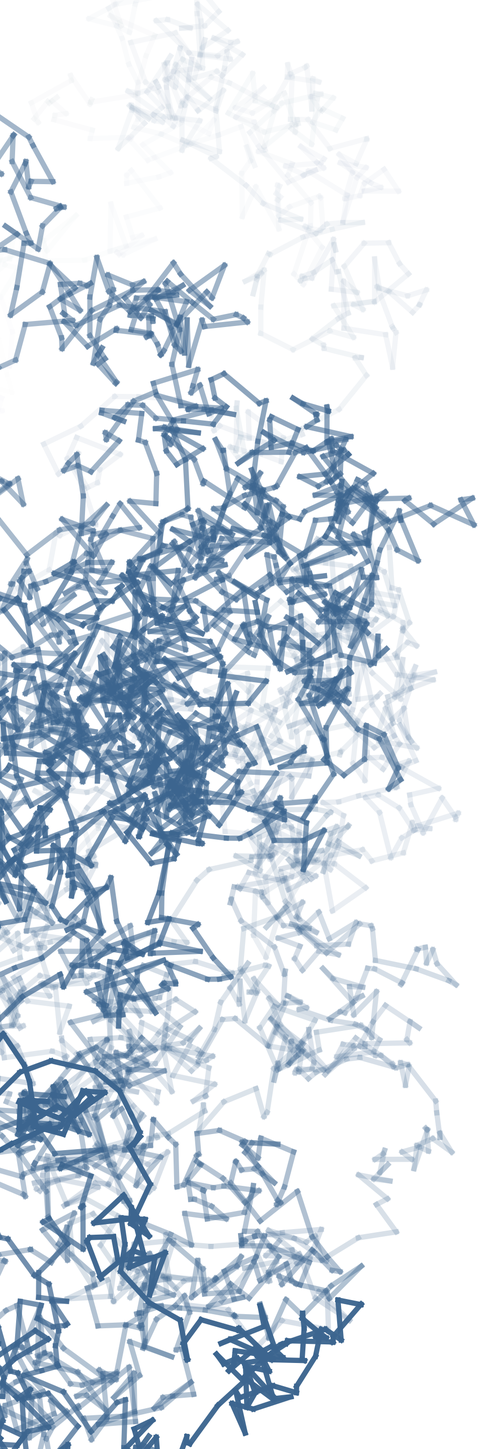


GENERATIVE AI FOR DRUG DISCOVERY: DREAMING UP NEW MOLECULES

Gerrit Großmann

09.11.2024

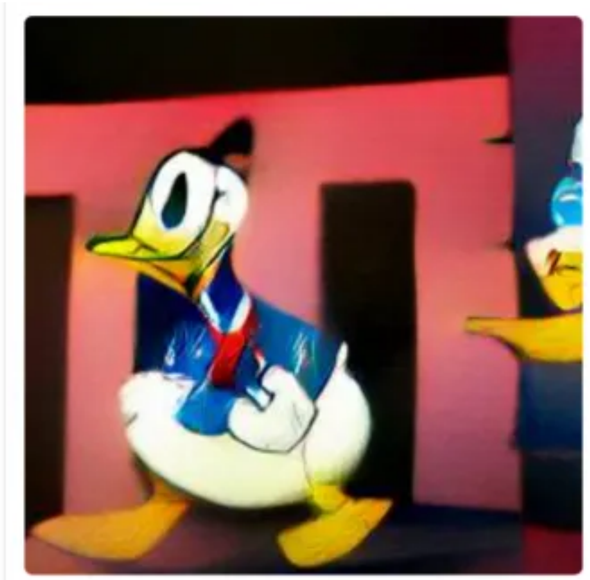
gerritgrossmann.de/#cerfa

dfki ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for Artificial Intelligence

CERFA

Motivation

2022



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

2023

2024

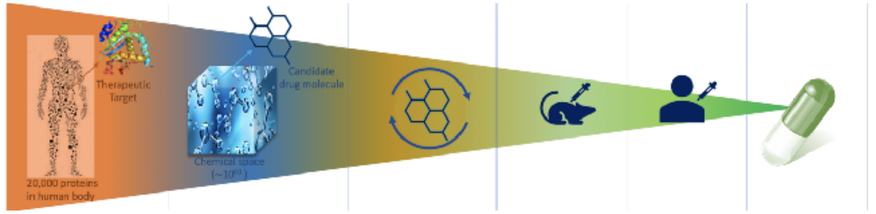
youtube.com/watch?v=TU1gMlol0kc&ab_channel=MagnaAI

Motivation

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

Jeroen Pixel twitter.com/pixelprotest_/status/1743227361888186773

?



linkedin.com/pulse/ai-drug-discovery-vijay-morampudi

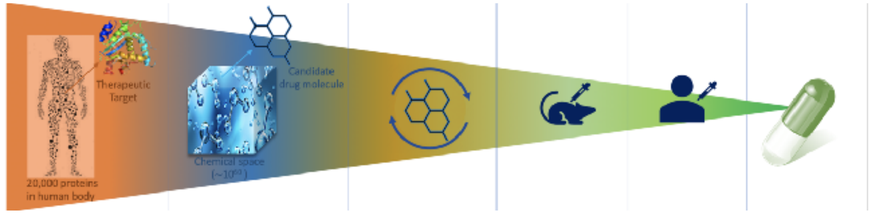
Motivation

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

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

Jeroen Pixel twitter.com/pixelprotest_/status/1743227361888186773

?



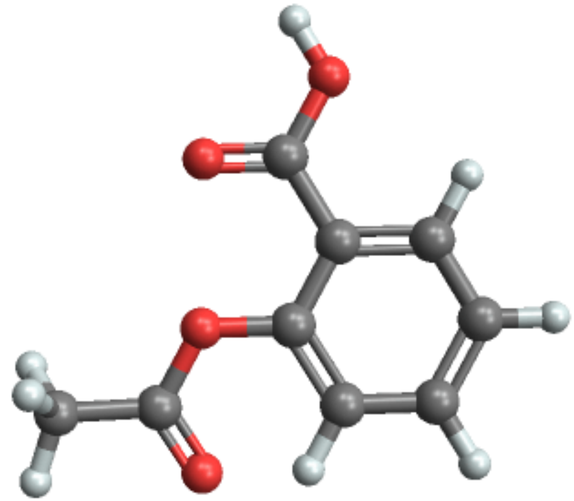
linkedin.com/pulse/ai-drug-discovery-vijay-morampudi

Motivation

Jeroen Pixel 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

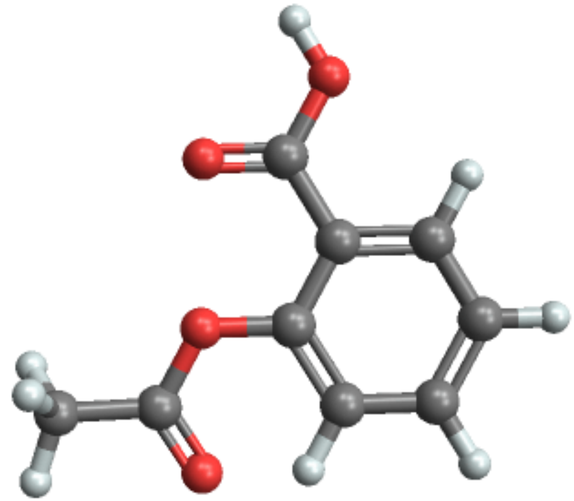


Motivation

Jeroen Pixel twitter.com/pixelprotest_/status/1743227361888186773

- High Dimensionality
- Large amount of data
- Complex manifold
- World model
- Conditional generation
- Ethically dubious potential

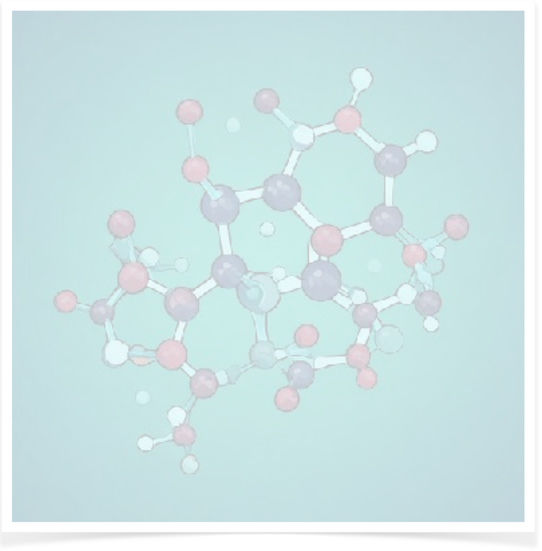
- Domain knowledge
- 3D objects
- Evaluation
- Discrete variables



Outline

Drug discovery in a nutshell

Foundations of diffusion models

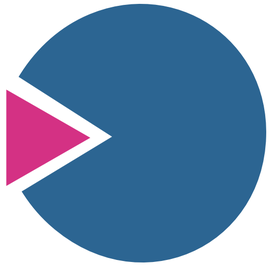


Dreaming up useful molecules

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

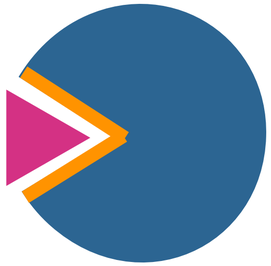


protein
enzyme
receptor
drug target

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



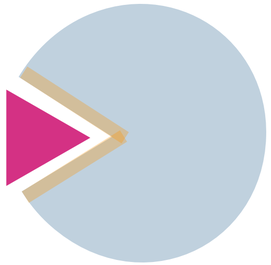
protein
enzyme
receptor
drug target

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**.

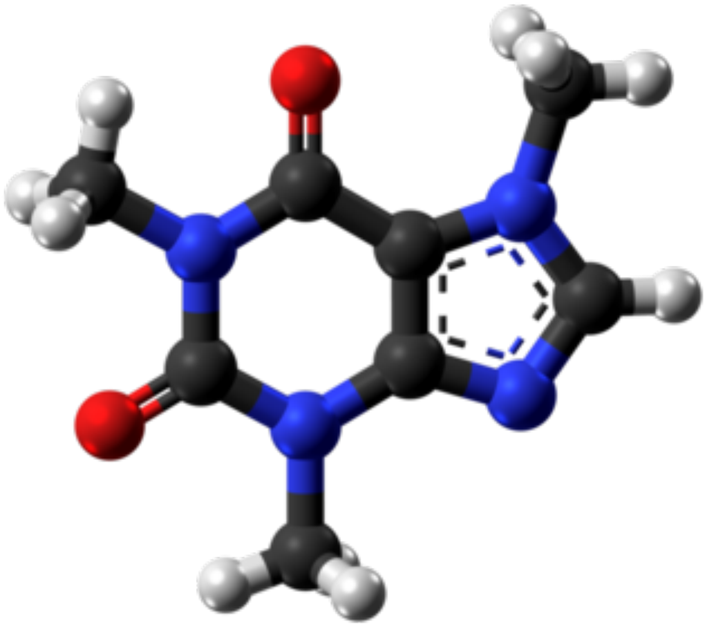
small molecule
drug
ligand
inhibitor



protein
enzyme
receptor
drug target

binding site
pocket

What is a (Small) Molecule



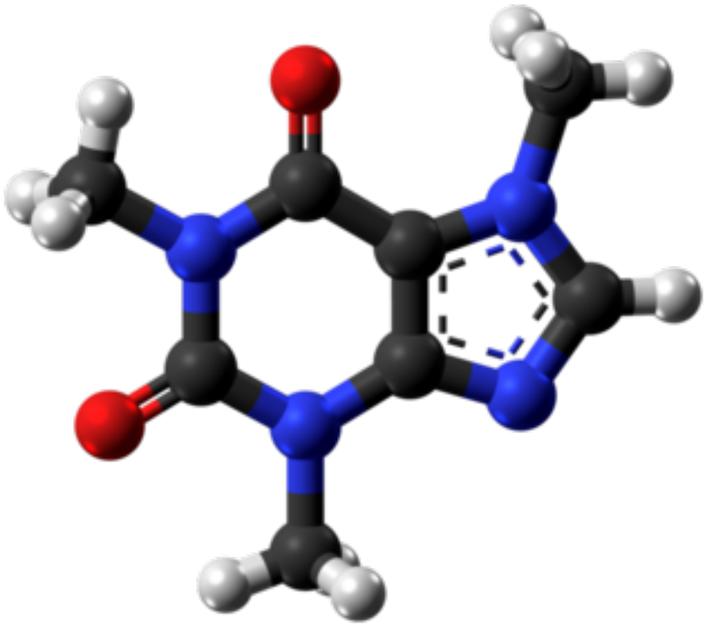
Molecule

- 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



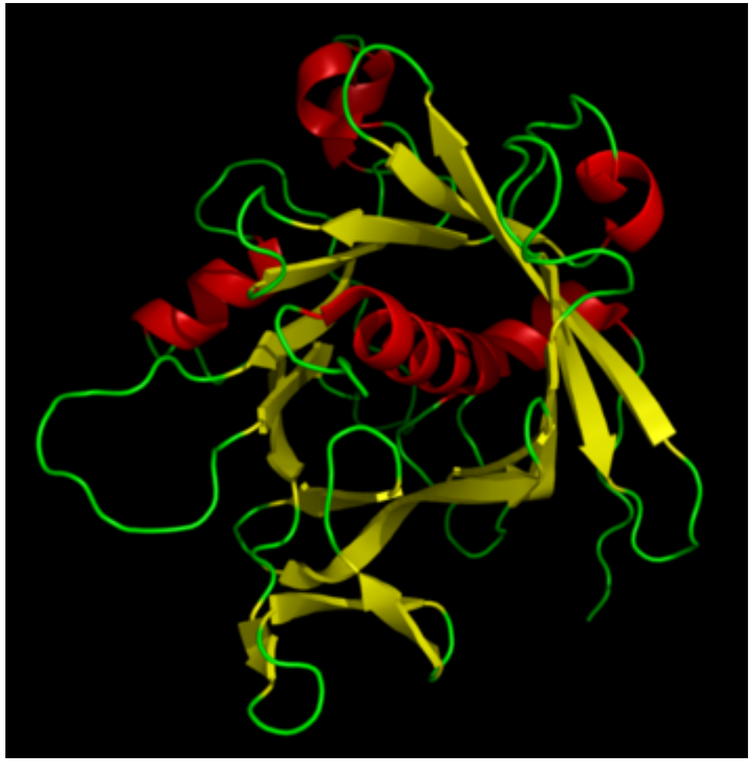
Molecule

- 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?



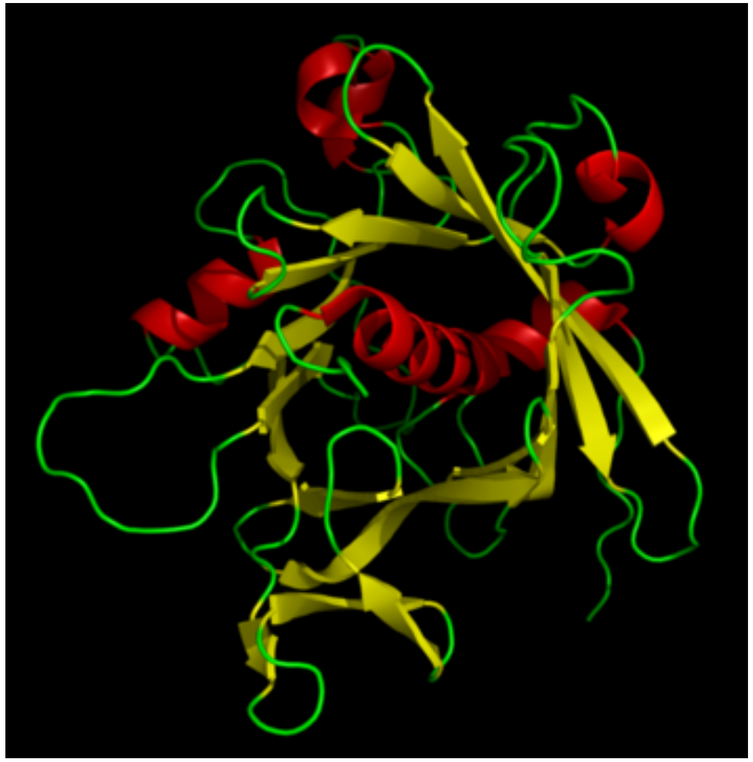
Protein

- 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

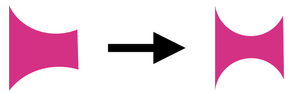
Druggable Target

- 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

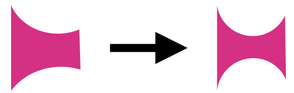
Human bias

Ultra-large virtual molecular libraries throw open chemical space (Nature)

AI to the Rescue

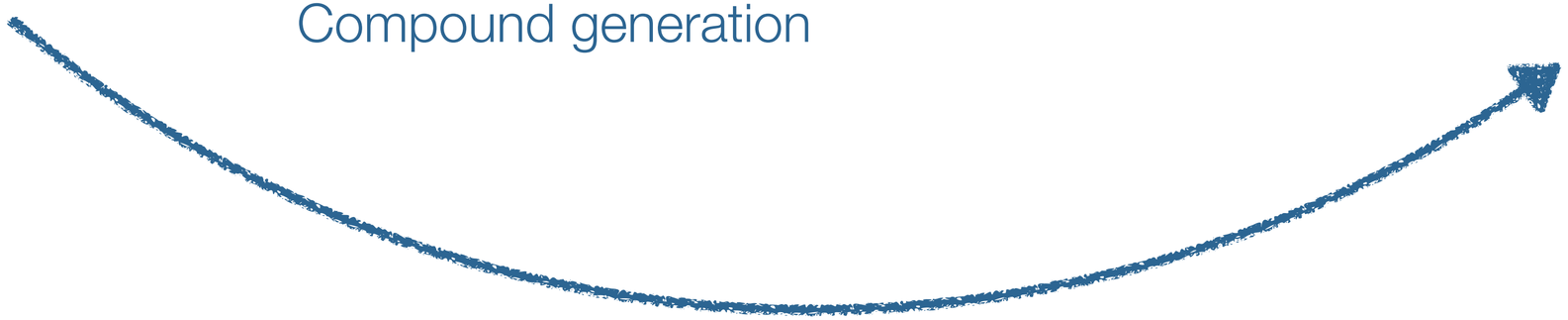
Identifying a **protein** that is involved in a disease process and can be **targeted** by a **drug** to produce a therapeutic effect.

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

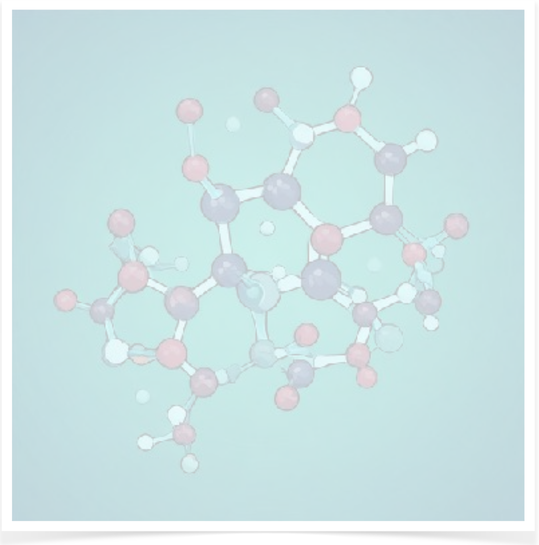


Generate novel chemical structures that are optimized for binding to the target.

Outline

Drug discovery in a nutshell

Foundations of diffusion models



Dreaming up useful molecules

Probabilistic Diffusion

Nucleus ☕ ✅ Ⓜ
@EsotericCofe

hardest LLM paper vs easiest diffusion paper

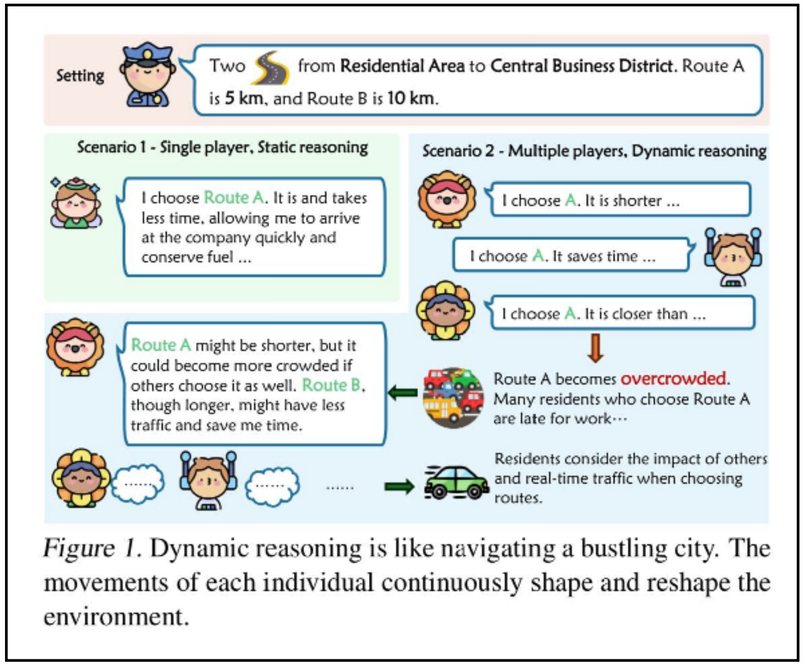


Figure 1. Dynamic reasoning is like navigating a bustling city. The movements of each individual continuously shape and reshape the environment.

twitter.com/EsotericCofe/status/1777280241884377474

$$L_{\text{DPO-Diffusion}}(\theta) \le -\mathbb{E}_{\substack{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),\\ \boldsymbol{x}_{t\ 1,t}^w\sim p_\theta(\boldsymbol{x}_{t\ 1,t}^w|\boldsymbol{x}_0^w),\\ \boldsymbol{x}_{t\ 1,t}^l\sim p_\theta(\boldsymbol{x}_{t\ 1,t}^l|\boldsymbol{x}_0^l)}}$$
$$\log\sigma\left(\beta T\log\frac{p_\theta(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)}-\beta T\log\frac{p_\theta(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l)}\right) \tag{12}$$

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

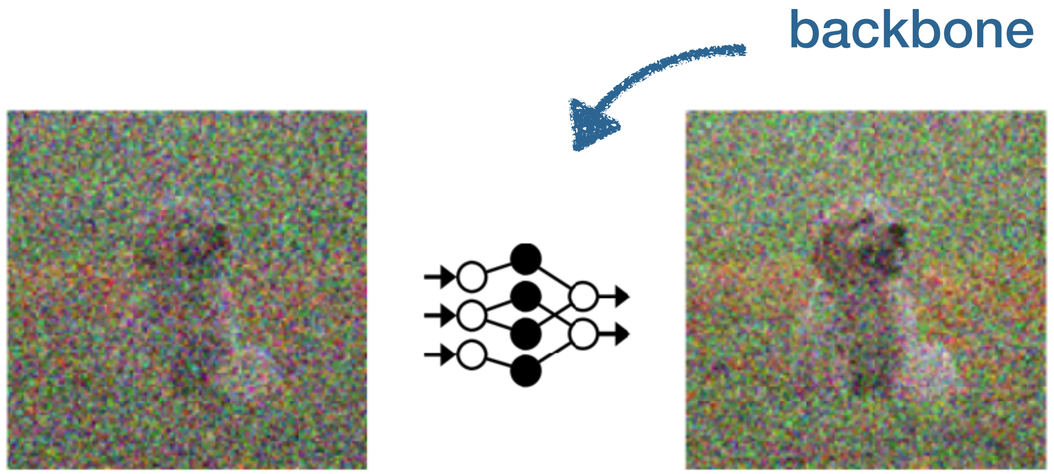
$$\begin{aligned}L(\theta)=-\mathbb{E}_{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D},t\sim\mathcal{U}\{0,T\},\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t^w|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t^l|\boldsymbol{x}_0^l)}\\ \log\sigma(-\beta T(\\ +\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_{0,t}^w)\|p_\theta(\boldsym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

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

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

where $\boldsymbol{x}_t^*=\alpha_t\boldsymbol{x}_0^*+\sigma_t\boldsymbol{\epsilon}^*,\boldsymbol{\epsilon}^*\sim\mathcal{N}(0,I)$ is a draw from $q(\boldsymbol{x}_t^*|\boldsymbol{x}_0^*)$ (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)

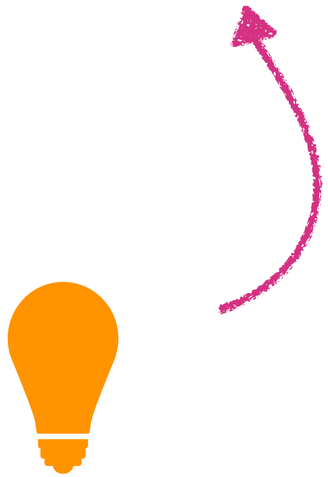
2.

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

Probabilistic Diffusion

1.

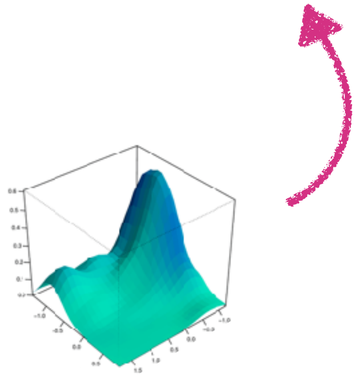
Non-trainable part that **removes** information
(stochastically or deterministically)



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).

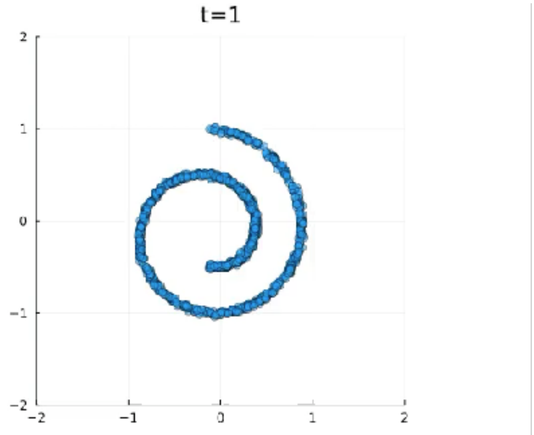
Probabilistic Diffusion

1.

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

2.

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

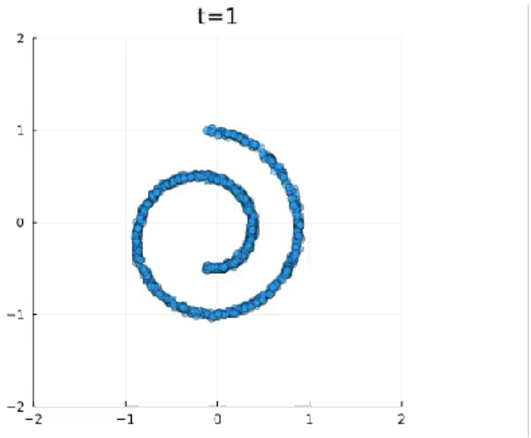


Obvious way: Degrade a sample by **adding Gaussian noise** to it.

Probabilistic Diffusion

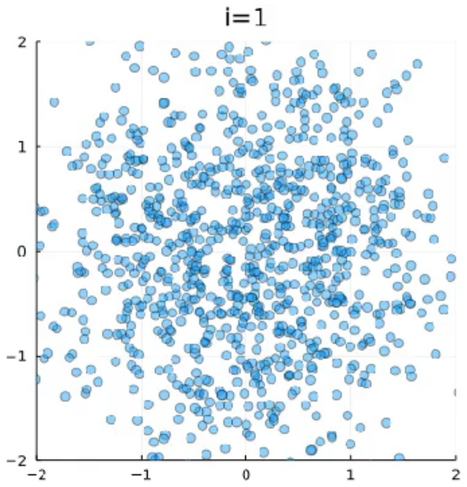
1.

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.

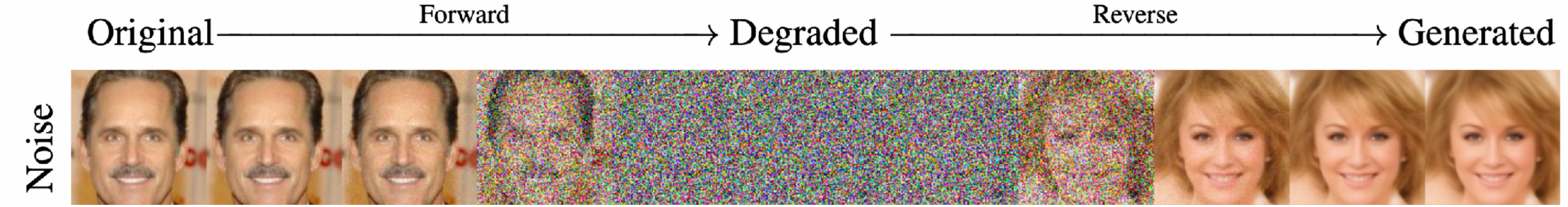
Probabilistic Diffusion

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.

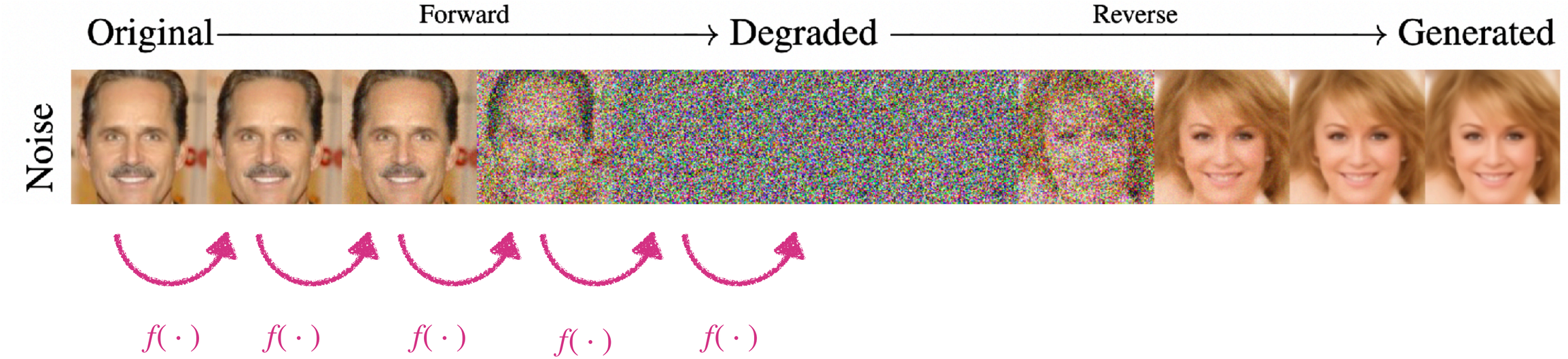
Probabilistic Diffusion

1.

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

2.

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



Step-by-step (typically 1000)

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

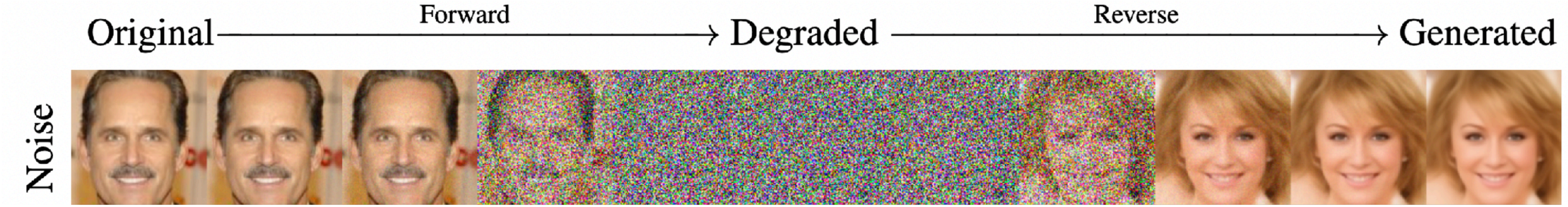
Probabilistic Diffusion

1.

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

2.

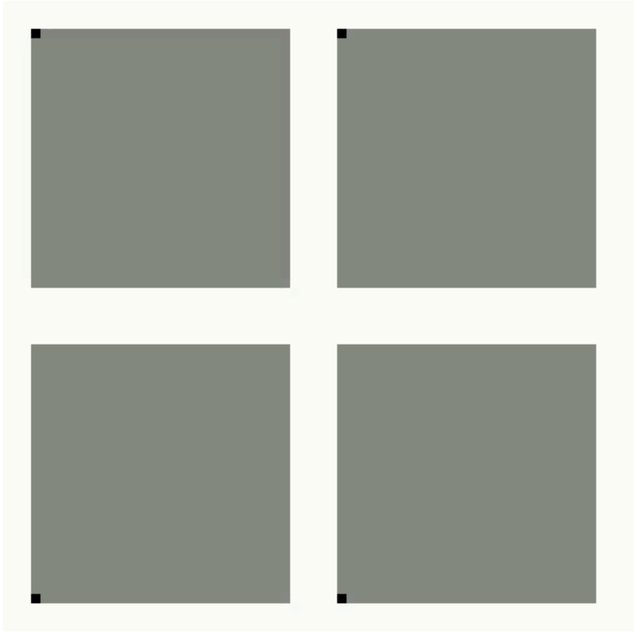
Trainable part that **reconstructs** information
(typically with a stochastic component)



Step-by-step (typically 1000)

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

Probabilistic Diffusion



ajayjain.github.io/lmconv/

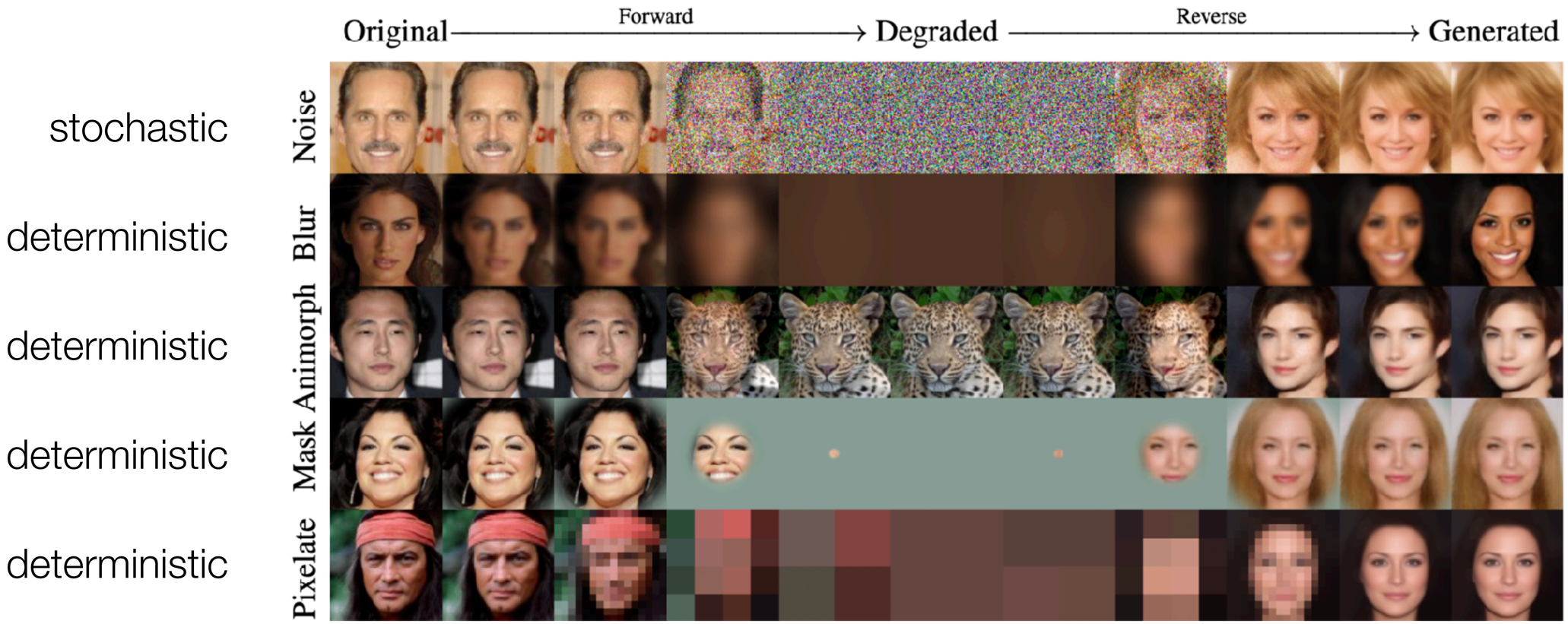
“What if I the forward process masks one dimension in each step?“

“You get an autoregressive model!“

Probabilistic Diffusion

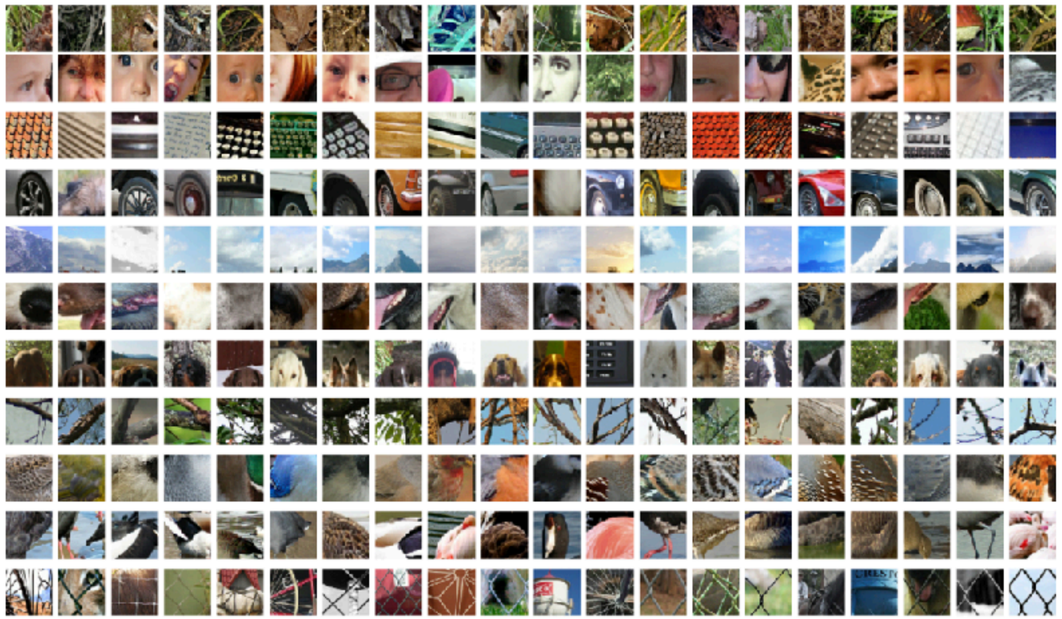
Non-trainable part that **removes** information
(stochastically or deterministically)

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



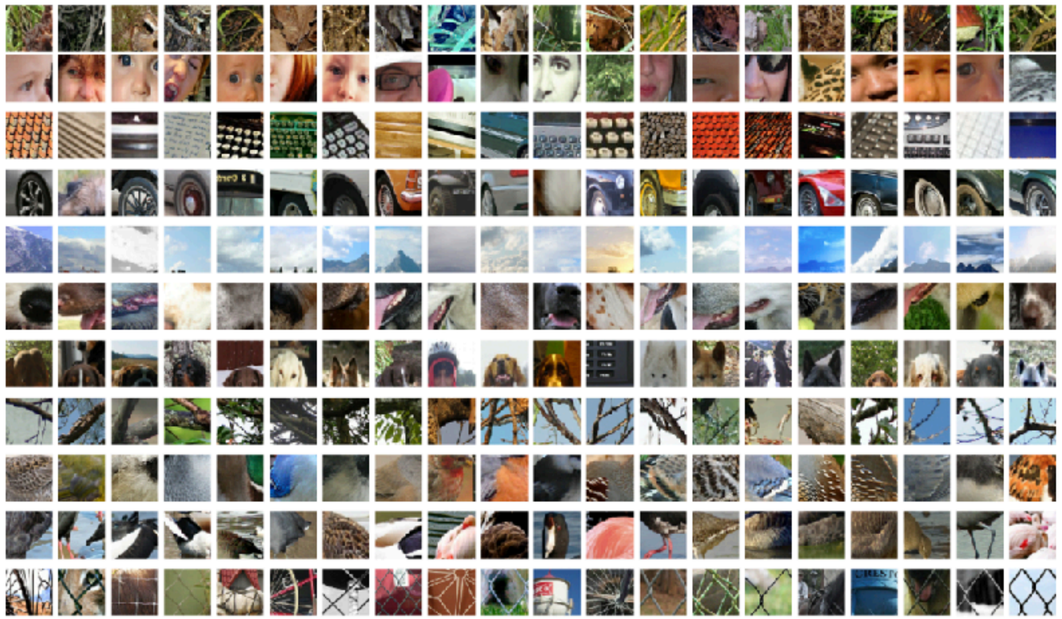
Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise (Bansal et al.)

Conditional Sampling



semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7

Conditional Sampling



semanticscholar.org/paper/Representation-Learning-with-Contrastive-Predictive-Oord-Li/b227f3e4c0dc96e5ac5426b85485a70f2175a205/figure/7

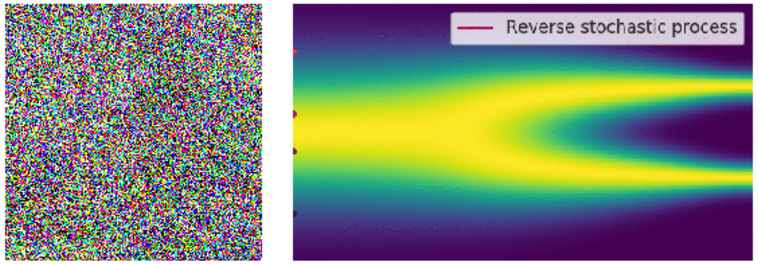
A panda on a motorcycle with a red sombrero and a cigar, oil painting.

Diffusion in Literature

Denoising (DDPM)

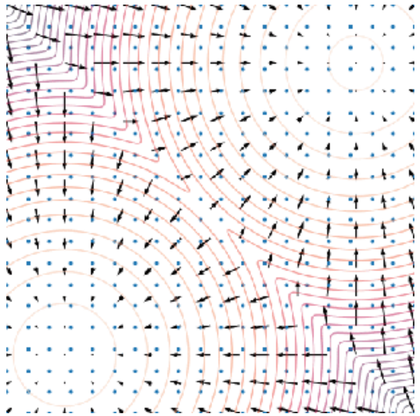
Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise (Bansal et al.)

SDE



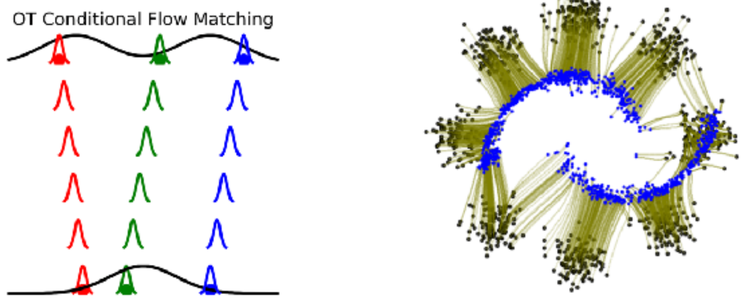
yang-song.net/blog/2021/score/

Score Matching



yang-song.net/blog/2021/score/

Flow Matching



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

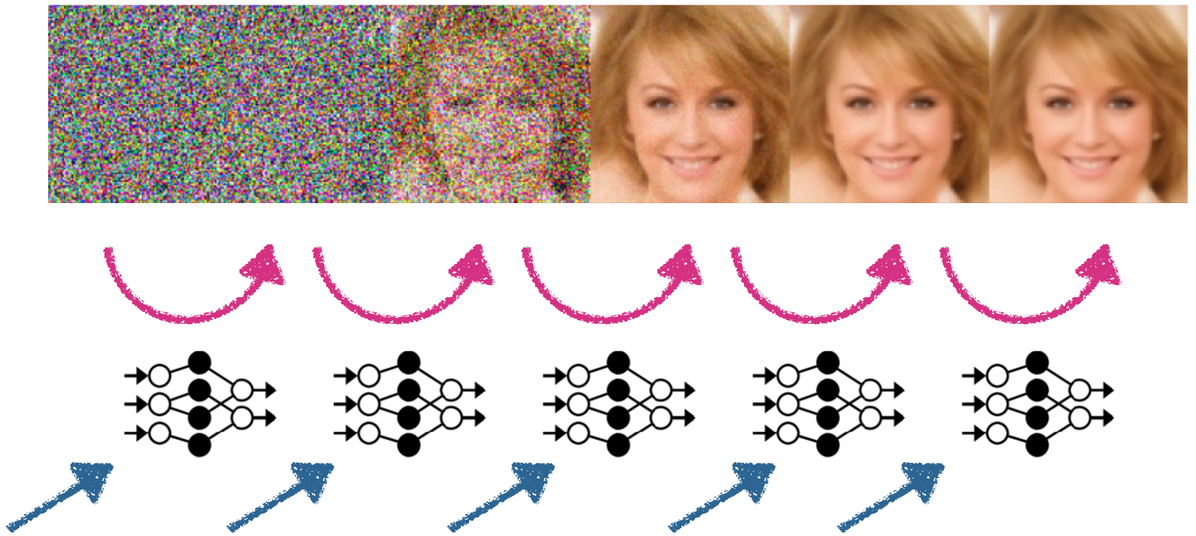
Conditional Sampling

1.

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

2.

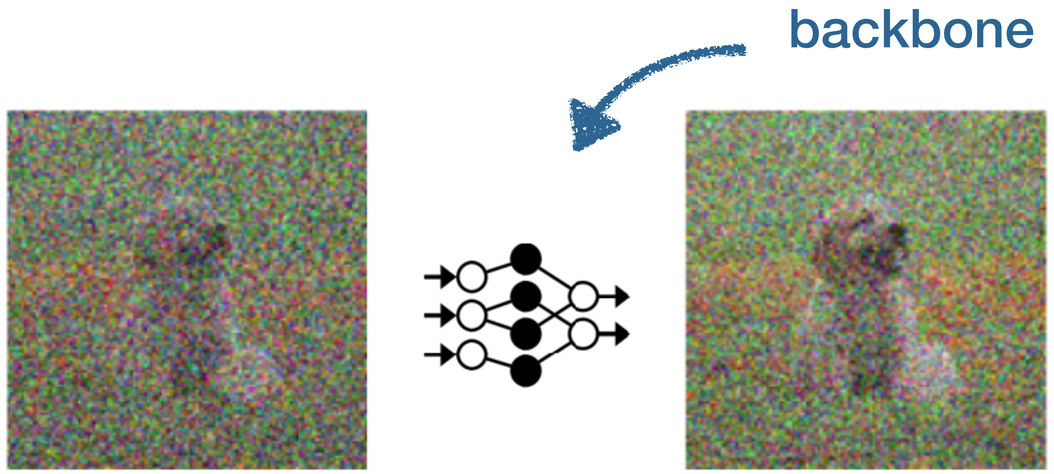
Trainable part that **reconstructs** information
(typically with a stochastic component)



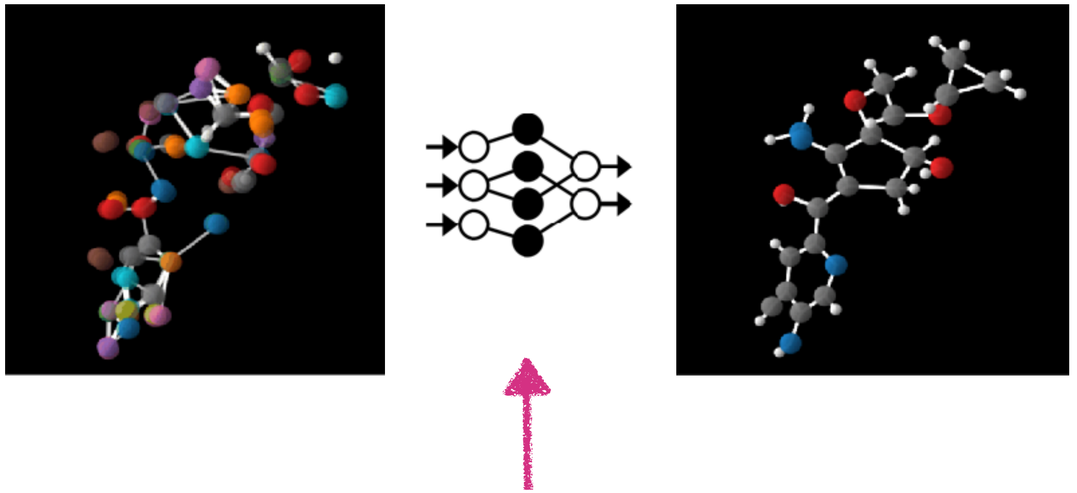
condition as conditional input

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

Diffusion



Diffusion

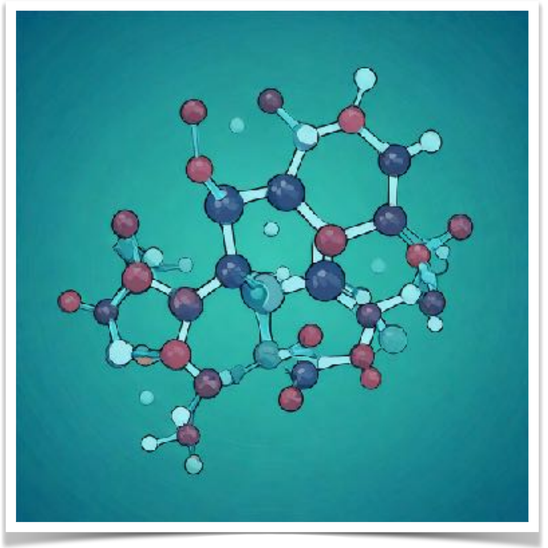


Some Flavour of Graph Neural Networks

Outline

Drug discovery in a nutshell

Foundations of diffusion models



Dreaming up useful molecules

Molecule Representations

Text-based

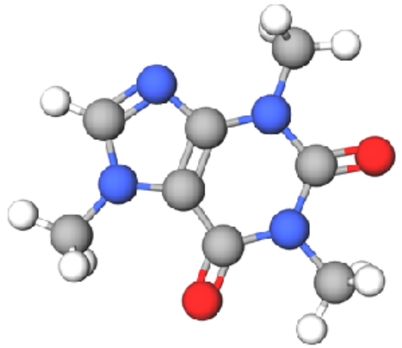
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

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, …

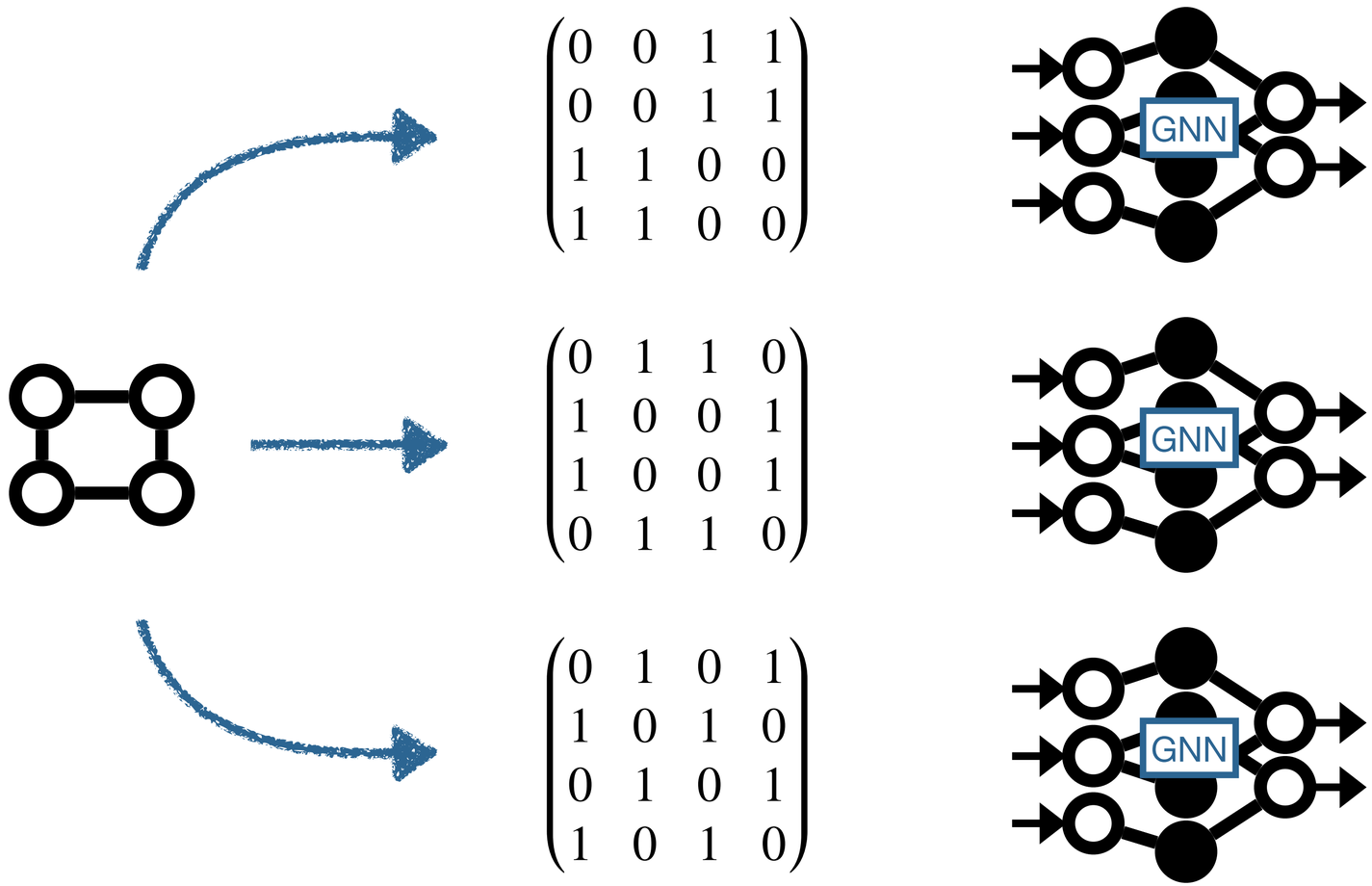
Graph-based

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

Point Cloud

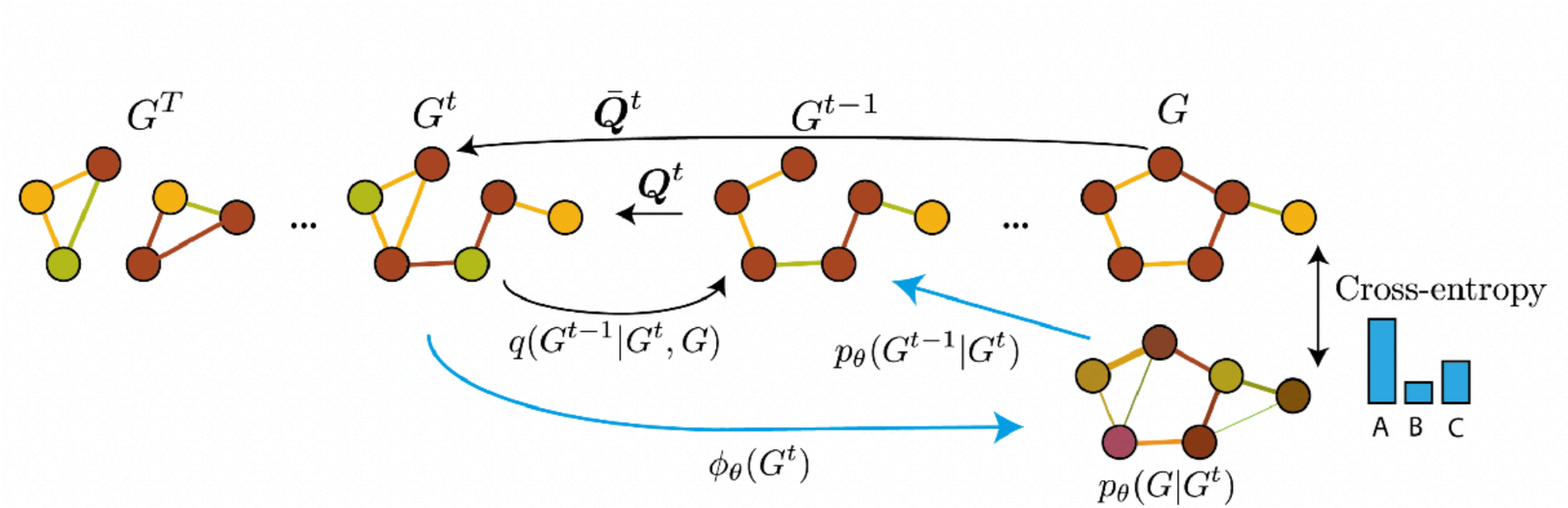
- Nodes in the graph have additional **3D spatial annotations.**
- Unconnected points are represented without explicit edge (bond) information.

Permutation Invariance



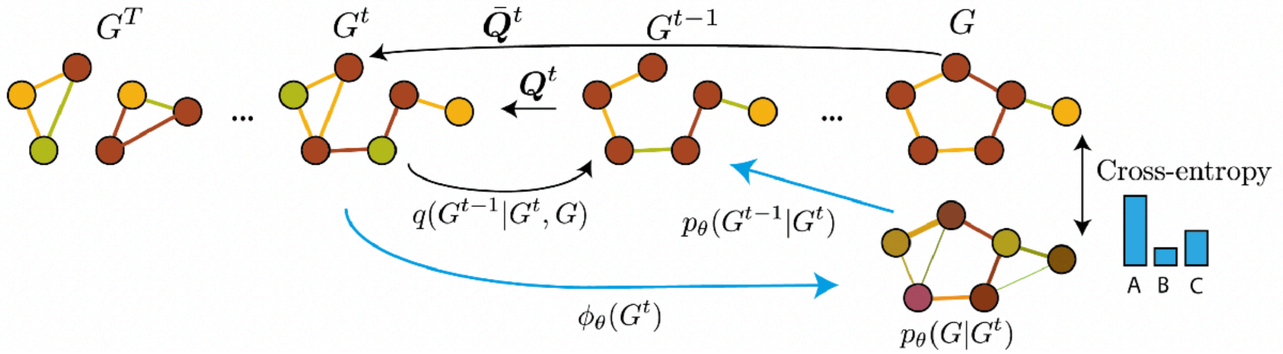
Guaranteed to produce the same results.

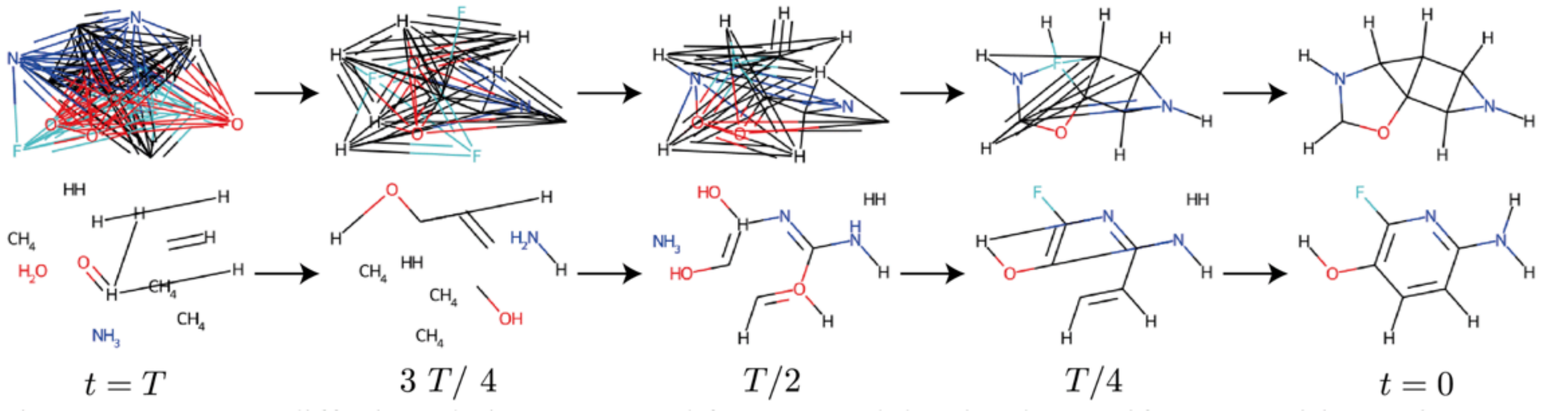
DiGress



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

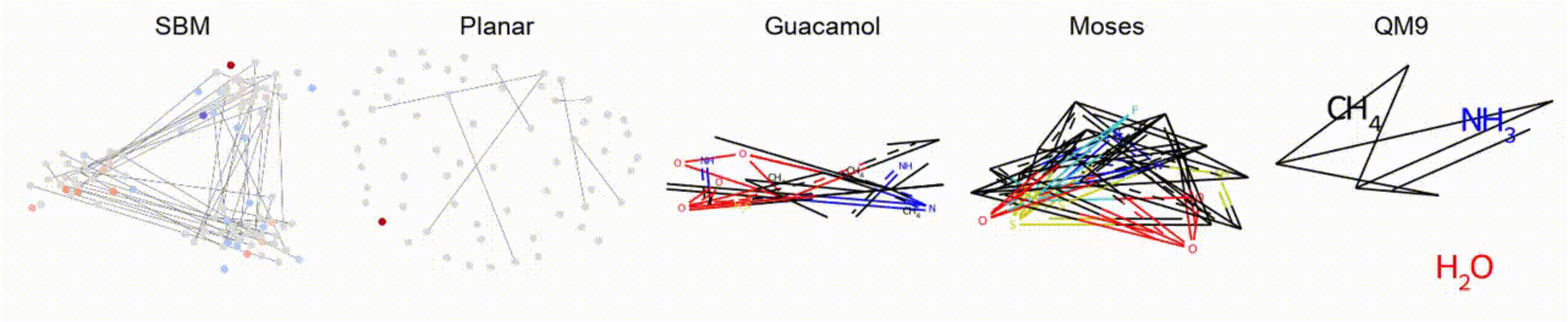
DiGress





DiGress: Discrete Denoising diffusion for graph generation (Vignac et al.)

DiGress



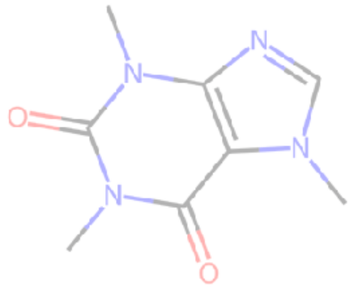
Molecule Representations

Text-based

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

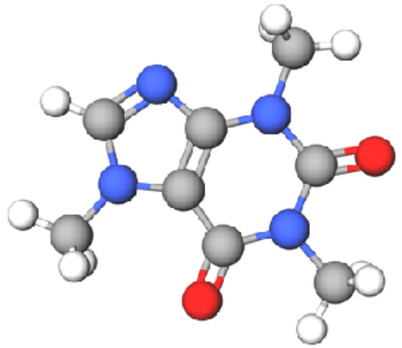
Smiles, Selfies, InChI, …

Graph-based



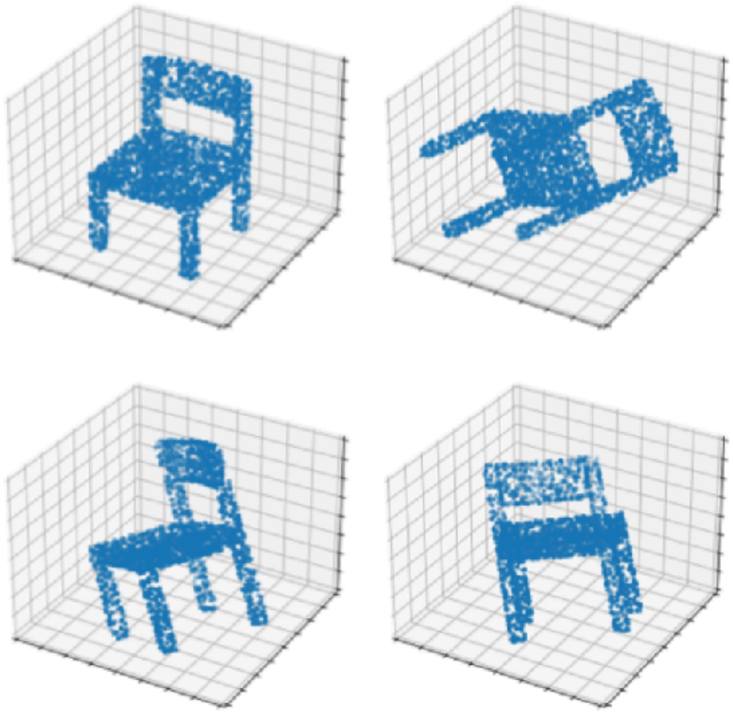
Nodes = Atoms
Node feature = element
Edges = Covalent bonds
Edge feature = Bond type

Point Cloud



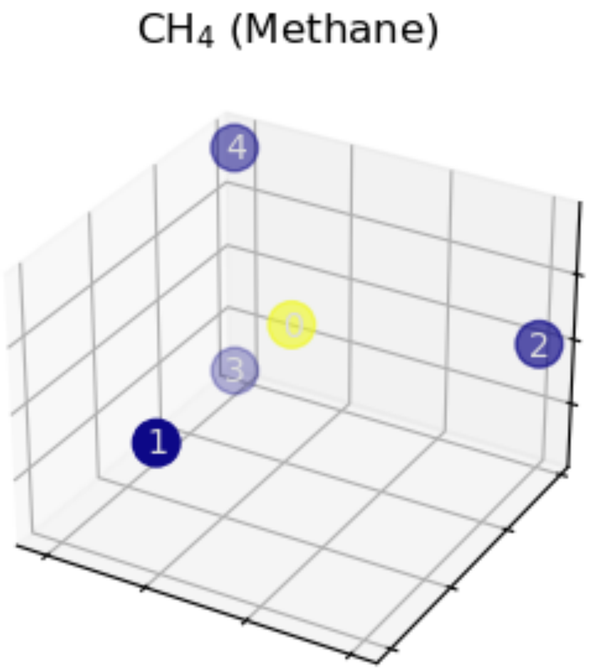
- Nodes in the graph have additional **3D spatial annotations.**
- Unconnected points are represented without explicit edge (bond) information.

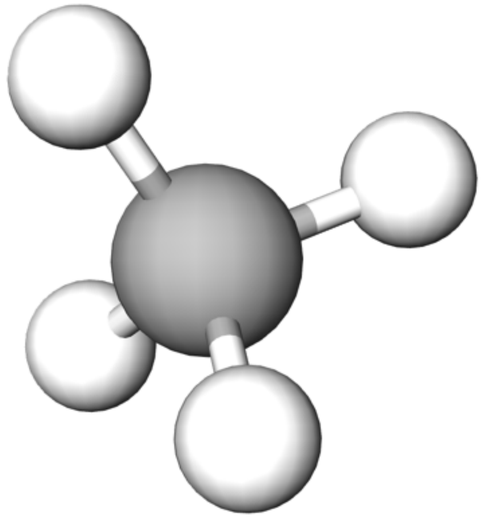
Point Clouds



https://www.helixon.com/wp-content/uploads/2022/09/Equivariant-Point-Cloud-Analysis-via-Learning-Orientations-for-Message-Passing.pdf

Permutation Invariance

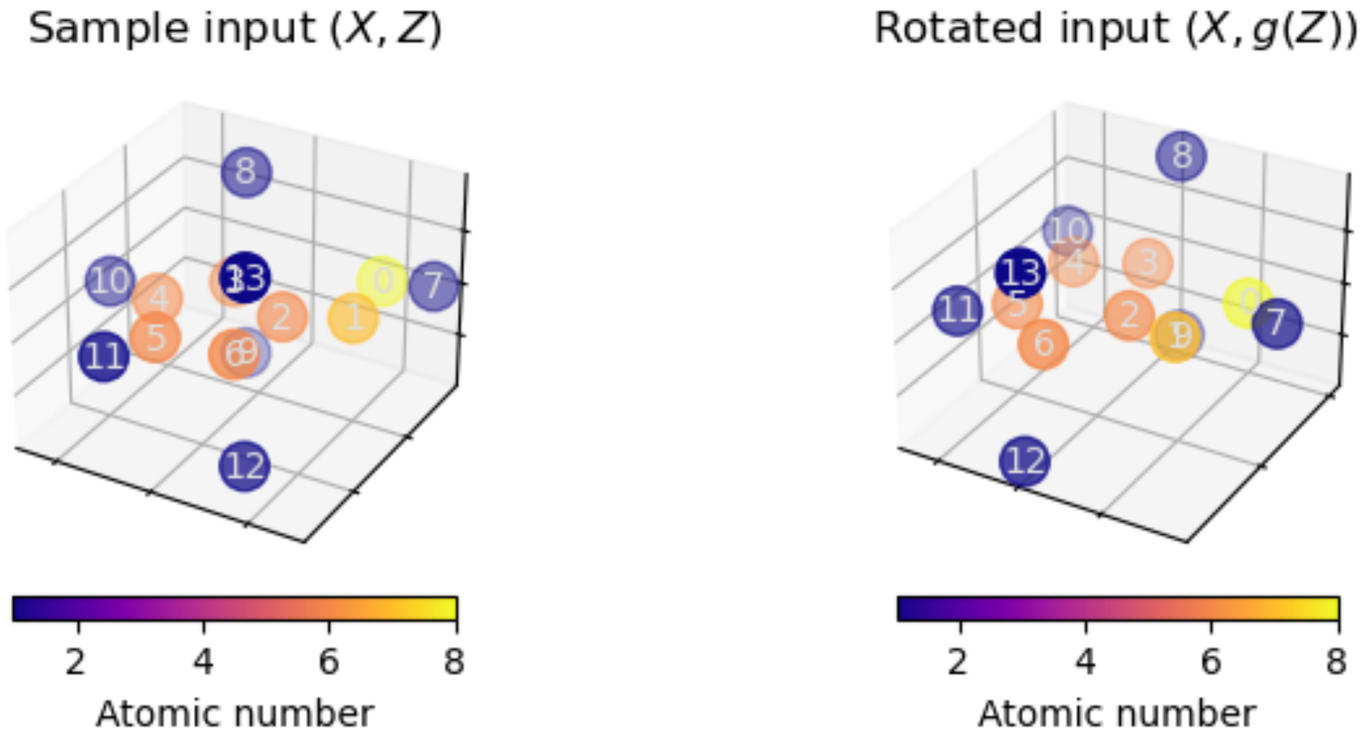




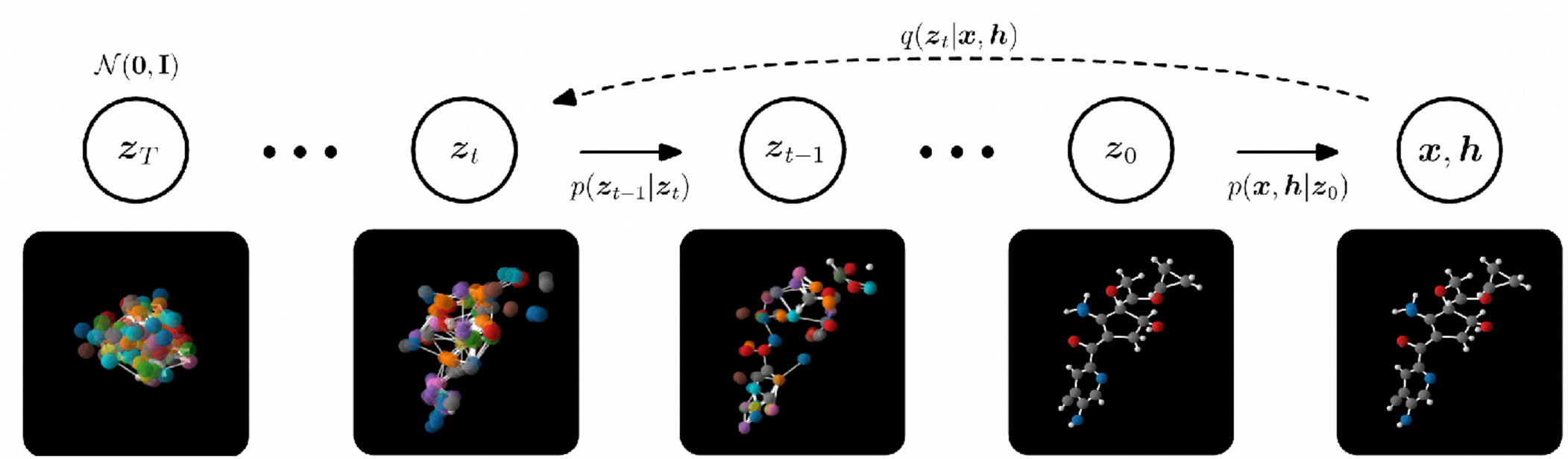
Hydrogen **Carbon**

https://projects.volkamerlab.org/teachopencadd

Permutation Invariance

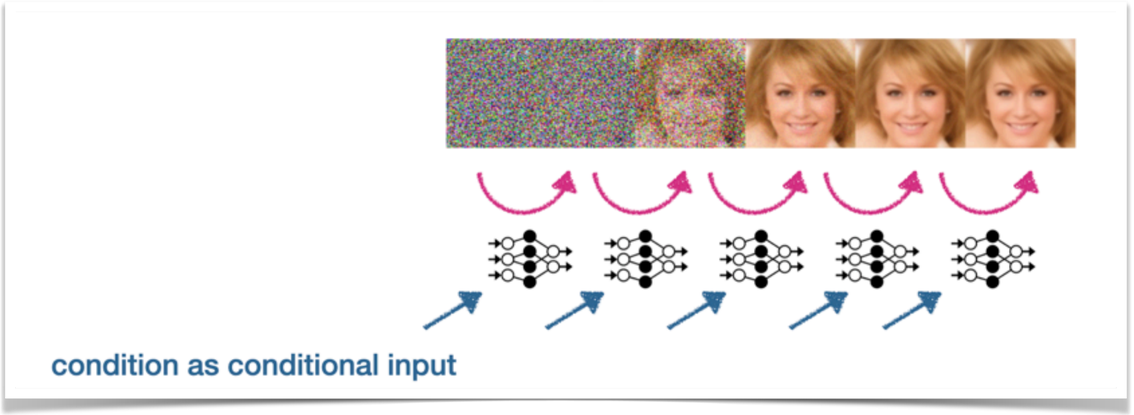


Molecule Generation in 3D



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

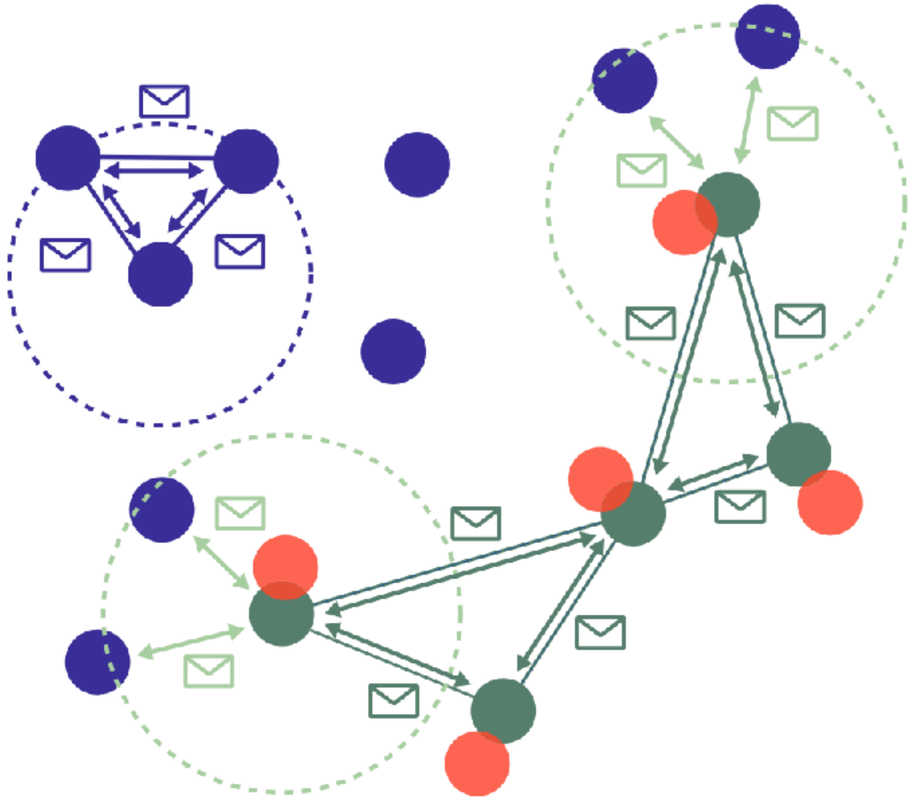
Pocket Conditioning

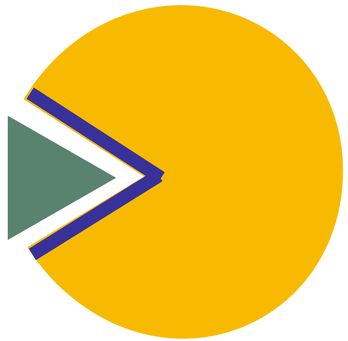


It's a Match!

Autoregressive fragment-based diffusion for pocket-aware ligand design (Ghorbani et al.)

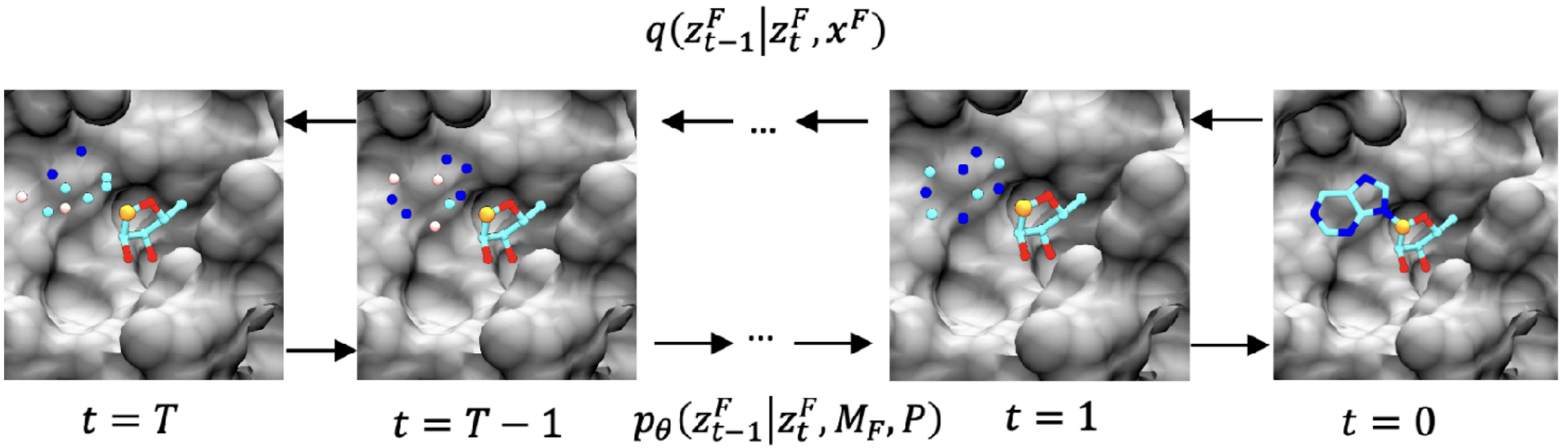
Pocket Conditioning





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

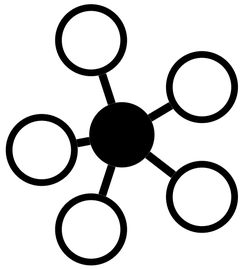
Pocket Conditioning



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

Evaluation

- **Validity:** Percentage of molecules with correct **valency**.



(Carbon atom with 5 hydrogen bonds)

- **Novelty:** Percentage of molecules **not in training dataset**

Train set:

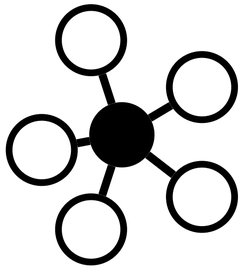
Generated Molecule:

- **Uniqueness:** Percentage of generated unique molecule

Generated Molecules:

Evaluation

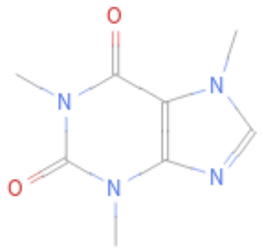
- **Validity:** Percentage of molecules with correct **valency**.



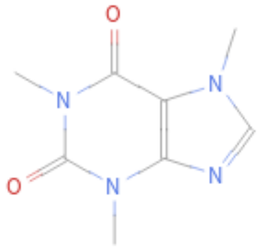
(Carbon atom with 5 hydrogen bonds)

- **Novelty:** Percentage of molecules **not in training dataset**

Train set:



Generated Molecule:

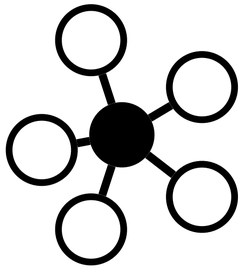


- **Uniqueness:** Percentage of generated unique molecule

Generated Molecules:

Evaluation

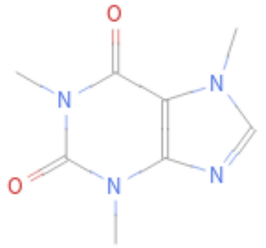
- **Validity:** Percentage of molecules with correct **valency**.



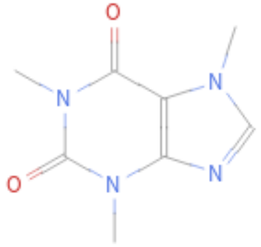
(Carbon atom with 5 hydrogen bonds)

- **Novelty:** Percentage of molecules **not in training dataset**

Train set:

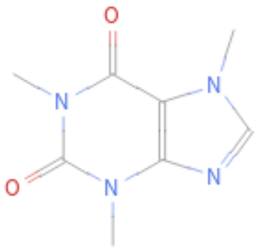


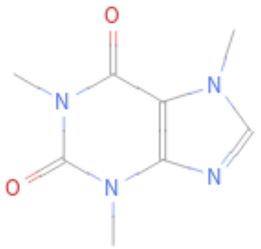
Generated Molecule:



- **Uniqueness:** Percentage of generated unique molecule

Generated Molecules:



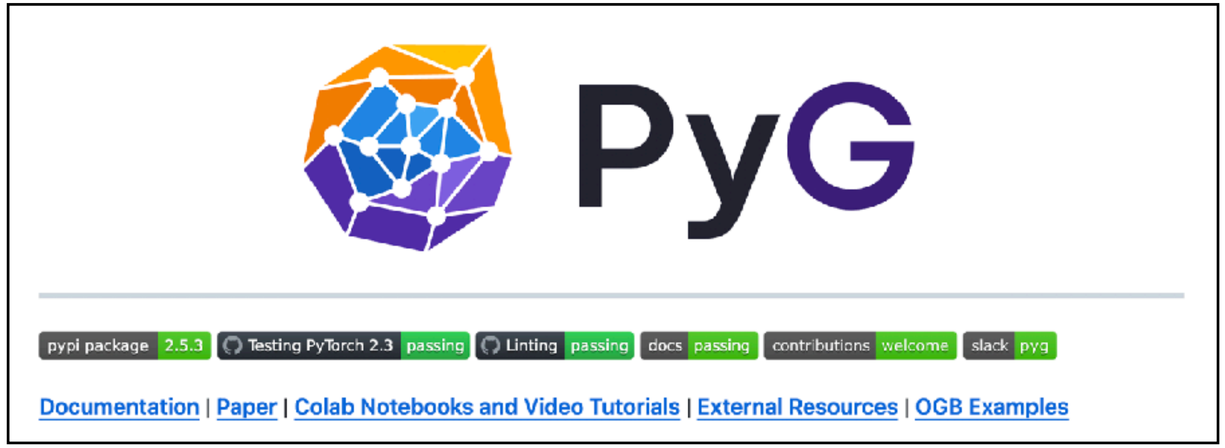


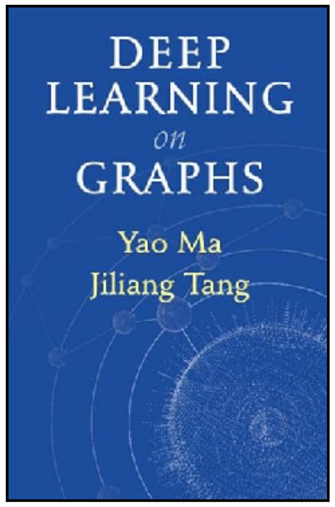
Evaluation

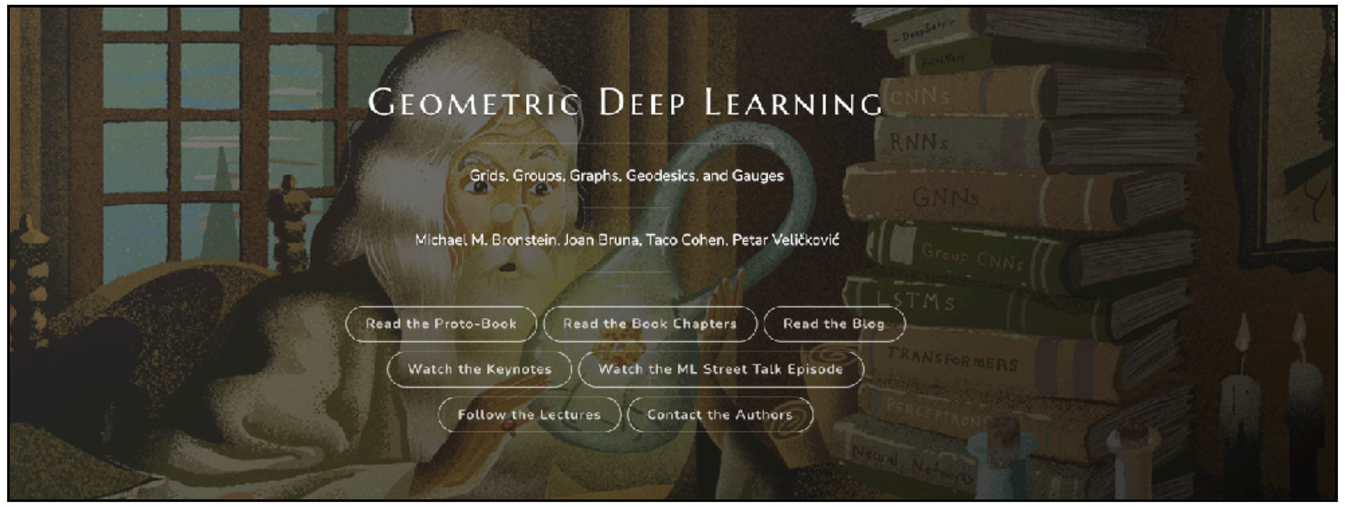
- **Validity:** Percentage of molecules with correct **valency**.
- **Novelty:** Percentage of molecules **not in training dataset**
- **Uniqueness:** Percentage of generated unique molecule

- **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

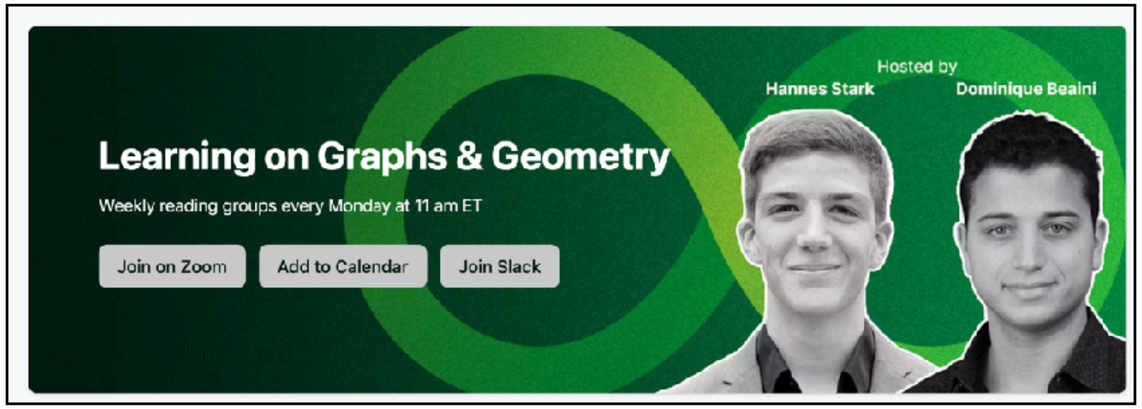
Getting Started

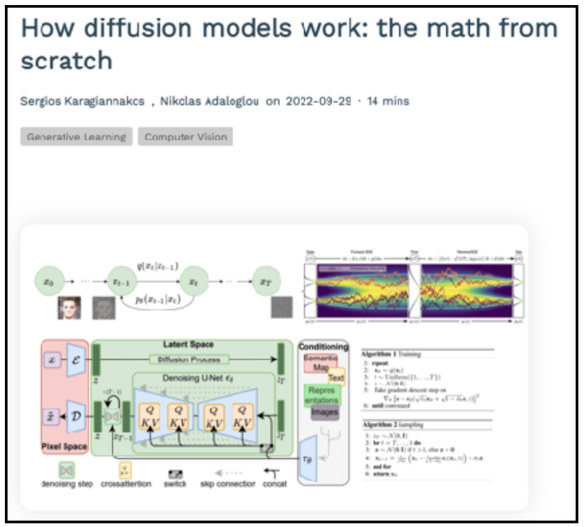


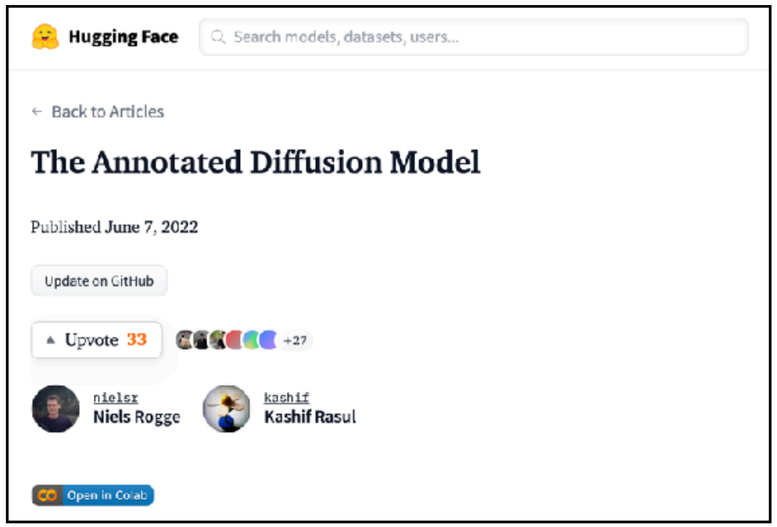


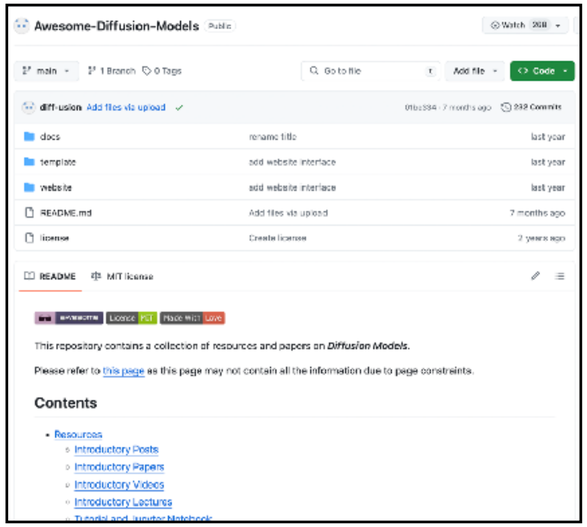


Diffusion Models

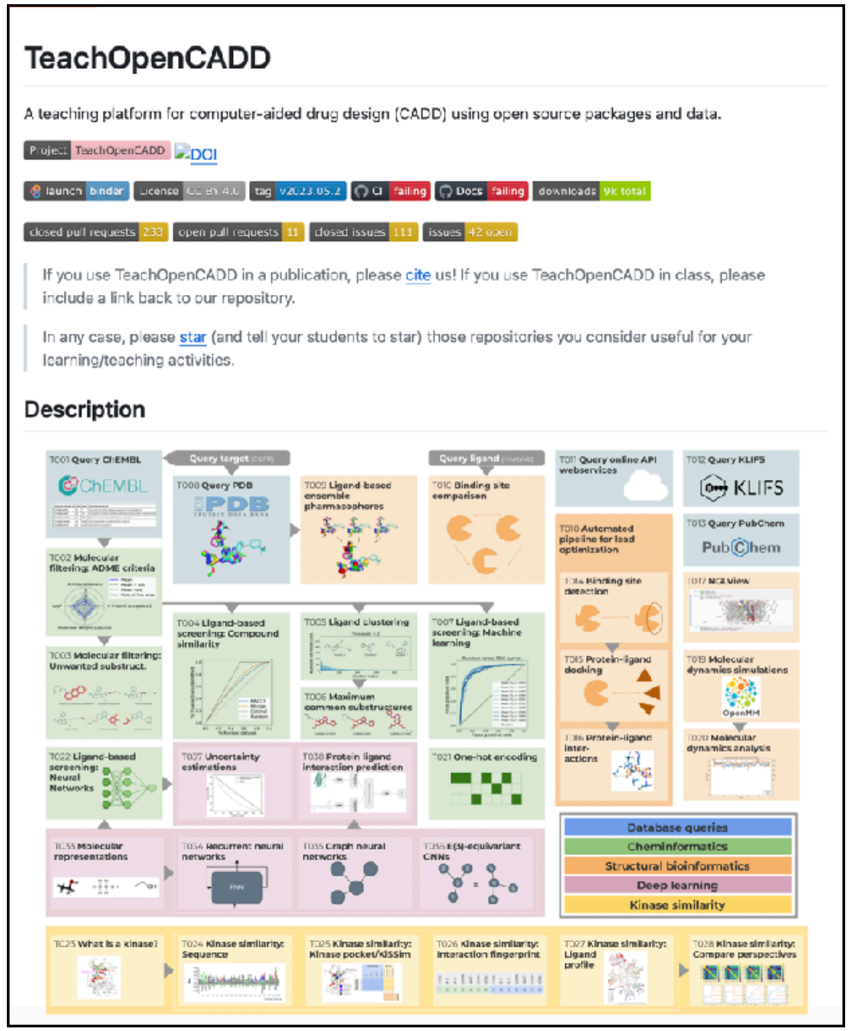






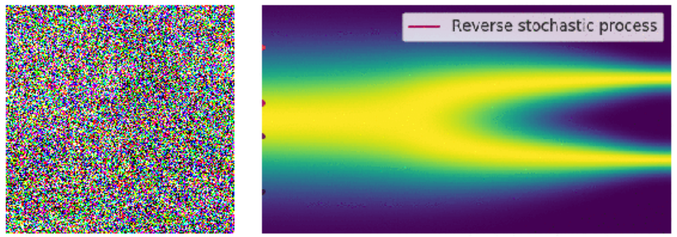


Computational Drug Discovery

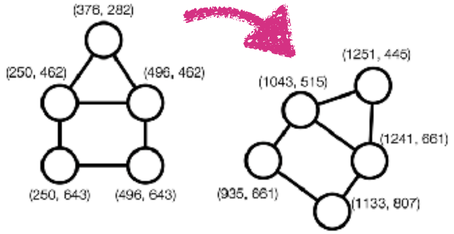


Summary

Method:

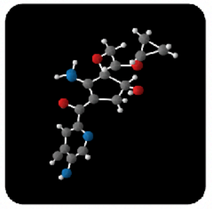


Diffusion Models

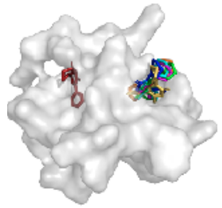


(Geometric) Graph Neural Networks

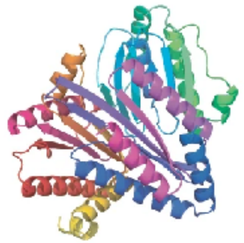
Applications:



Drug design

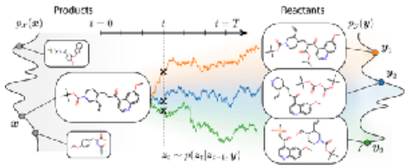


Docking



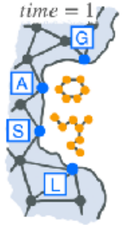
Structure Prediction

Future Directions:



Domain knowledge

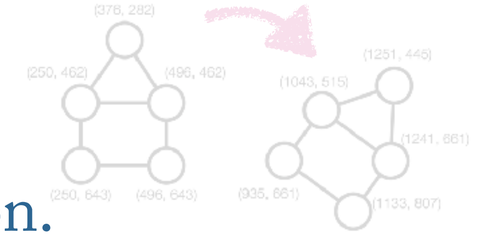
Benchmarks & Software



Better conditioning

Summary

Method:



Thank you four your attention.

Diffusion Models

(Geometric) Graph Neural Networks

Slides and Literature:
gerritgrossmann.de/#cerfa

Applications:

Drug design

Docking

Structure Prediction

Questions? Ideas? Want to collaborate?
gerrit.grossmann@dfki.de

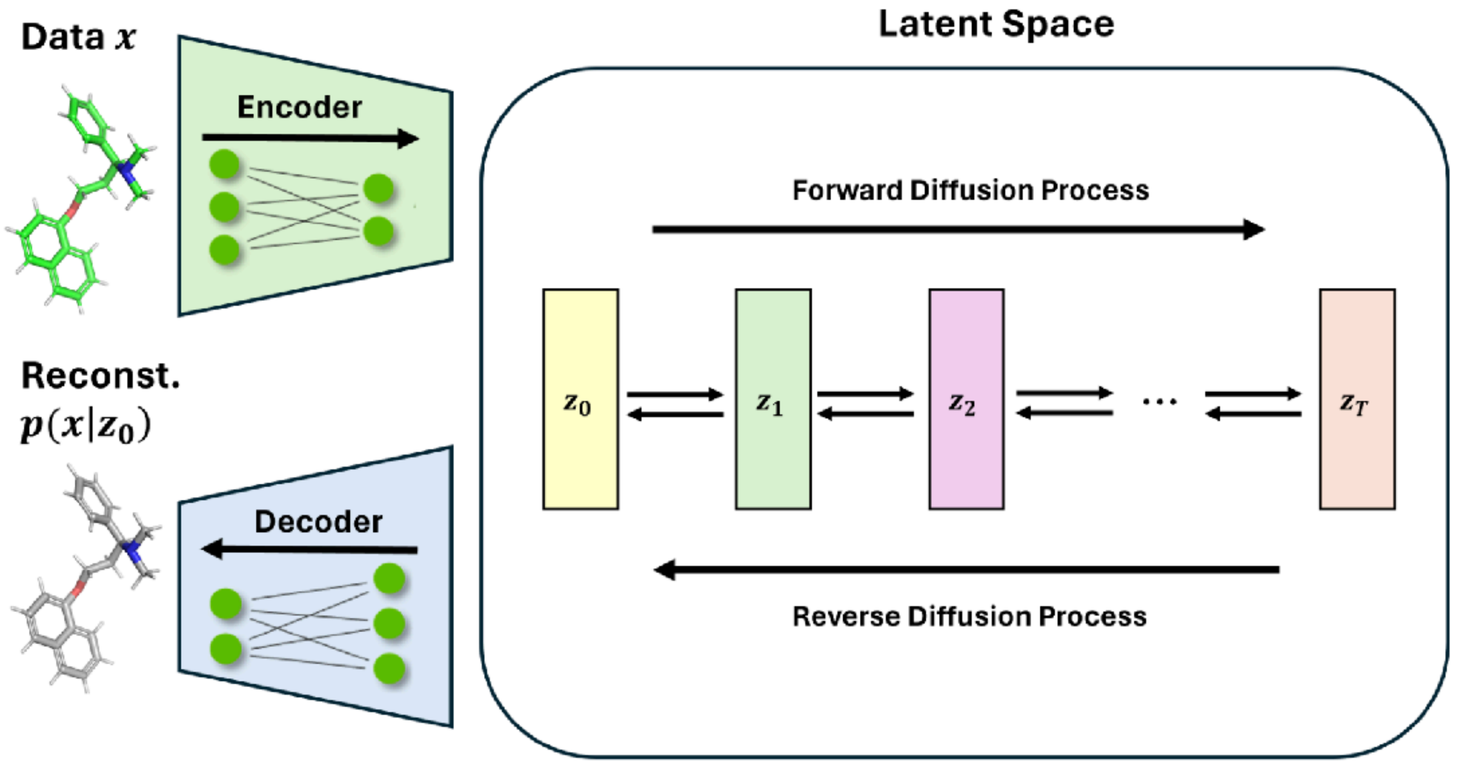
Future Directions:

Domain knowledge

Benchmarks & Software

Better conditioning

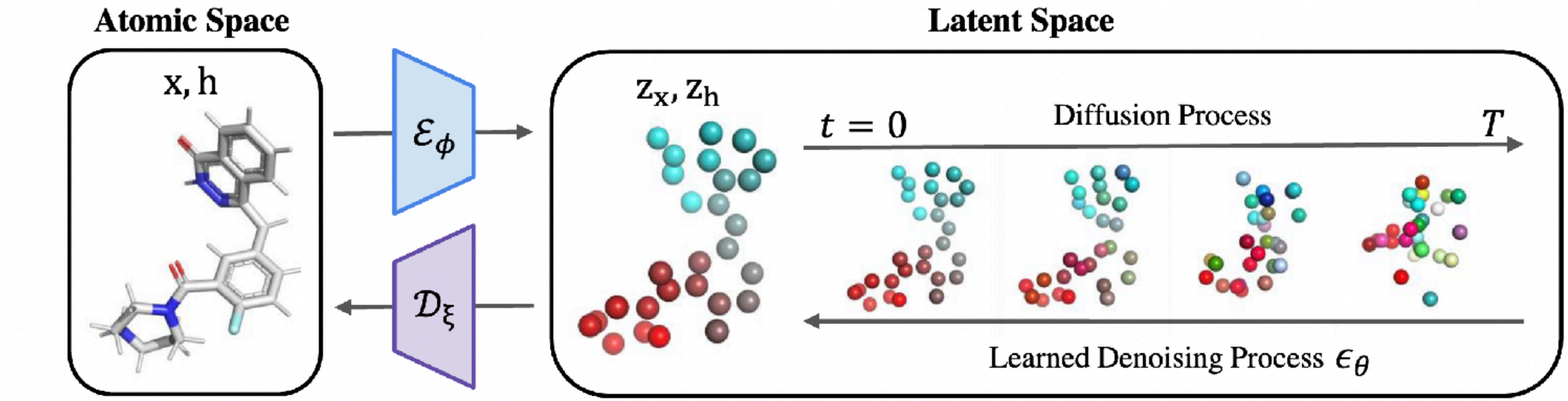
Latent Diffusion



Diffusion Models in De Novo Drug Design (Alakhdar et al.)

Geometric Latent Diffusion Models for 3D Molecule Generation (Xu et al.)

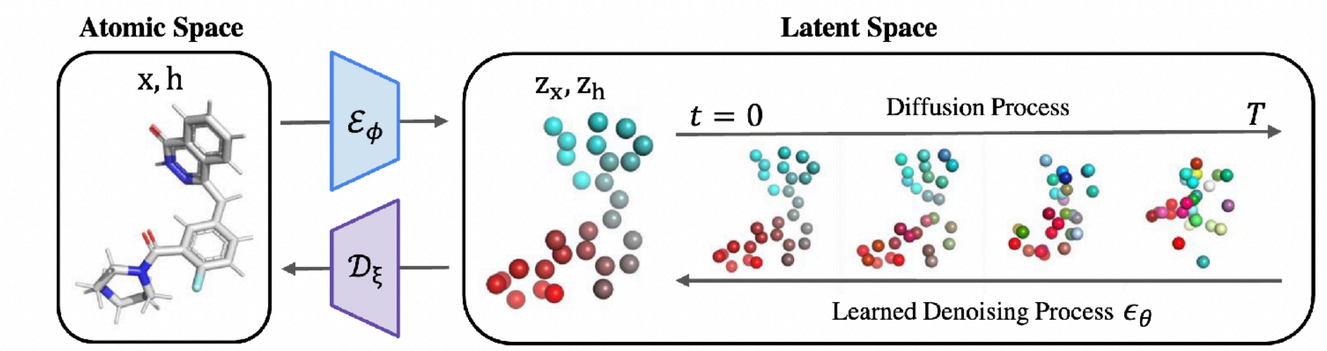
Latent Diffusion

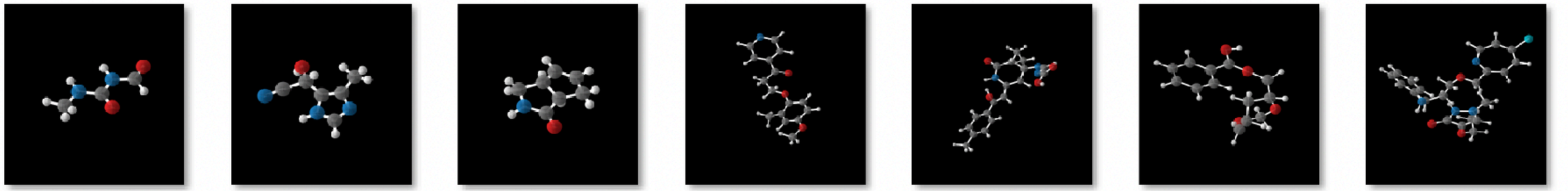


Diffusion Models in De Novo Drug Design (Alakhdar et al.)

Geometric Latent Diffusion Models for 3D Molecule Generation (Xu et al.)

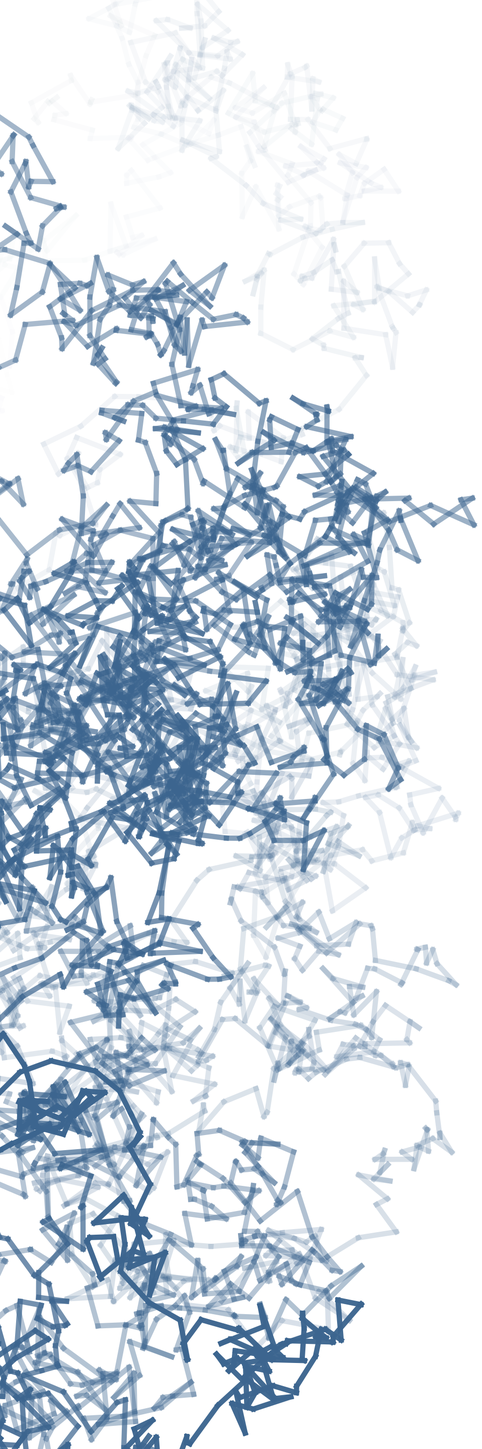
Latent Diffusion





Diffusion Models in De Novo Drug Design (Alakhdar et al.)

Geometric Latent Diffusion Models for 3D Molecule Generation (Xu et al.)



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.

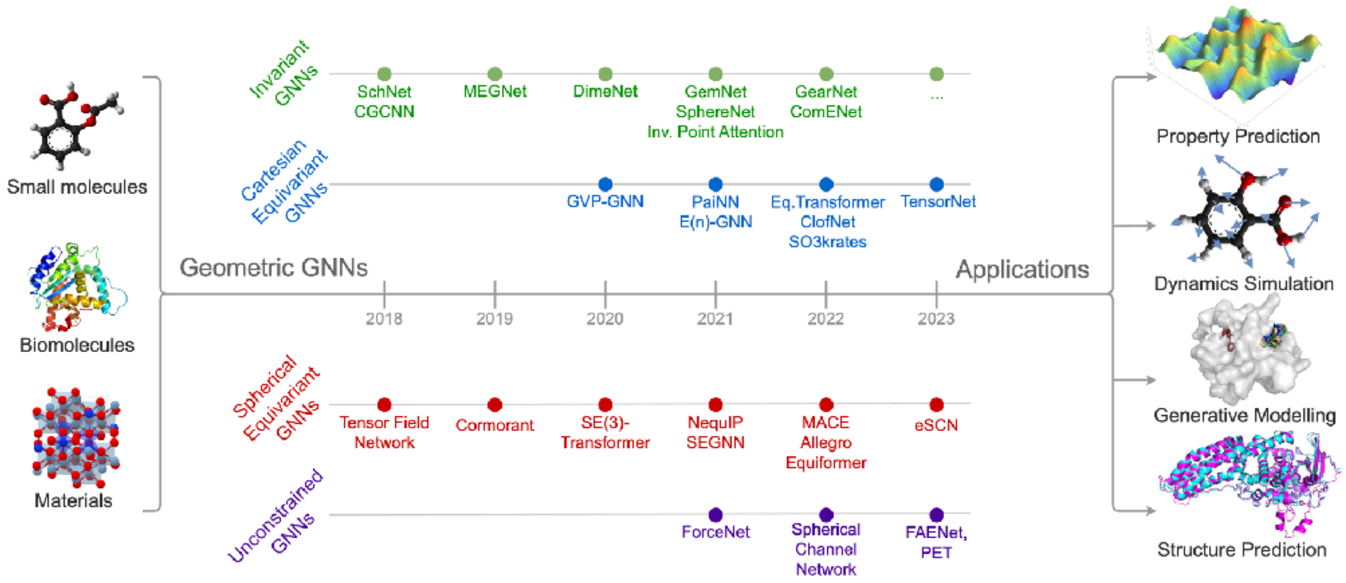
Geometric GNNs: Key Concepts

- **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.)