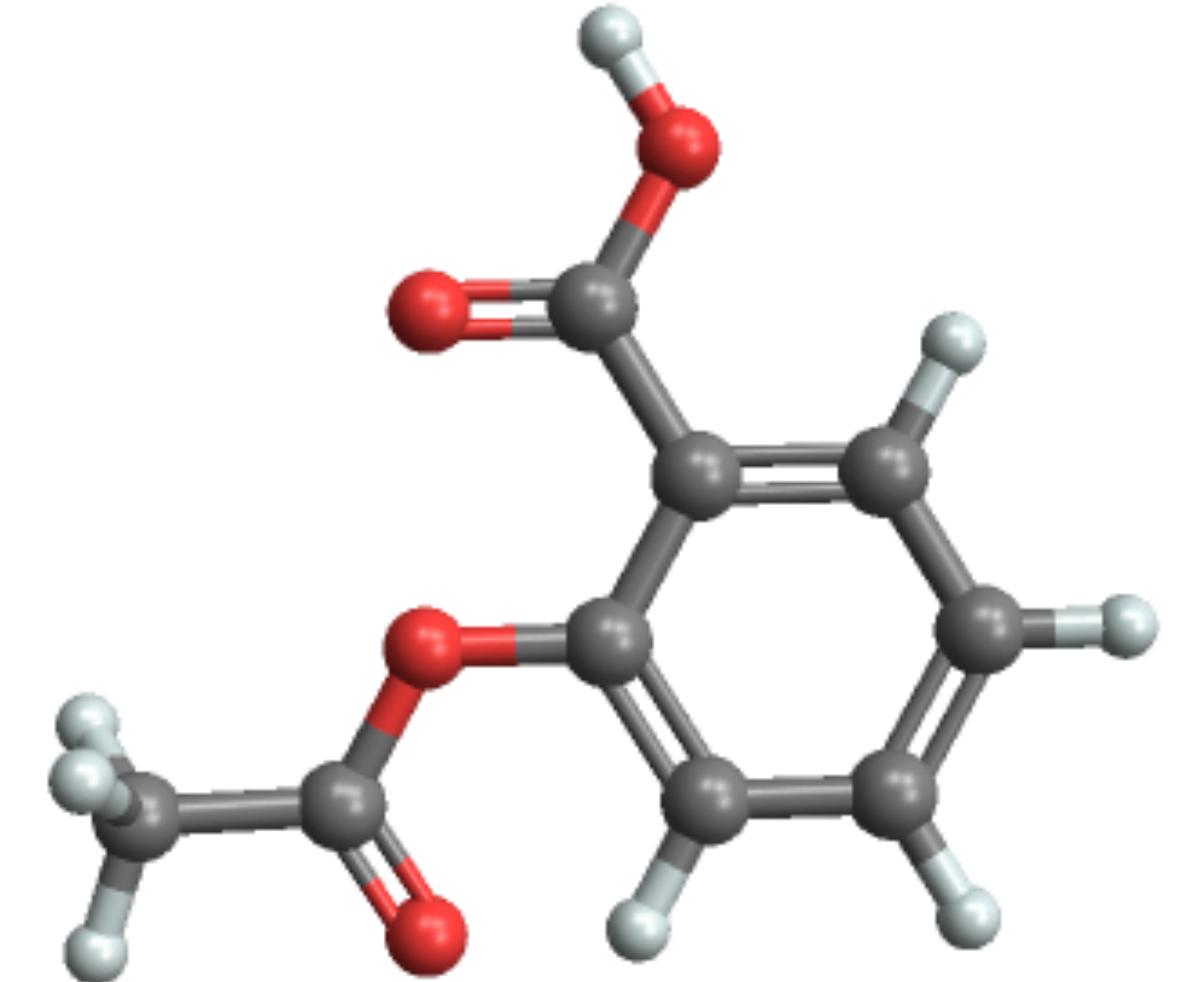
Jeroen Pixel [twitter.com/pixelprotest\\_/status/1743227361888186773](https://twitter.com/pixelprotest_/status/1743227361888186773)

# Motivation

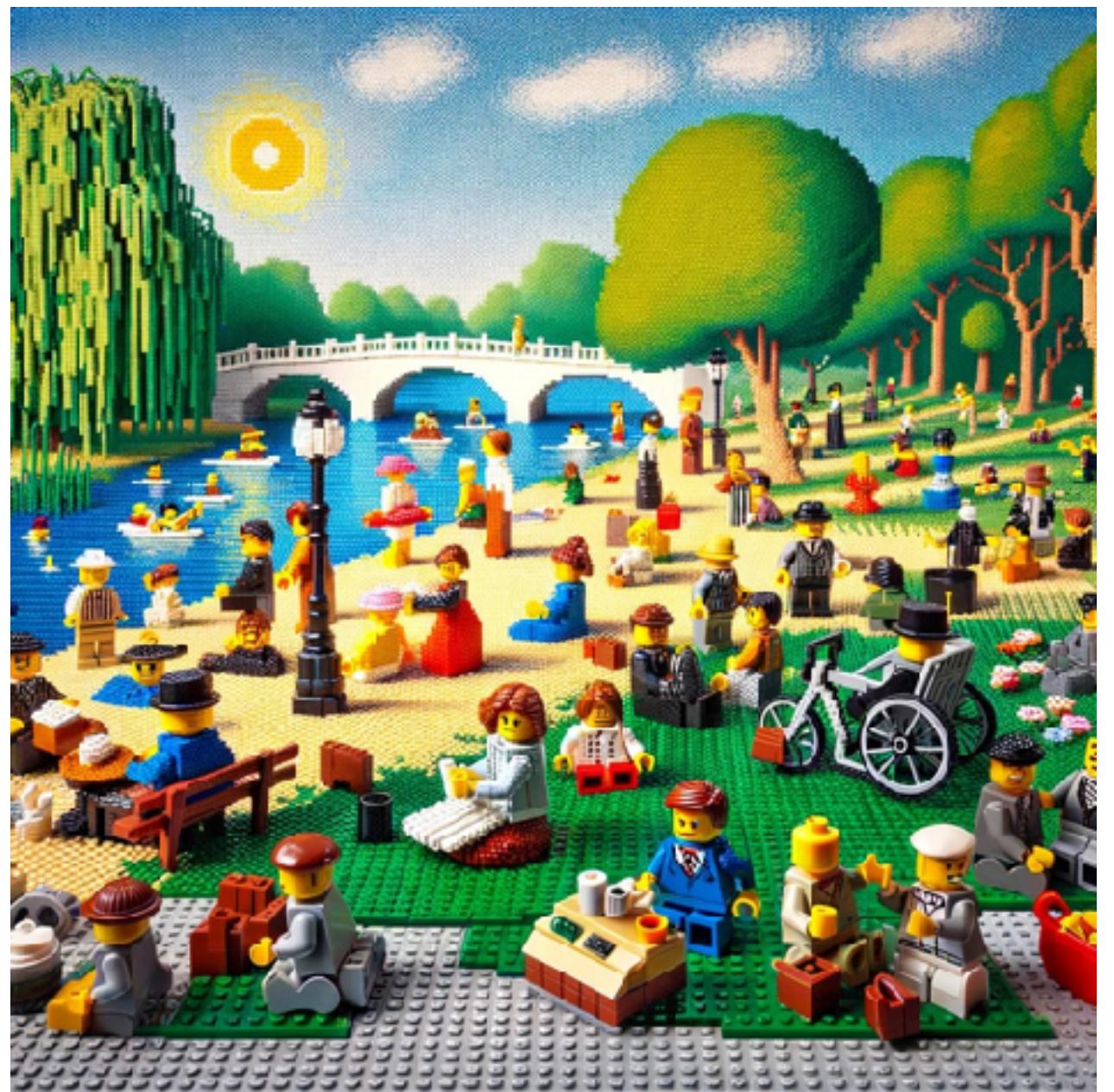


Jeroen Pixel [twitter.com/pixelprotest\\_/status/1743227361888186773](https://twitter.com/pixelprotest_/status/1743227361888186773)

- High Dimensionality
  - Large amount of data
  - Complex manifold
  - World model
  - Conditional generation
  - Ethically dubious potential
- 
- Domain knowledge
  - 3D objects
  - Evaluation
  - Discrete variables

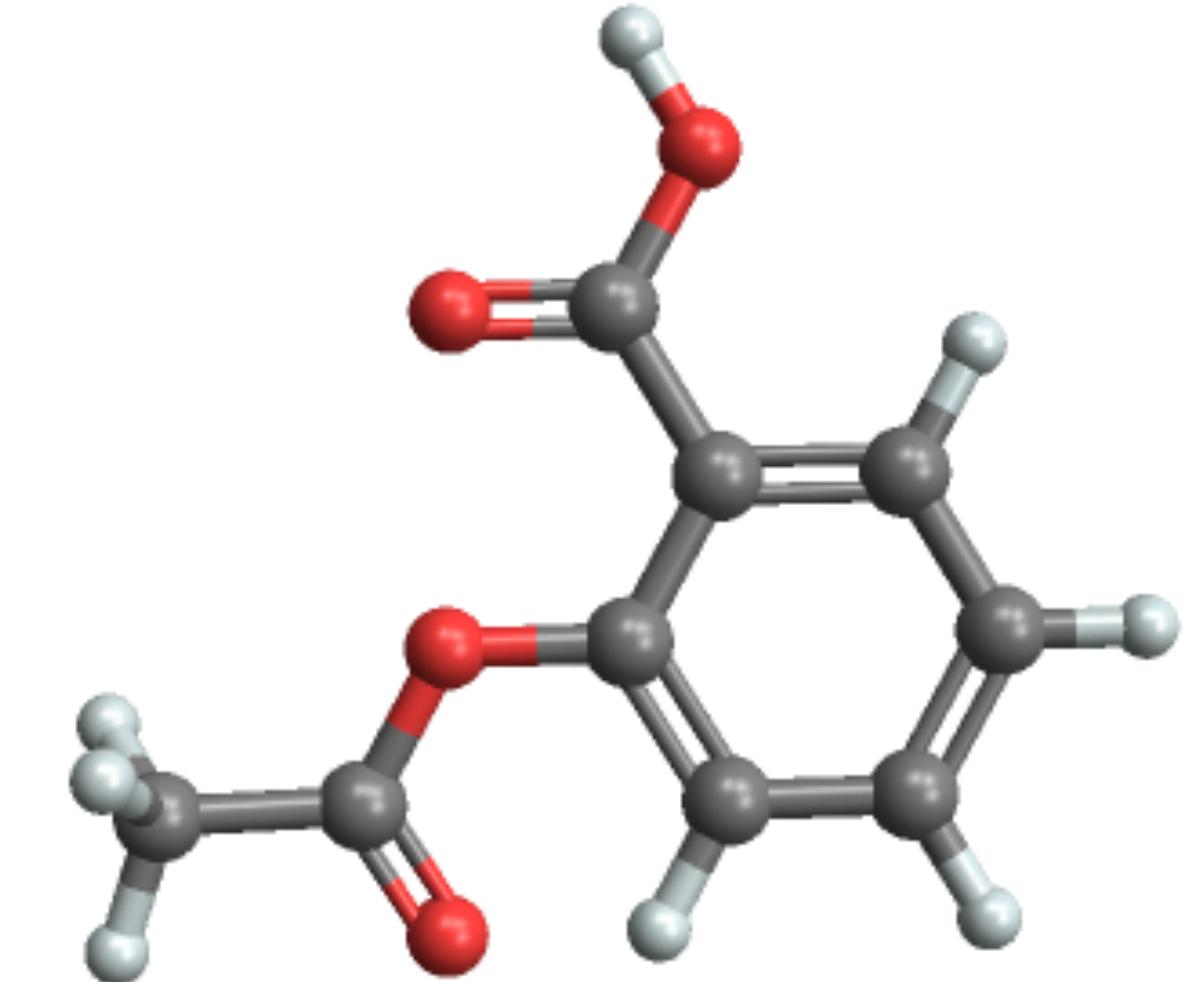


# Motivation



Jeroen Pixel [twitter.com/pixelprotest\\_/status/1743227361888186773](https://twitter.com/pixelprotest_/status/1743227361888186773)

- High Dimensionality
- Large amount of data
- Complex manifold
- World model
- Conditional generation
- Ethically dubious potential
- Domain knowledge
- 3D objects
- Evaluation
- Discrete variables



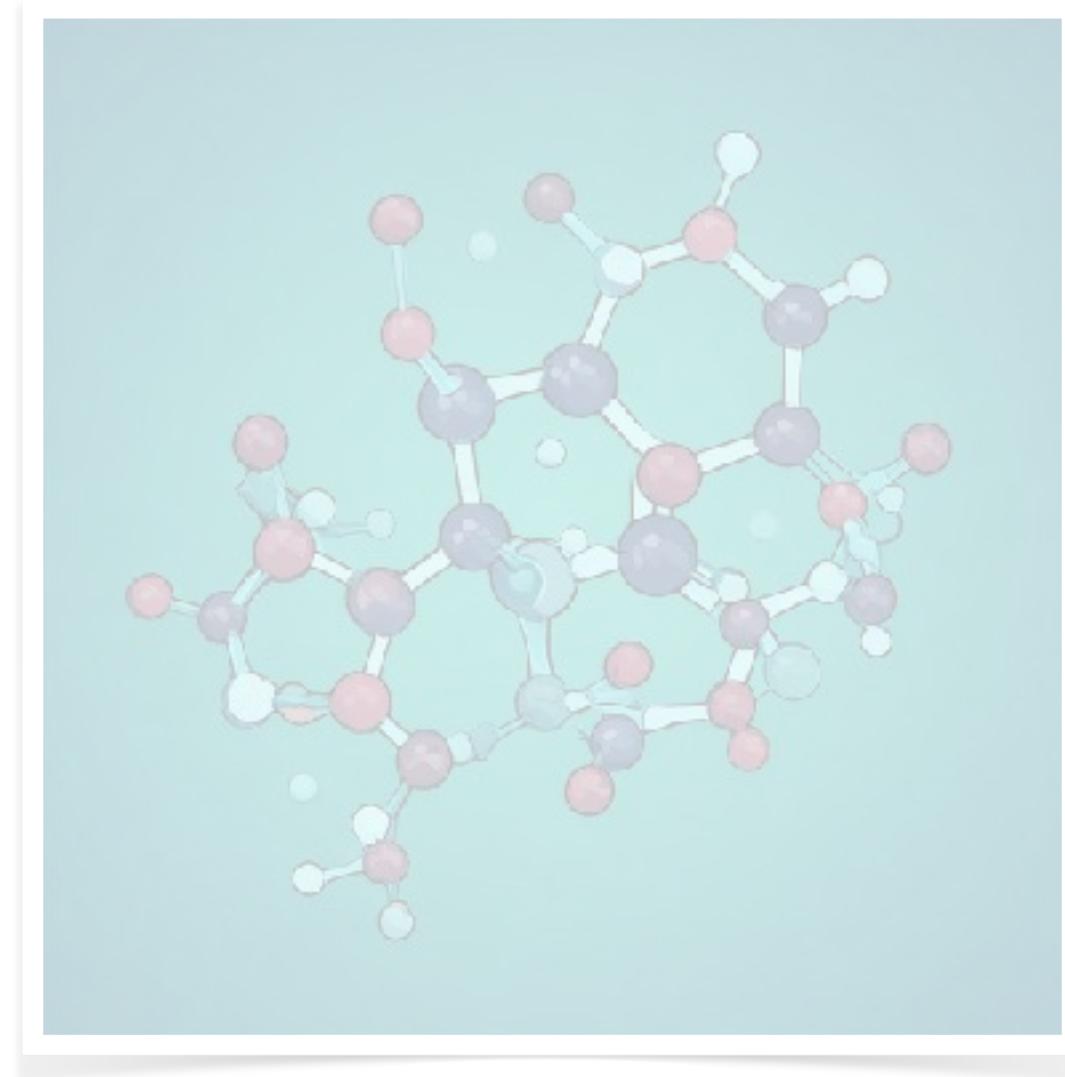
# Outline



Drug discovery in a nutshell



Foundations of diffusion models

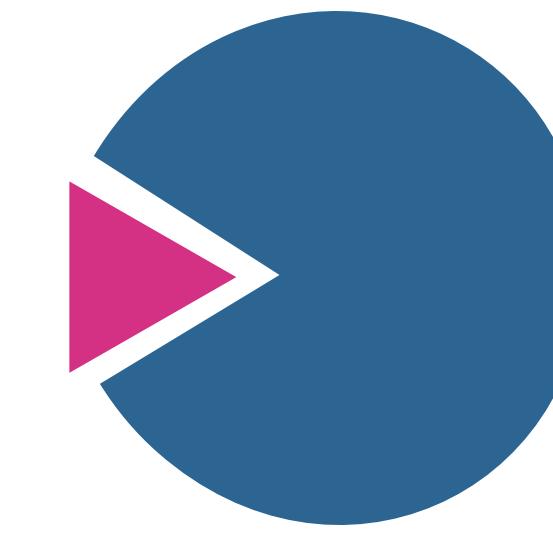


Dreaming up useful molecules

# How Drugs Work

Most medical intervention are based on the fact that  
a **small molecule** fits into a **large molecule**.

**small molecule**  
**drug**  
**ligand**  
**inhibitor**

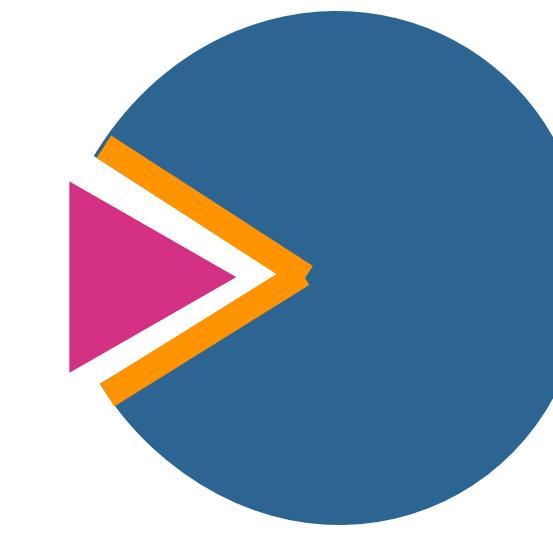


**protein**  
**enzyme**  
**receptor**  
**drug target**

# How Drugs Work

Most medical intervention are based on the fact that  
a **small molecule** fits into a **large molecule**.

**small molecule**  
**drug**  
**ligand**  
**inhibitor**



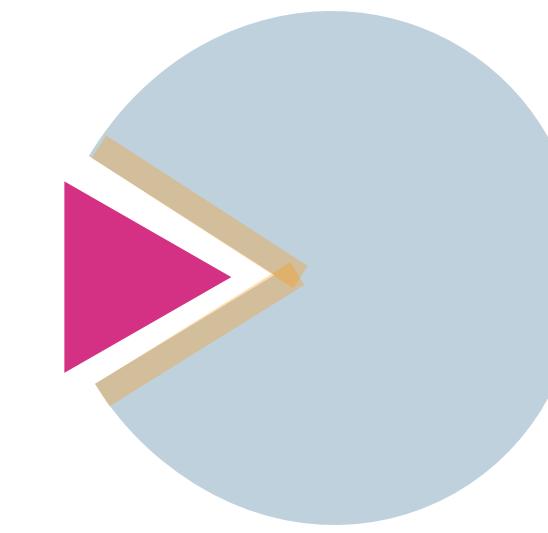
**protein**  
**enzyme**  
**receptor**  
**drug target**

**binding site**  
**pocket**

# What is a Small Molecule

Most medical intervention are based on the fact that  
a **small molecule** fits into a **large molecule**.

**small molecule**  
**drug**  
**ligand**  
**inhibitor**



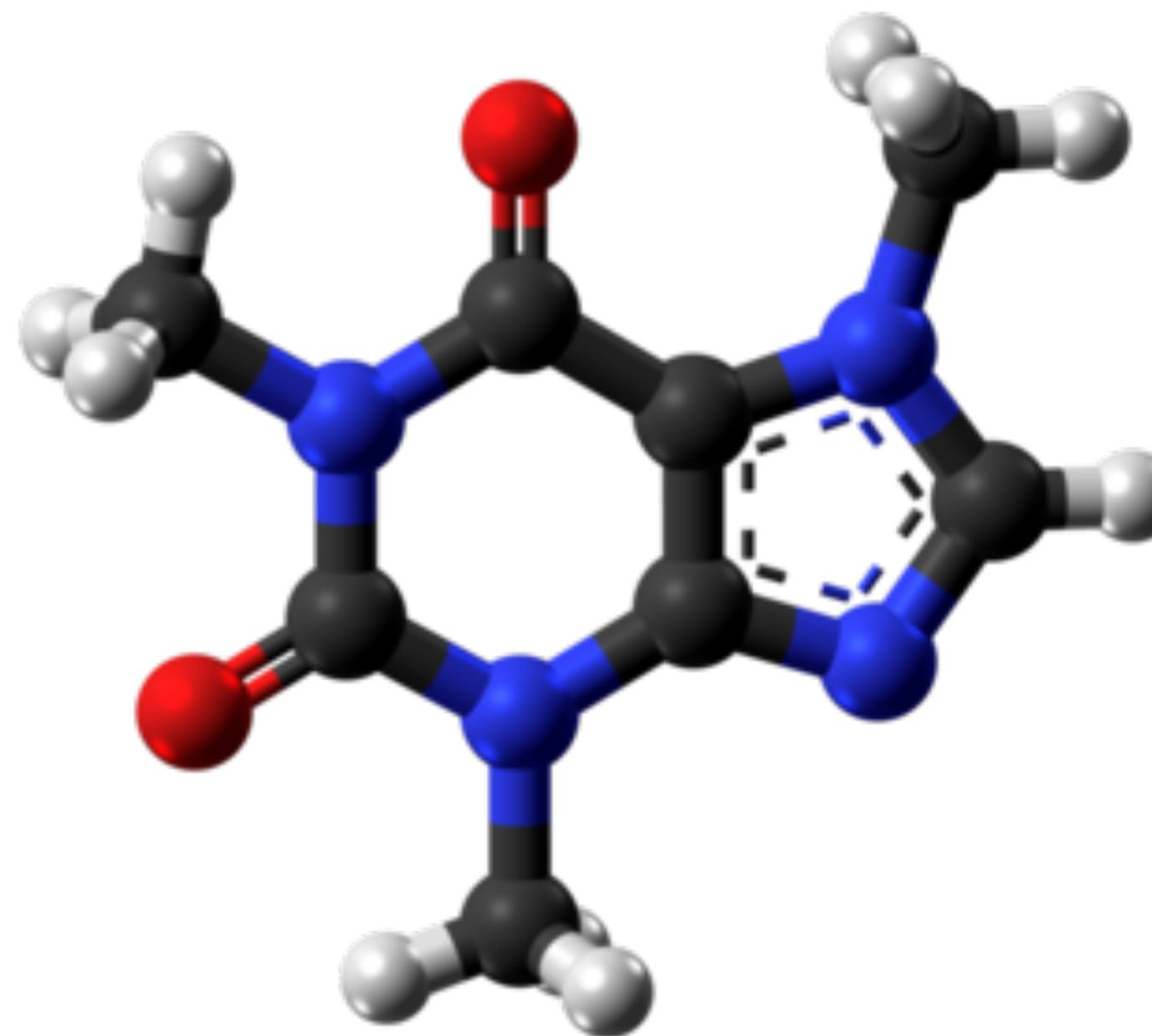
**protein**  
**enzyme**  
**receptor**  
**drug target**

**binding site**  
**pocket**

# What is a (Small) Molecule

## Molecule

- group of atoms connected through chemical bonds
- electrostatic force shapes geometry
- properties emerge from its constituent atoms
- dynamic and complex three-dimensional objects



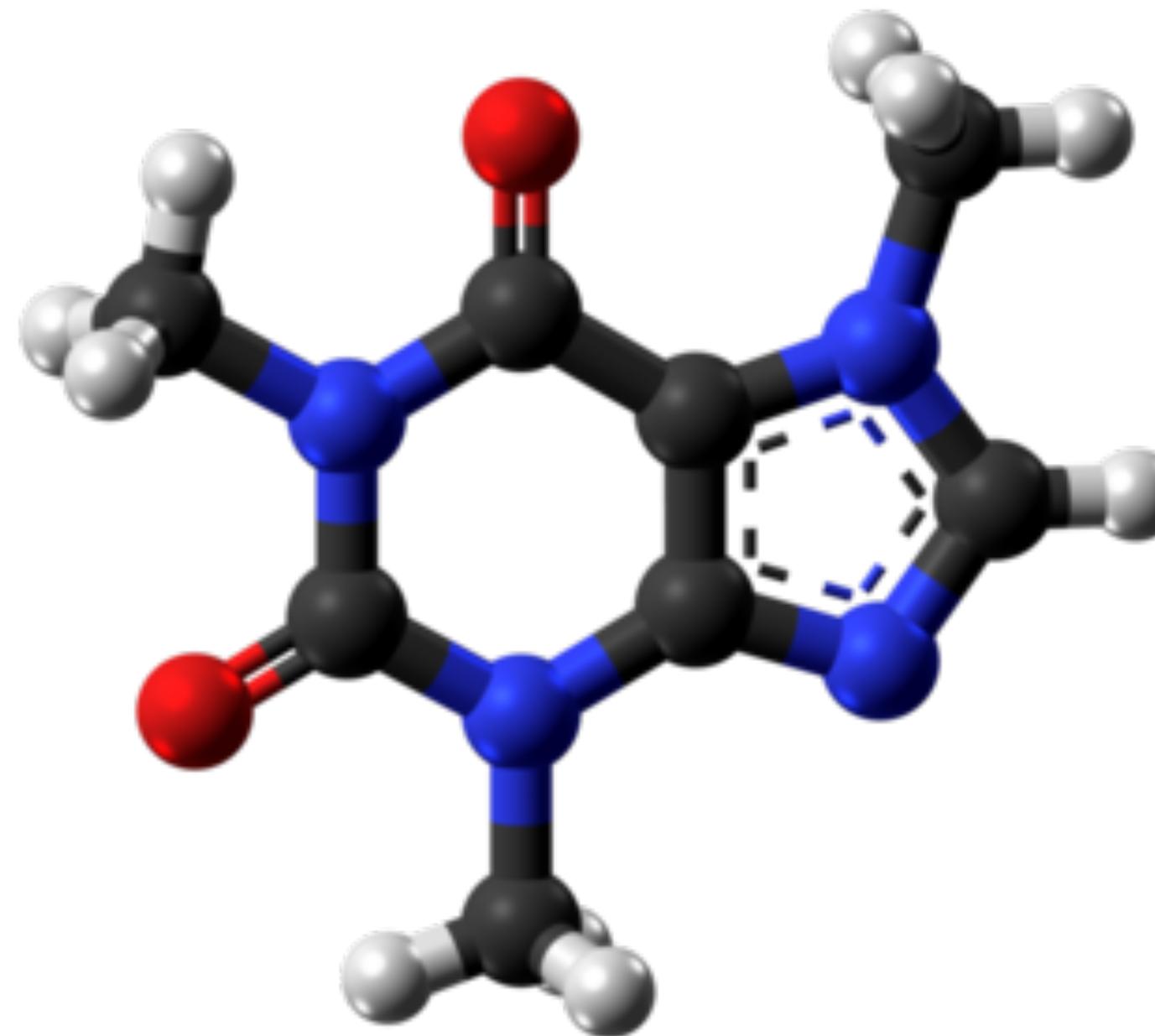
## Small and organic molecule

- typically < 100 atoms
- carbon-hydrogen or carbon-carbon bonds
- modulates a biological process
- druglikeness (bioavailability, solubility, stability)

# What is a (Small) Molecule

## Molecule

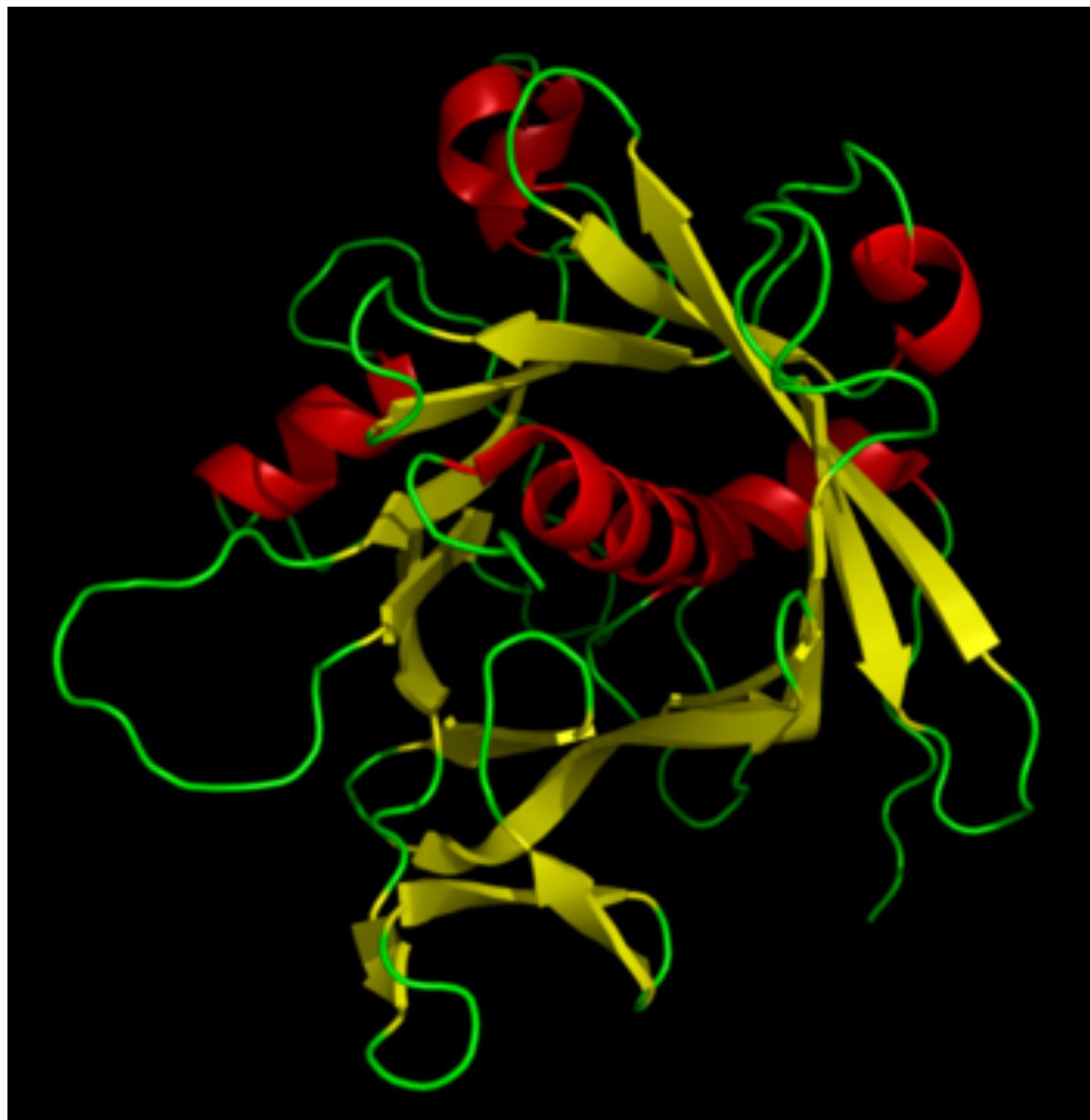
- group of atoms connected through chemical bonds
- electrostatic force shapes geometry
- properties emerge from its constituent atoms
- dynamic and complex three-dimensional objects



## Small and organic molecule

- typically < 100 atoms
- carbon-hydrogen or carbon-carbon bonds
- modulates a biological process
- druglikeness (bioavailability, solubility, stability)

# What is a Protein?



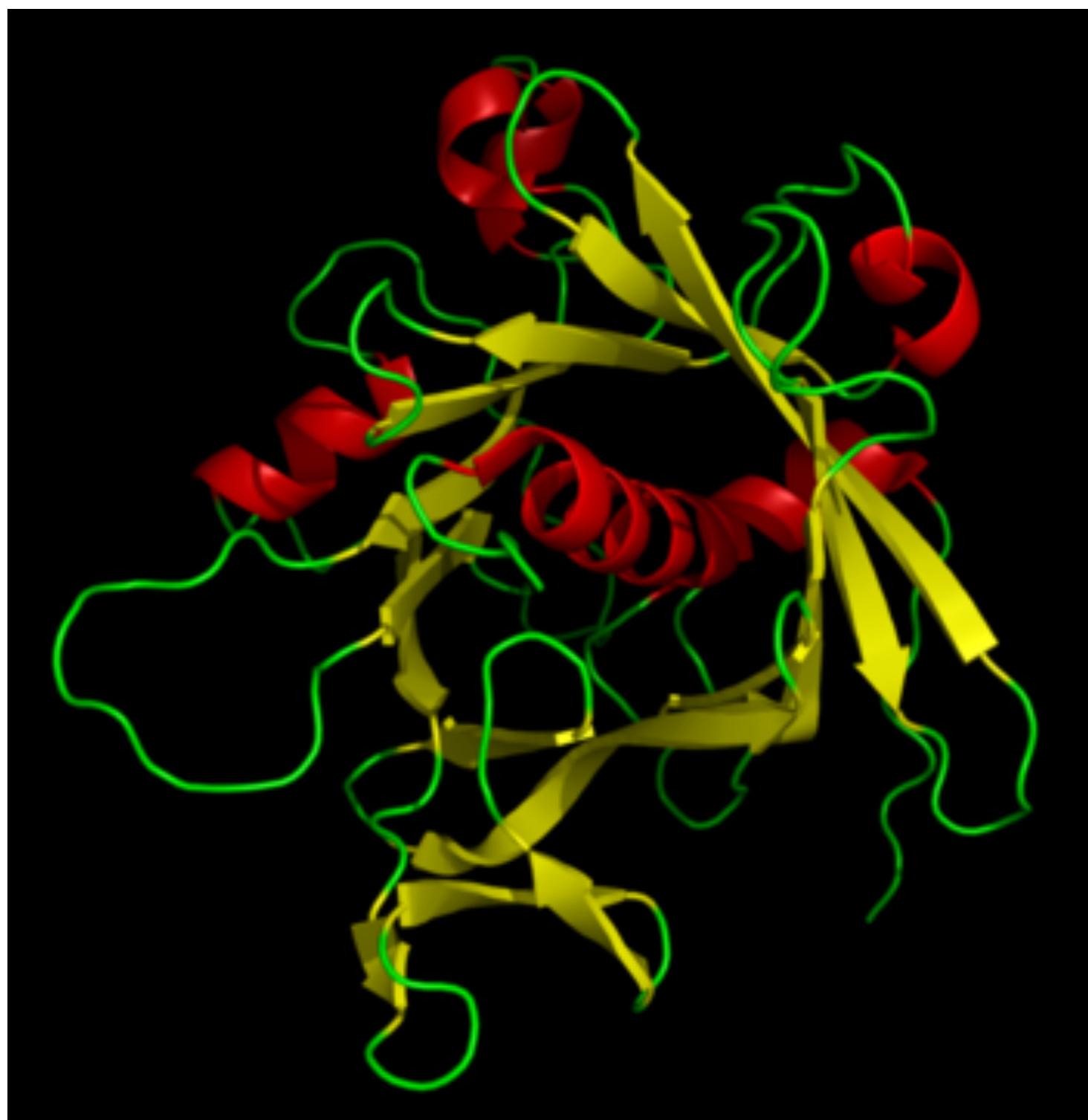
## Protein

- large, complex molecules
- several thousand atoms
- chain of amino acids
- forms 3D structure in a highly non-trivial way

## Druggable Target

- presence of a binding site
- minimize off-target effects
- modifiable activity:
  - inhibit or activate enzyme activity
  - activate or block receptor

# What is a Protein?



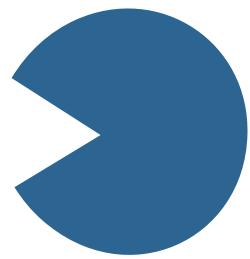
## Protein

- large, complex molecules
- several thousand atoms
- chain of amino acids
- forms 3D structure in a highly non-trivial way

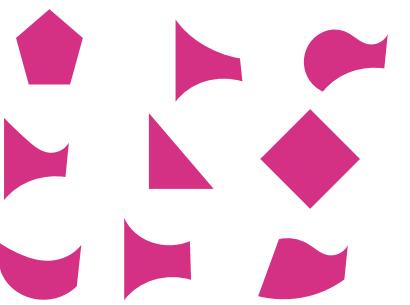
## Druggable Target

- presence of a binding site
- minimize off-target effects
- modifiable activity:
  - inhibit or activate enzyme activity
  - activate or block receptor

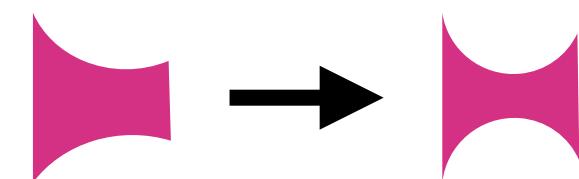
# Pipeline



Identifying a **protein** that is involved in a disease process and can be **targeted** by a **drug** to produce a therapeutic effect.



Use algorithms (docking + filtering) to **screen** virtual libraries of **small molecules** and predict which ones are likely to bind to the target.



Refine **hits** to optimize their properties for use as drugs.

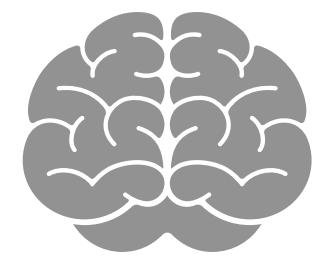
$10^{60}$

different drug-like molecules

**Vast chemical space**

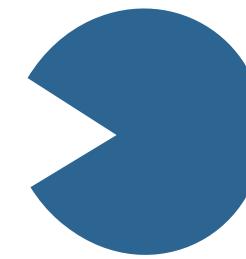


**Expensive**

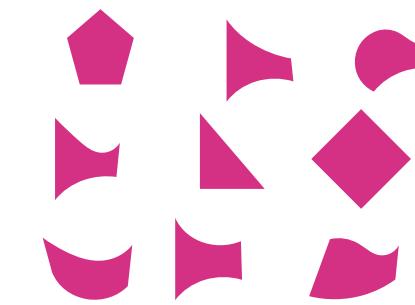


**Human bias**

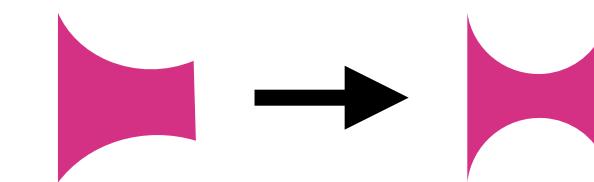
# AI to the Rescue



Identifying a **protein** that is involved in a disease process and can be **targeted** by a **drug** to produce a therapeutic effect.



Use algorithms (docking + filtering) to **screen** virtual libraries of **small molecules** and predict which ones are likely to bind to the target.



Refine **hits** to optimize their properties for use as drugs.



Compound generation

Generate novel chemical structures that are optimized for binding to the target.

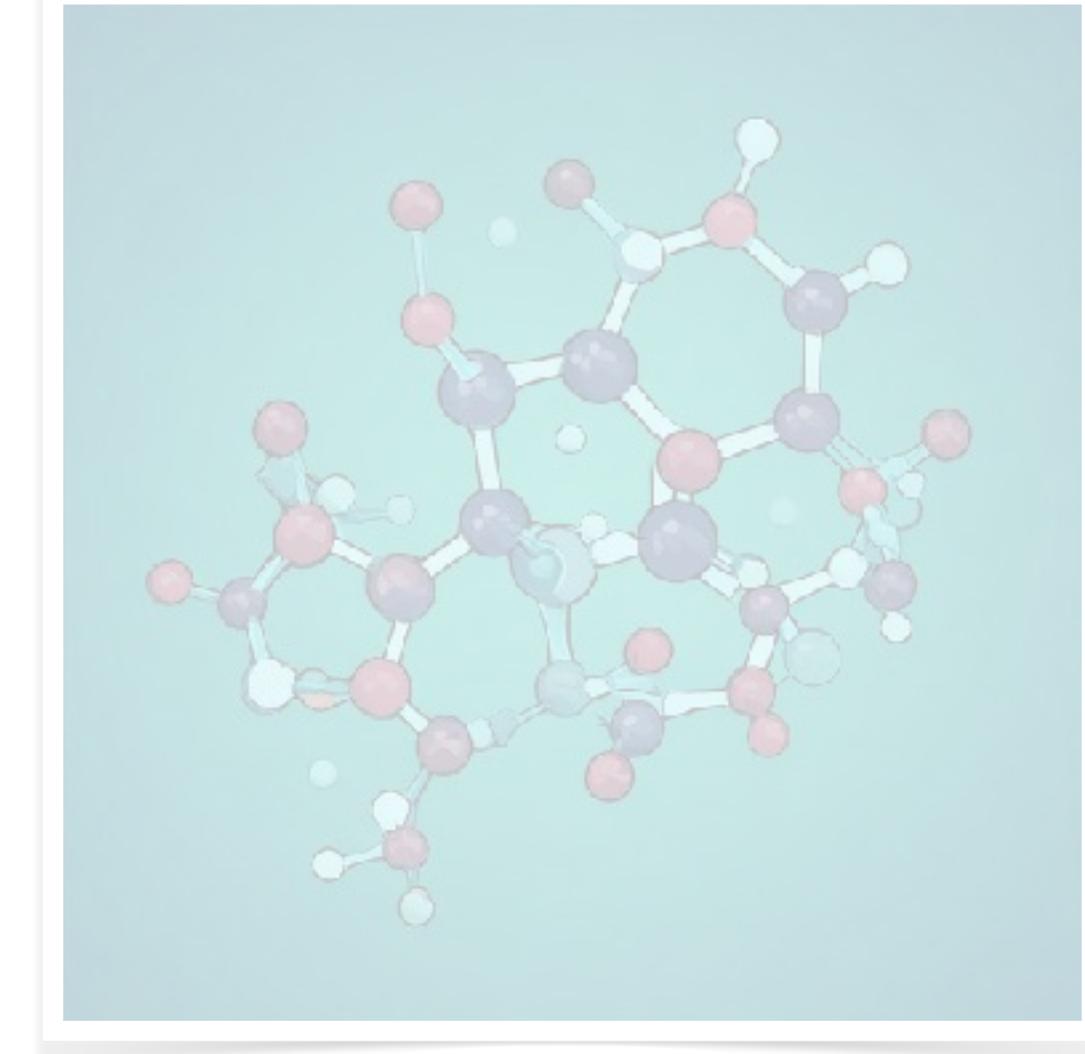
# Outline



Drug discovery in a nutshell



Foundations of diffusion models



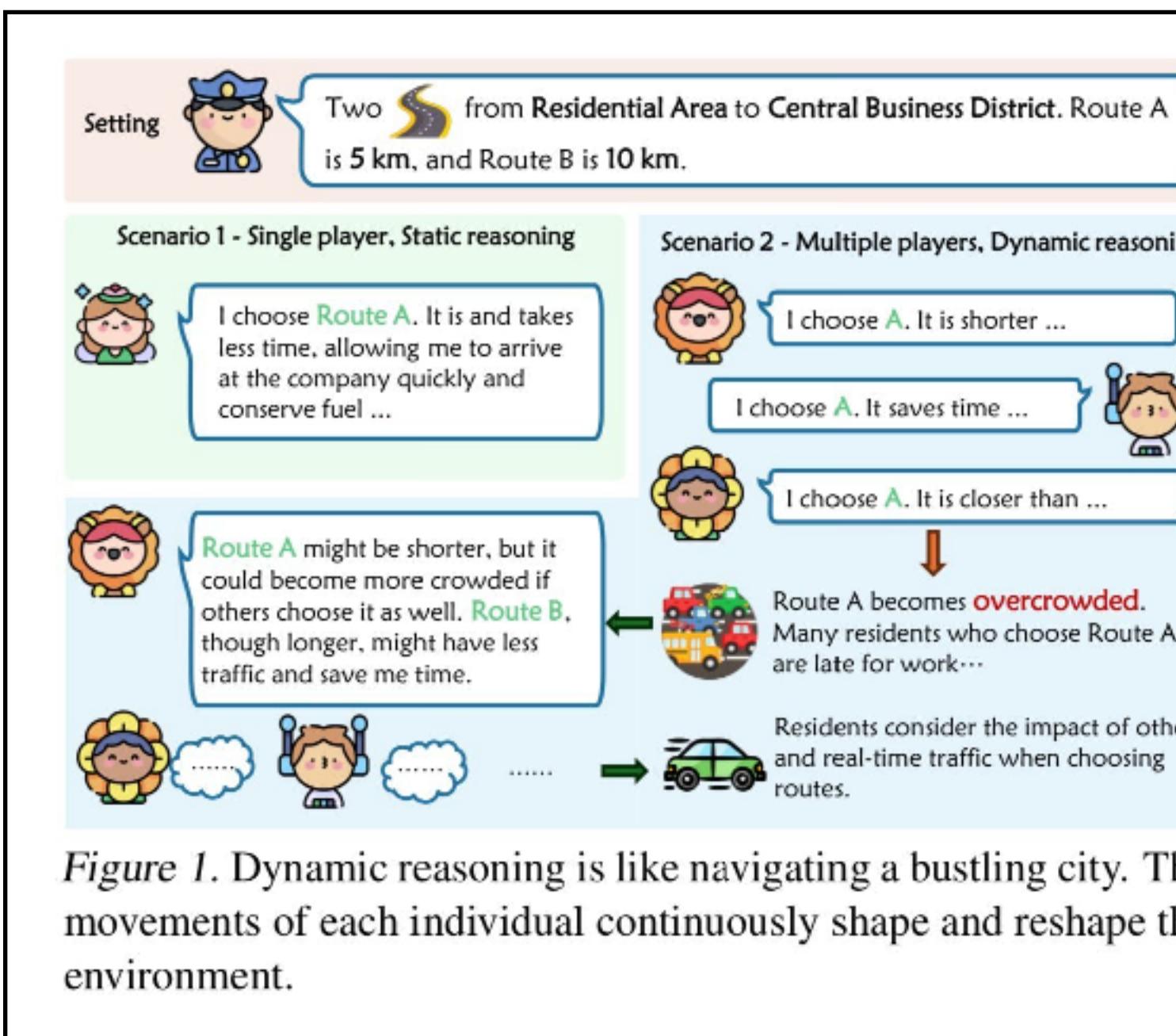
Dreaming up useful molecules

# Probabilistic Diffusion



Nucleus ☕ ✅ M  
@EsotericCofe

hardest LLM paper vs easiest diffusion paper



[twitter.com/EsotericCofe/status/1777280241884377474](https://twitter.com/EsotericCofe/status/1777280241884377474)

$$\begin{aligned} L\text{-DPO-Diffusion}(\theta) \leq & -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0, T),} \\ & \mathbf{x}_{t-1, t}^w \sim p_\theta(\mathbf{x}_{t-1, t}^w | \mathbf{x}_0^w), \\ & \mathbf{x}_{t-1, t}^l \sim p_\theta(\mathbf{x}_{t-1, t}^l | \mathbf{x}_0^l)} \\ & \log \sigma \left( \beta T \log \frac{p_\theta(\mathbf{x}_{t-1}^w | \mathbf{x}_t^w)}{p_{\text{ref}}(\mathbf{x}_{t-1}^w | \mathbf{x}_t^w)} - \beta T \log \frac{p_\theta(\mathbf{x}_{t-1}^l | \mathbf{x}_t^l)}{p_{\text{ref}}(\mathbf{x}_{t-1}^l | \mathbf{x}_t^l)} \right) \end{aligned} \quad (12)$$

Efficient training via gradient descent is now possible. However, sampling from reverse joint  $p_\theta(\mathbf{x}_{t-1}, \mathbf{x}_t | \mathbf{x}_0, c)$  is still intractable and  $r$  of Eq. (9) has an expectation over  $p_\theta(\mathbf{x}_{1:T} | \mathbf{x}_0)$ . So we approximate the reverse process  $p_\theta(\mathbf{x}_{1:T} | \mathbf{x}_0)$  with the forward  $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$  (an alternative scheme in Supp. S2). With some algebra, this yields:

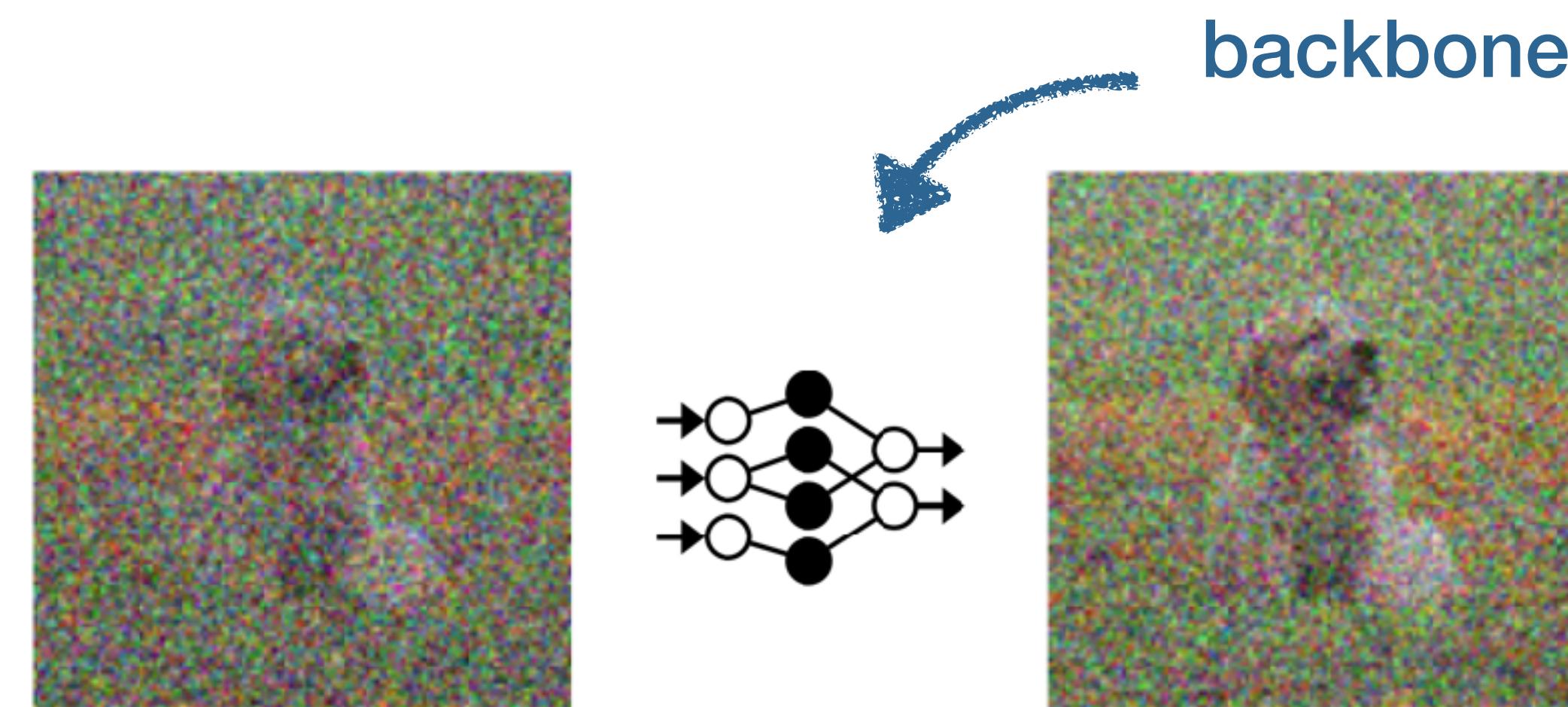
$$\begin{aligned} L(\theta) = & -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0, T), \mathbf{x}_t^w \sim q(\mathbf{x}_t^w | \mathbf{x}_0^w), \mathbf{x}_t^l \sim q(\mathbf{x}_t^l | \mathbf{x}_0^l)} \\ & \log \sigma(-\beta T) \\ & + \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1}^w | \mathbf{x}_{0,t}^w) \| p_\theta(\mathbf{x}_{t-1}^w | \mathbf{x}_t^w)) \\ & - \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1}^l | \mathbf{x}_{0,t}^l) \| p_{\text{ref}}(\mathbf{x}_{t-1}^l | \mathbf{x}_t^l)) \\ & - \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1}^l | \mathbf{x}_{0,t}^l) \| p_\theta(\mathbf{x}_{t-1}^l | \mathbf{x}_t^l)) \\ & + \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1}^l | \mathbf{x}_{0,t}^l) \| p_{\text{ref}}(\mathbf{x}_{t-1}^l | \mathbf{x}_t^l)). \end{aligned} \quad (13)$$

Using Eq. (1) and algebra, the above loss simplifies to:

$$\begin{aligned} L(\theta) = & -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0, T), \mathbf{x}_t^w \sim q(\mathbf{x}_t^w | \mathbf{x}_0^w), \mathbf{x}_t^l \sim q(\mathbf{x}_t^l | \mathbf{x}_0^l)} \\ & \log \sigma(-\beta T \omega(\lambda_t)) \\ & \| \boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^w, t) \|_2^2 - \| \boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\mathbf{x}_t^w, t) \|_2^2 \\ & - (\| \boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^l, t) \|_2^2 - \| \boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\mathbf{x}_t^l, t) \|_2^2) \end{aligned} \quad (14)$$

where  $\mathbf{x}_t^* = \alpha_t \mathbf{x}_0^* + \sigma_t \boldsymbol{\epsilon}^*$ ,  $\boldsymbol{\epsilon}^* \sim \mathcal{N}(0, I)$  is a draw from  $q(\mathbf{x}_t^* | \mathbf{x}_0^*)$  (Eq. (2)).  $\lambda_t = \alpha_t^2 / \sigma_t^2$  is the signal-to-noise ratio,

# Diffusion



# Probabilistic Diffusion

1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)

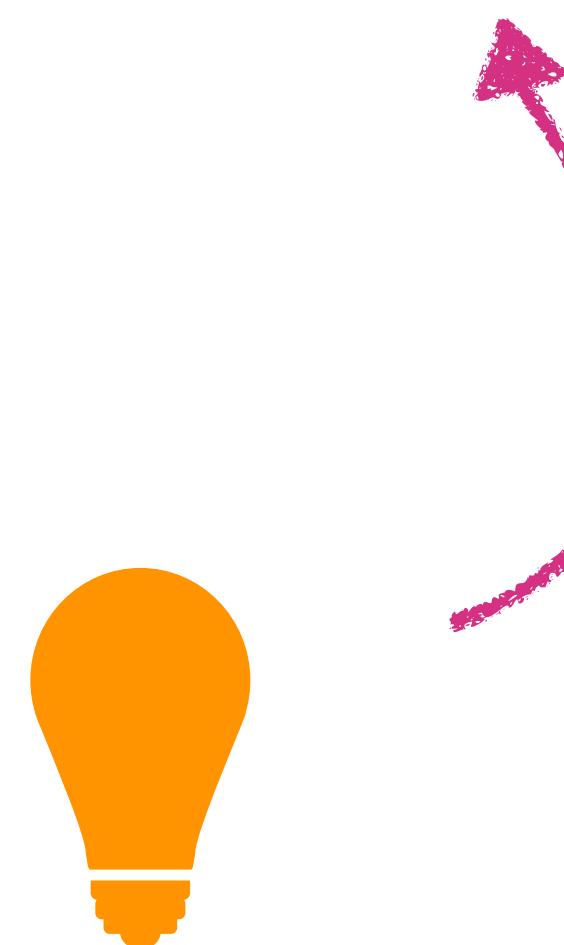
2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)

# Probabilistic Diffusion

1.

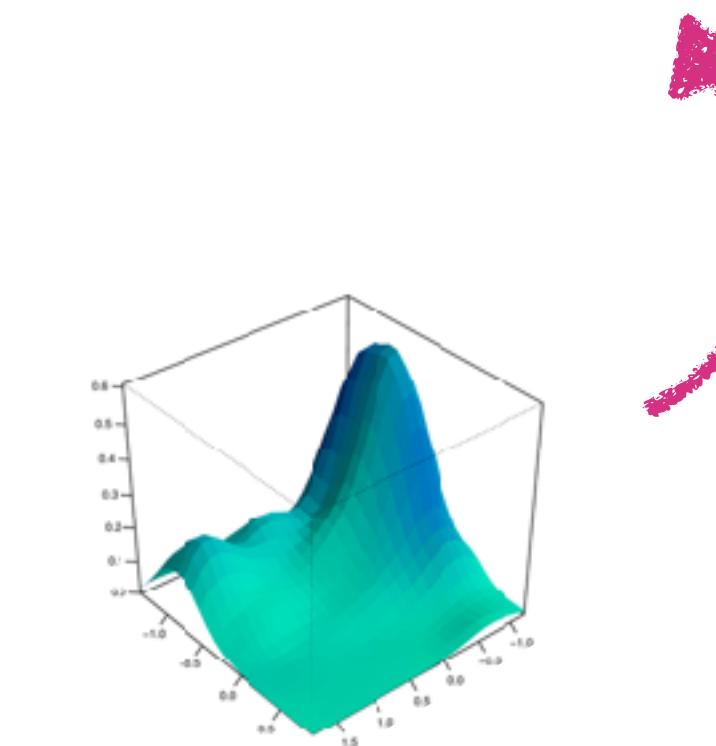
**Non-trainable** part that **removes** information  
(stochastically or deterministically)



This is a trick to guide the training.

2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



This is where we learn the distribution  
(implicitly represented by NN weights).

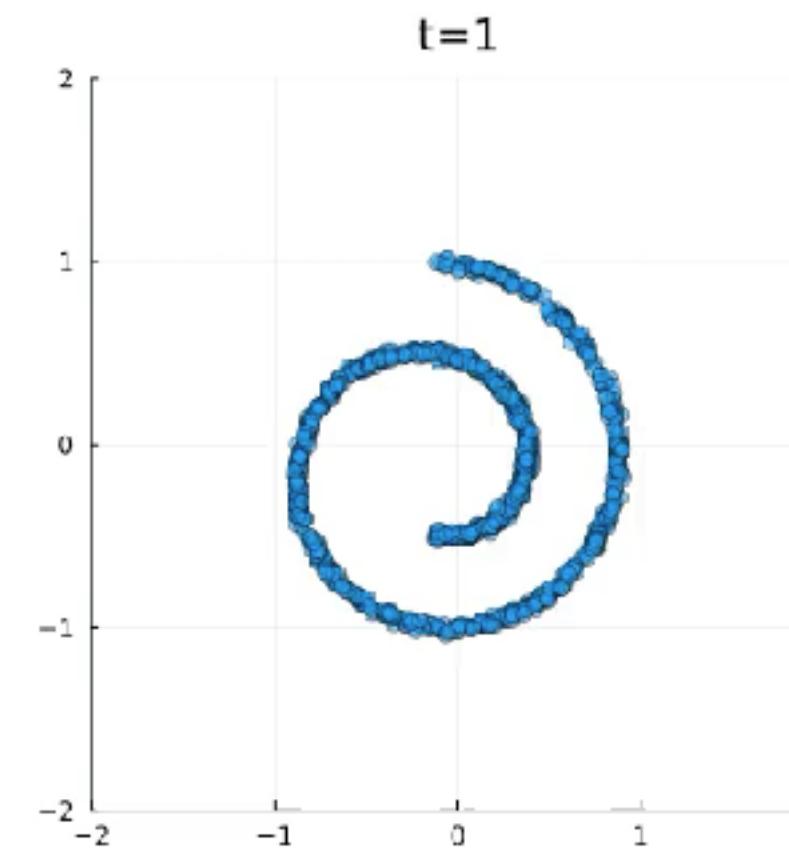
# Probabilistic Diffusion

1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)

2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)

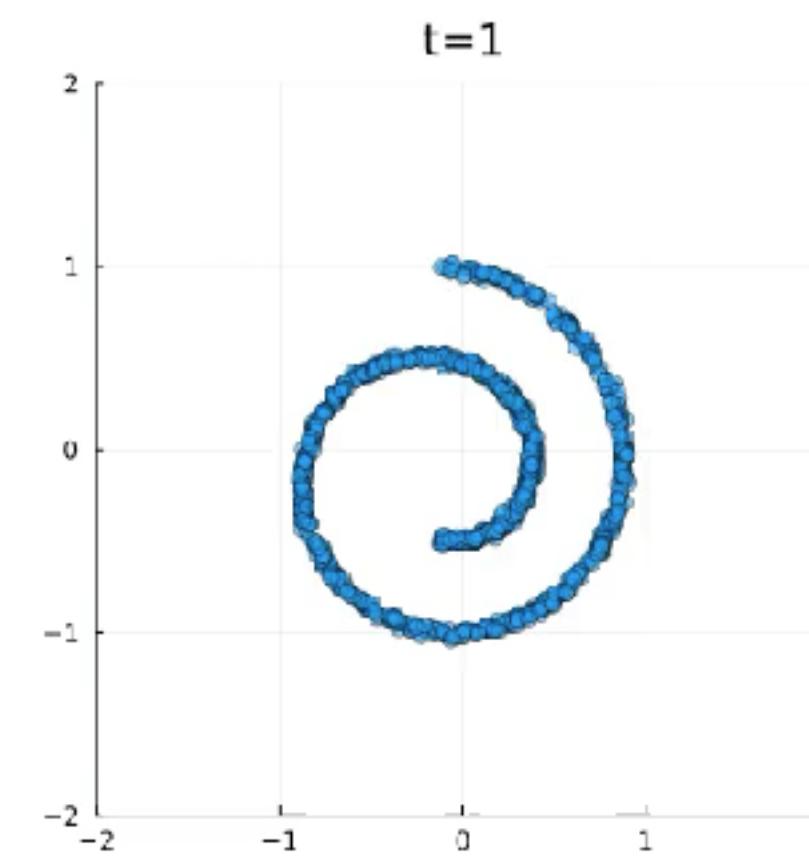


Obvious way: Degrade a sample by **adding Gaussian noise** to it.

# Probabilistic Diffusion

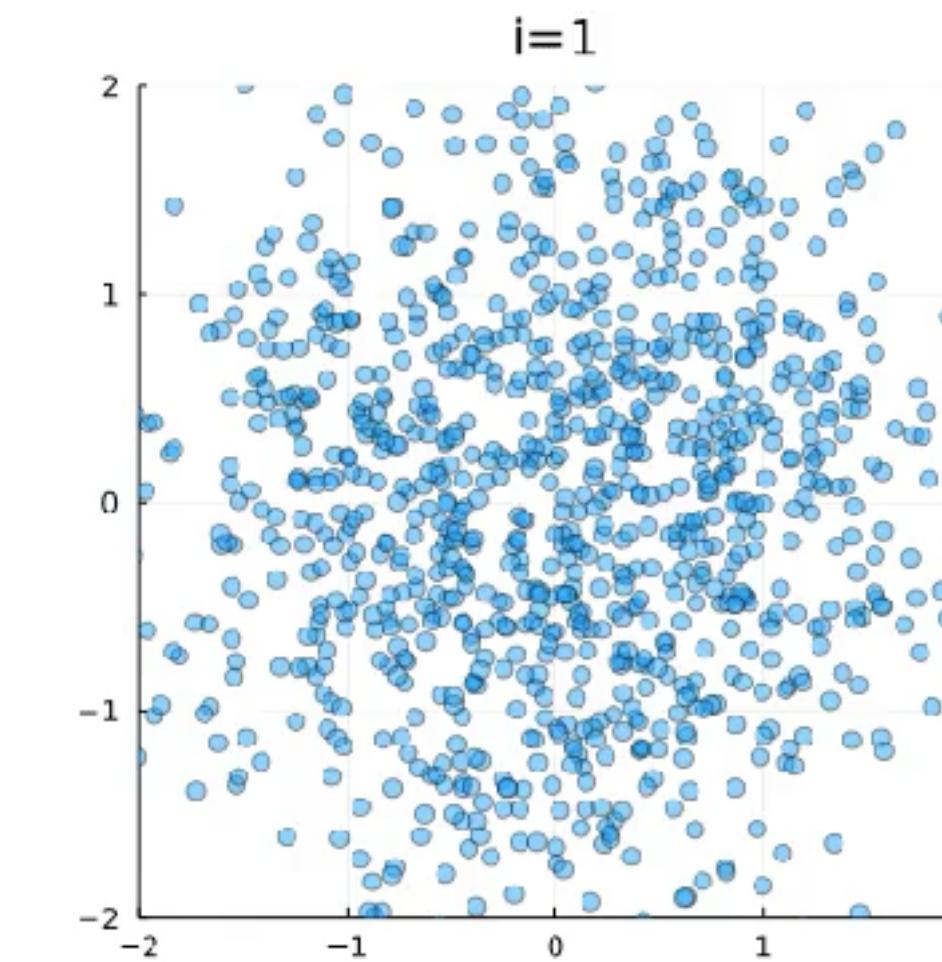
1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)



2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



Obvious way: Degrade a sample by **adding Gaussian noise** to it.

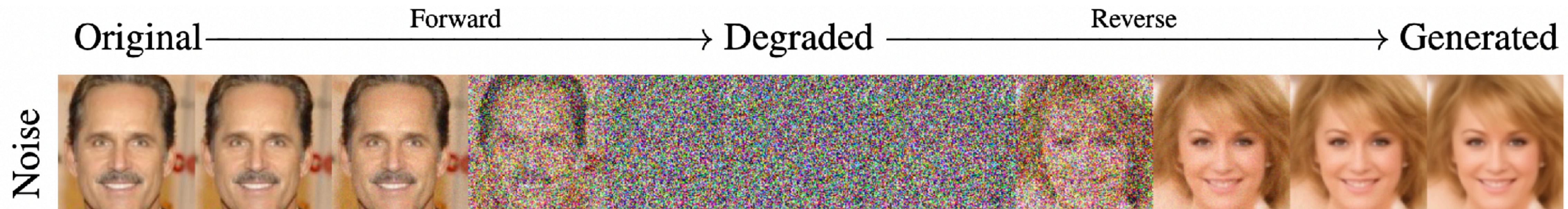
# Probabilistic Diffusion

1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)

2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



Obvious way: Degrade a sample by **adding Gaussian noise** to it.

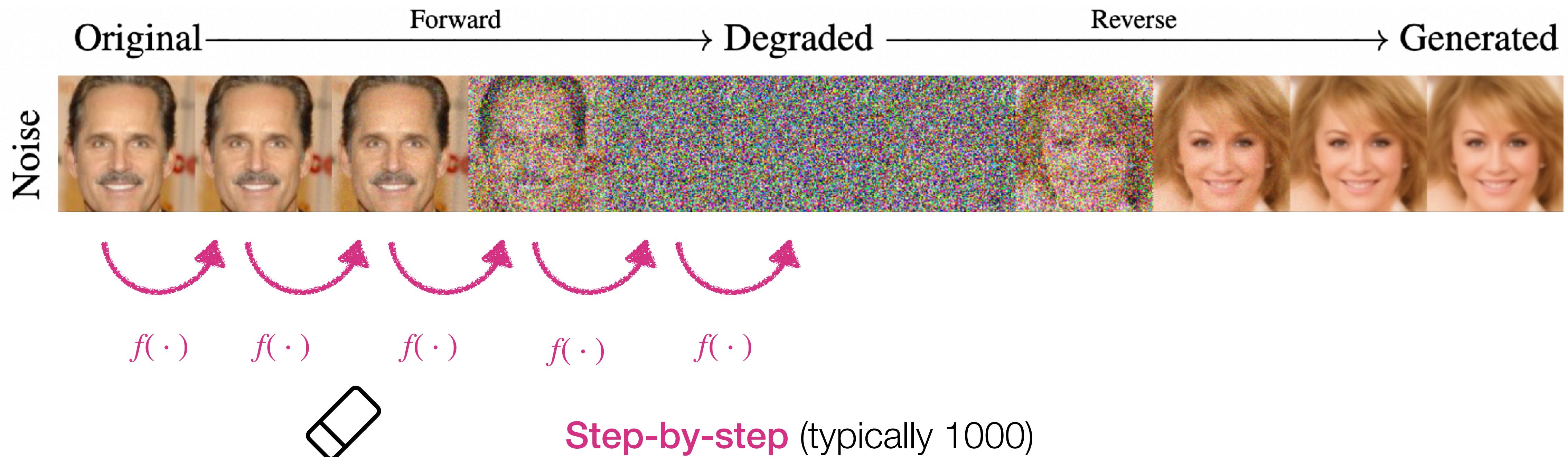
# Probabilistic Diffusion

1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)

2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



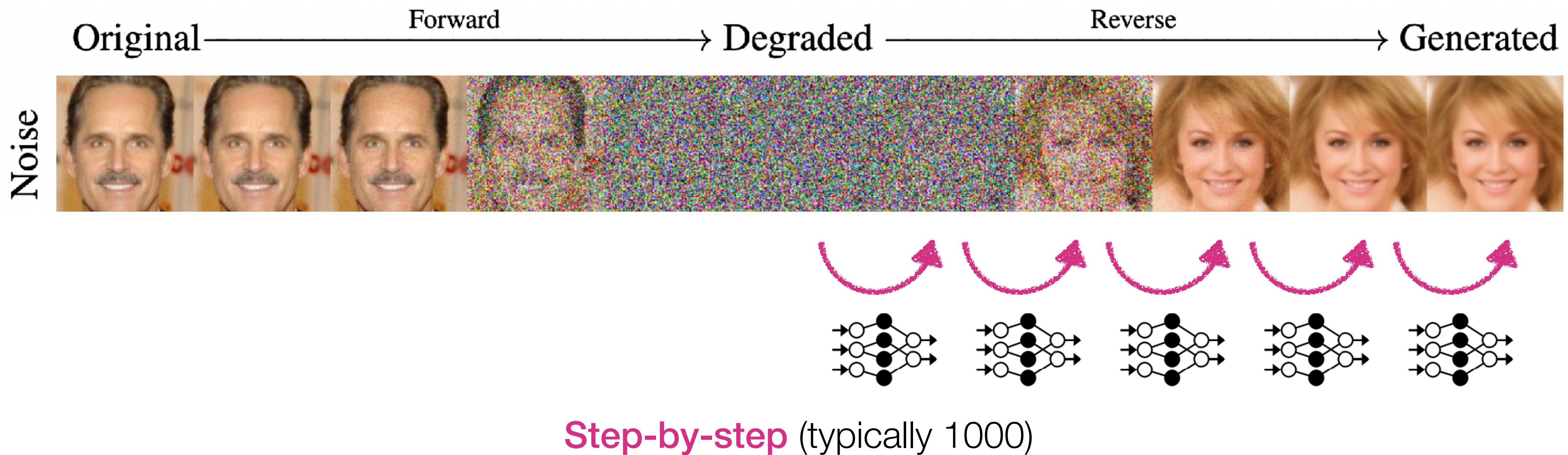
# Probabilistic Diffusion

1.

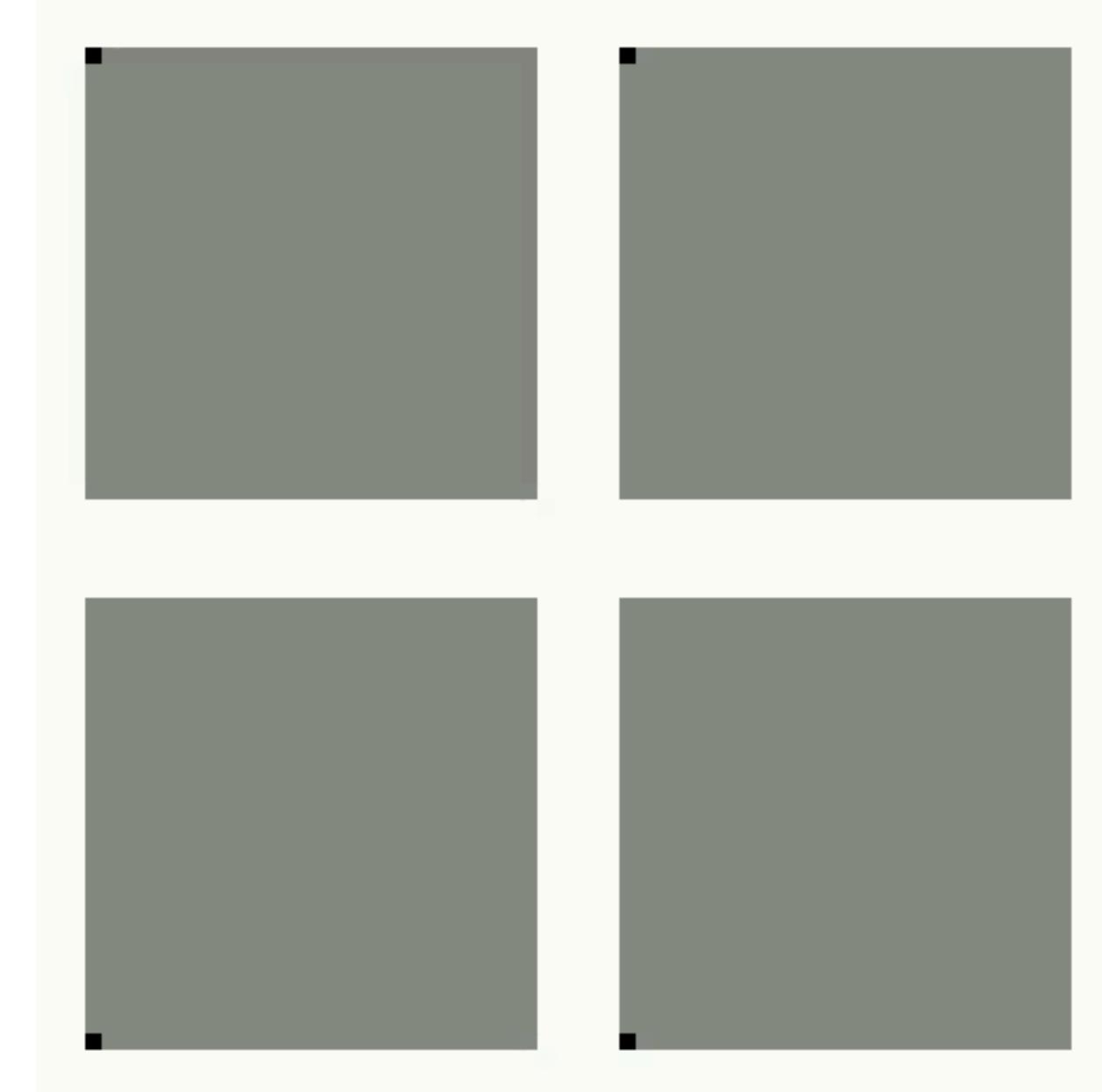
**Non-trainable** part that **removes** information  
(stochastically or deterministically)

2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



# Probabilistic Diffusion



[ajayjain.github.io/lmconv/](http://ajayjain.github.io/lmconv/)

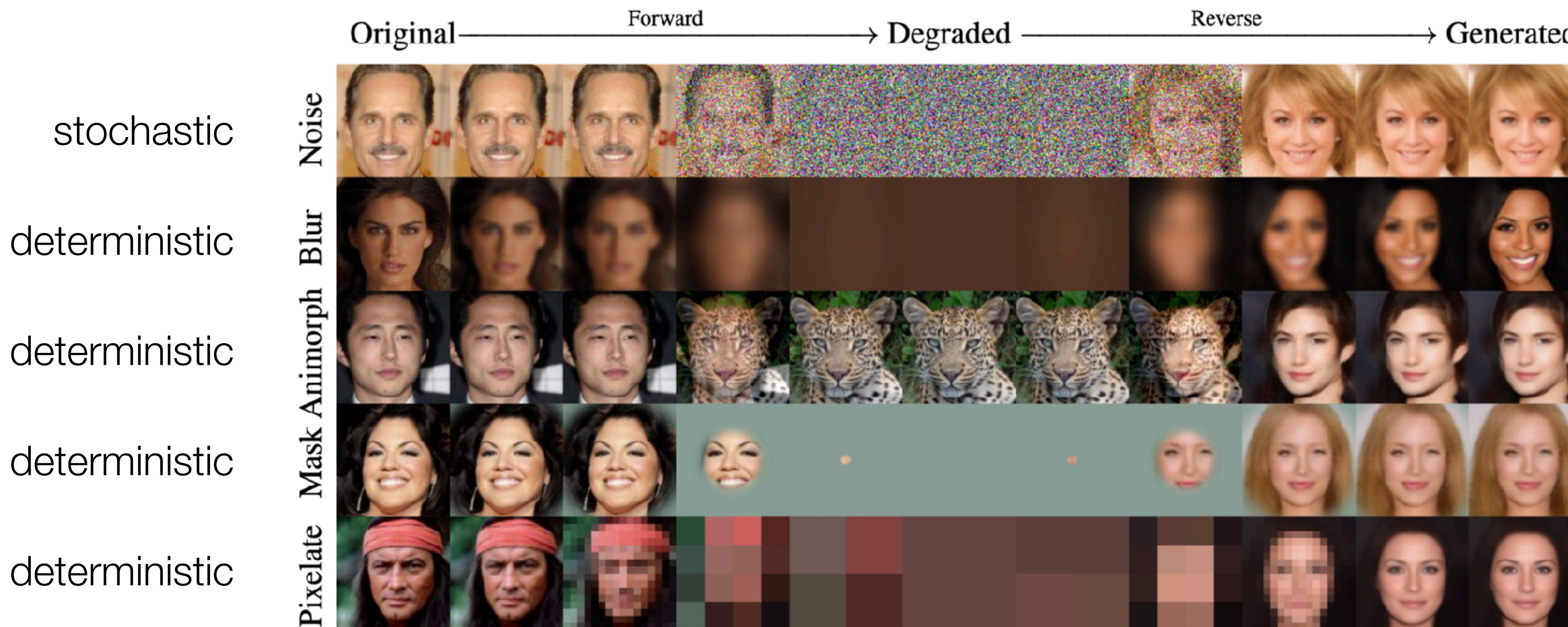
“What if the forward process masks one dimension in each step?”

“You get an autoregressive model!”

# Probabilistic Diffusion

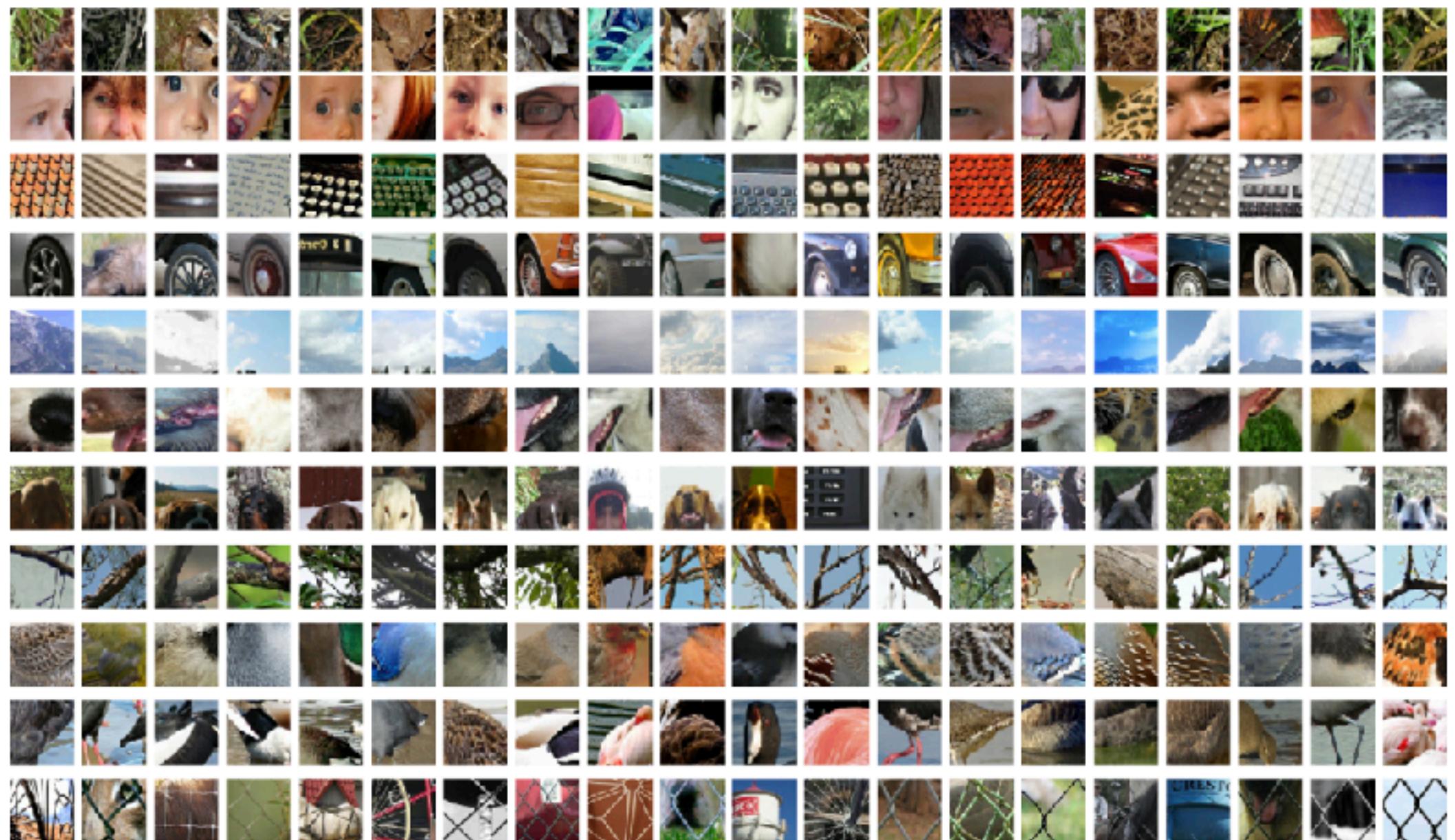
**Non-trainable** part that **removes** information  
(stochastically or deterministically)

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)



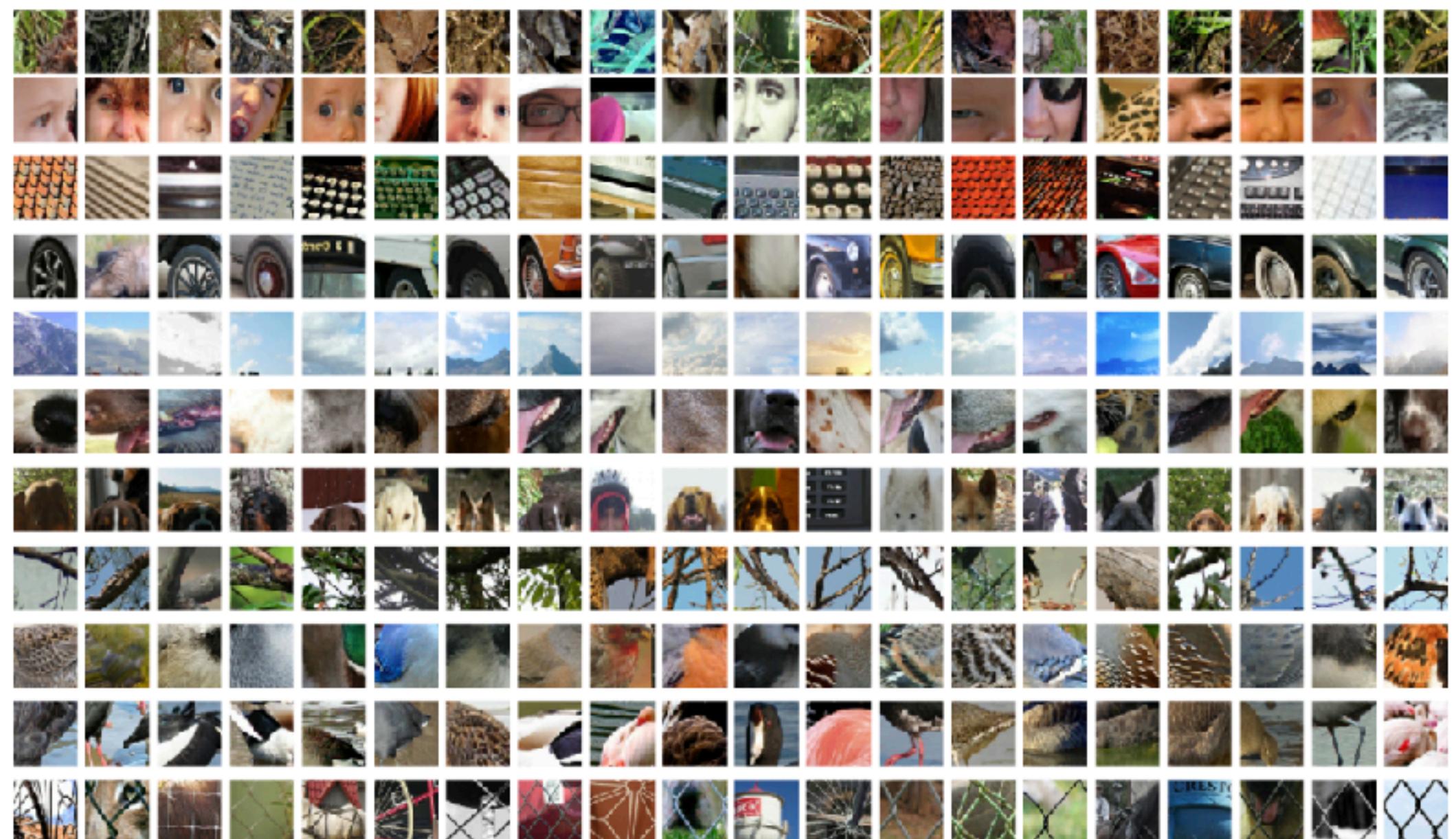
Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise (Bansal et al.)

# Conditional Sampling



[semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7](https://semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7)

# Conditional Sampling



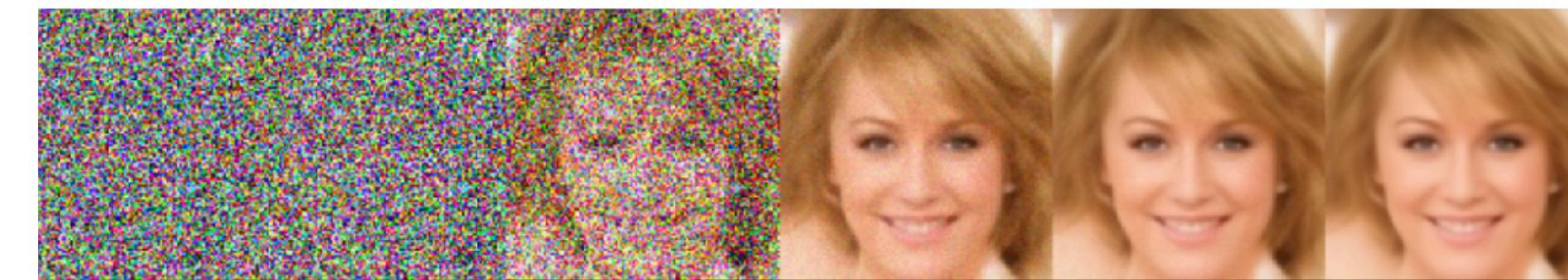
A panda on a motorcycle with a red sombrero and a cigar, oil painting.



[semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7](https://semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7)

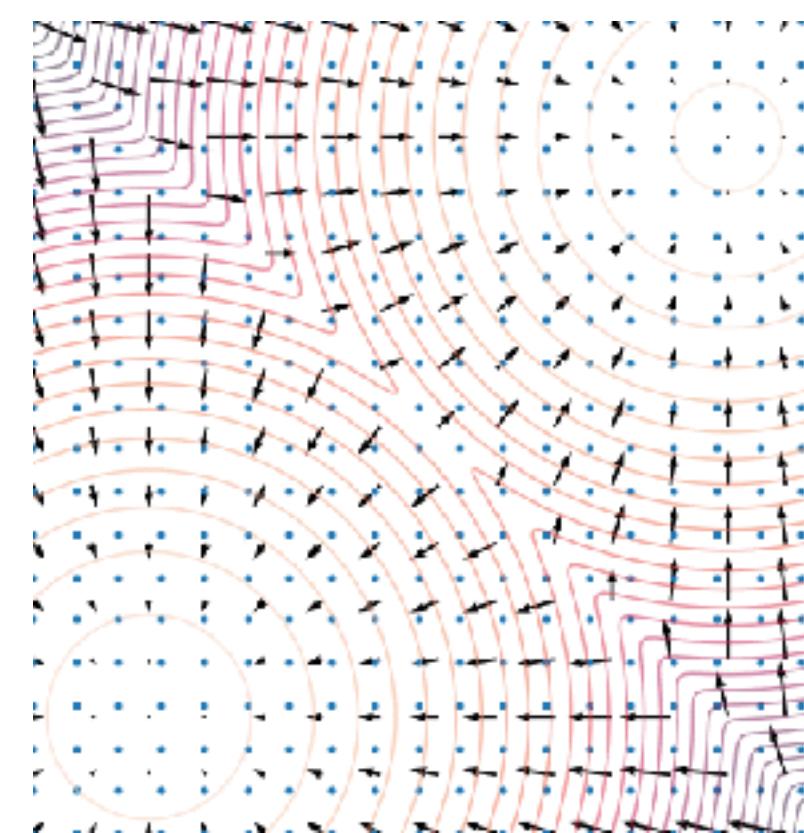
# Diffusion in Literature

## Denoising (DDPM)

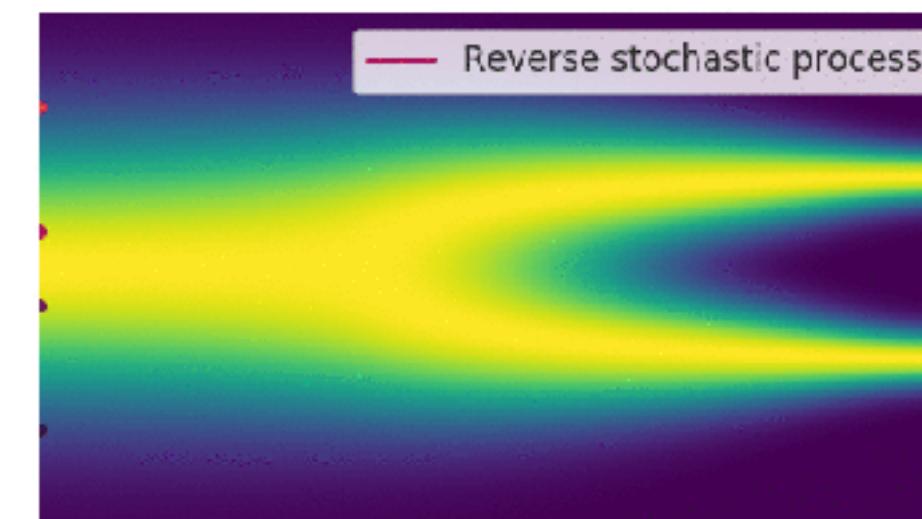
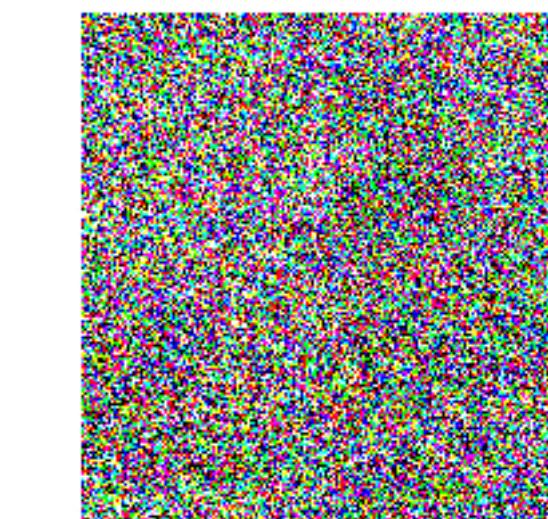


Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise (Bansal et al.)

## SDE

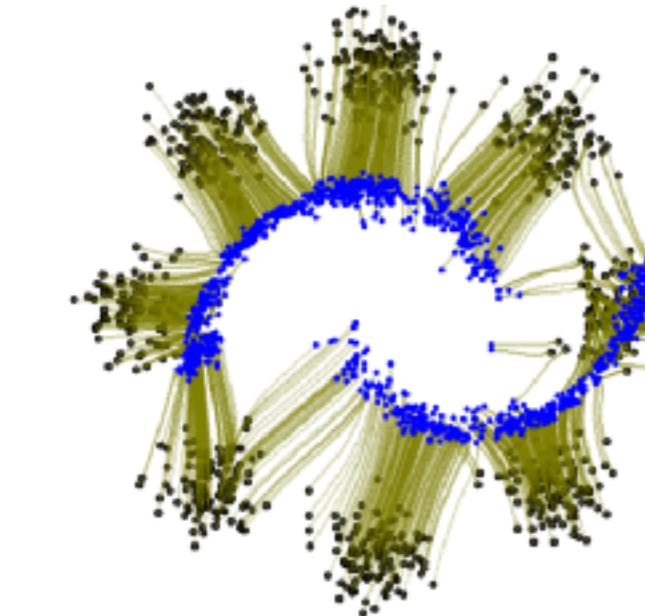
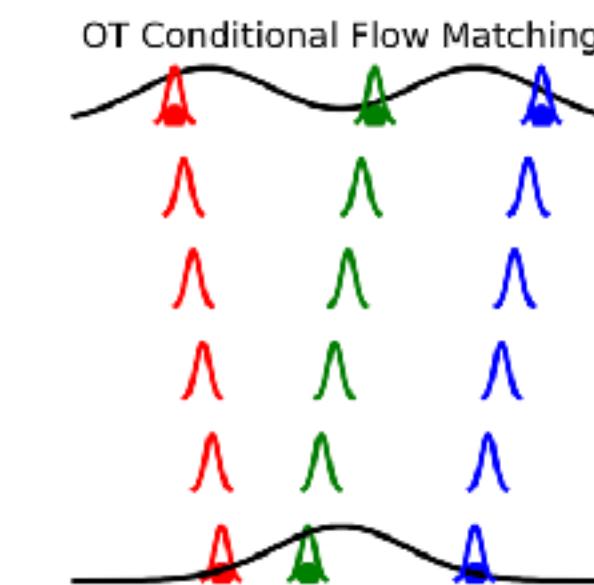


## Score Matching



[yang-song.net/blog/2021/score/](http://yang-song.net/blog/2021/score/)

## Flow Matching



Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport (Tong et al.)

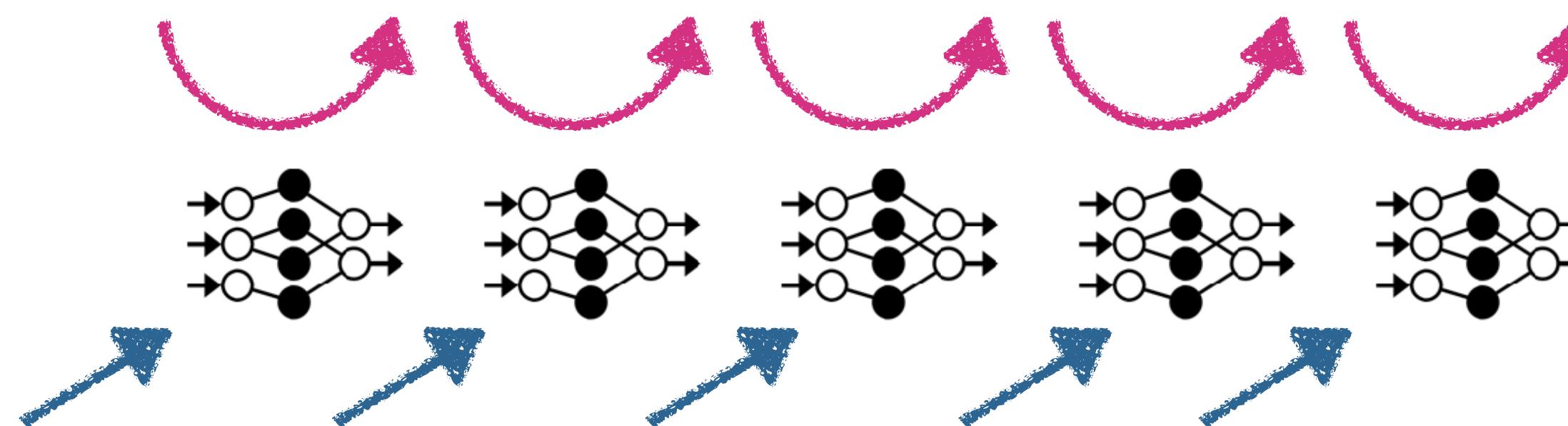
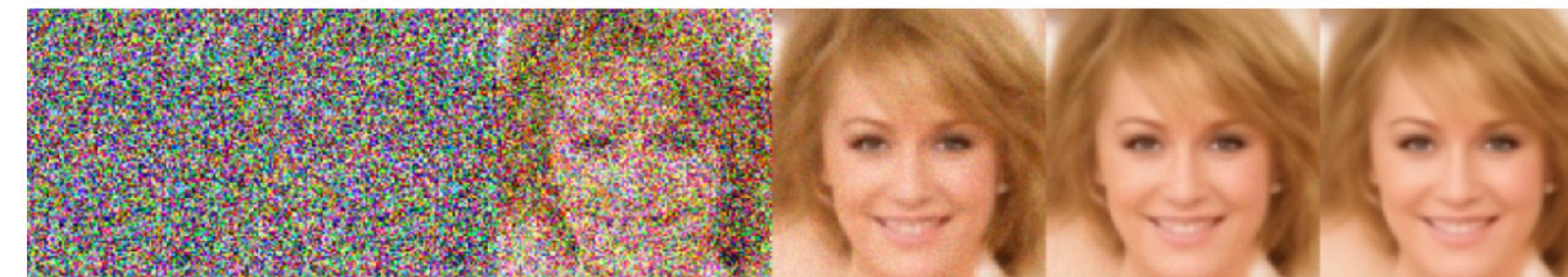
# Conditional Sampling

1.

**Non-trainable** part that **removes** information  
(stochastically or deterministically)

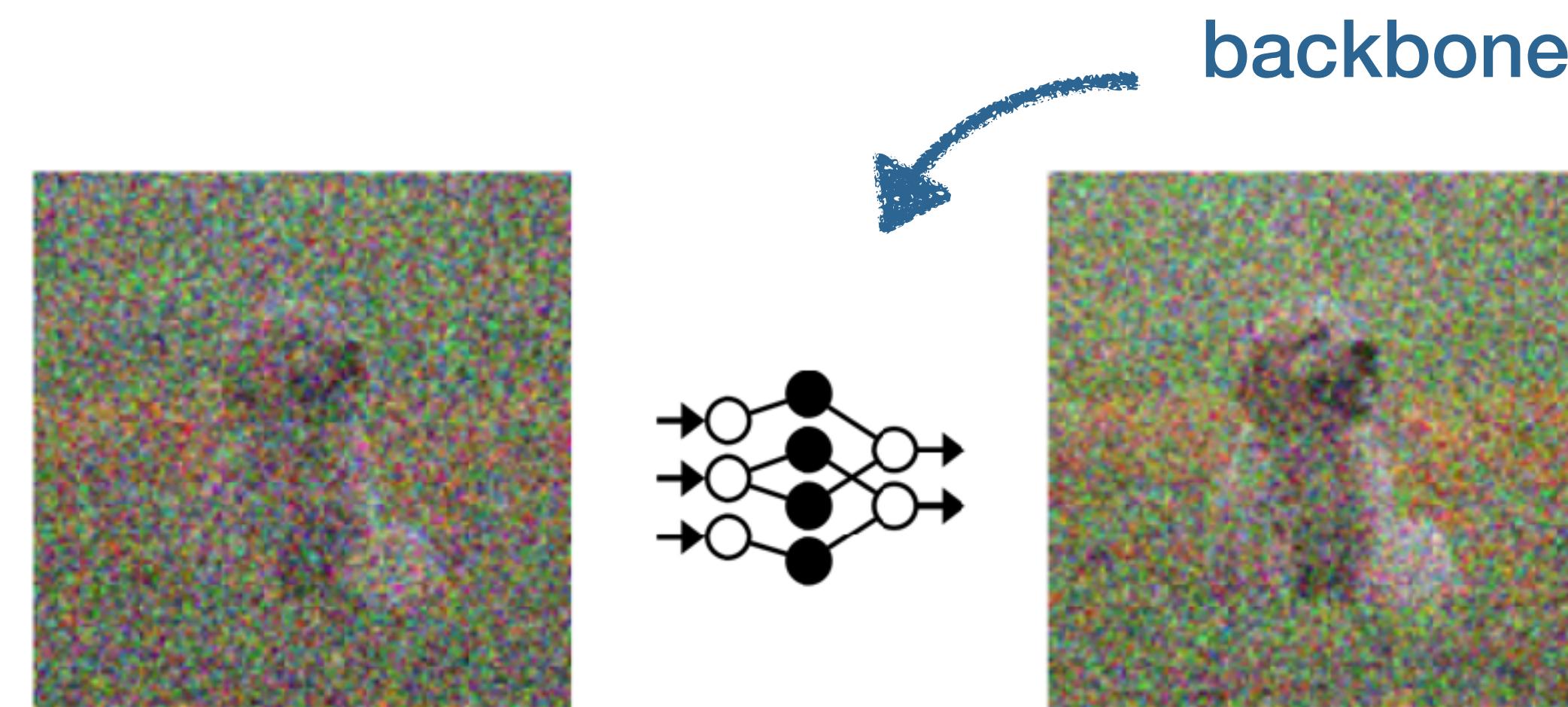
2.

**Trainable** part that **reconstructs** information  
(typically with a stochastic component)

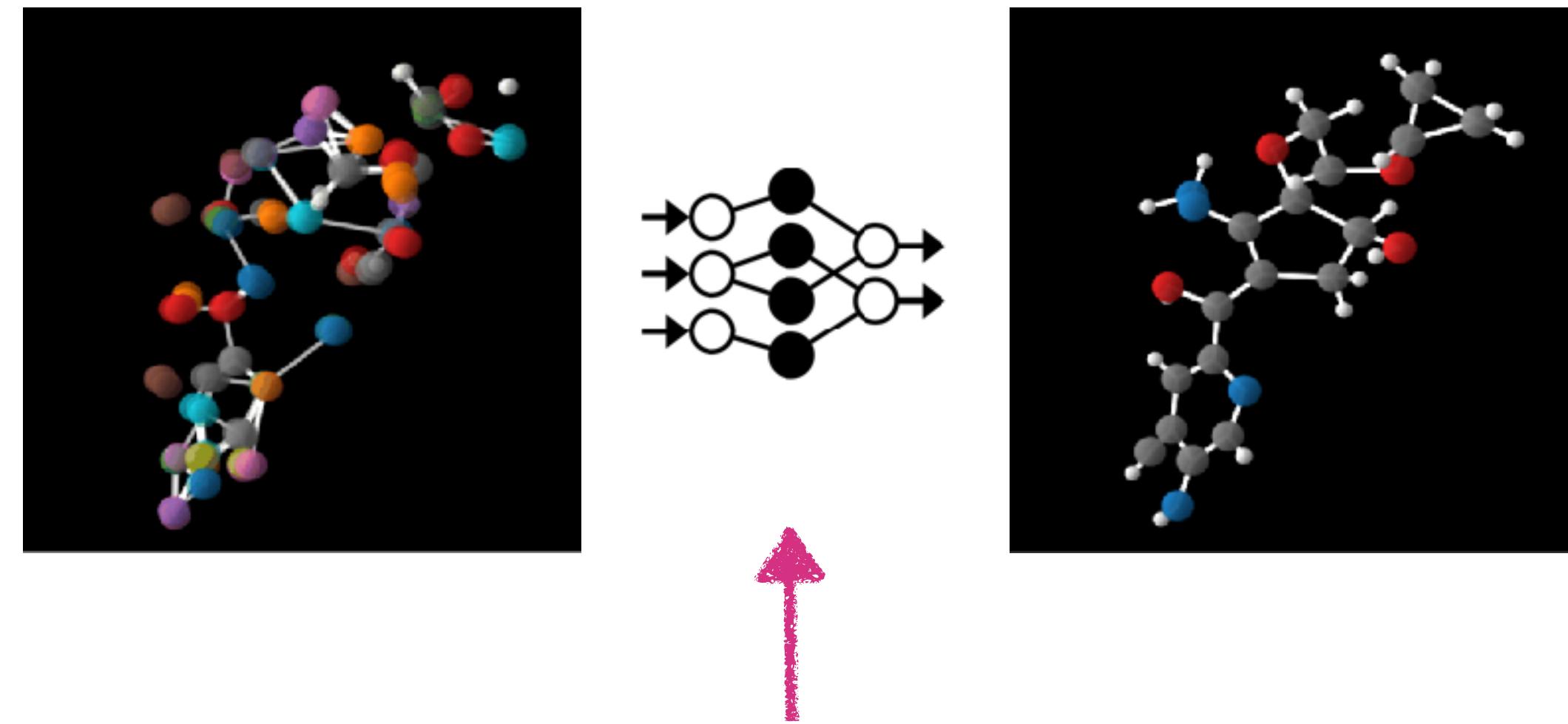


**condition as conditional input**

# Diffusion



# Diffusion



Some Flavour of Graph Neural Networks

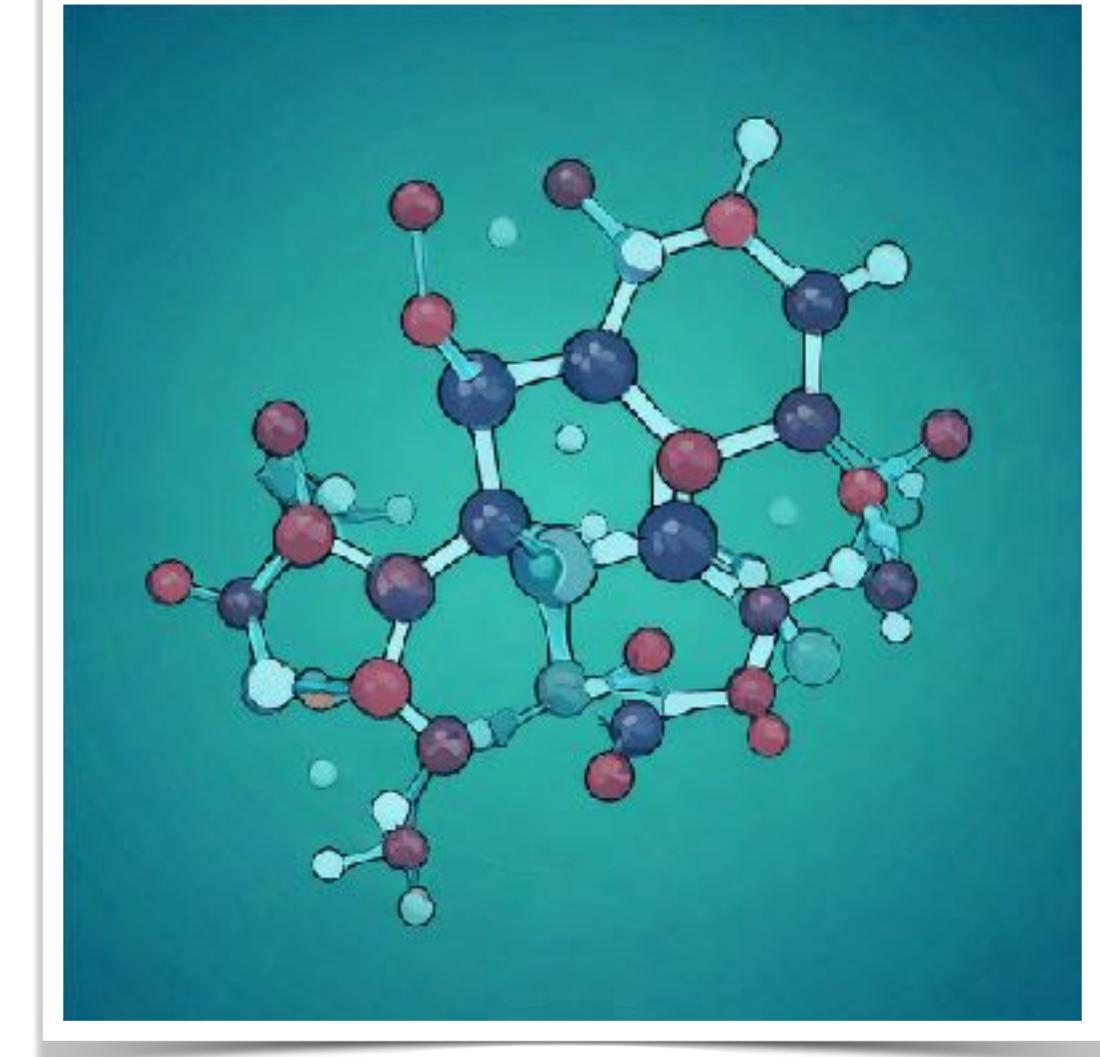
# Outline



Drug discovery in a nutshell



Foundations of diffusion models



Dreaming up useful molecules

# Molecule Representations

Text-based

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

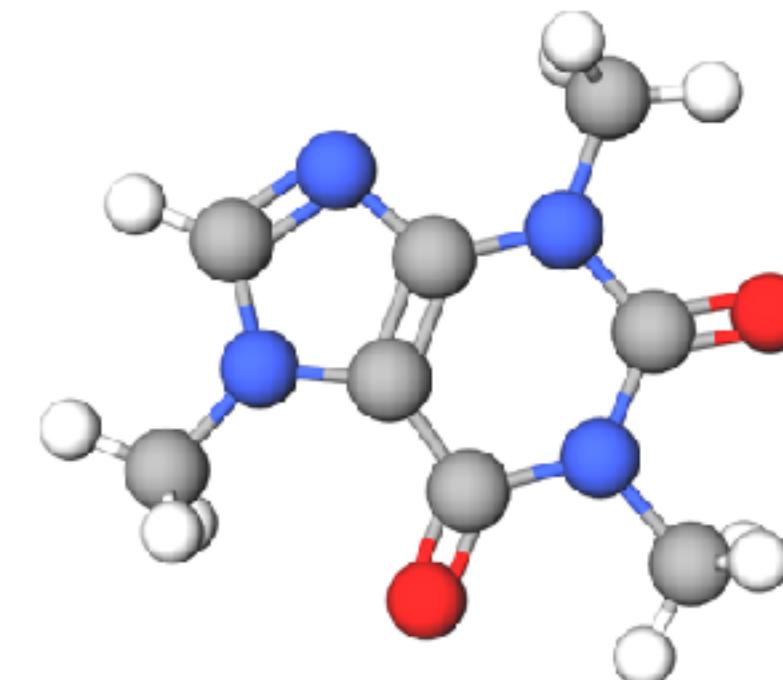
Smiles, Selfies, InChI, ...

Graph-based



Nodes = Atoms  
Node feature = element  
Edges = Covalent bonds  
Edge feature = Bond type

Point Cloud



- Nodes in the graph have additional **3D spatial annotations**.
- Unconnected points are represented without explicit edge (bond) information.

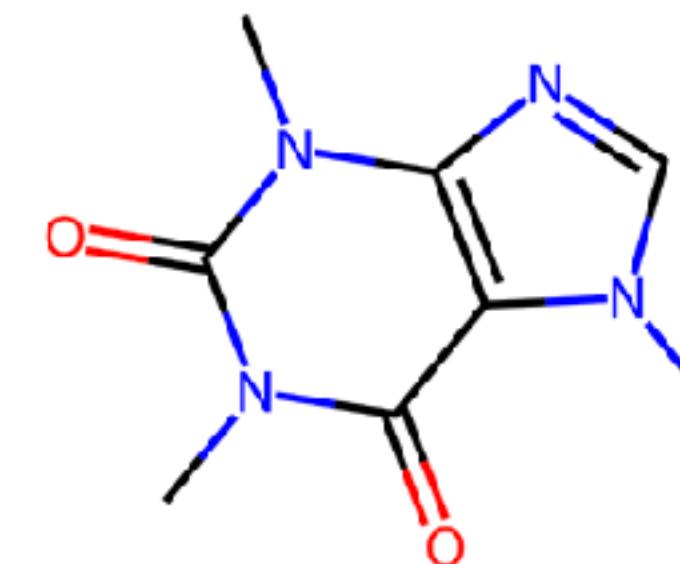
# Molecule Representations

Text-based

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

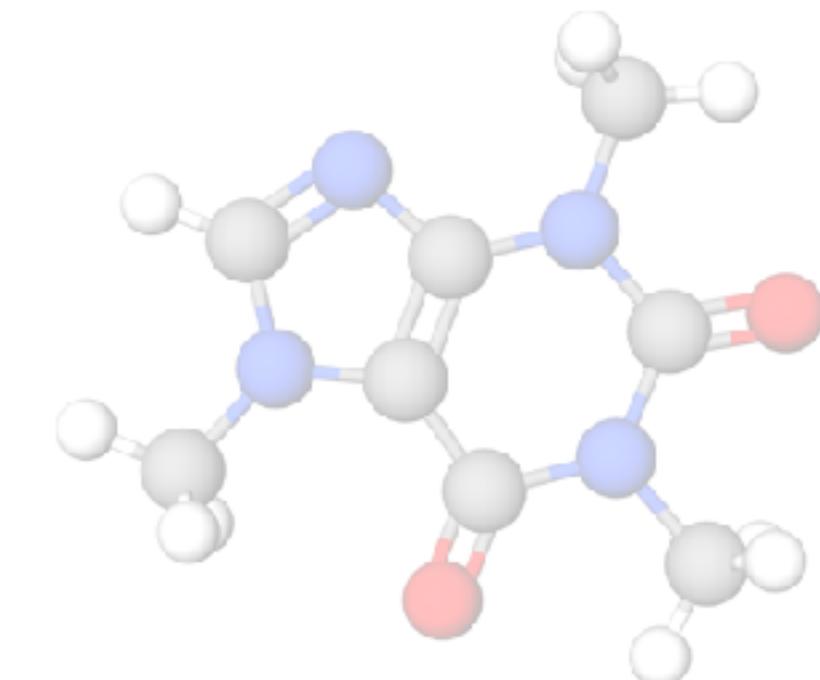
Smiles, Selfies, InChI, ...

Graph-based



Nodes = Atoms  
Node feature = element  
Edges = Covalent bonds  
Edge feature = Bond type

Point Cloud



- Nodes in the graph have additional **3D spatial annotations**.
- Unconnected points are represented without explicit edge (bond) information.

# Permutation Invariance

The diagram illustrates three different permutations of a simple 2x2 graph structure (two nodes connected by two parallel edges) and their corresponding 4x4 adjacency matrices.

Top row:

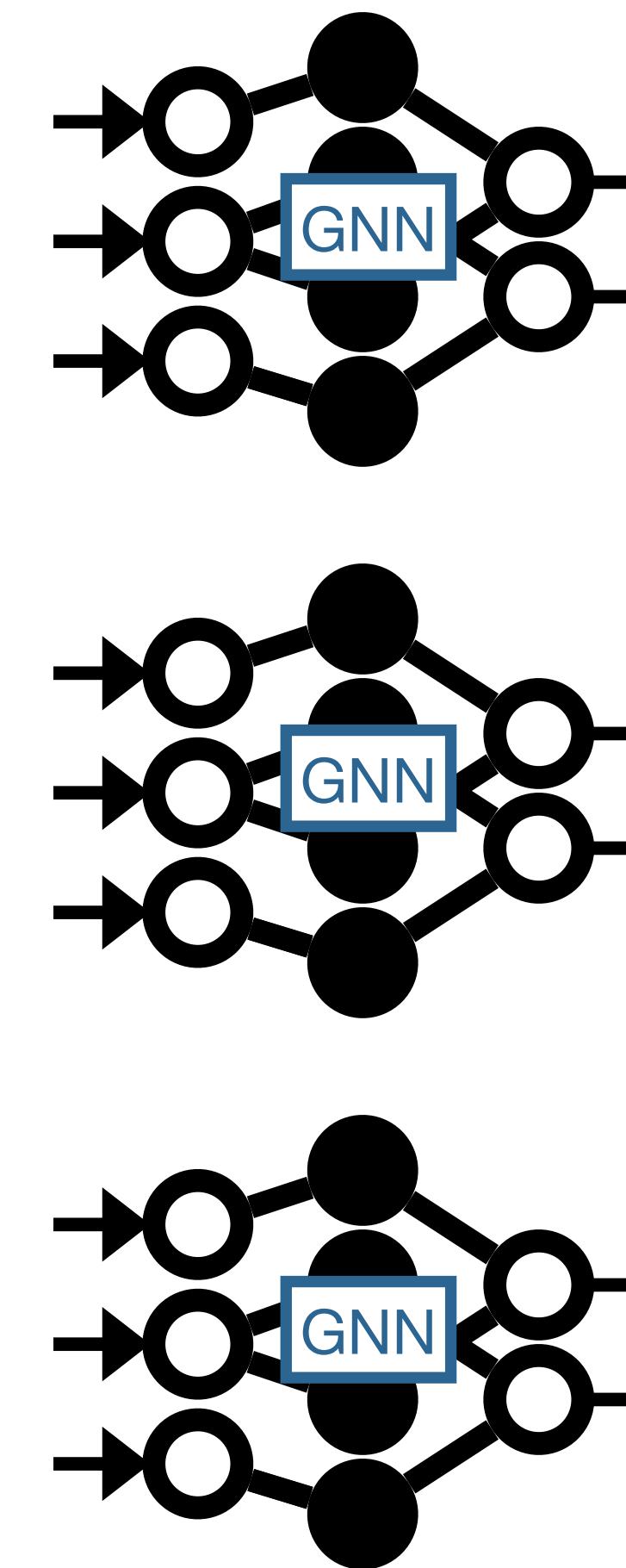
- Original graph: Two nodes connected by two parallel edges.
- Matrix:  $\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$

Middle row:

- Permuted graph 1: Nodes swapped vertically.
- Matrix:  $\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$

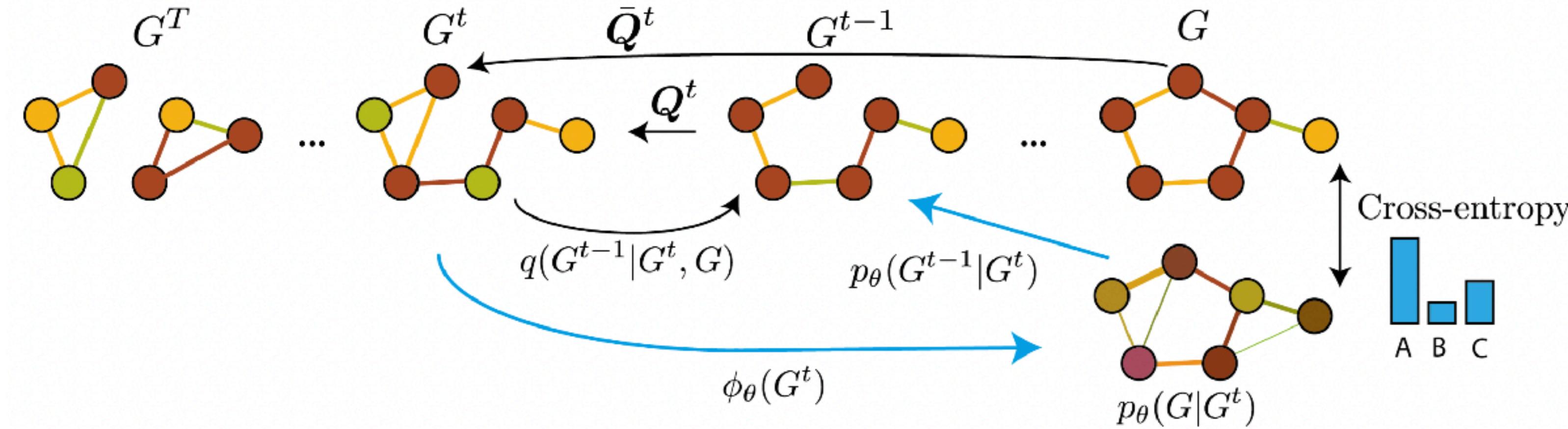
Bottom row:

- Permuted graph 2: Nodes swapped horizontally.
- Matrix:  $\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$



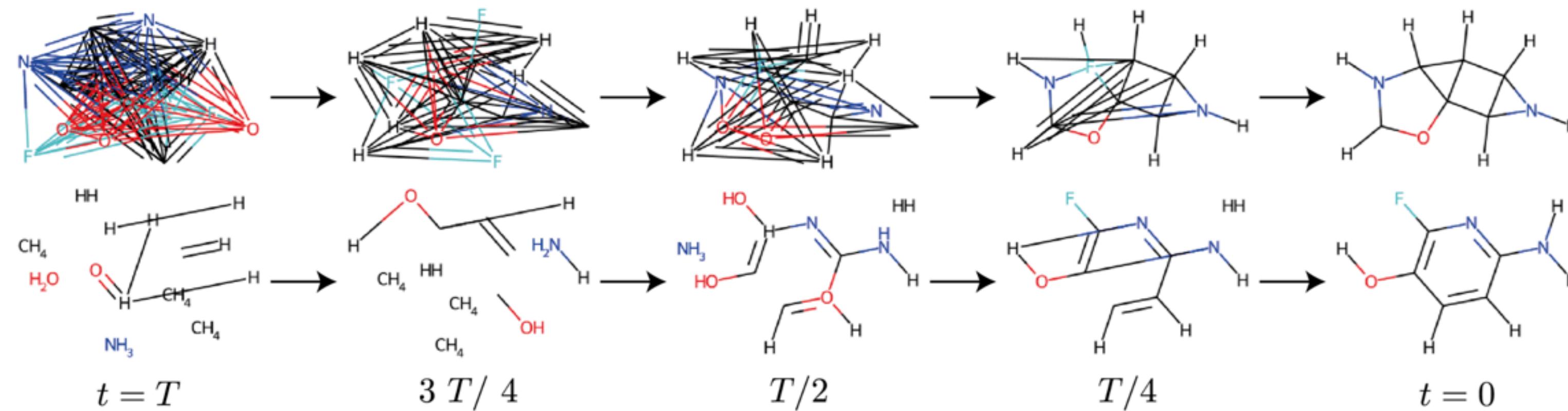
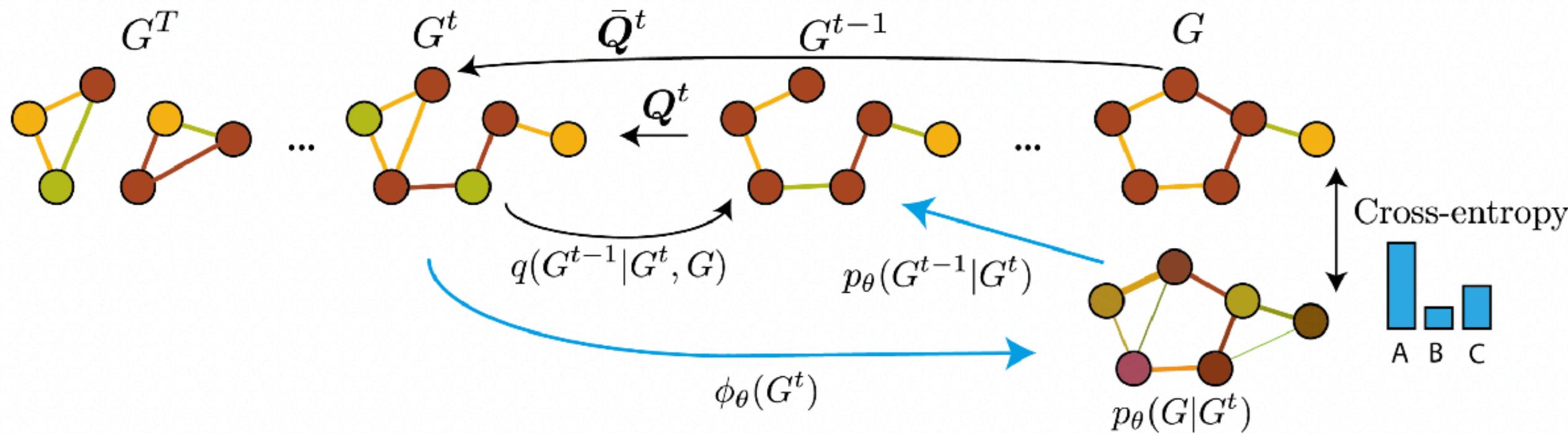
Guaranteed to  
produce the  
same results.

# DiGress

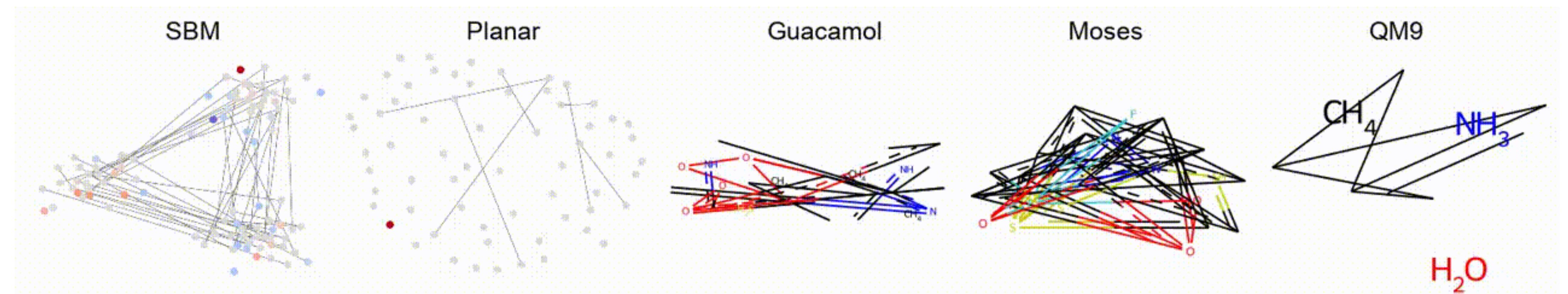


- Discrete diffusion by removing and adding edges
- Discrete jumps in categories (elements)

# DiGress



# DiGress



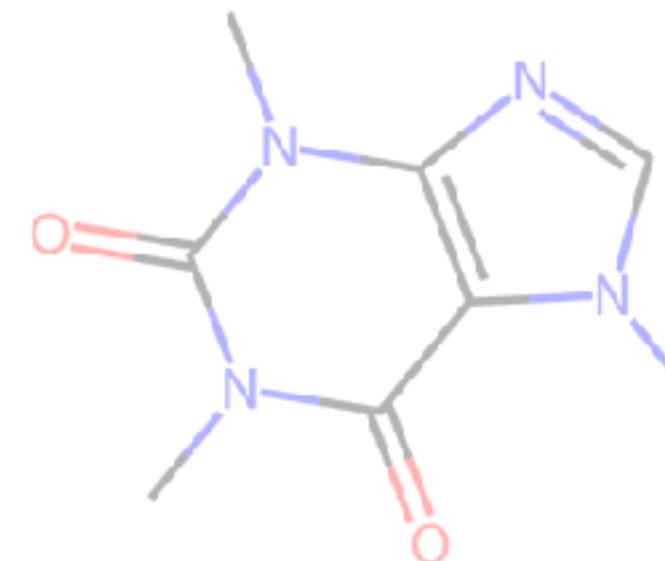
# Molecule Representations

Text-based

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

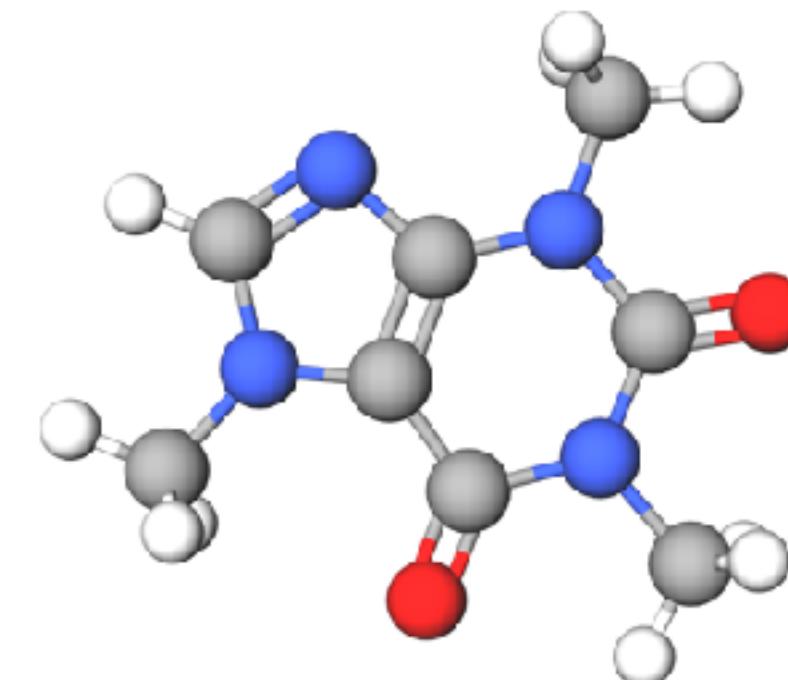
Smiles, Selfies, InChI, ...

Graph-based



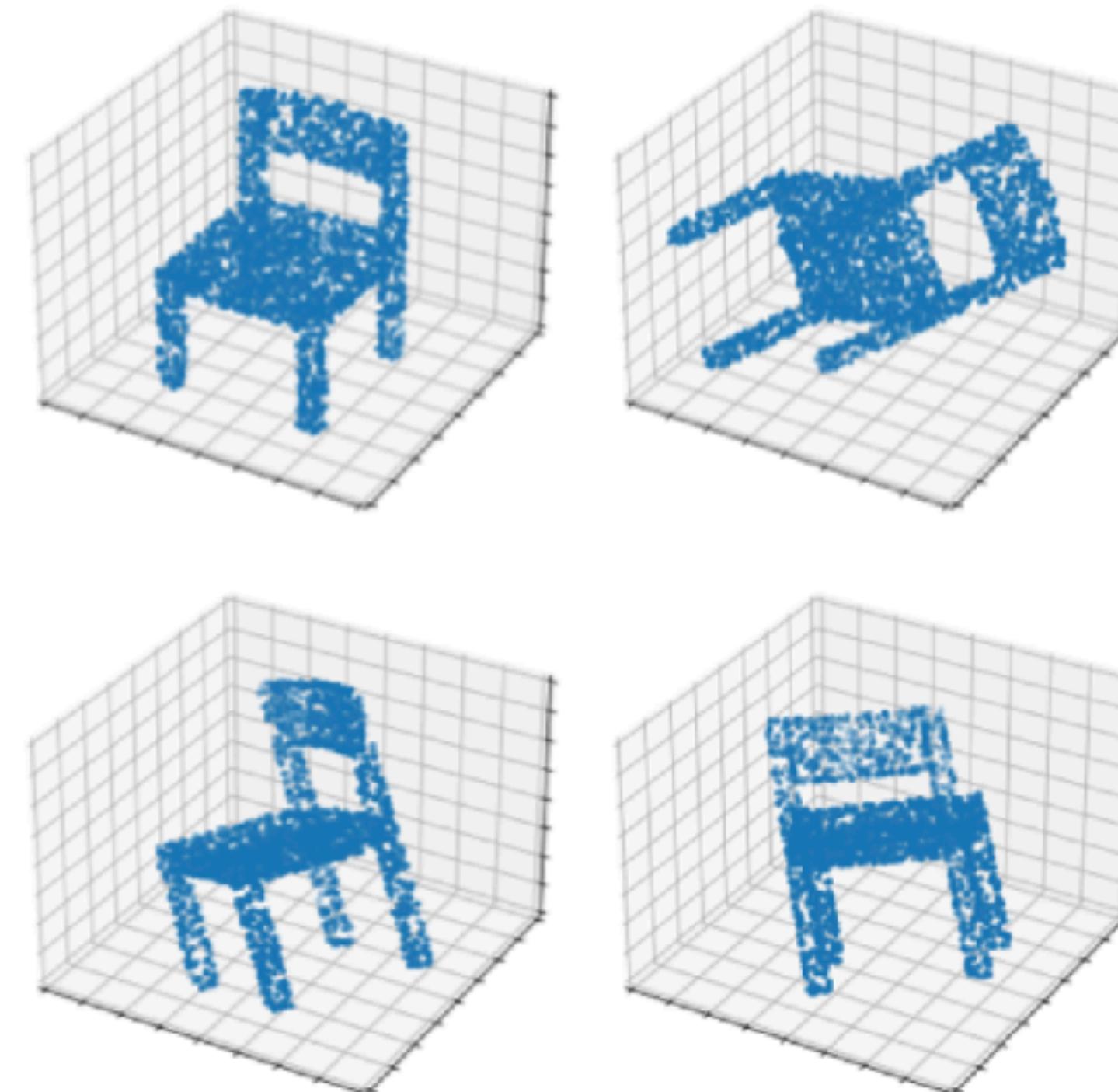
Nodes = Atoms  
Node feature = element  
Edges = Covalent bonds  
Edge feature = Bond type

Point Cloud



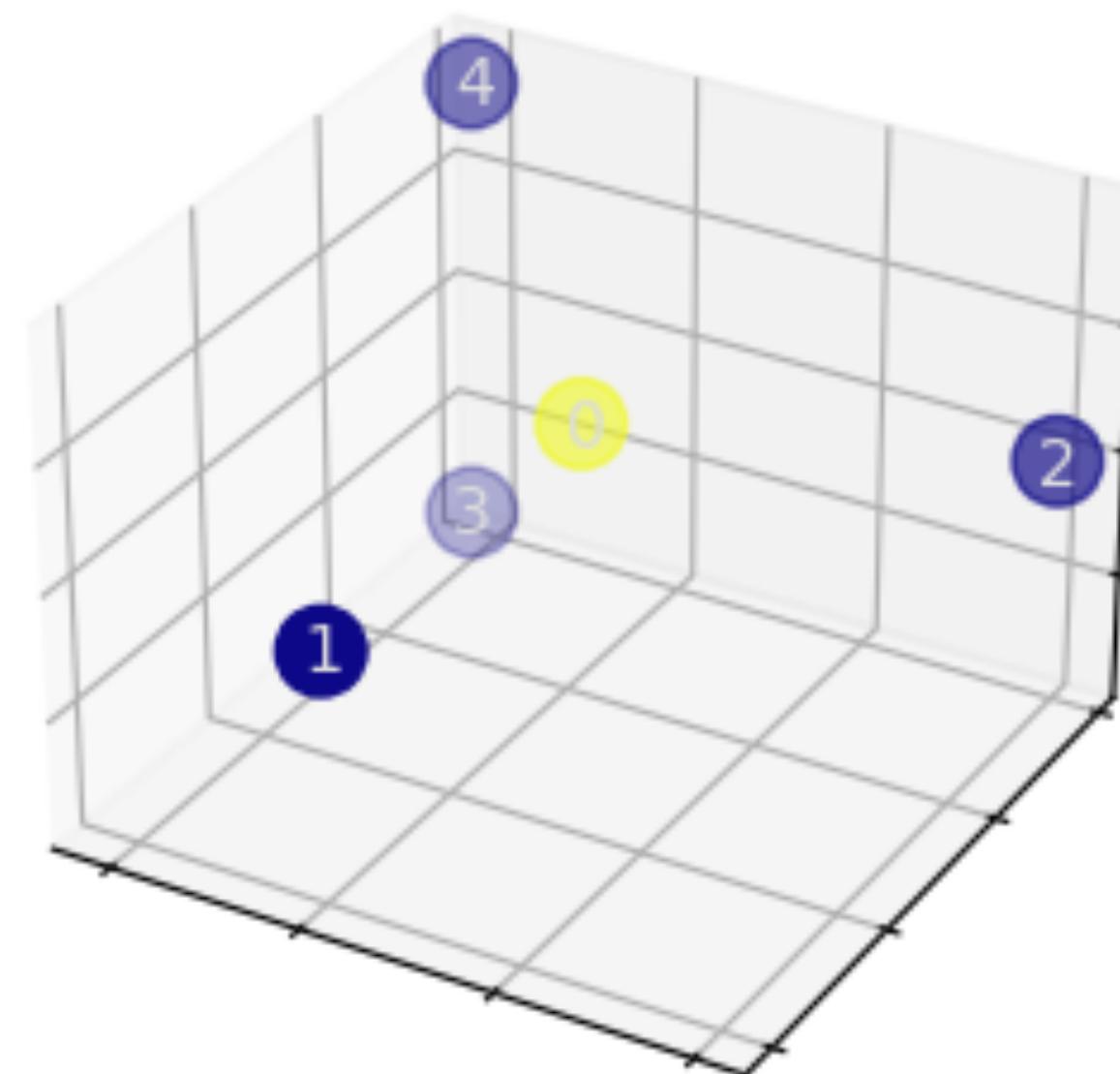
- Nodes in the graph have additional **3D spatial annotations**.
- Unconnected points are represented without explicit edge (bond) information.

# Point Clouds



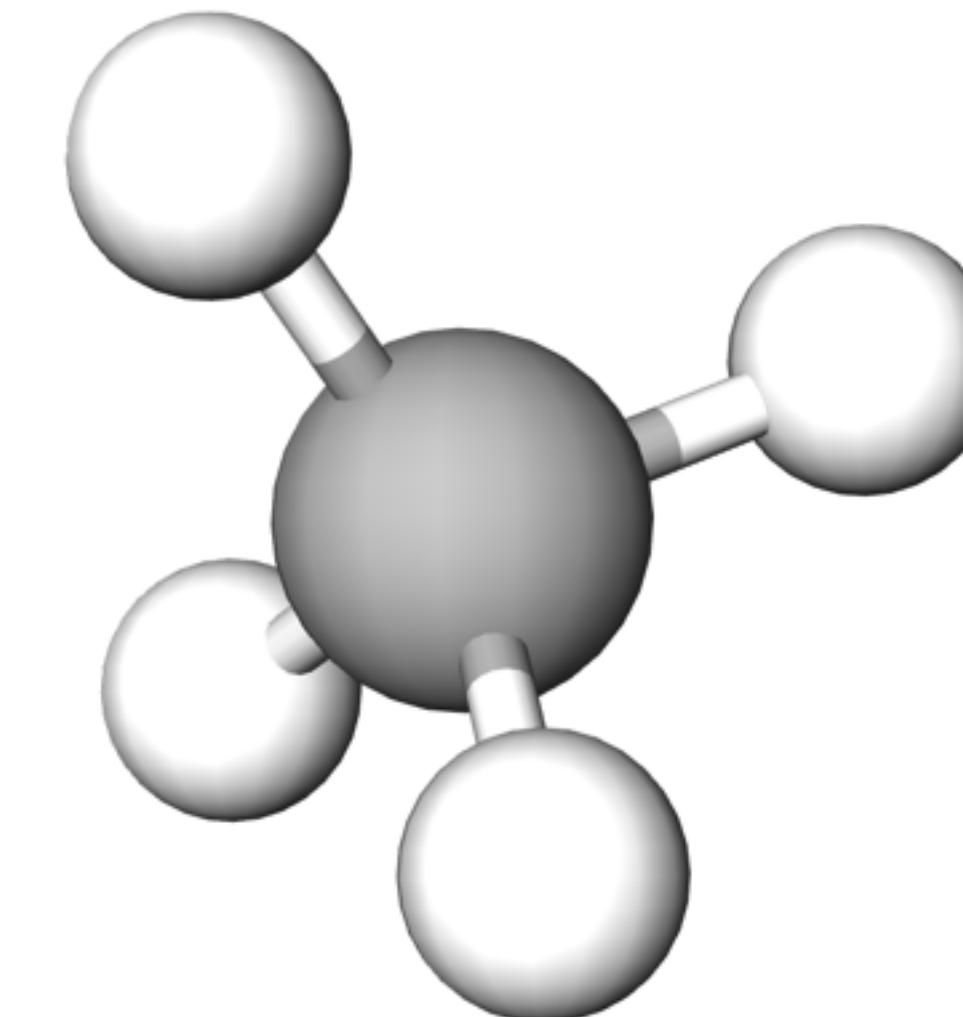
# Permutation Invariance

$\text{CH}_4$  (Methane)



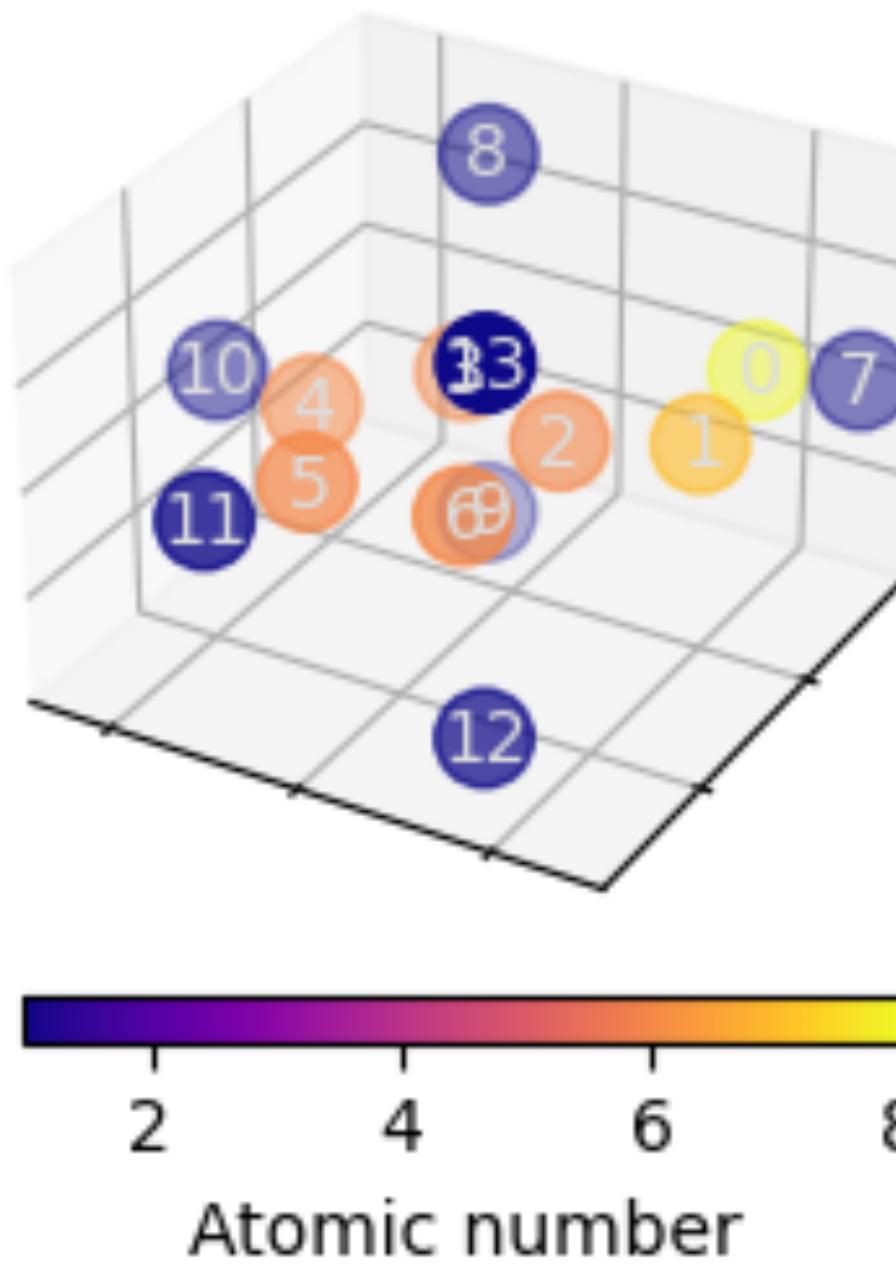
Hydrogen

Carbon

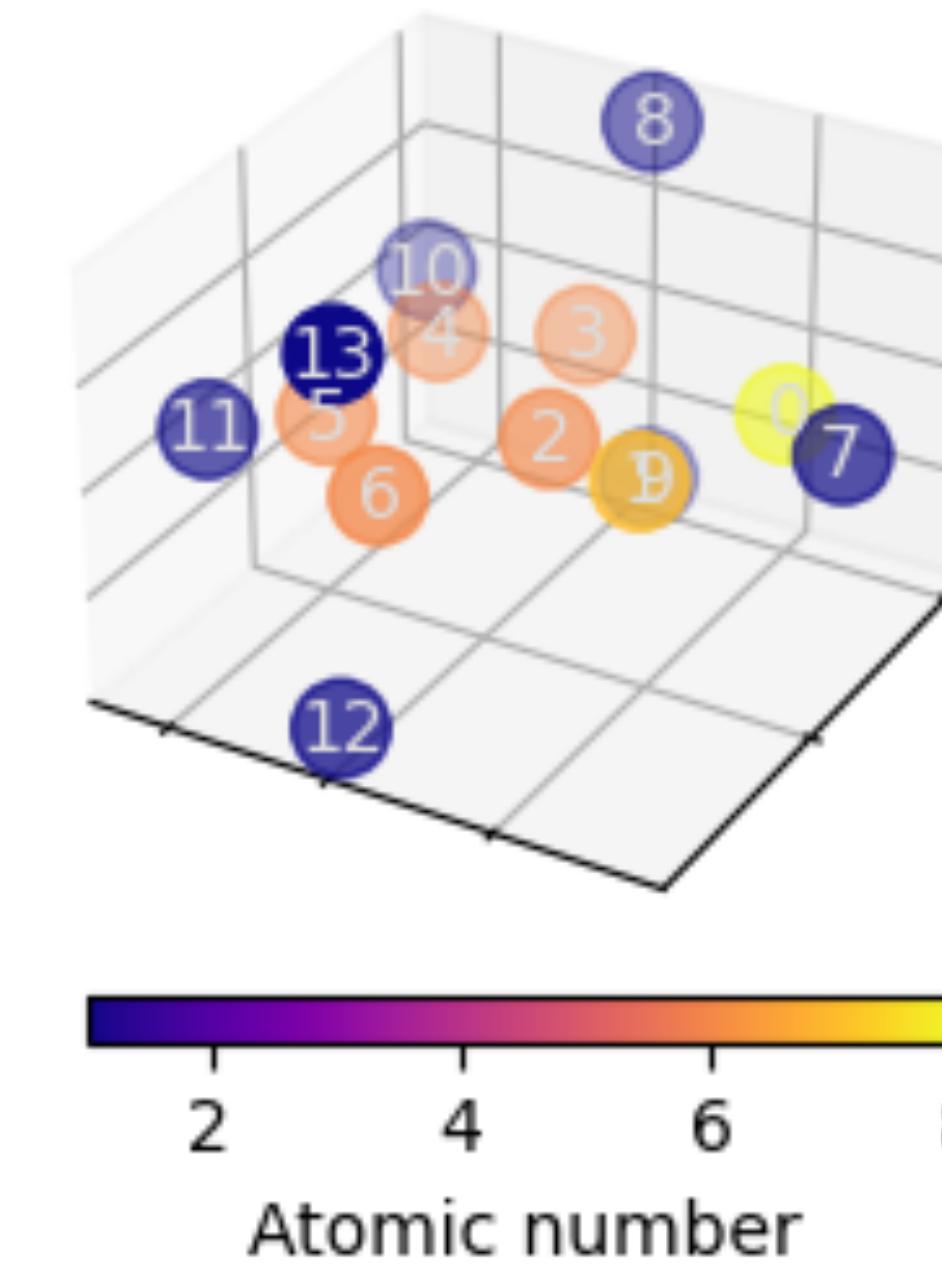


# Permutation Invariance

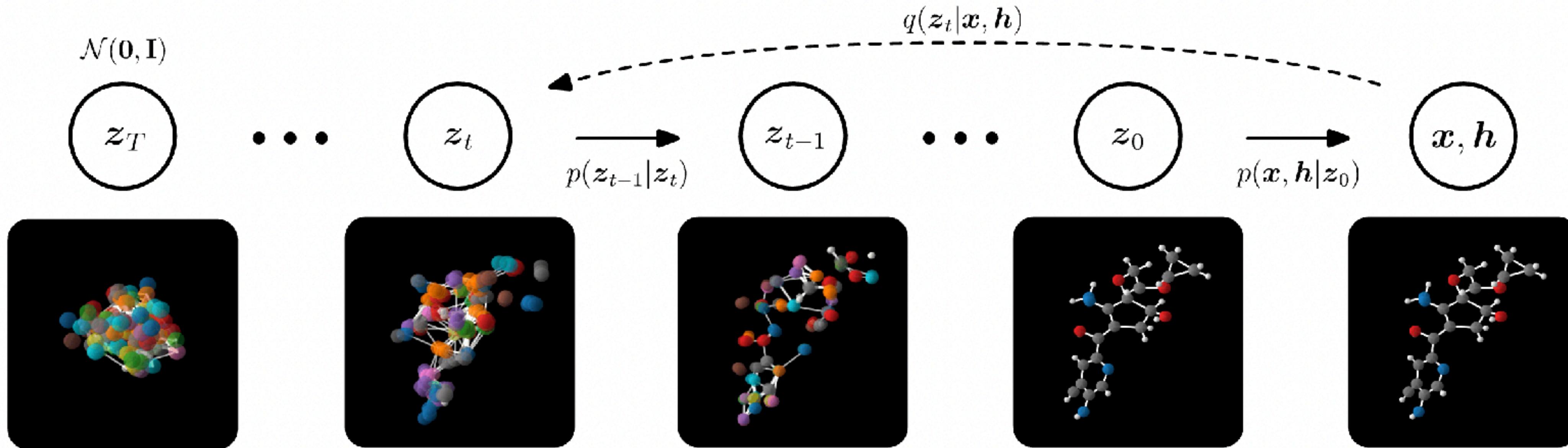
Sample input ( $X, Z$ )



Rotated input ( $X, g(Z)$ )

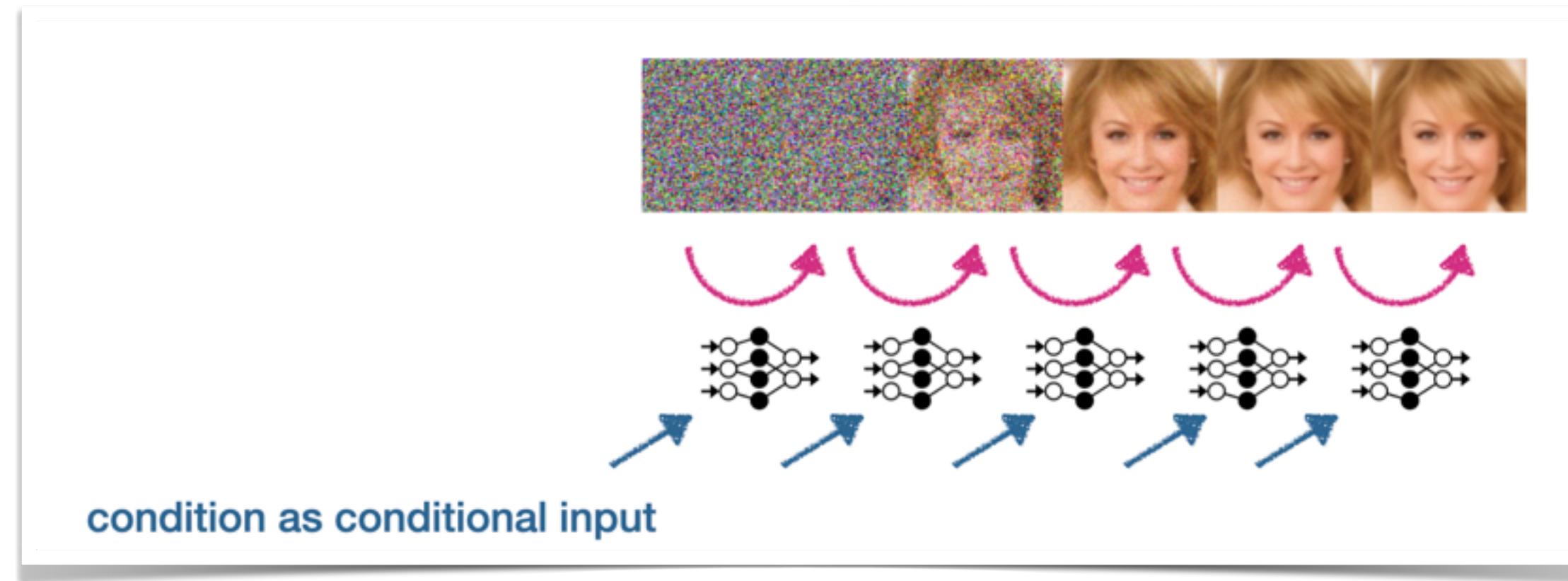


# Molecule Generation in 3D



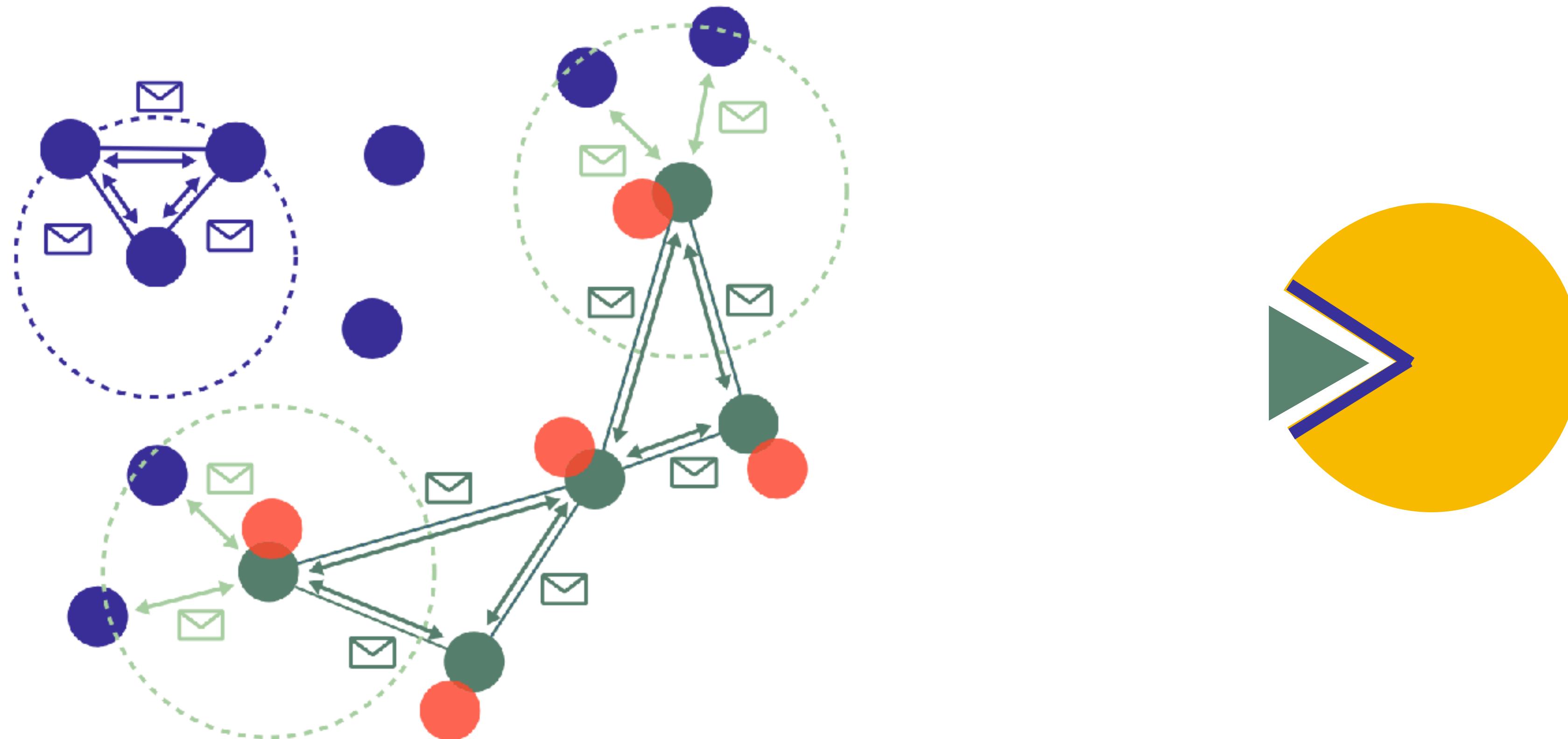
- First ones to achieve SOTA results for 3D molecule generation using diffusion

# Pocket Conditioning



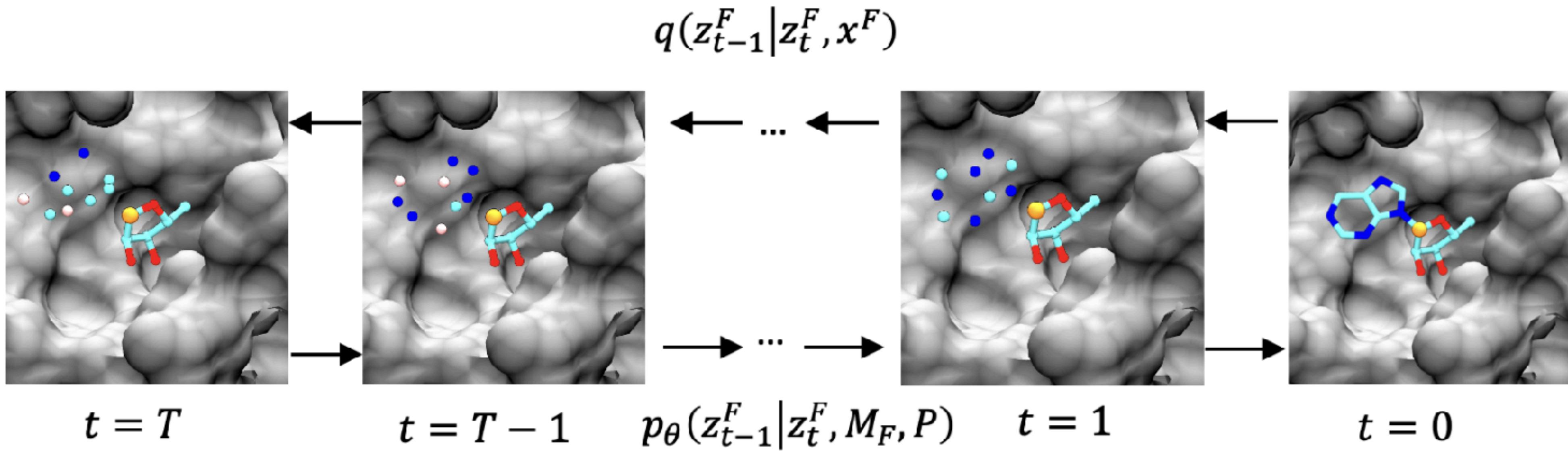
*It's a Match!*

# Pocket Conditioning



PILOT: Equivariant diffusion for pocket conditioned de novo ligand generation with multi-objective guidance via importance sampling (Cremer et al.)

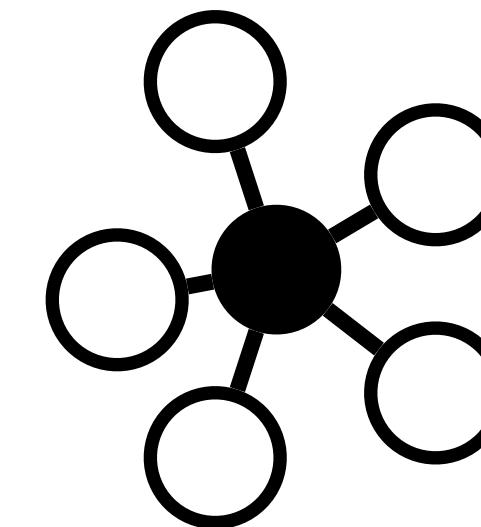
# Pocket Conditioning



- Conditioned generation on protein pockets
- Protein pocket is also processed as point cloud

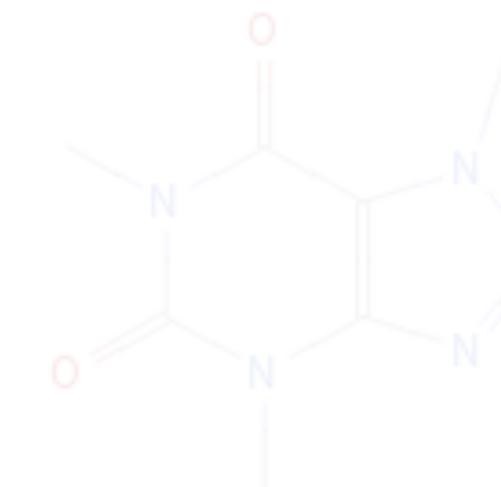
# Evaluation

- **Validity:** Percentage of molecules with correct **valency**.
- **Novelty:** Percentage of molecules **not in training dataset**
- **Uniqueness:** Percentage of generated unique molecule

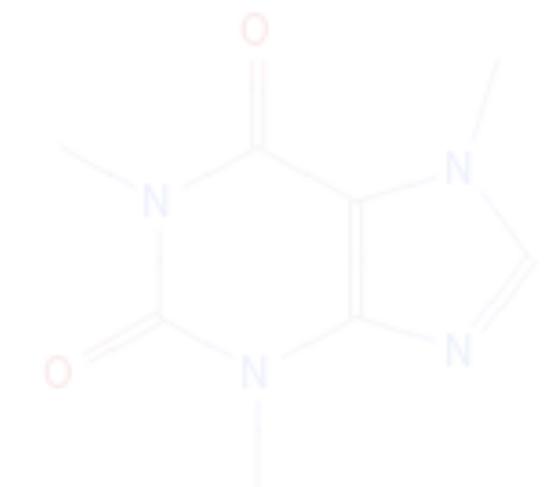


(Carbon atom with 5 hydrogen bonds)

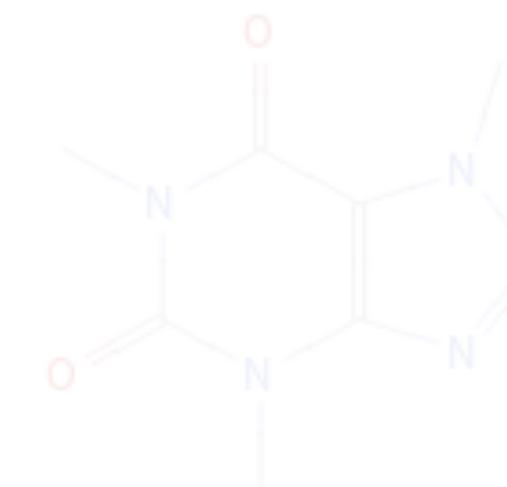
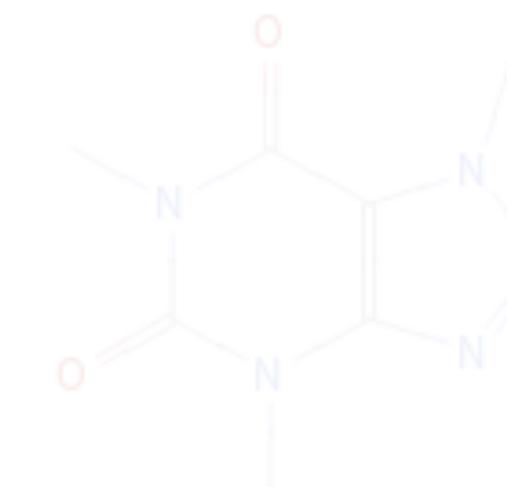
Train set:



Generated Molecule:

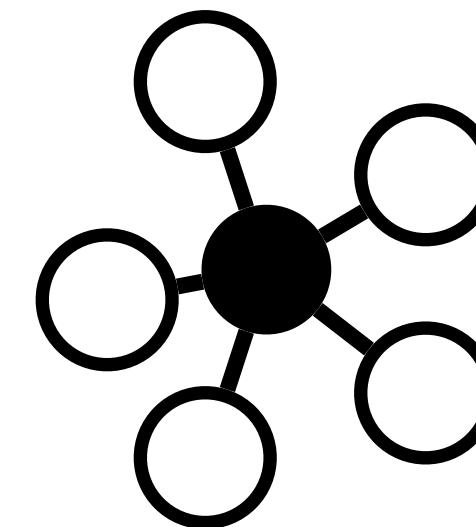


Generated Molecules:



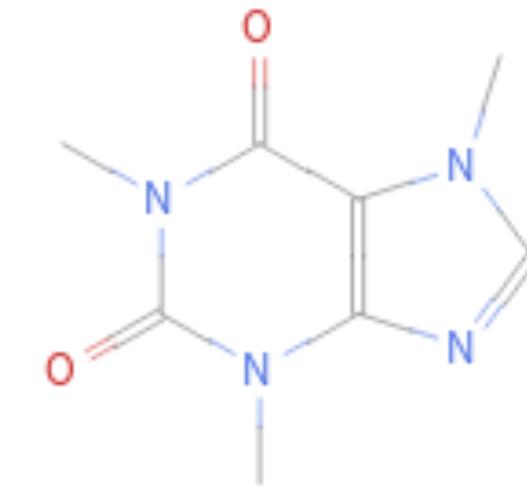
# Evaluation

- **Validity:** Percentage of molecules with correct **valency**.
- **Novelty:** Percentage of molecules **not in training dataset**
- **Uniqueness:** Percentage of generated unique molecule

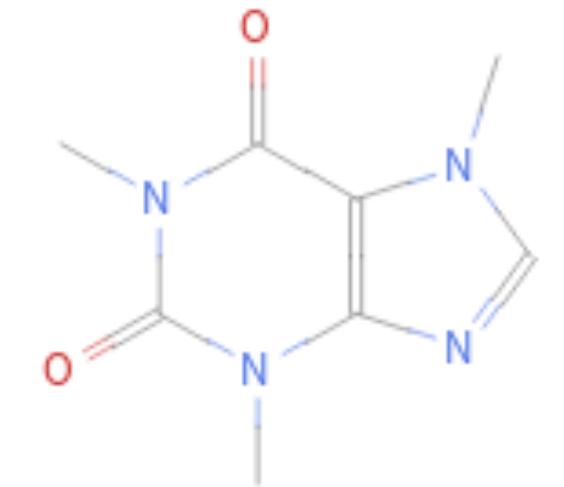


(Carbon atom with 5 hydrogen bonds)

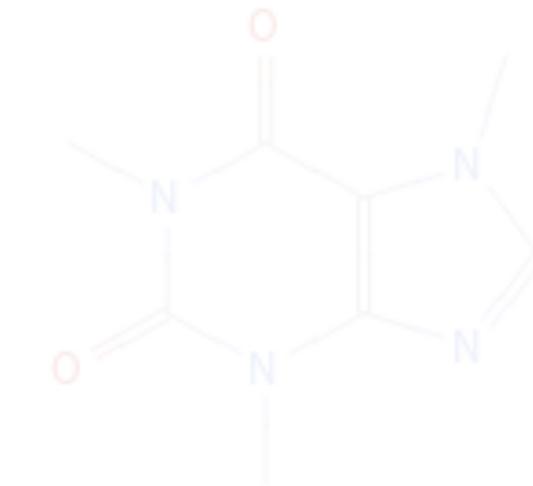
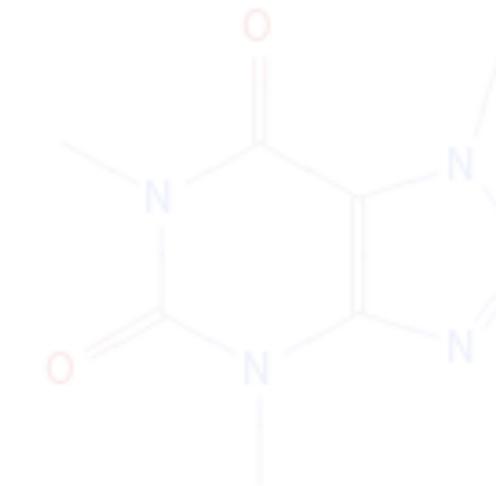
Train set:



Generated Molecule:

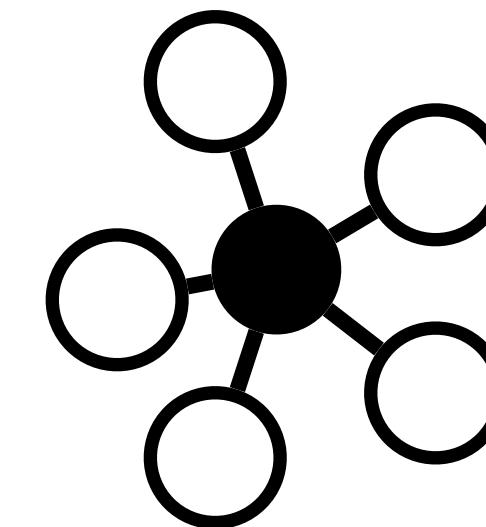


Generated Molecules:



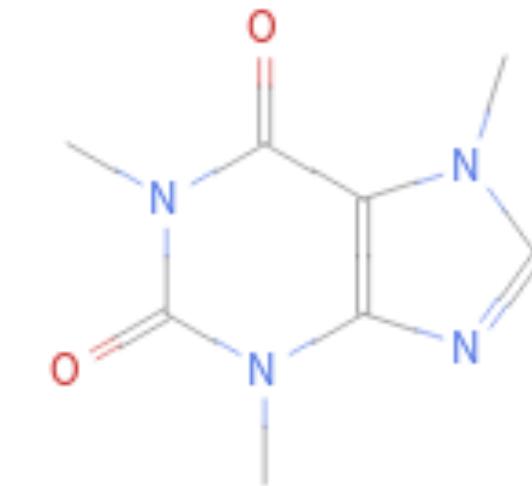
# Evaluation

- **Validity:** Percentage of molecules with correct **valency**.
- **Novelty:** Percentage of molecules **not in training dataset**
- **Uniqueness:** Percentage of generated unique molecule

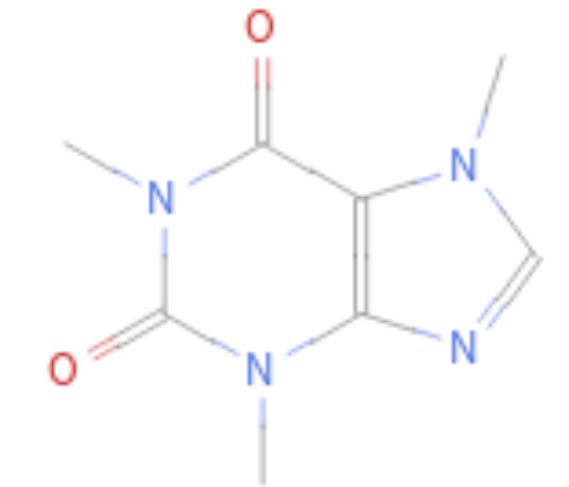


(Carbon atom with 5 hydrogen bonds)

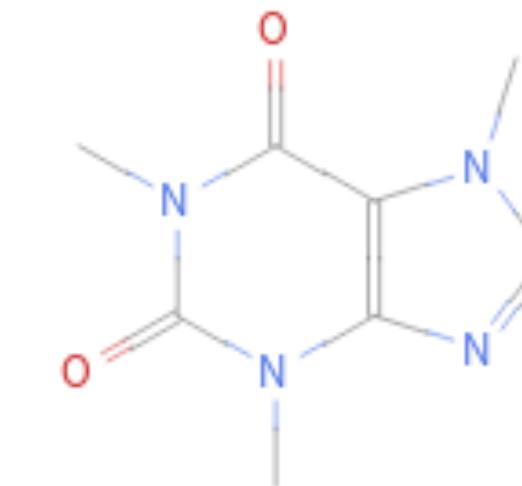
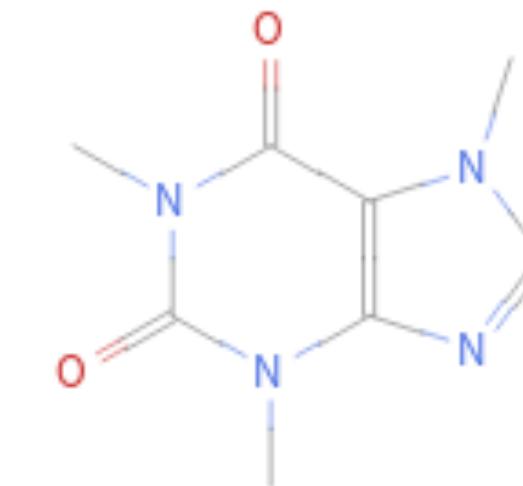
Train set:



Generated Molecule:



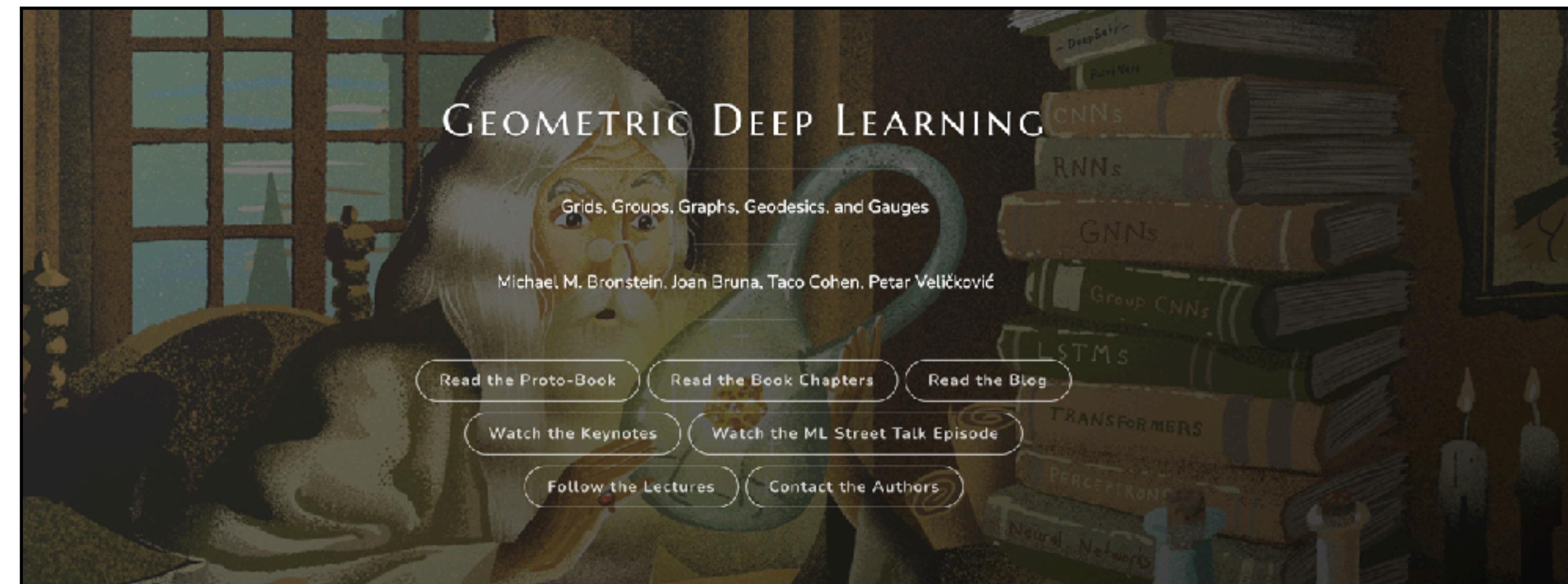
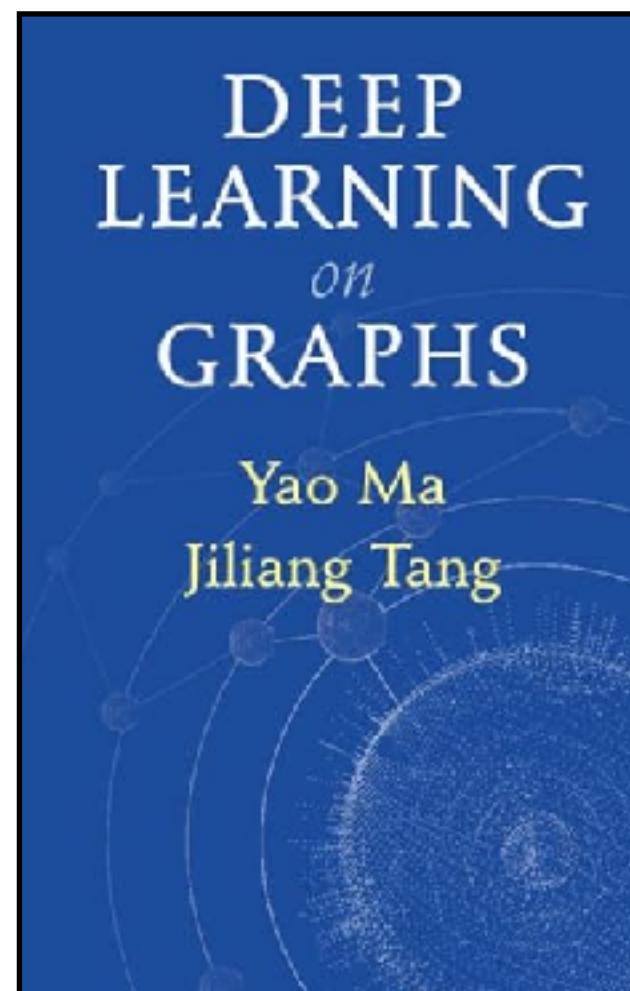
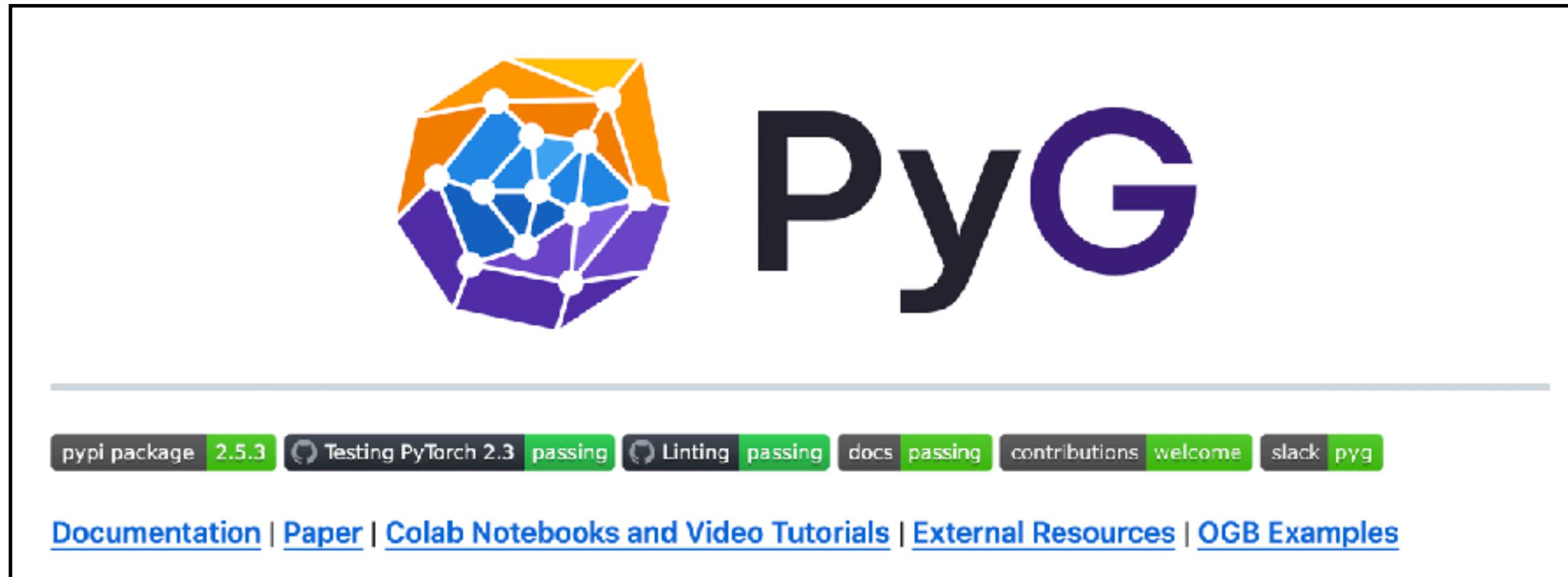
Generated Molecules:



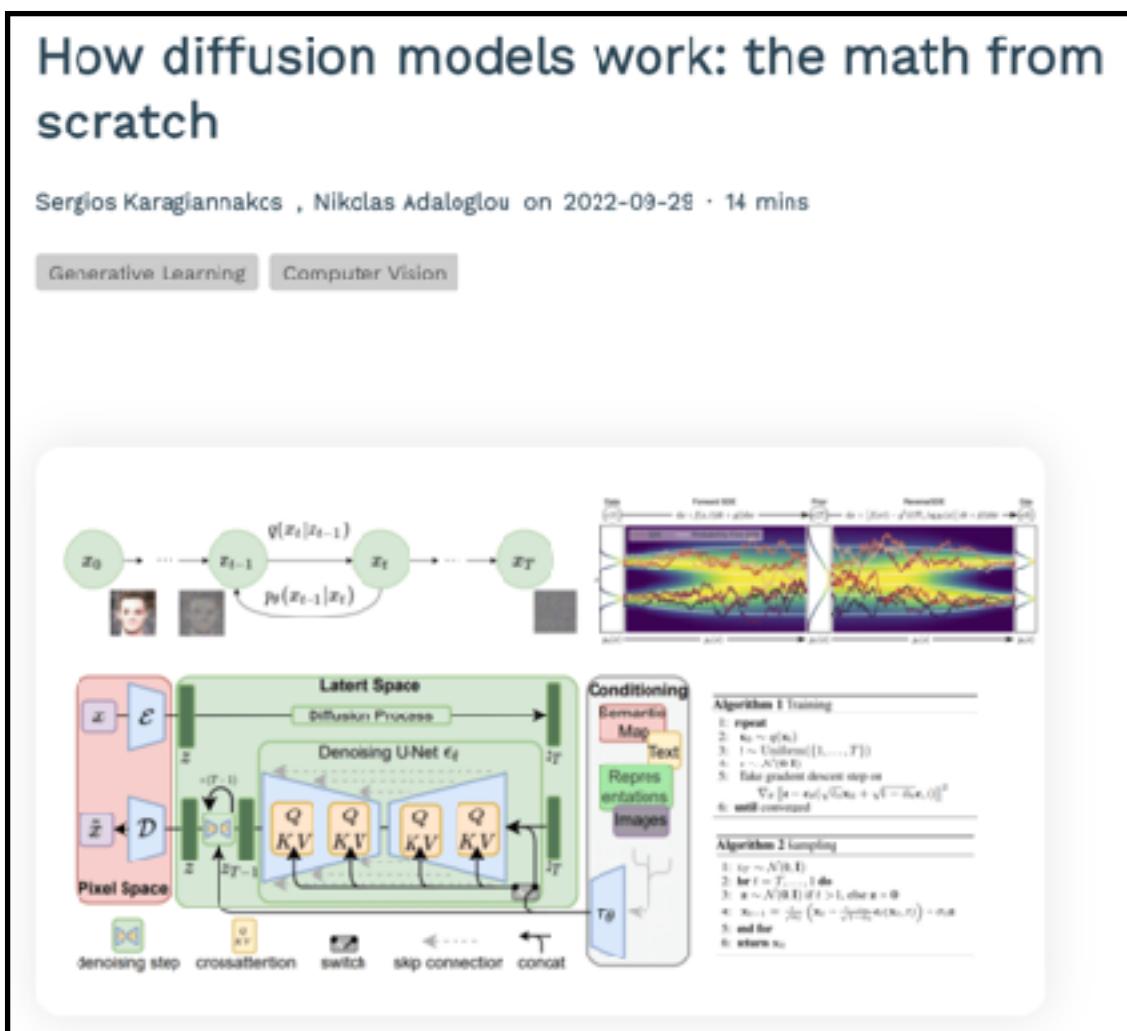
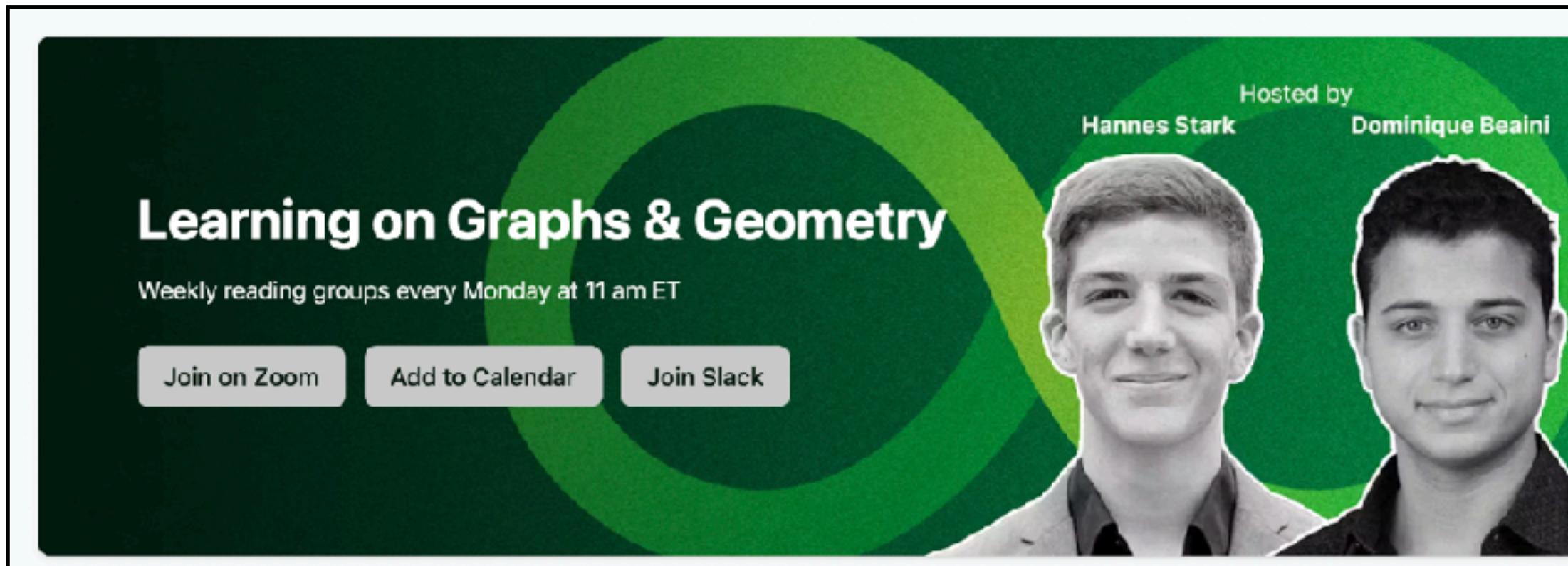
# Evaluation

- **Validity:** Percentage of molecules with correct **valency**.
- **Novelty:** Percentage of molecules **not in training dataset**
- **Uniqueness:** Percentage of generated unique molecule
  
- **Lipophilicity Assessment:** Measures absorption
- **Quantitative Estimate of Druglikeness:** Based on many molecular properties
- **Synthetic Accessibility Score:** Fragment scores + complexity penalty
- **Lipinski Rule of Five:** Drug-likeness based on molecular structure

# Getting Started



# Diffusion Models



Hugging Face

← Back to Articles

## The Annotated Diffusion Model

Published June 7, 2022

Update on GitHub

▲ Upvote 33 +27

nielsx Niels Rogge kashif Kashif Rasul

Open in Colab

Awesome-Diffusion-Models Public

diffusion Add files via upload 0 Tags 01b3394 · 7 months ago 932 Commits

docs rename title last year

template add website interface last year

website add website interface last year

README.md Add files via upload 7 months ago

license Create license 2 years ago

README MIT license

View raw License PDF Place with raw

This repository contains a collection of resources and papers on Diffusion Models.

Please refer to [this page](#) as this page may not contain all the information due to page constraints.

Contents

- Resources
  - Introductory Posts
  - Introductory Papers
  - Introductory Videos
  - Introductory Lectures
  - Tutorial and Jupyter Notebooks

# Computational Drug Discovery

**TeachOpenCADD**

A teaching platform for computer-aided drug design (CADD) using open source packages and data.

Project TeachOpenCADD 

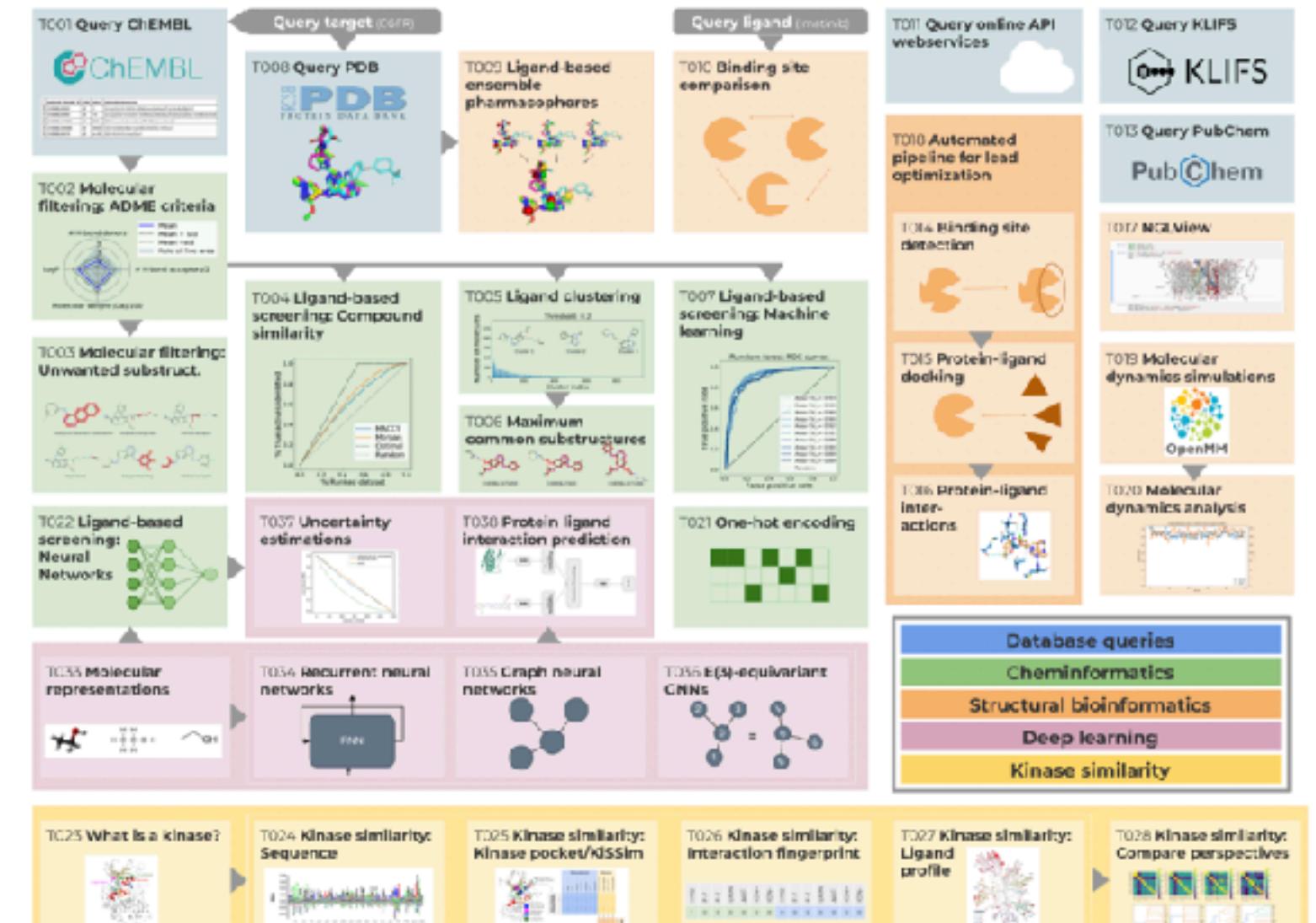
  License CC BY 4.0 tag v2023.05.2   downloads 9k total

closed pull requests 233 | open pull requests 11 | closed issues 111 | issues 42 open

If you use TeachOpenCADD in a publication, please [cite](#) us! If you use TeachOpenCADD in class, please include a link back to our repository.

In any case, please [star](#) (and tell your students to star) those repositories you consider useful for your learning/teaching activities.

### Description



The diagram illustrates the TeachOpenCADD workflow, organized into several main sections:

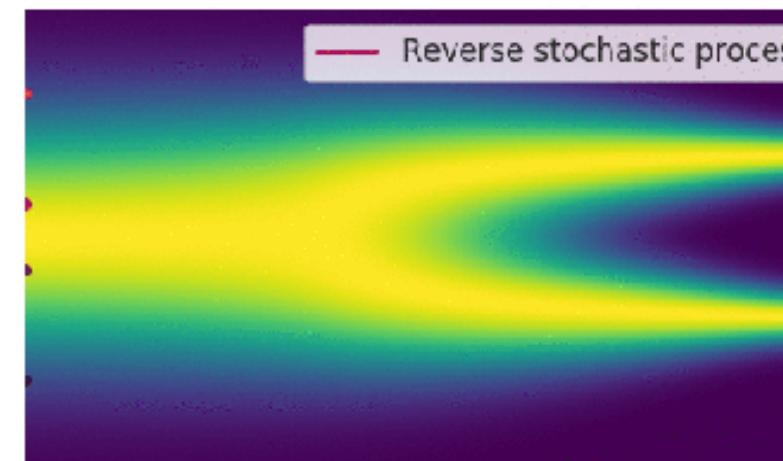
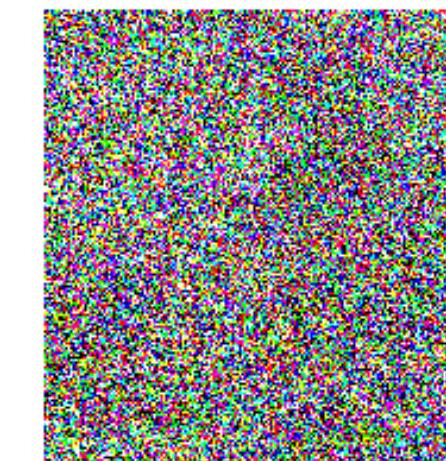
- Input and Database Queries:** T001 Query CHEMBL, T008 Query PDB, T011 Query online API webservices, T012 Query KLIFS, T013 Query PubChem, T014 Binding site detection, T017 MC3 View.
- Molecular Filtering:** T002 Molecular filtering: ADME criteria, T003 Molecular filtering: Unwanted substr., T004 Ligand-based screening: Compound similarity, T005 Ligand clustering, T006 Maximum common substructures, T007 Ligand-based screening: Machine learning, T010 Automated pipeline for lead optimization, T015 Protein-ligand docking, T019 Molecular dynamics simulations, T020 Molecular dynamics analysis.
- Machine Learning and Representations:** T022 Ligand-based screening: Neural Networks, T023 Molecular representations, T024 Recurrent neural networks, T025 Graph neural networks, T026 E(3)-equivariant CNNs.
- Kinase Similarity Analysis:** T027 Kinase similarity: Ligand profile, T028 Kinase similarity: Compare perspectives.
- Structural Bioinformatics and Cheminformatics:** T009 Ligand-based ensemble pharmacophores, T016 One-hot encoding, T021 Kinase similarity: Sequence, T022 Kinase similarity: Kinase pocket/kS3M, T023 Kinase similarity: Interaction fingerprint.

Below the main flowchart, there is a legend categorizing the components:

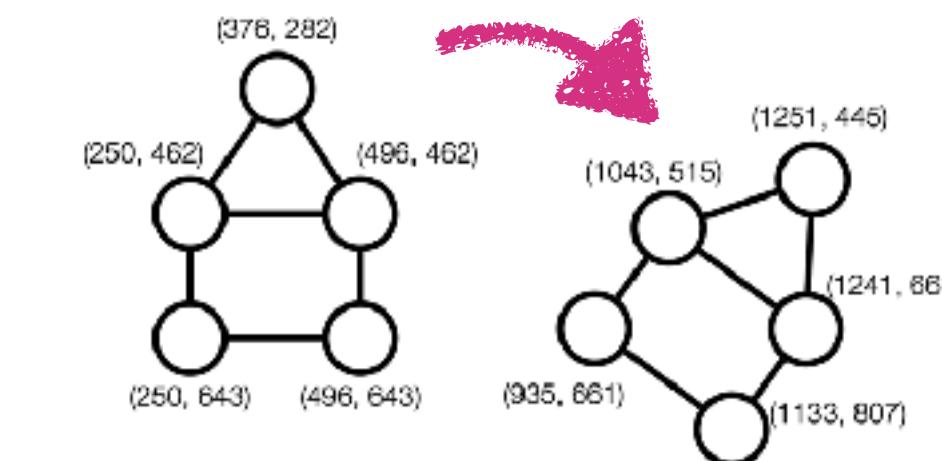
- Database queries
- Cheminformatics
- Structural bioinformatics
- Deep learning
- Kinase similarity

# Summary

Method:

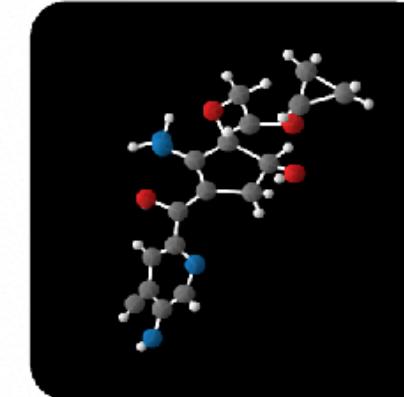


Diffusion Models

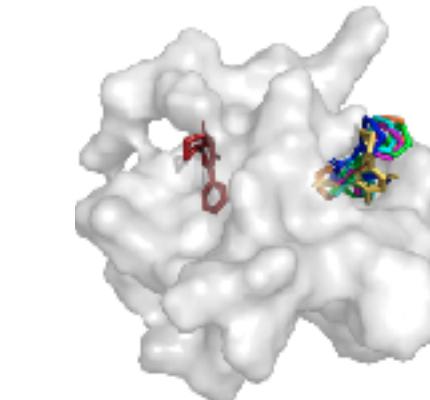


(Geometric) Graph Neural Networks

Applications:



Drug design

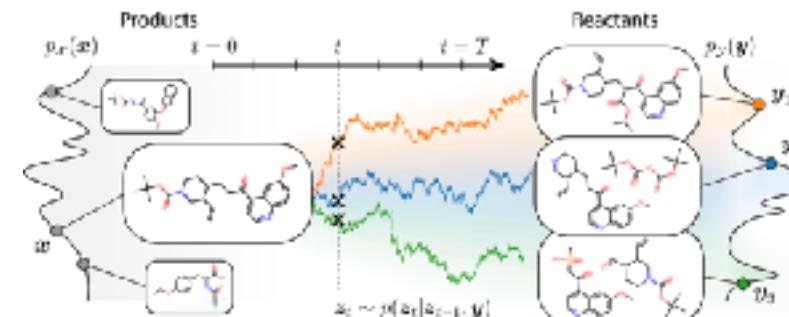


Docking



Structure Prediction

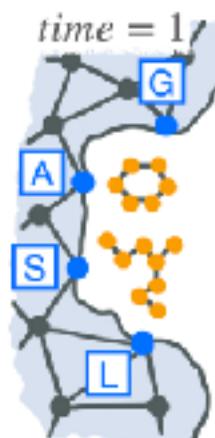
Future Directions:



Domain knowledge



Benchmarks & Software



time = 1  
Better conditioning

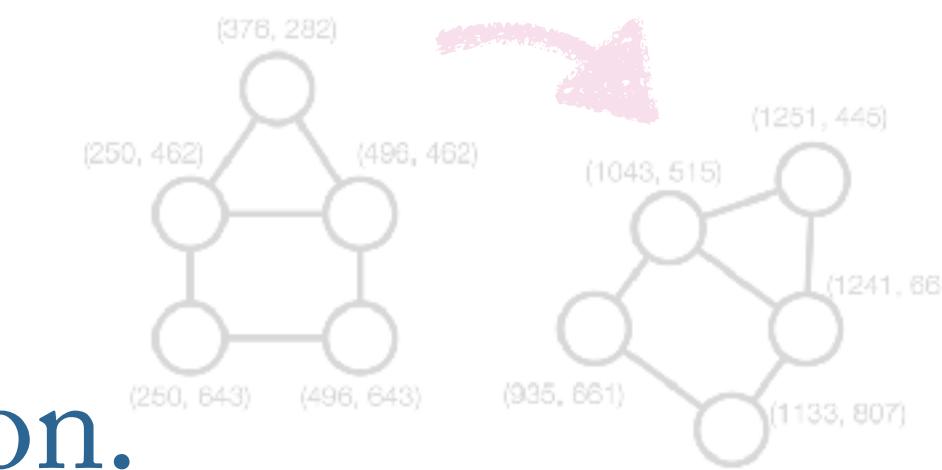
# Summary

Method:



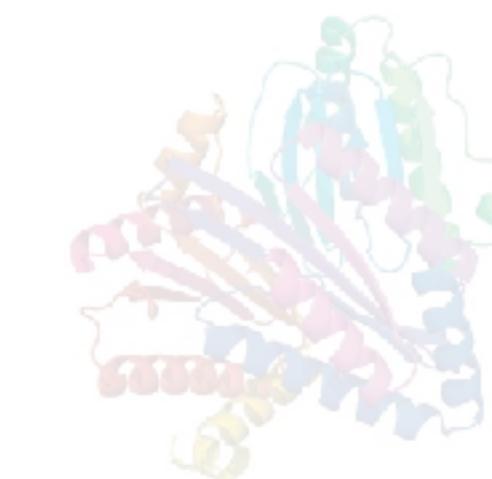
Thank you for your attention.

Diffusion Models



(Geometric) Graph Neural Networks

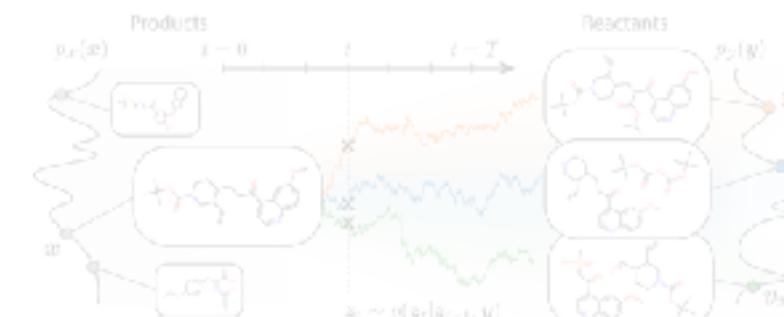
Applications:



Drug design? Docking? Structure Prediction?

[gerritgrossmann.de/#cerfa](http://gerritgrossmann.de/#cerfa)

Future Directions:



Domain knowledge



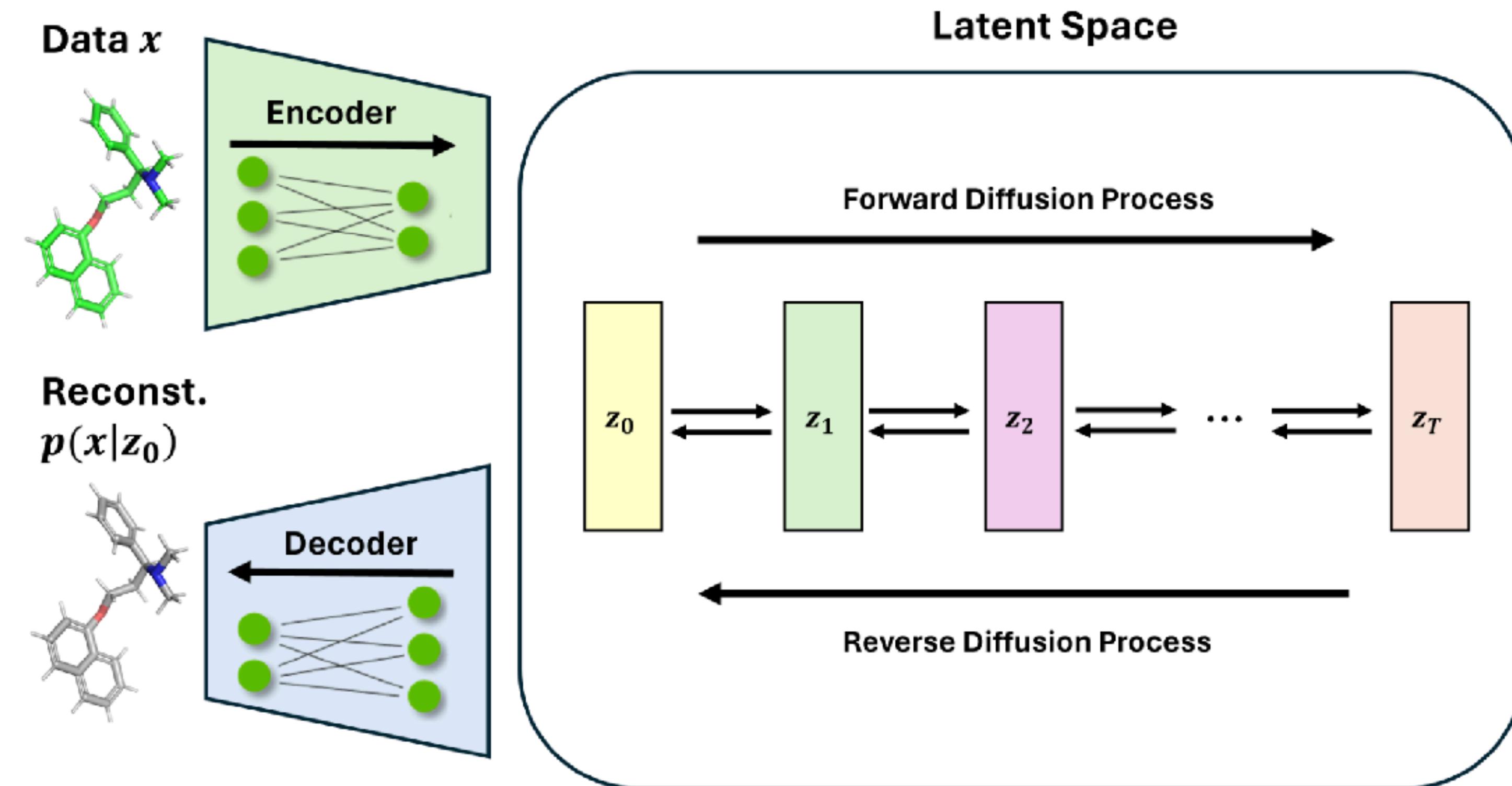
[pytorch-geometric.readthedocs.io/en/latest/](https://pytorch-geometric.readthedocs.io/en/latest/)

Documentation | Paper | Colab Notebooks and Video Tutorials | External Resources | OGB Examples

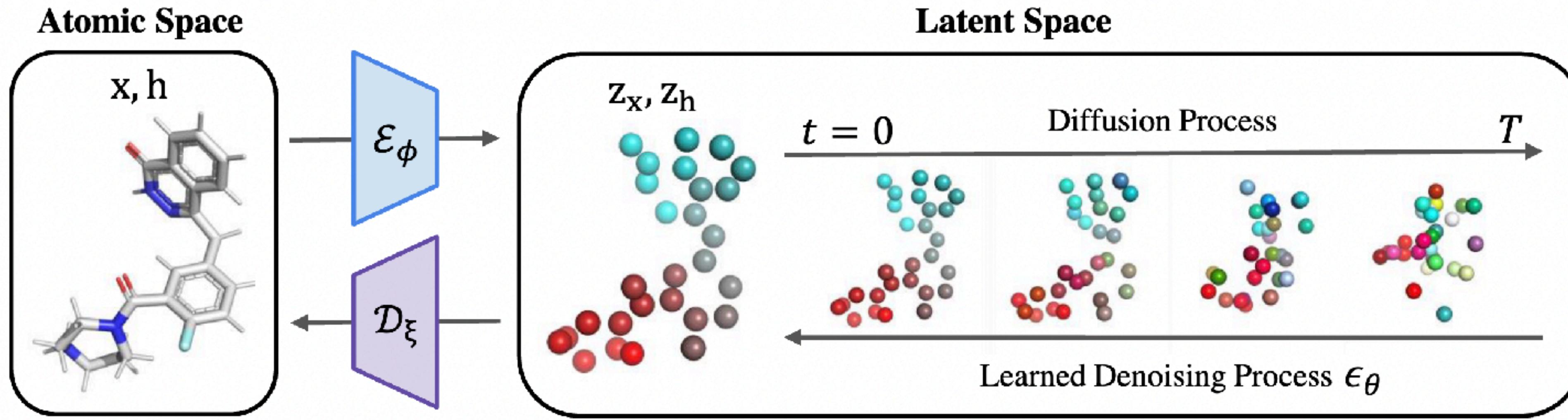


Better conditioning

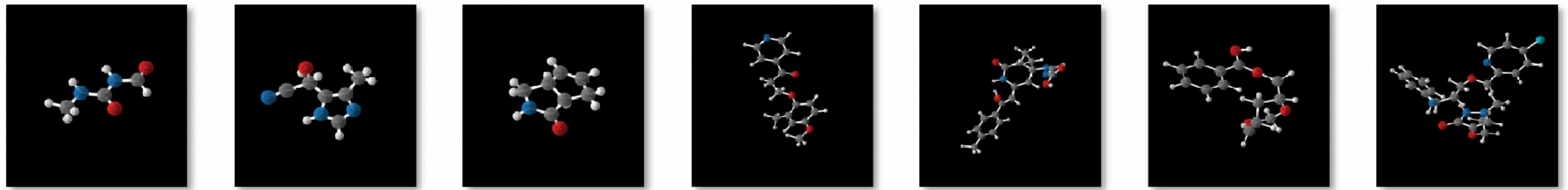
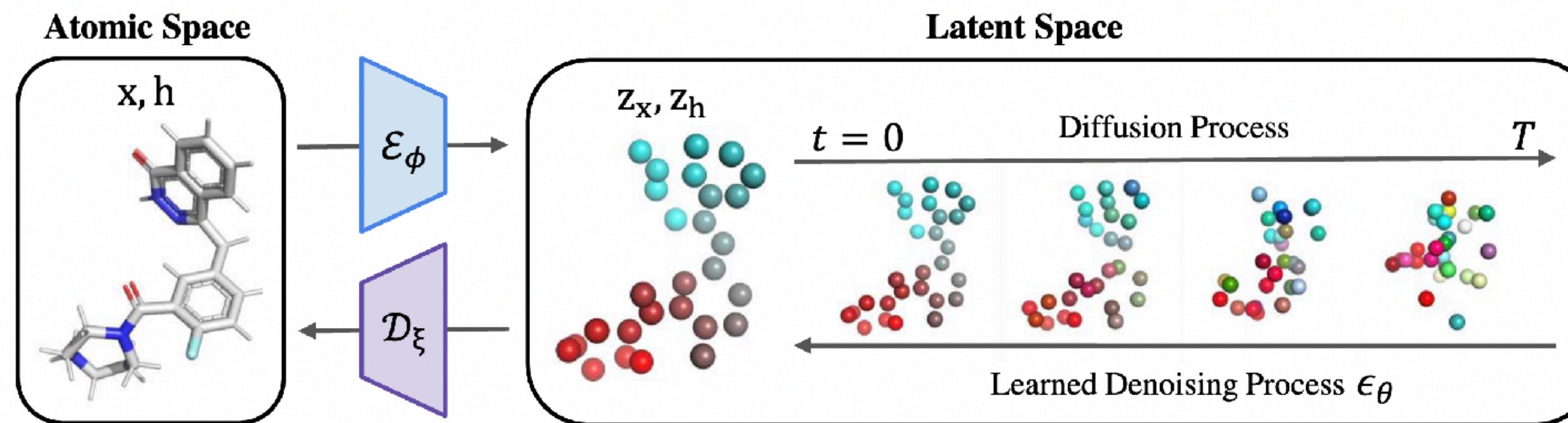
# Latent Diffusion

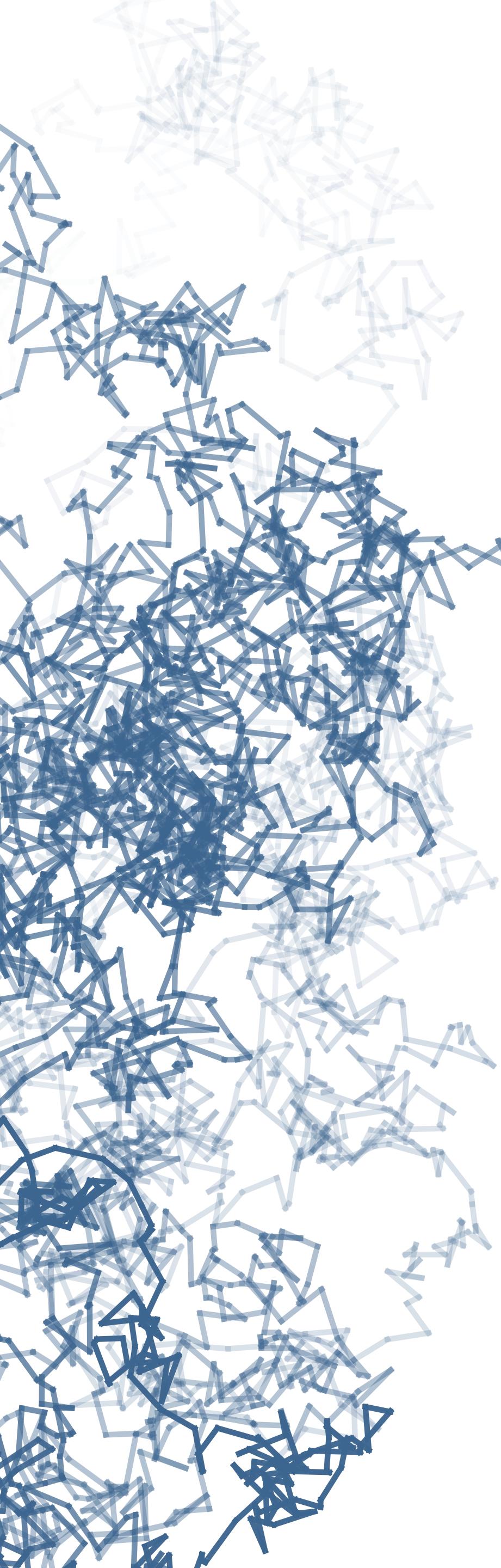


# Latent Diffusion



# Latent Diffusion



The background features a complex, abstract geometric pattern composed of numerous thin, light blue and white lines forming intricate shapes like triangles and hexagons, creating a sense of depth and motion.

## Part VI

# CONCLUDING REMAKES

# Challenges & Opportunities

The proteins we know have emerged over 3.5 billion years of evolutionary optimisation.

...

Since AlphaFold2 is trained on products of successful evolutionary optimisation, it is provided with an extremely strong signal

...

The positions of building blocks within a molecule are often explained not by function but by what is possible to synthesise with standard chemical processes

...

[Generative models] most likely to learn **synthesis biases** rather than function.

Michael Bronstein - The Road to Biology 2.0 Will Pass Through Black-Box Data

# Challenges & Opportunities

The proteins we know have emerged over 3.5 billion years of evolutionary optimisation.

...

Since AlphaFold2 is trained on products of successful evolutionary optimisation, it is provided with an extremely strong signal

...

The positions of building blocks within a molecule are often explained not by function but by what is possible to synthesise with standard chemical processes

...

[Generative models] most likely to learn **synthesis biases** rather than function.

Michael Bronstein - The Road to Biology 2.0 Will Pass Through Black-Box Data

**In other words:** Molecular generative models cannot reveal substantial insights from data like *ChatGPT* with *language* or *AlphaFold* with *proteins*.

# Challenges & Opportunities

The proteins we know have emerged over 3.5 billion years of evolutionary optimisation.

...

Since AlphaFold2 is trained on products of successful evolutionary optimisation, it is provided with an extremely strong signal

...

The positions of building blocks within a molecule are often explained not by function but by what is possible to synthesise with standard chemical processes

...

[Generative models] most likely to learn **synthesis biases** rather than function.

Michael Bronstein - The Road to Biology 2.0 Will Pass Through Black-Box Data

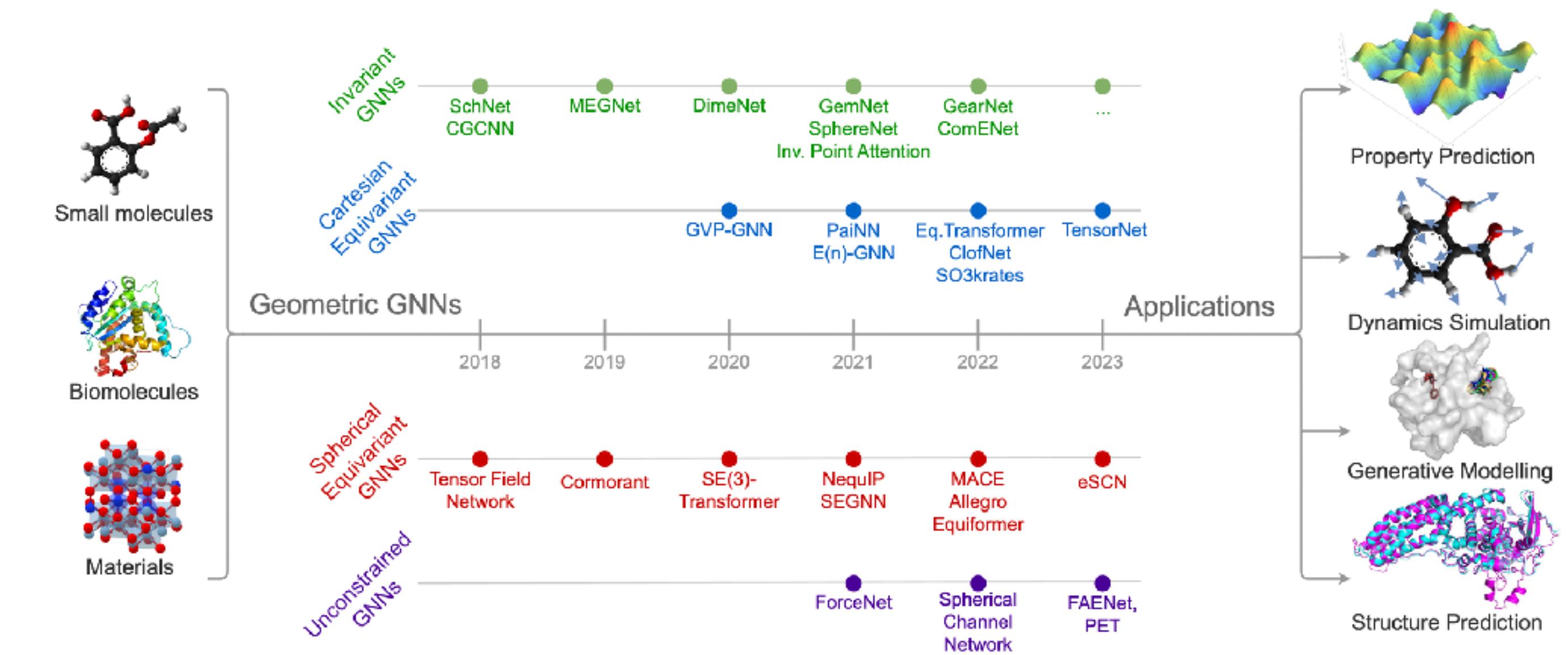
**In other words (II):** Training on QM9 to improve drugs is like learning on randomly generated text-sequences to learn about human communication.

# Challenges & Opportunities

- **Benchmarks:** Existing benchmarks may not adequately evaluate model performance (in the domain of drug discovery).
- **Synthesizability:** Difficulty in generating molecules that are synthetically feasible.
- **Conditional Generation:** Current backbone NNs struggle with conditional molecule generation.
- **Best Practices:** Many arbitrary trick and hacks to make diffusion models and GNNs better.
- **Interpretability:** Models lack interpretability, making it hard to understand the reasoning behind predictions (e.g., why should this molecule fit into this pocket).
- **Synthesis biases:** No foundation model in sight.

# Geometric GNNs: Key Concepts

- **Integration of Geometry:**  
Add 3D (or 3D) coordinates to node feature.
- **Equivariance:** given **rotation** and **translation**
- **Enhanced Message Passing:**  
Use geometric attributes like distances and angles.
- **Applications:**  
Spatial reasoning, like molecular modeling, 3D object recognition, and physical simulations.



A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems (Duval et al.)