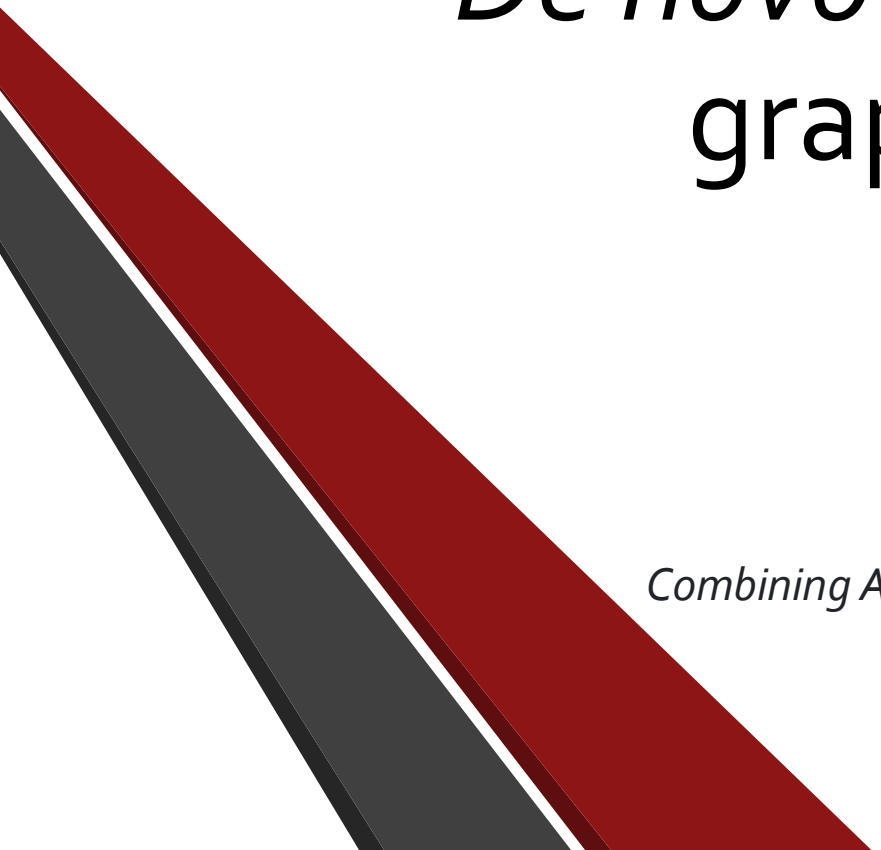


De novo molecular design using graph neural networks

Carlo Abate

Combining AI and physical modeling for contemporary simulations
CECAM-EPFL Workshop – Lausanne

Outline

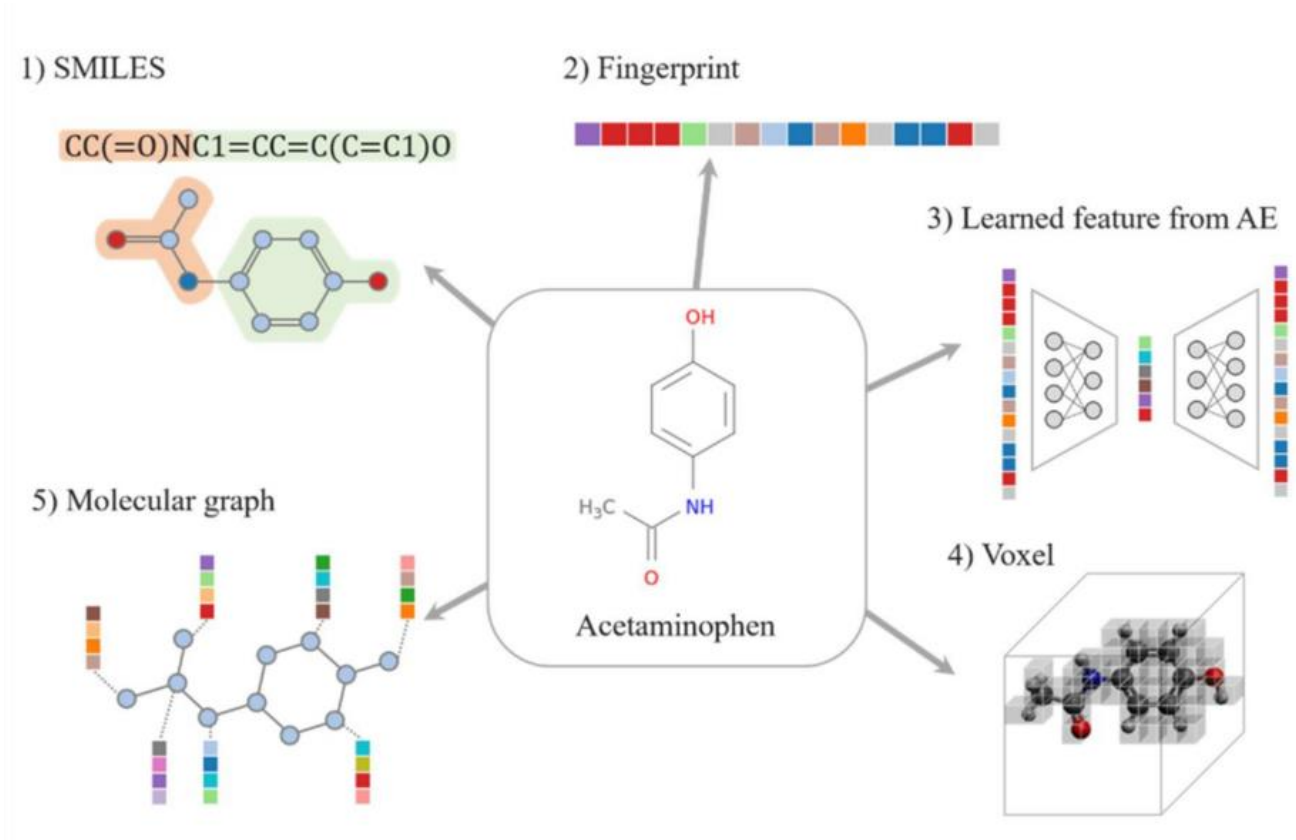
- Part I: Introduction to GNNs for conditional *de novo* drug design
- Part II: AMCG: A dual Atomic-Molecular Conditional Generator

Outline

- Part I: Introduction to GNNs for conditional *de novo* drug design
- Part II: AMCG: A dual Atomic-Molecular Conditional Generator

Molecule -> Graph

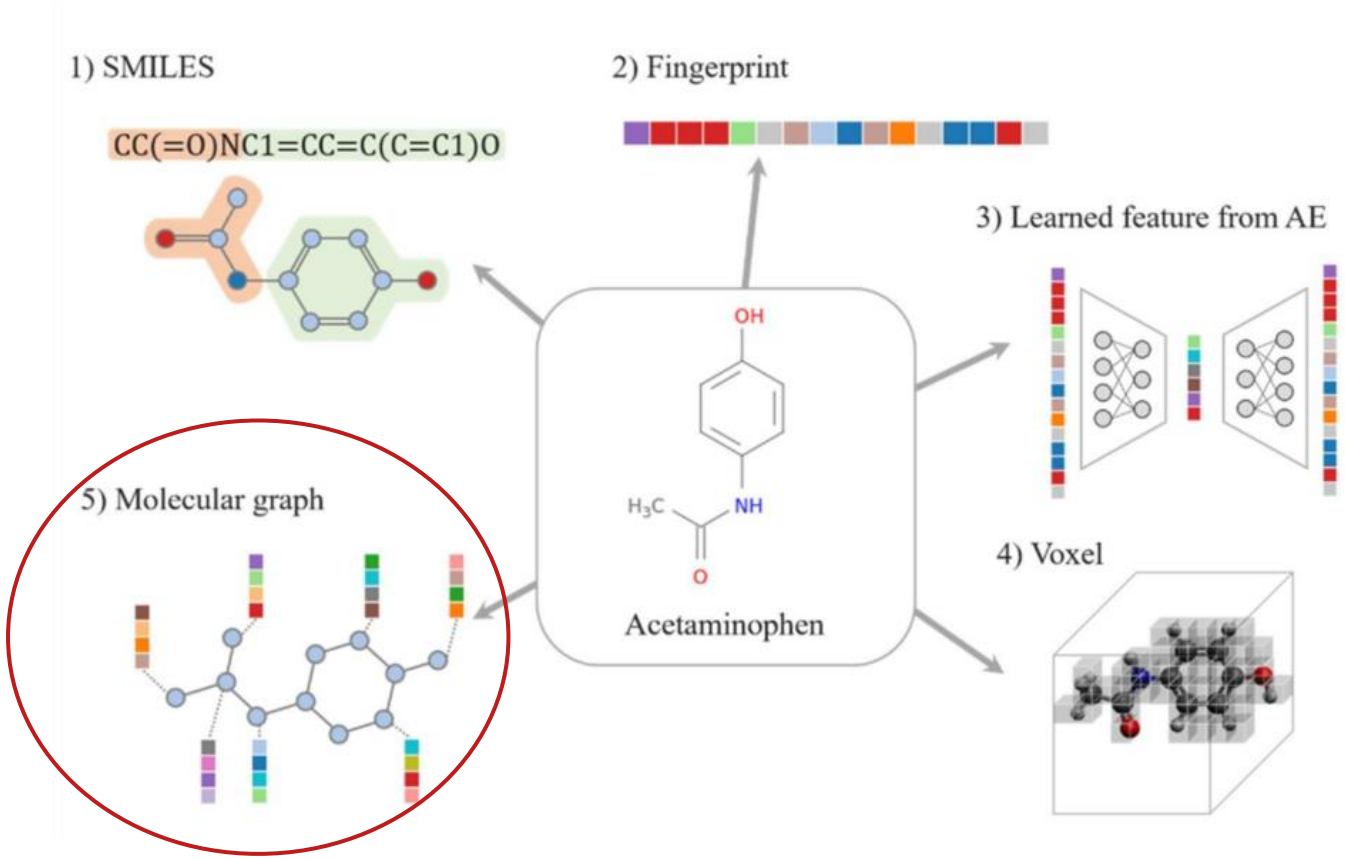
There are several ways to represent a molecule



Molecule -> Graph

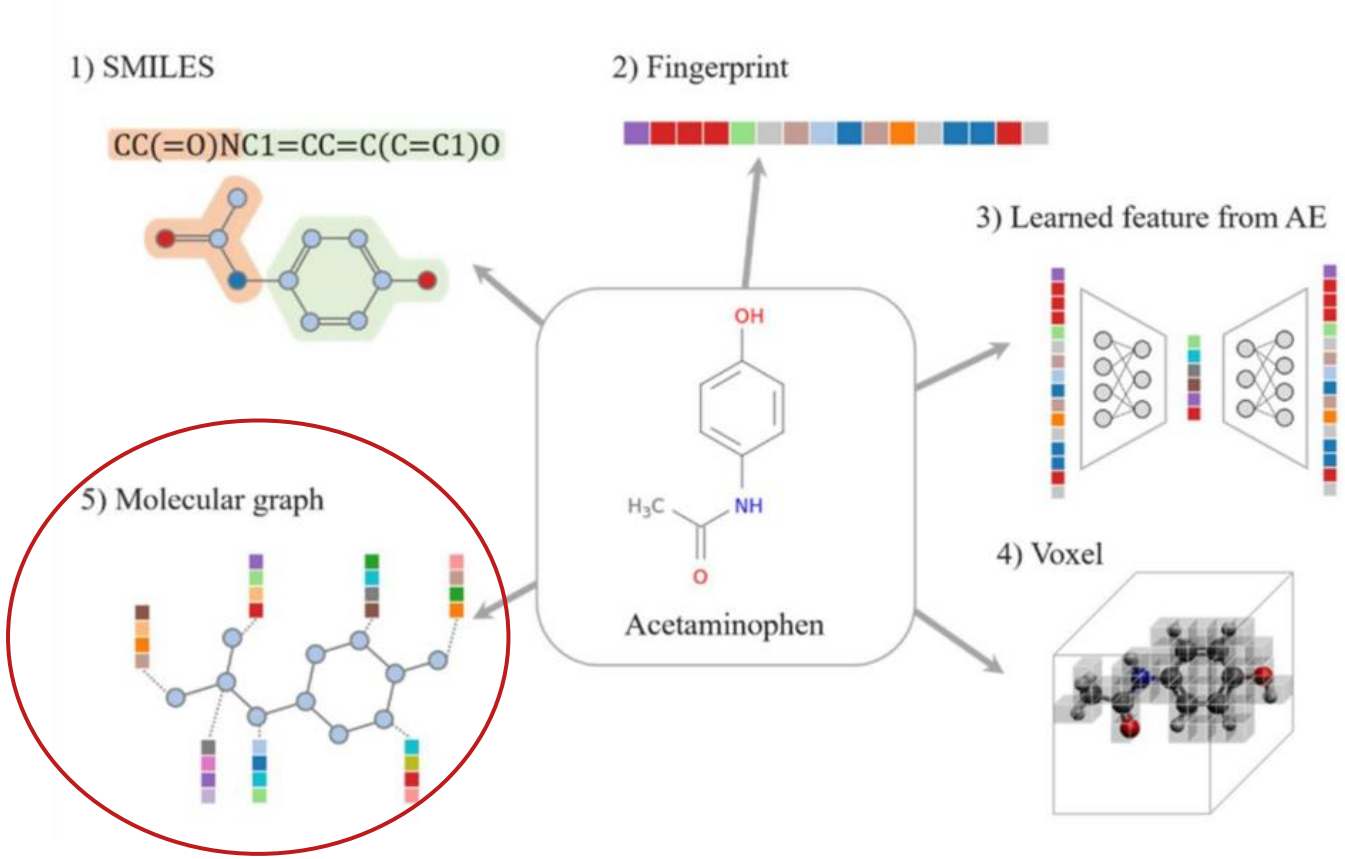
There are several ways to represent a molecule

Let's focus on molecular graphs



Molecule -> Graph

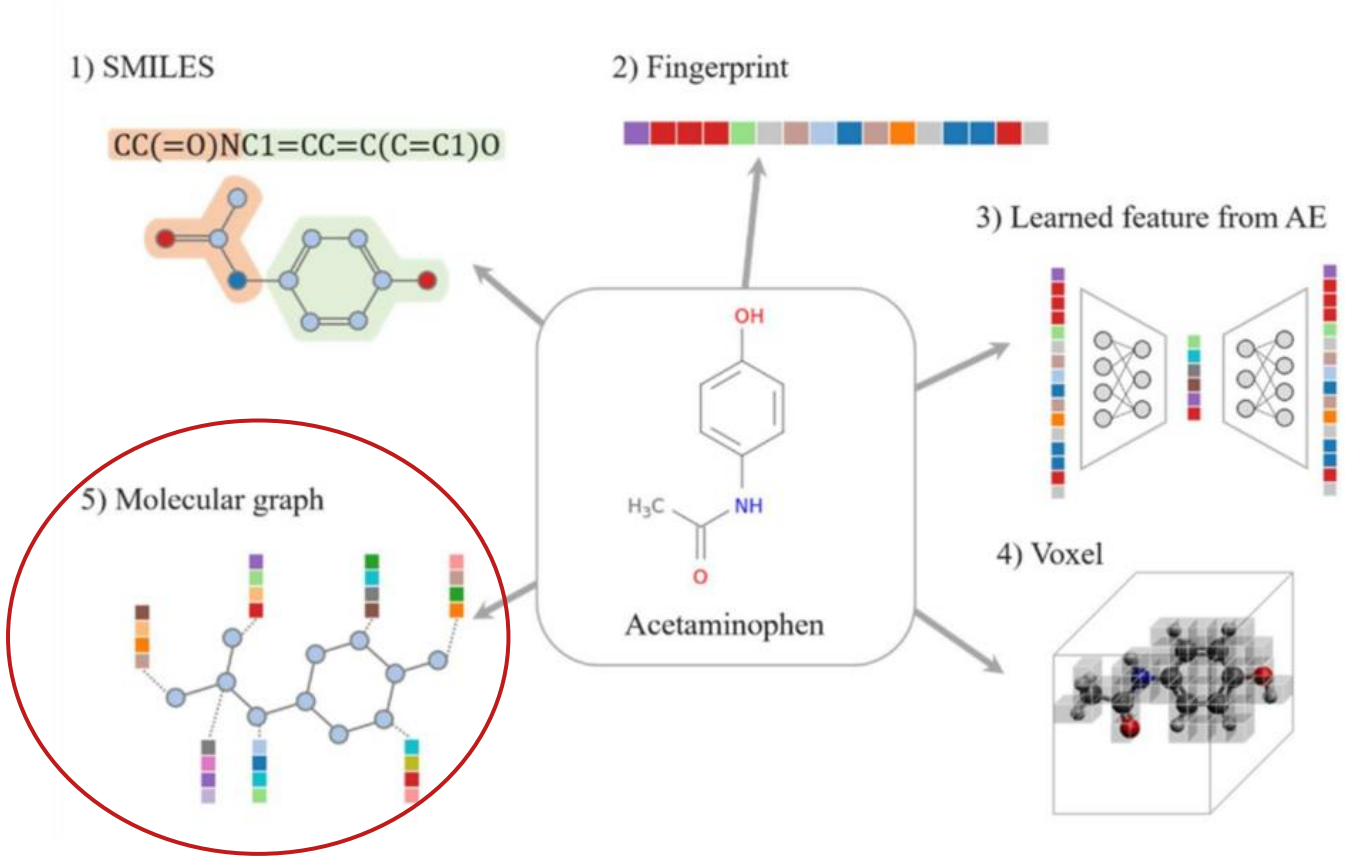
A graph is a pair $G=(V,E)$ where V is a set whose elements are called *vertices* and E is a set of pairs of vertices whose elements are called *edges*



Molecule -> Graph

A graph is a pair $G=(V,E)$ where V is a set whose elements are called *vertices* and E is a set of pairs of vertices whose elements are called *edges*

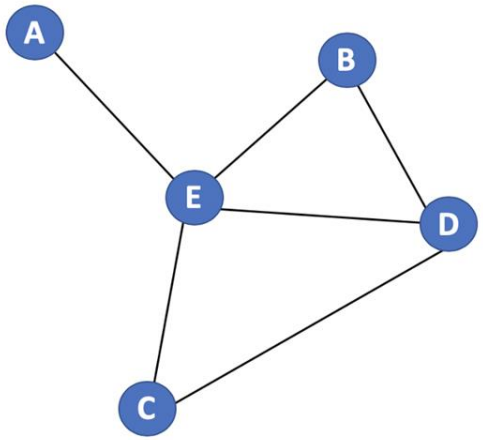
A molecule can be naturally represented as a graph in which the set of nodes is the set of atoms and the set of edges is the set of bonds



Molecule -> Graph

Each node is equipped with a *feature vector*

A graph can be seen as a pair of matrices (F, A) where F is the *feature matrix* (containing single node information) and A is the *adjacency matrix* (containing neighbourhood information)



Graph G

	A	B	C	D	E
A	0	0	0	0	1
B	0	0	0	1	1
C	0	0	0	1	1
D	0	1	1	0	1
E	1	1	1	1	0

Adjacency matrix A

	A	B	C	D	E
A	1	0	0	0	0
B	0	2	0	0	0
C	0	0	2	0	0
D	0	0	0	3	0
E	0	0	0	0	4

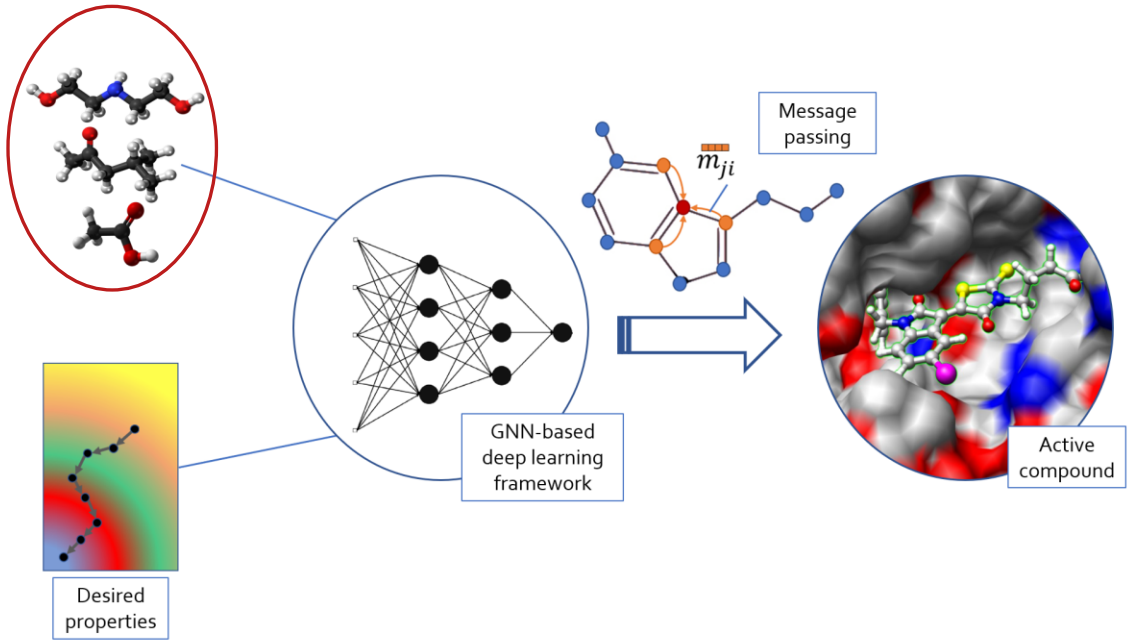
Degree matrix D

A	-1.1	3.2	4.2
B	0.4	5.1	-1.2
C	1.2	1.3	2.1
D	1.4	-1.2	2.5
E	1.4	2.5	4.5

Feature vector X

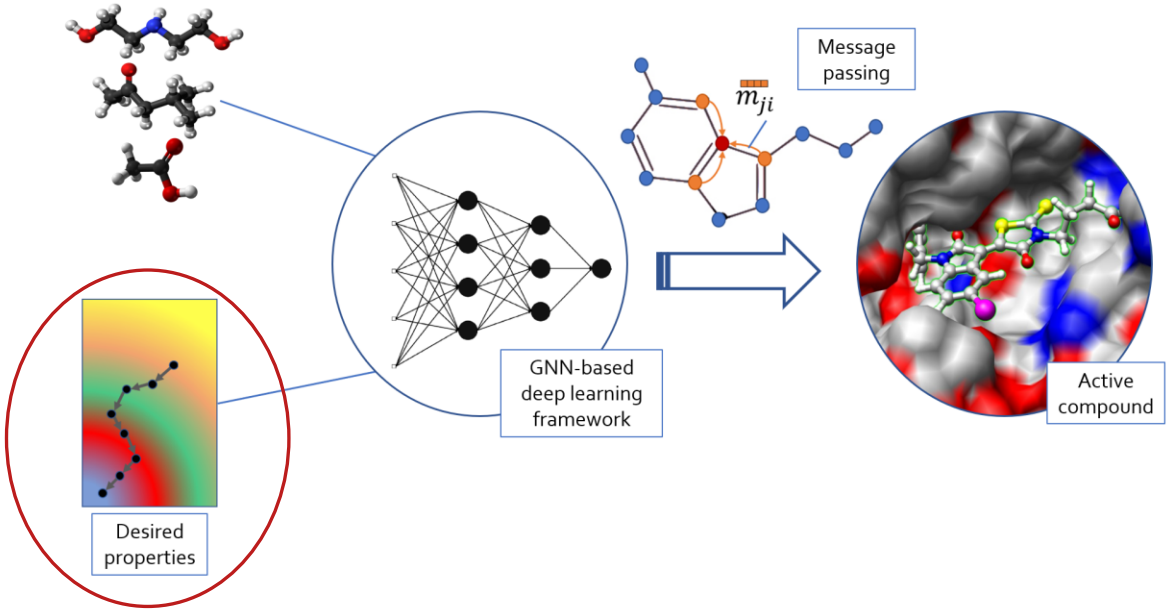
A GNN-based approach

The model is fed with a molecular dataset and learns its latent distribution (i.e. the generated molecules can reasonably belong to the given dataset)



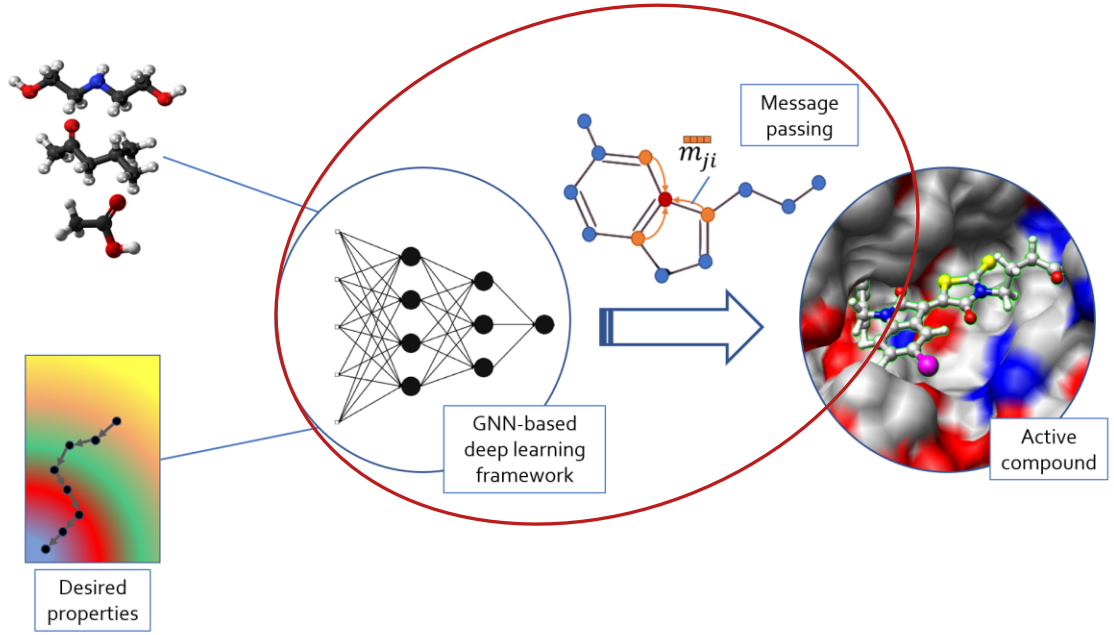
A GNN-based approach

The generation is conditioned towards the optimization of desired (numerical) properties (such as QED, SAS, biological activity)



A GNN-based approach

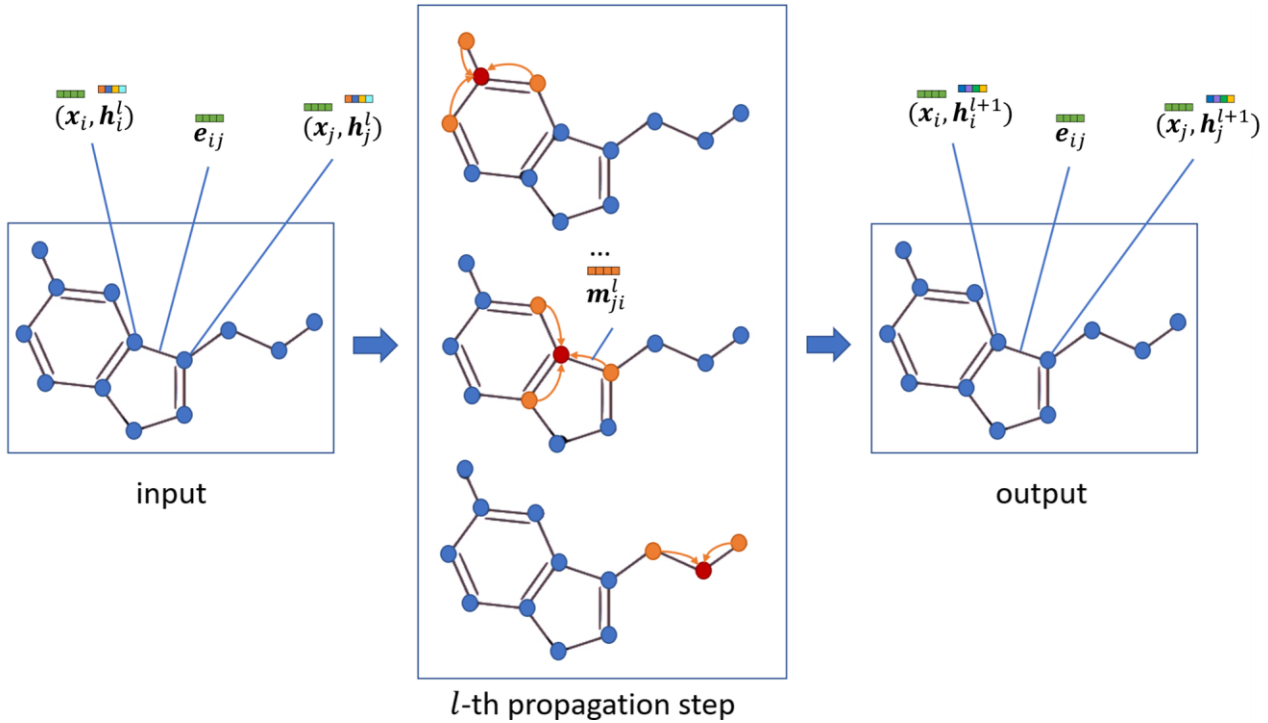
This is obtained using Graph Neural Networks which rely on message passing modules



Message Passing Neural Networks

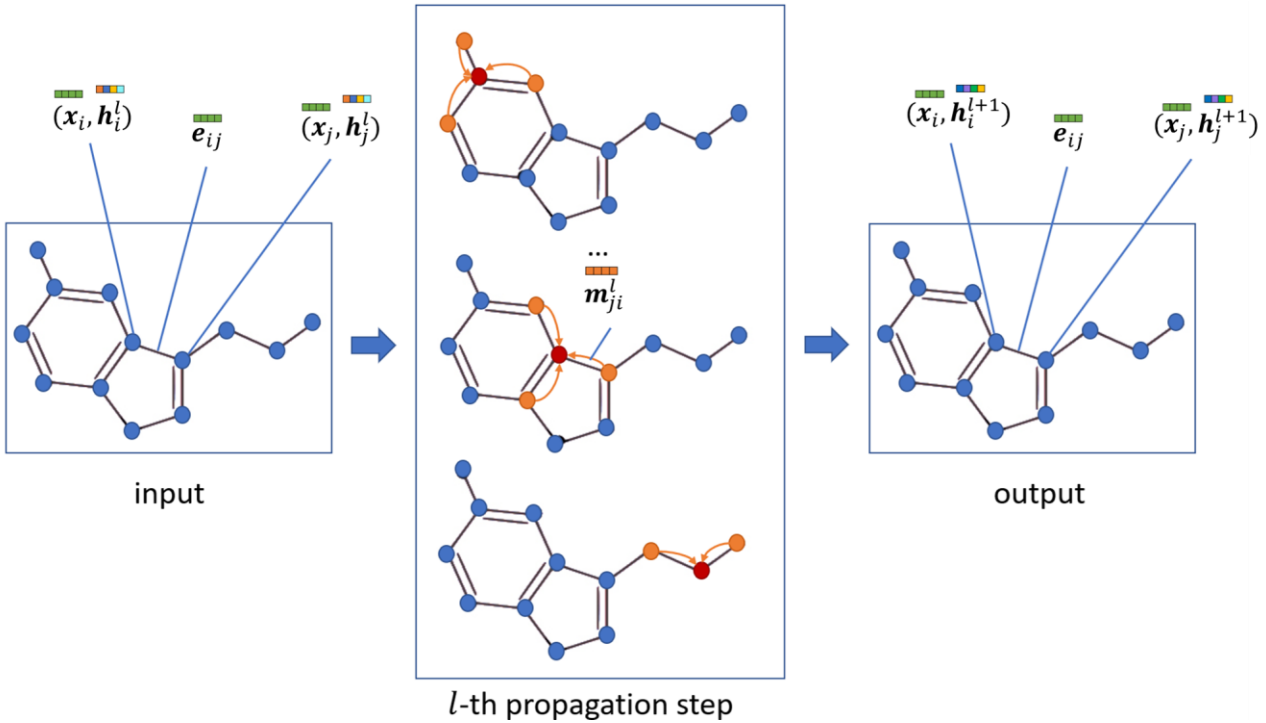
A general framework for GNNs

All the main currently available GNN-based generative methods for molecular graphs are MPNNs



Message Passing Neural Networks

The information h_i^{l+1} at each node at step $l+1$ is obtained by the information present in its neighbors at step l



Message Passing Neural Networks

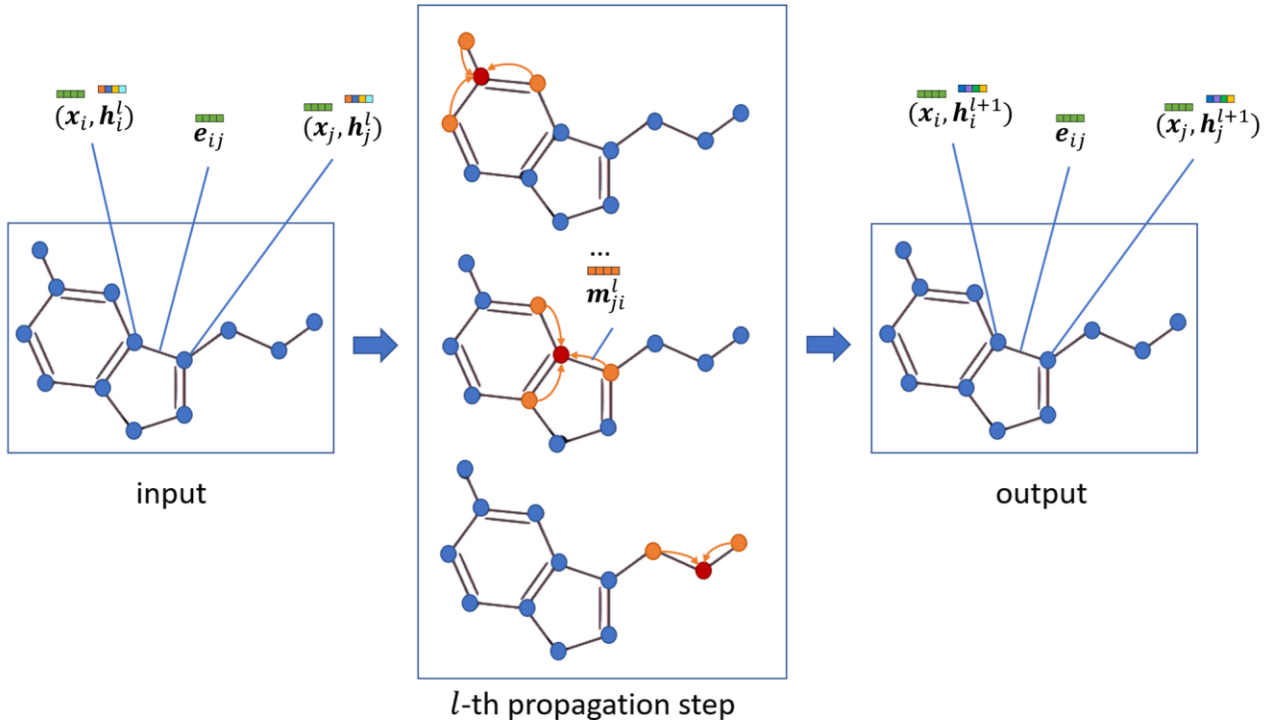
The information h_i^{l+1} at each node at step $l+1$ is obtained by the information present in its neighbors at step l

A message passing module updates the information as follows:

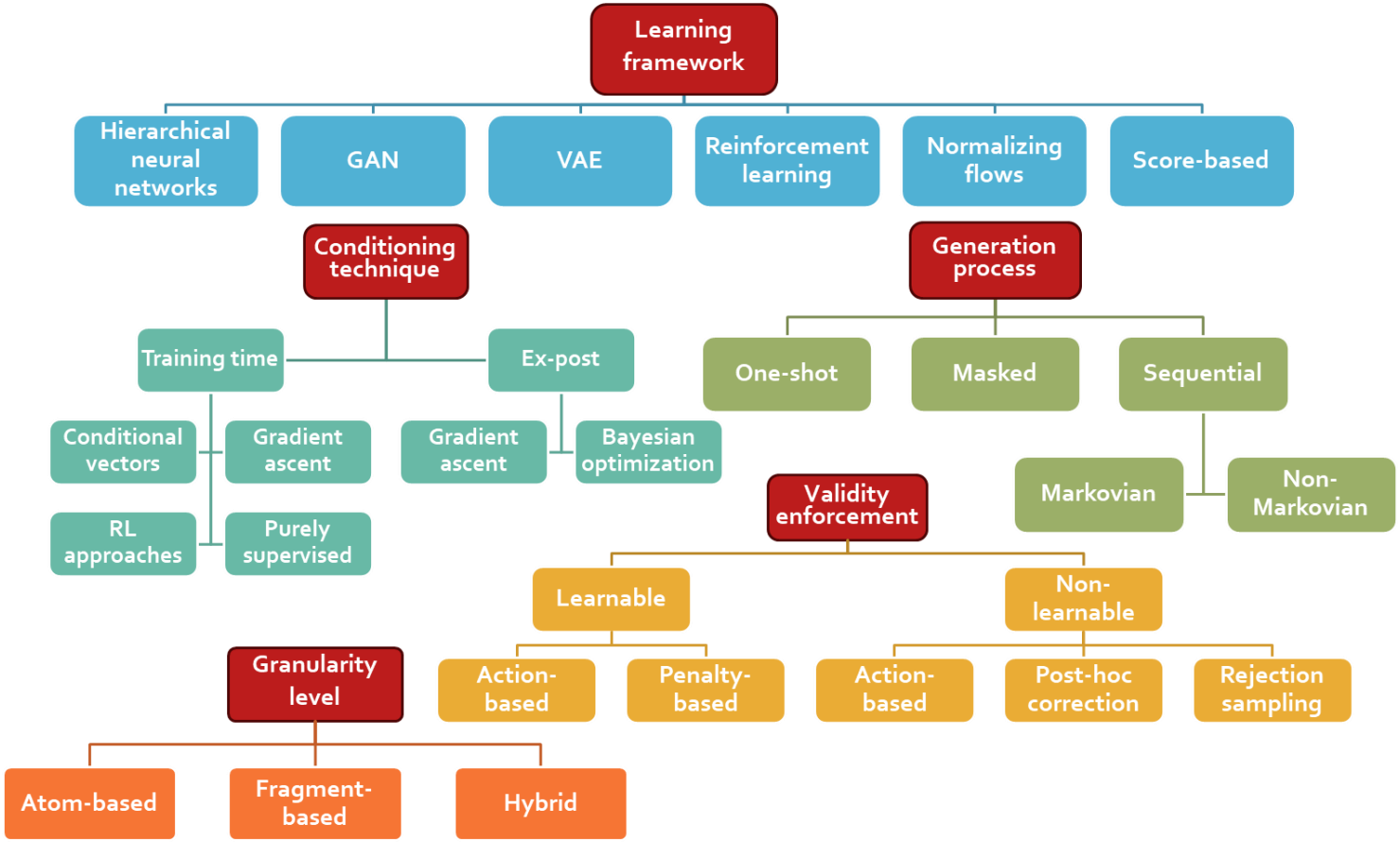
$$\boldsymbol{m}_{ji}^{l} = M_l(\boldsymbol{x}_i, \boldsymbol{x}_j, a_{ij}, \boldsymbol{e}_{ij}, \boldsymbol{h}_i^l, \boldsymbol{h}_j^l)$$

$$\boldsymbol{m}_i^{l} = A_l(\boldsymbol{m}_{ji}^l, v_j \in \mathcal{N}(v_i))$$

$$\boldsymbol{h}_i^{l+1} = U_l(\boldsymbol{x}_i, \boldsymbol{h}_i^l, \boldsymbol{m}_i^l)$$

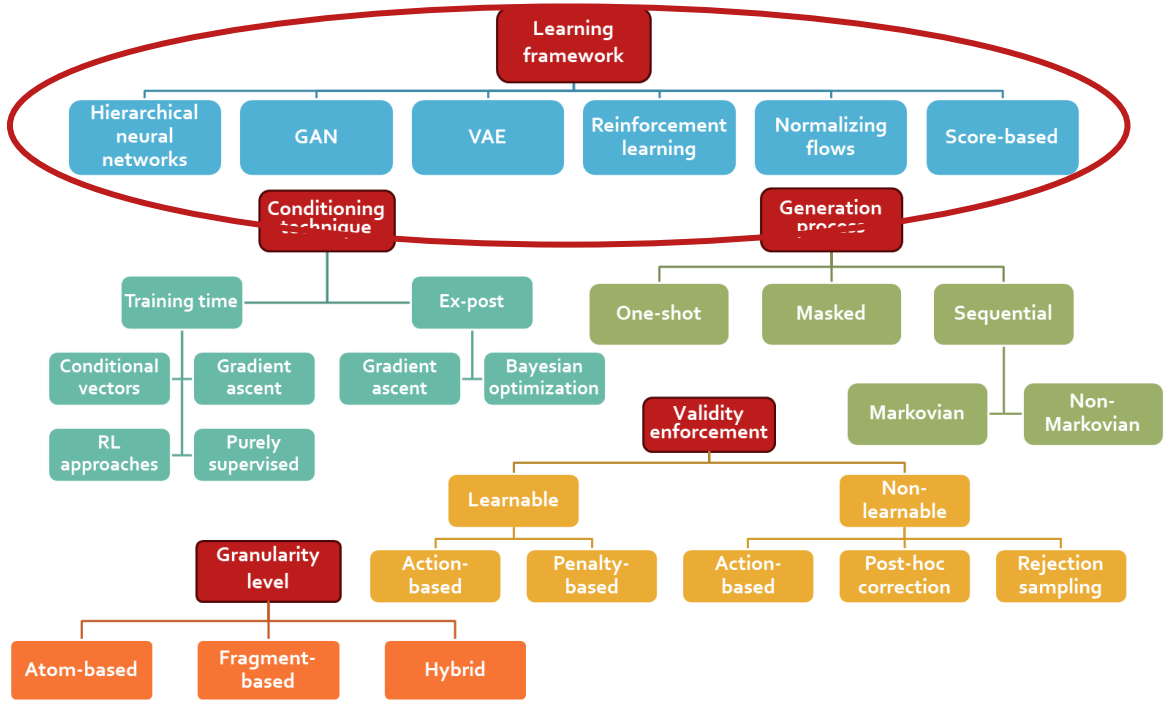


A GNN-based approach



Abate, C, Decherchi, S, Cavalli, A. Graph neural networks for conditional de novo drug design. *WIREs Comput Mol Sci*. 2023; 13(4):e1651. https://doi.org/10.1002/wcms.1651

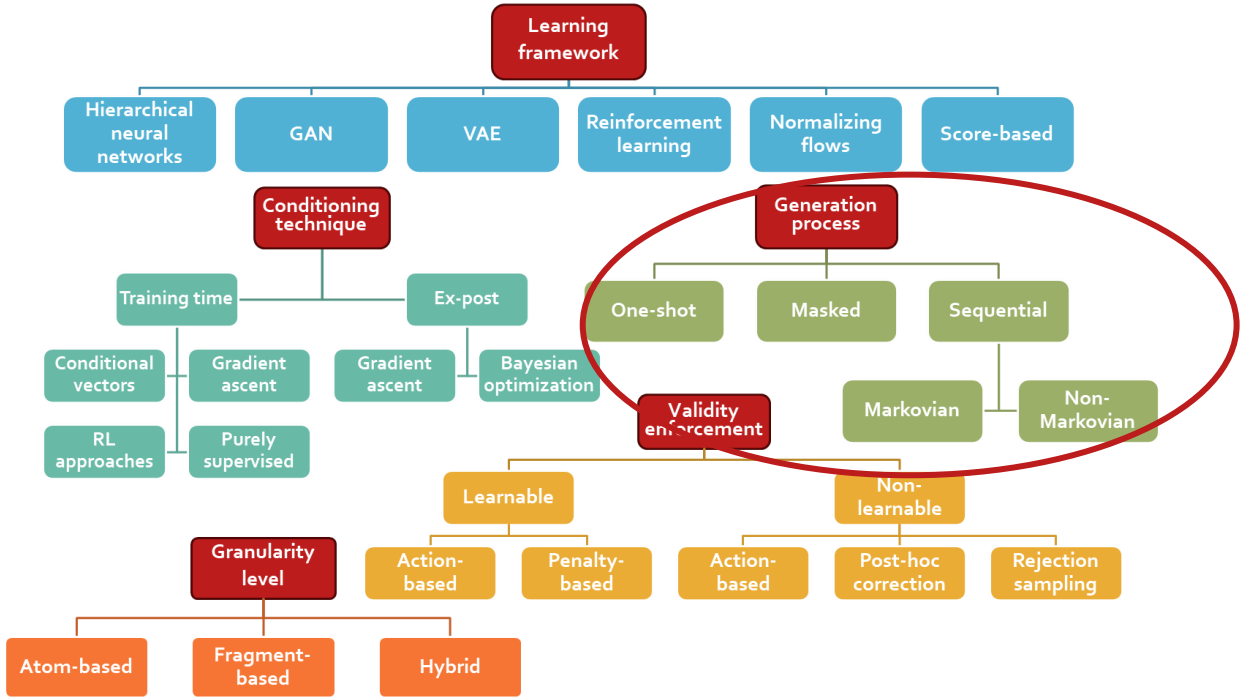
A GNN-based approach



It is the foundational architecture that defines how the model learns

Both derives from and impacts the other, task-specific modeling choices

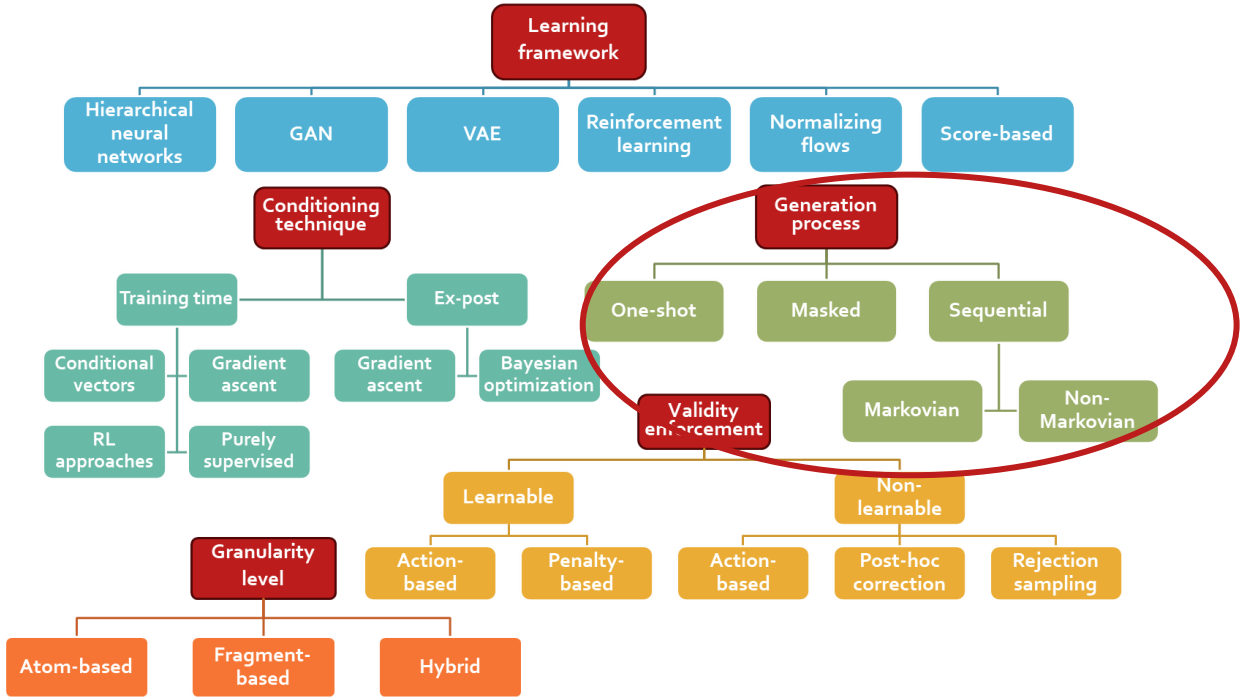
A GNN-based approach



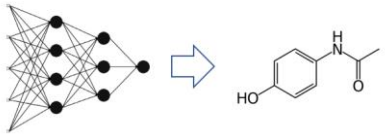
Determines the way the generation process is modeled

Offers a trade-off between generation speed and structural control

A GNN-based approach

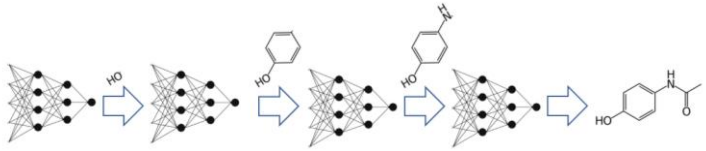


One-shot:



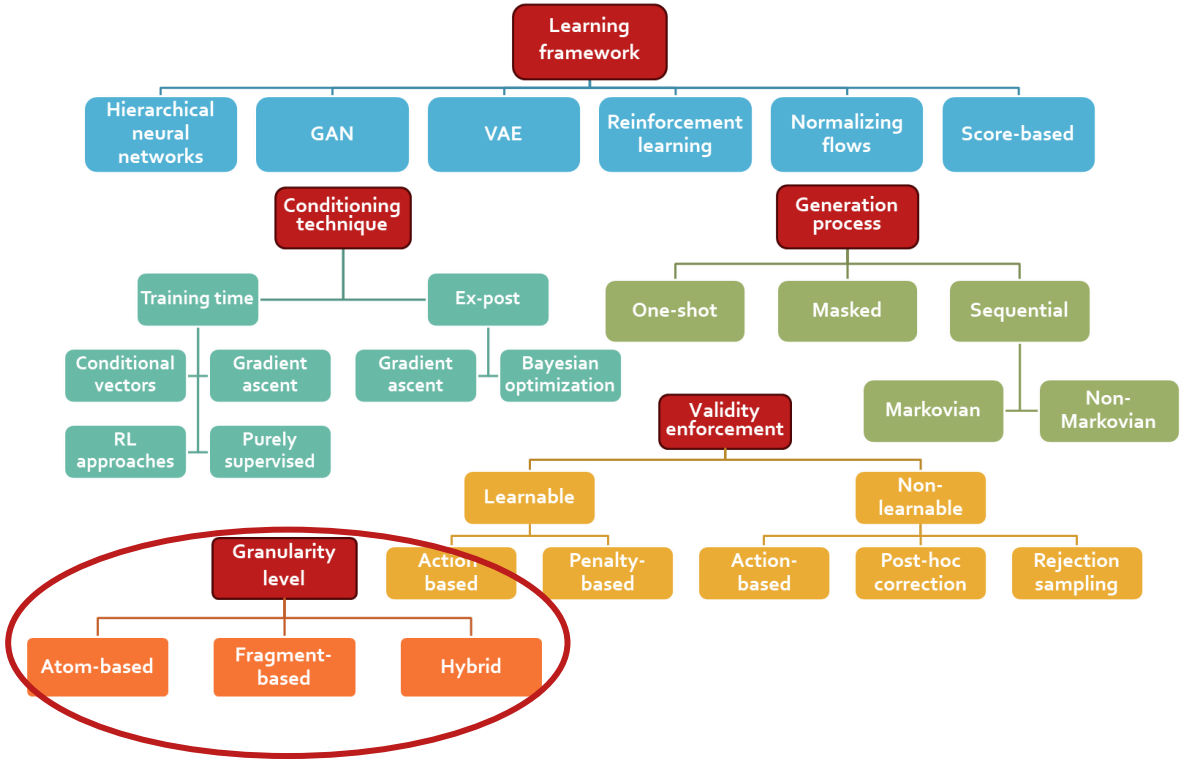
$$p_\theta(G) = p_\theta(G|\boldsymbol{z})$$

Sequential:



$$p_\theta(G) = \prod_t p_\theta(a_t|G_t, \ldots, G_0)$$

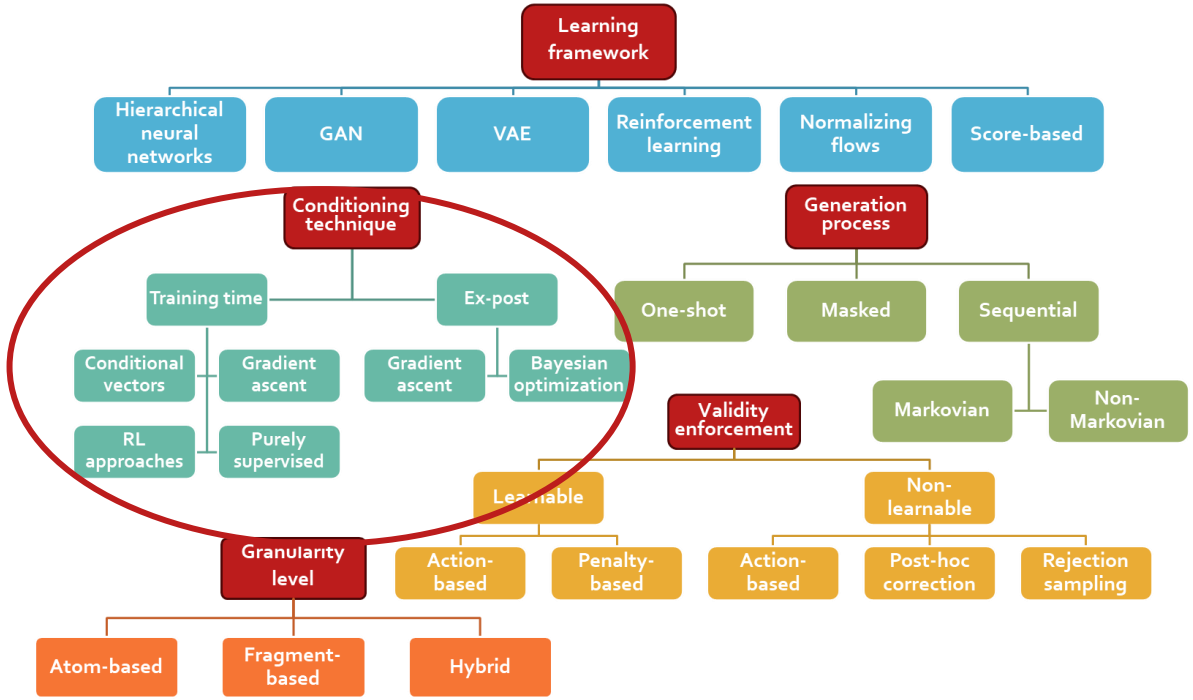
A GNN-based approach



Defines the meaning of every node of the graph

It affects how the model explores the chemical space and its ability to generate realistic structures.

A GNN-based approach



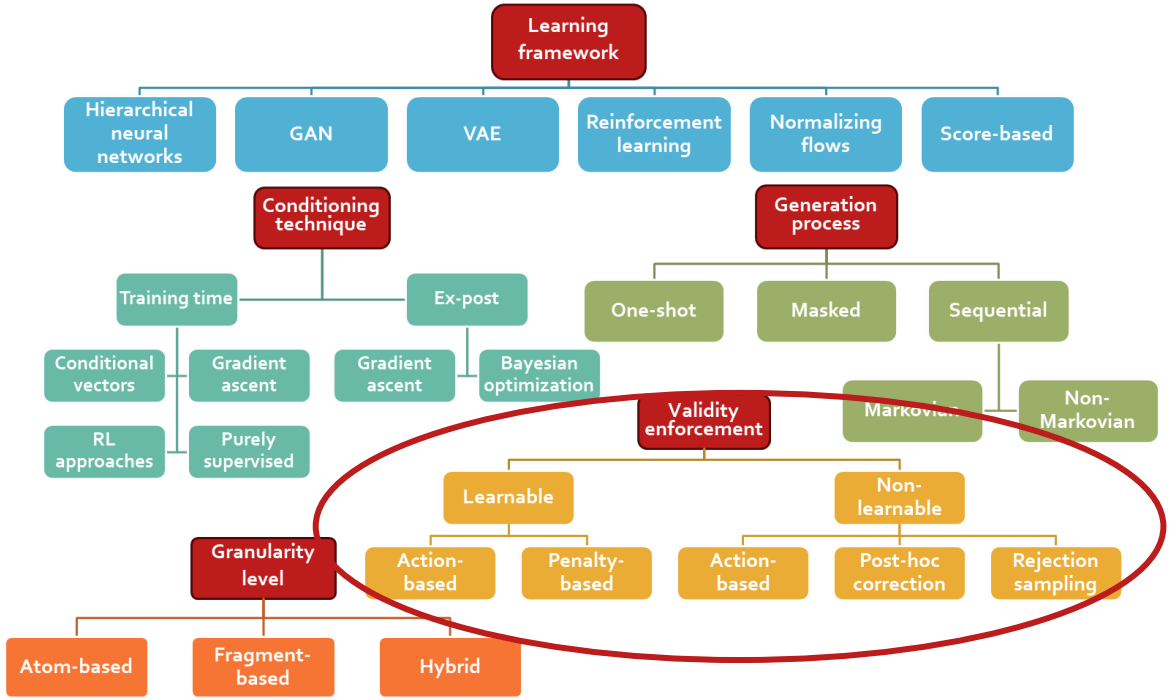
Training time:

The model learns to incorporate property objectives during training through conditional vectors, rewards, or auxiliary losses

Ex-post:

The generation is first performed unconditionally, then the latent space is navigated to find molecules with desired properties

A GNN-based approach



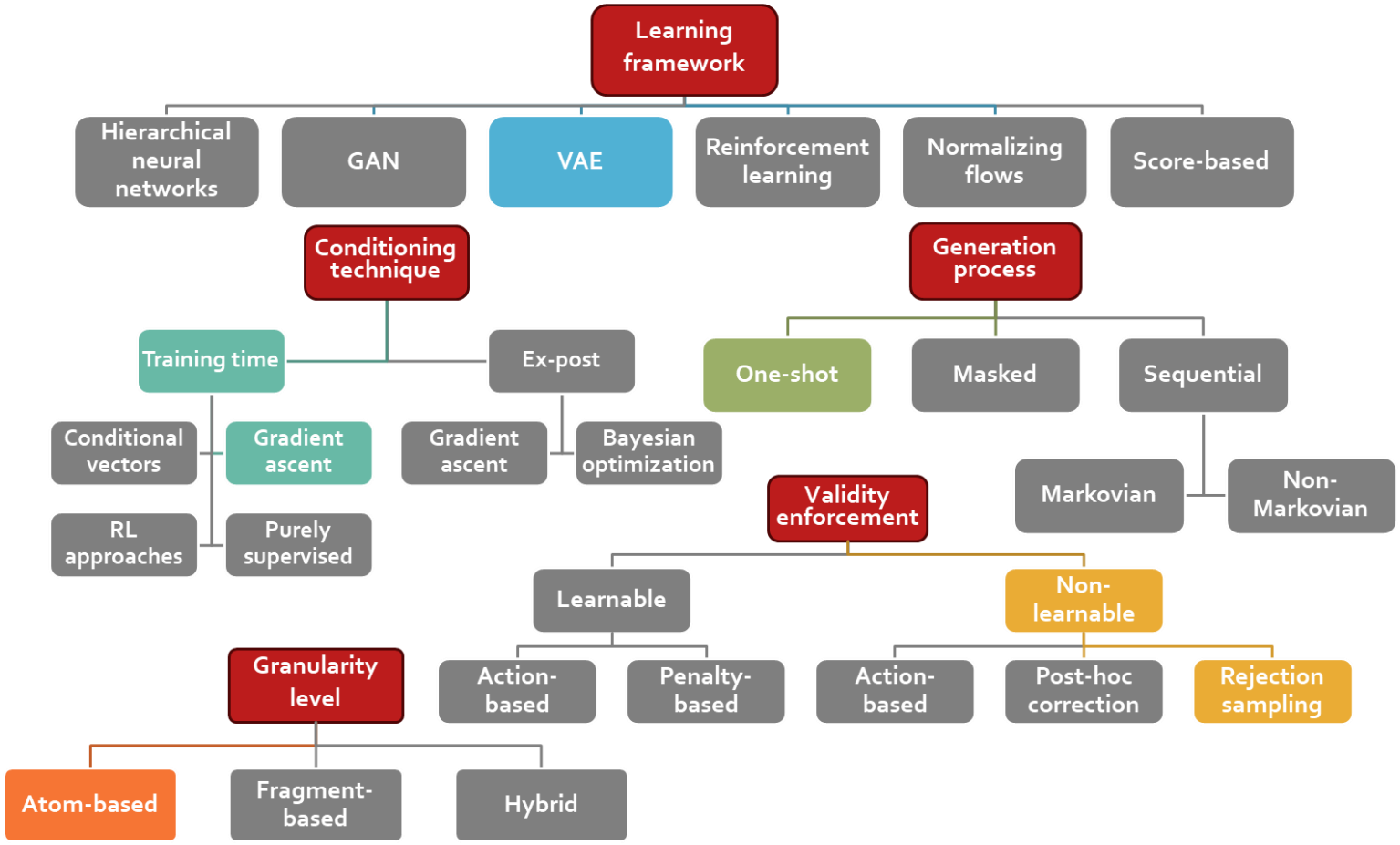
Ensures generated molecules obey chemical rules and constraints

Can be implemented through learning or post-processing approaches

Outline

- Part I: Introduction to GNNs for conditional de novo drug design
- Part II: AMCG: A dual Atomic-Molecular Conditional Generator

AMCG framework



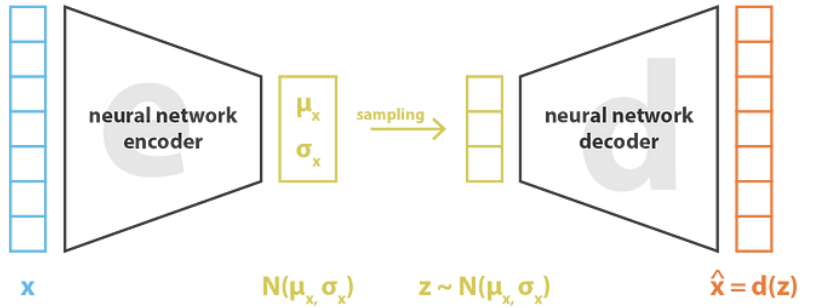
AMCG framework

Main idea

Navigable latent space («VAE-like») embedding for molecular graphs

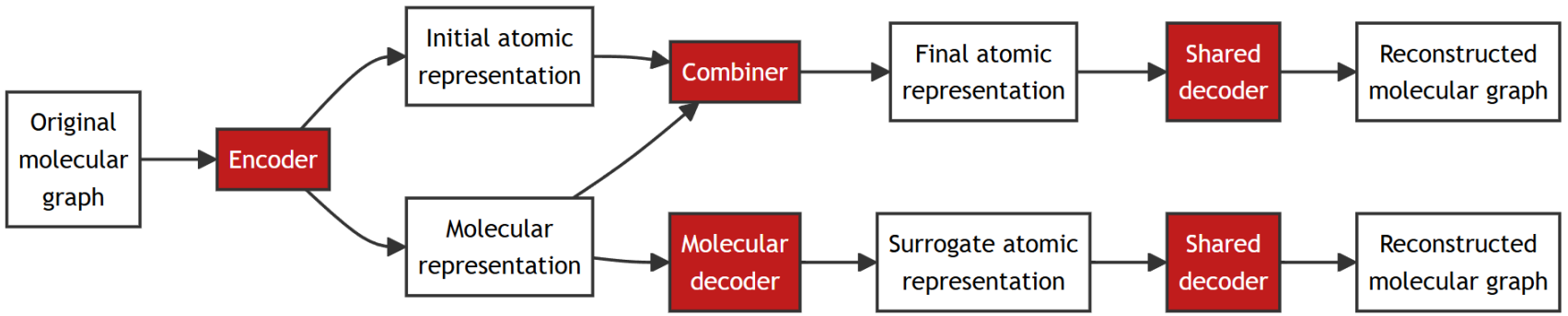
Features

- Modularity
- Scalability
- No max number of atoms
- Conditionability wrt atom types histogram
- Conditionability wrt molecular properties



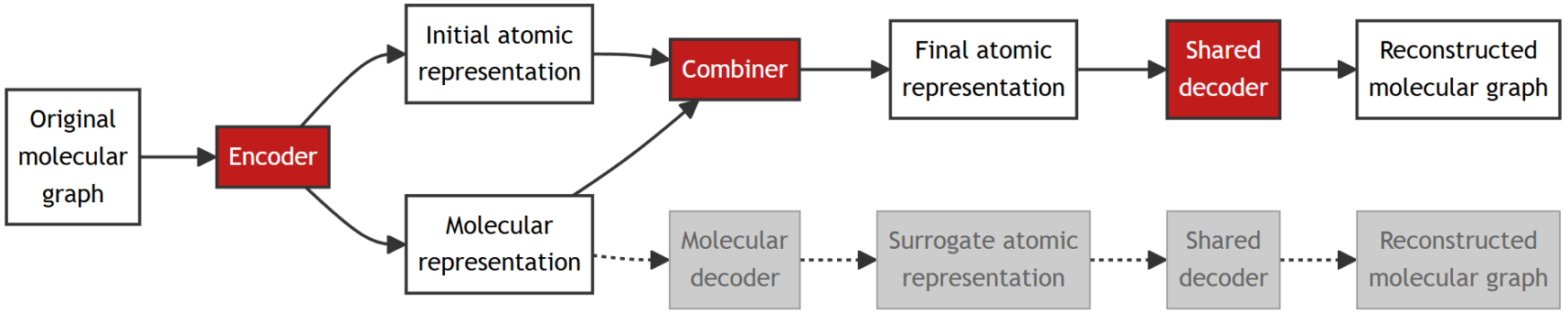
https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Training process



We make use of a *self-distilling* approach

Training process

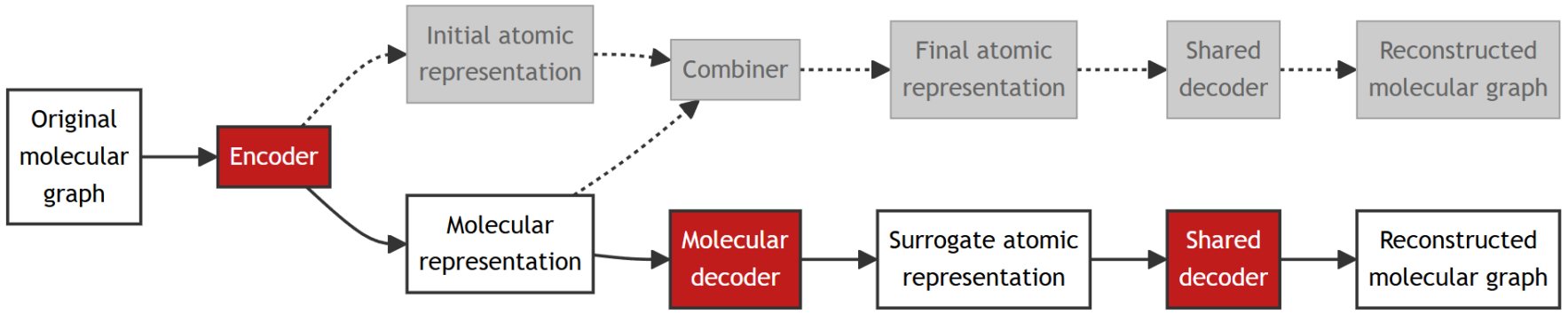


The *teacher* branch of the network uses all the available information

Quick learner - loss fast to converge

Hard to control

Training process

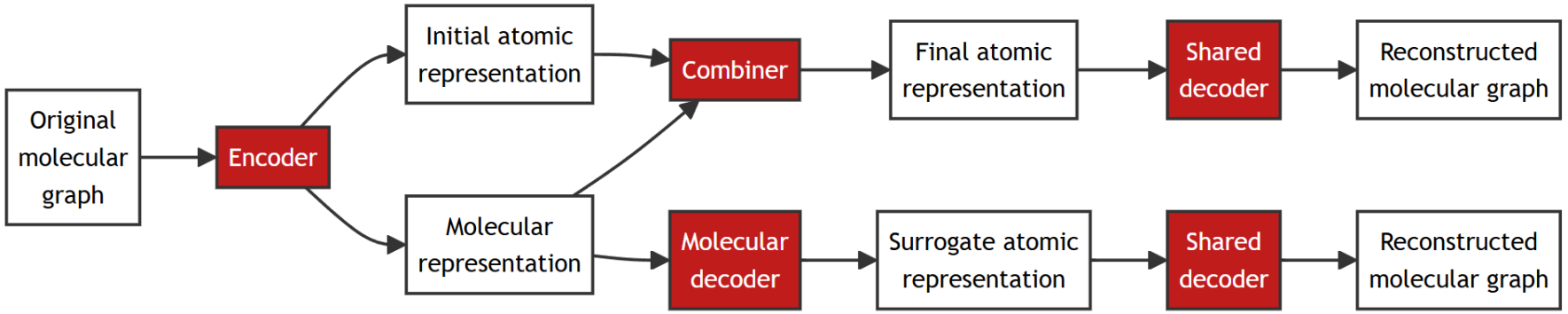


The *student* branch of the network uses only the molecular representation

Hard to train

Easy to control – standard conditioning techniques

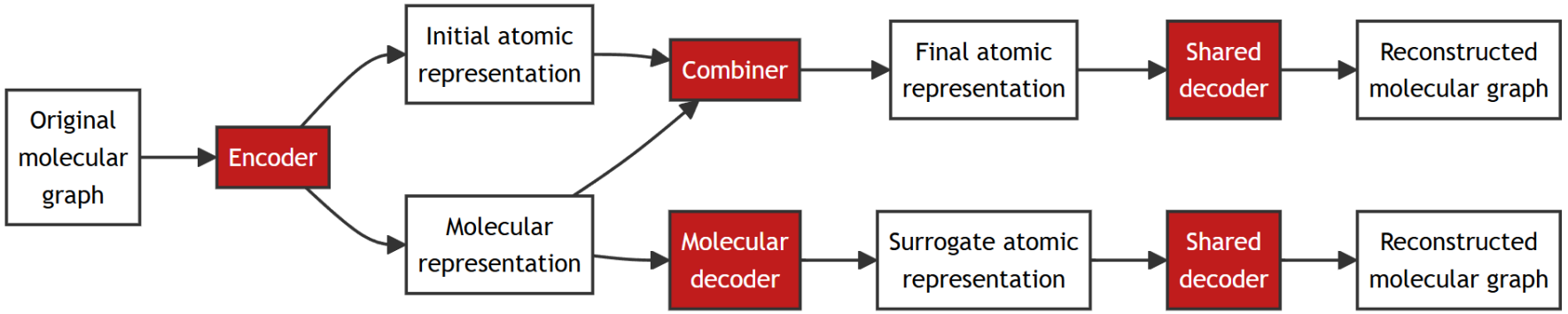
Training process



Main idea

- Using the (easy-to-train) teacher to guide the student model a molecular latent space

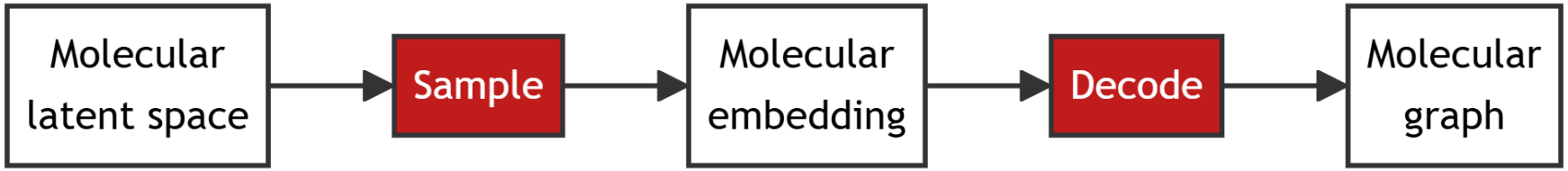
Training process



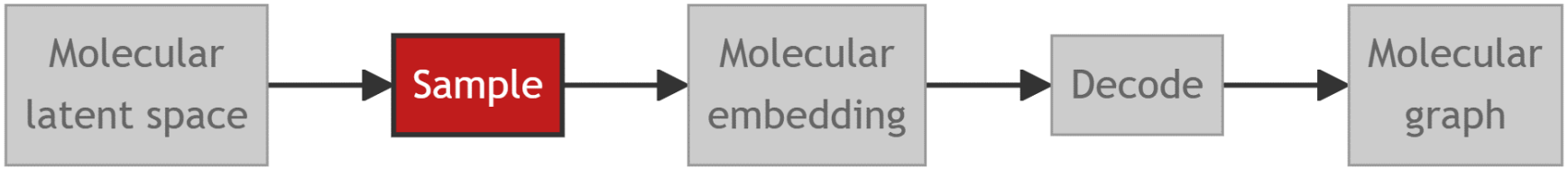
Main idea

- Using the (easy-to-train) teacher to guide the student model a molecular latent space
- Using the (easy-to-control) student to condition the generation towards the optimization of desired properties

Generation process



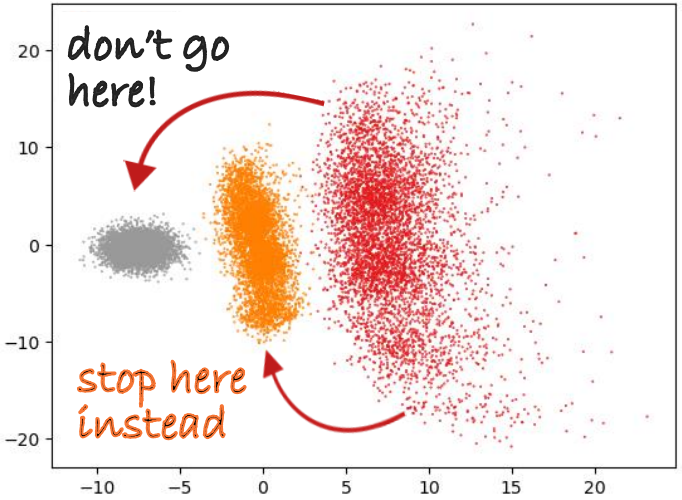
Generation process



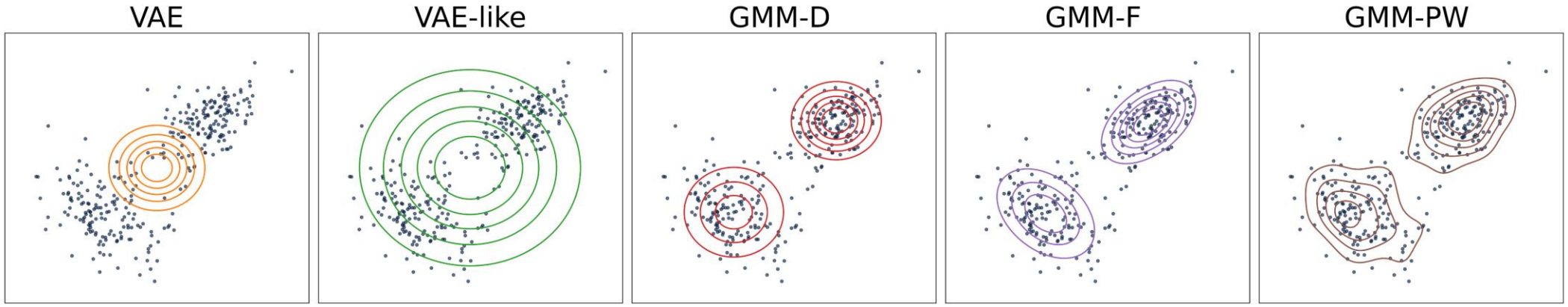
Classical VAEs sample from a unit Gaussian.

We utilize Gaussian Mixture Models (GMMs).

- Overall model easier to train
- Enables conservative and explorative generation
- Fast (with respect to diffusion in latent space)

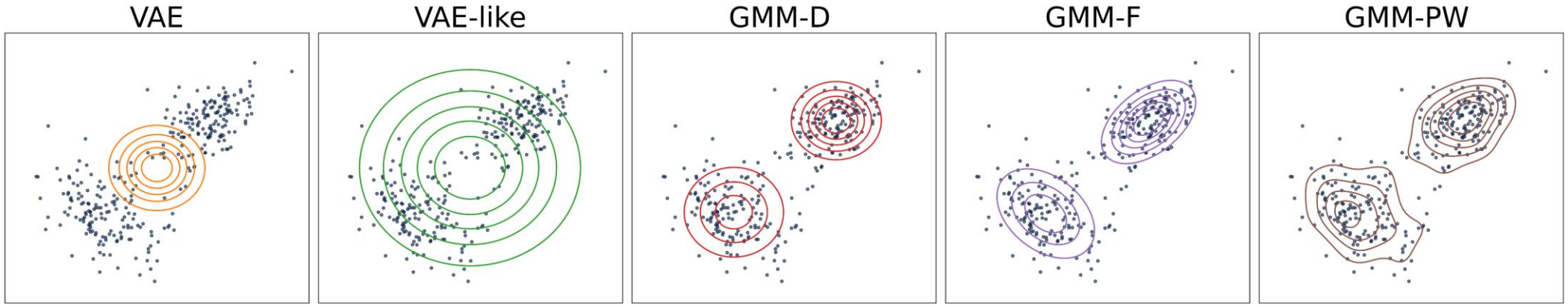


Sampling the right space



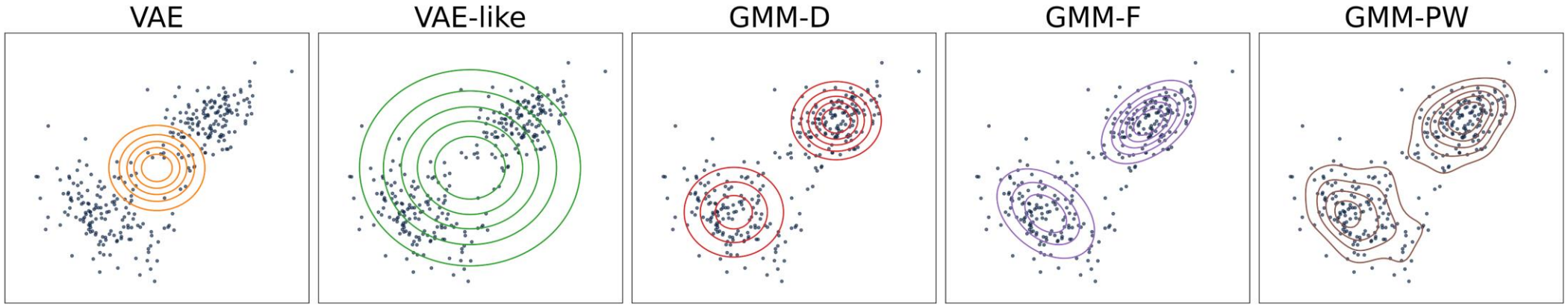
- VAE: unit Gaussian centered in 0

Sampling the right space



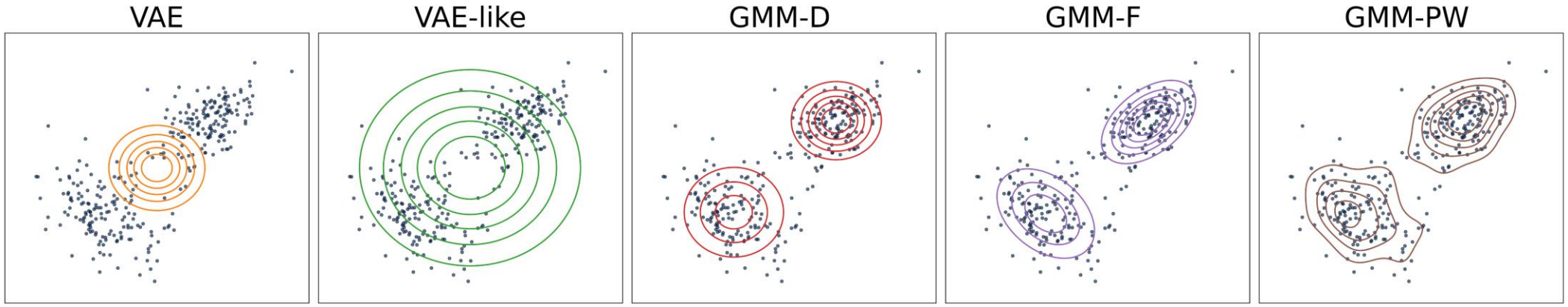
- VAE: unit Gaussian centered in 0
- VAE-like: single Gaussian centered in μ with diagonal covariance matrix

Sampling the right space



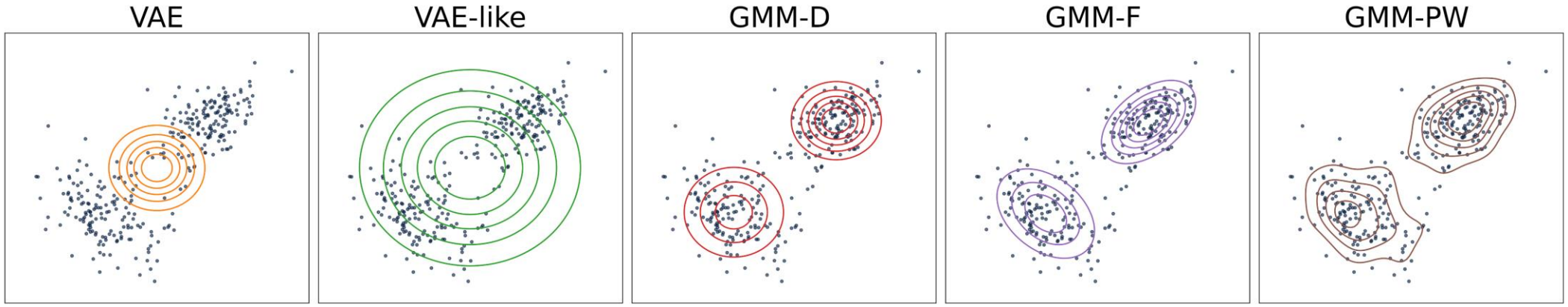
- VAE: unit Gaussian centered in 0
- VAE-like: single Gaussian centered in μ with diagonal covariance matrix
- GMM-D: combination of multiple Gaussians with diagonal covariance matrices

Sampling the right space



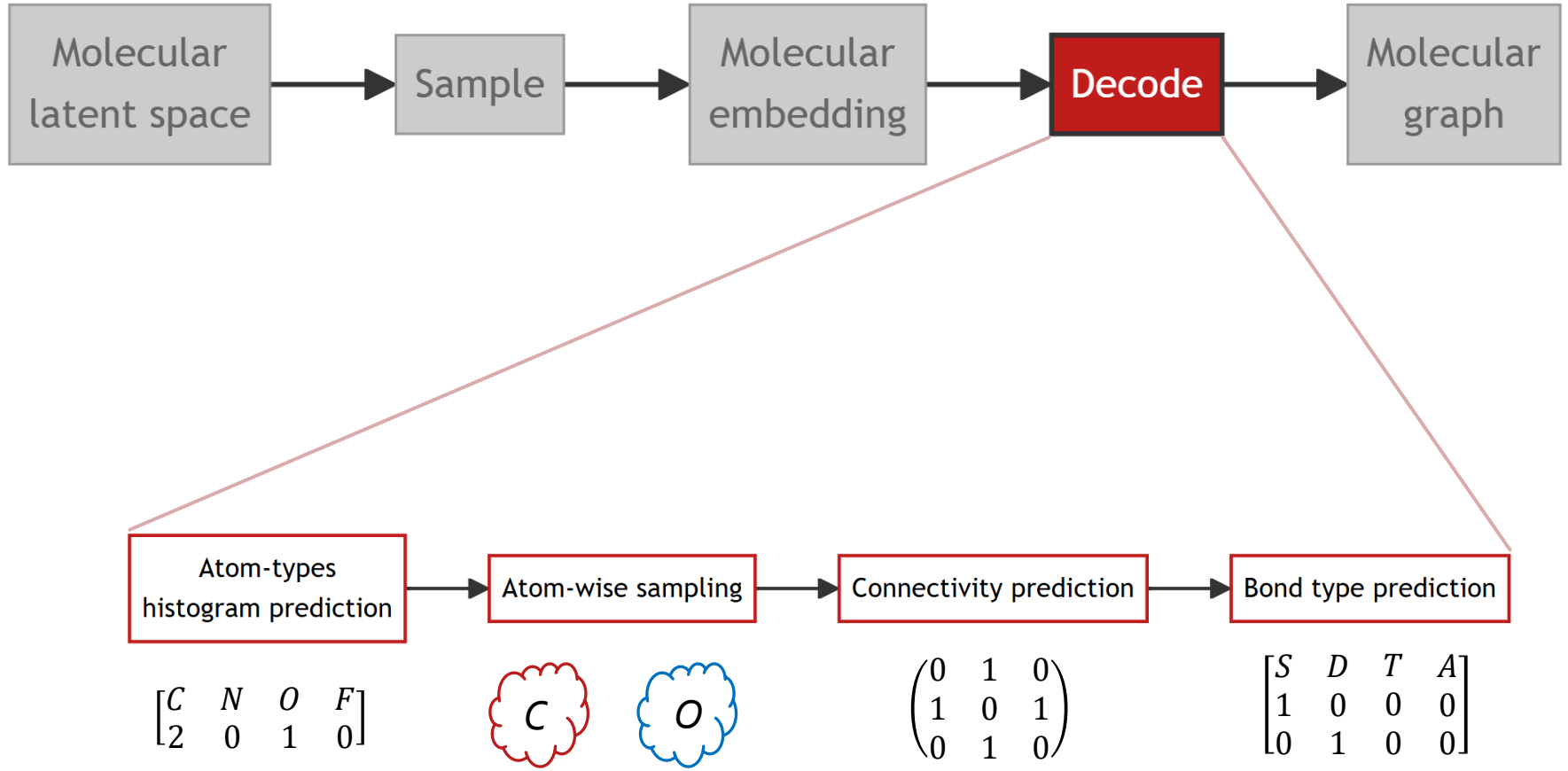
- VAE: unit Gaussian centered in 0
- VAE-like: single Gaussian centered in μ with diagonal covariance matrix
- GMM-D: combination of multiple Gaussians with diagonal covariance matrices
- GMM-F: combination of multiple Gaussians with full covariance matrices

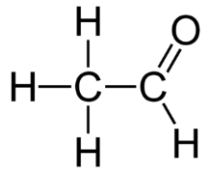
Sampling the right space



- VAE: unit Gaussian centered in 0
- VAE-like: single Gaussian centered in μ with diagonal covariance matrix
- GMM-D: combination of multiple Gaussians with diagonal covariance matrices
- GMM-F: combination of multiple Gaussians with full covariance matrices
- GMM-PW: combination of a Gaussian per data point, with diagonal covariance matrix

Generation process





Learning QM9 dataset

- ~130k small organic molecules
- 4 atom types (C, O, N, F)
- 58 atomic features
- 13 bond features
- 19 annotated properties

VUN assessment

By leveraging GMM priors, we are able to perform competitively or better than state-of-the-art latent variable models:

Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
MPG-VAE	-	0.9100	0.6800	0.540	0.3340
GraphNVP	-	0.8310	0.9920	0.582	0.4797
GRF	-	0.8450	0.6600	0.586	0.3268
GraphAF	1.000	0.6700	0.9451	0.8883	0.8395
MoFlow	1.000	0.8896	0.9853	0.9604	0.9462
GraphDF	1.000	0.8267	0.9762	0.9810	0.9576
Ours - VAE	1.000	0.4006	0.1293	0.8987	0.1162
Ours - VAE-like	1.000	0.5803	0.7756	0.8829	0.6848
Ours - GMM-F	1.000	0.4075	0.9428	0.8001	0.7543
Ours - GMM-D1	1.000	0.1653	0.9693	0.9640	0.9344
Ours - GMM-D2	1.000	0.0555	0.9982	0.9964	0.9946

VUN assessment

By leveraging GMM priors, we are able to perform competitively or better than state-of-the-art latent variable models:

Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
MPG-VAE	-	0.9100	0.6800	0.540	0.3340
GraphNVP	-	0.8310	0.9920	0.582	0.4797
GRF	-	0.8450	0.6600	0.586	0.3268
GraphAF	1.000	0.6700	0.9451	0.8883	0.8395
MoFlow	1.000	0.8896	0.9853	0.9604	0.9462
GraphDF	1.000	0.8267	0.9762	0.9810	0.9576
Ours - VAE	1.000	0.4006	0.1293	0.8987	0.1162
Ours - VAE-like	1.000	0.5803	0.7756	0.8829	0.6848
Ours - GMM-F	1.000	0.4075	0.9428	0.8001	0.7543
Ours - GMM-D1	1.000	0.1653	0.9693	0.9640	0.9344
Ours - GMM-D2	1.000	0.0555	0.9982	0.9964	0.9946

GMM priors help exploring the latent space

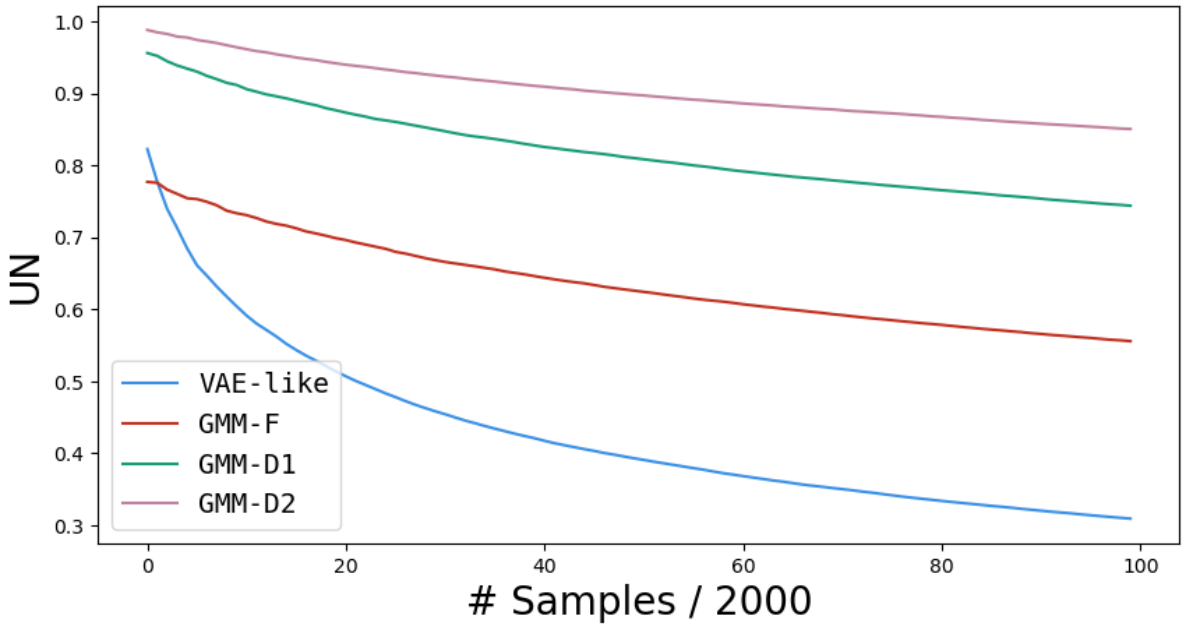
VUN assessment

By leveraging GMM priors, we are able to perform competitively or better than state-of-the-art latent variable models:

Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
MPG-VAE	-	0.9100	0.6800	0.540	0.3340
GraphNVP	-	0.8310	0.9920	0.582	0.4797
GRF	-	0.8450	0.6600	0.586	0.3268
GraphAF	1.000	0.6700	0.9451	0.8883	0.8395
MoFlow	1.000	0.8896	0.9853	0.9604	0.9462
GraphDF	1.000	0.8267	0.9762	0.9810	0.9576
Ours - VAE	1.000	0.4006	0.1293	0.8987	0.1162
Ours - VAE-like	1.000	0.5803	0.7756	0.8829	0.6848
Ours - GMM-F	1.000	0.4075	0.9428	0.8001	0.7543
Ours - GMM-D1	1.000	0.1653	0.9693	0.9640	0.9344
Ours - GMM-D2	1.000	0.0555	0.9982	0.9964	0.9946

Low validity rate → Fast resampling

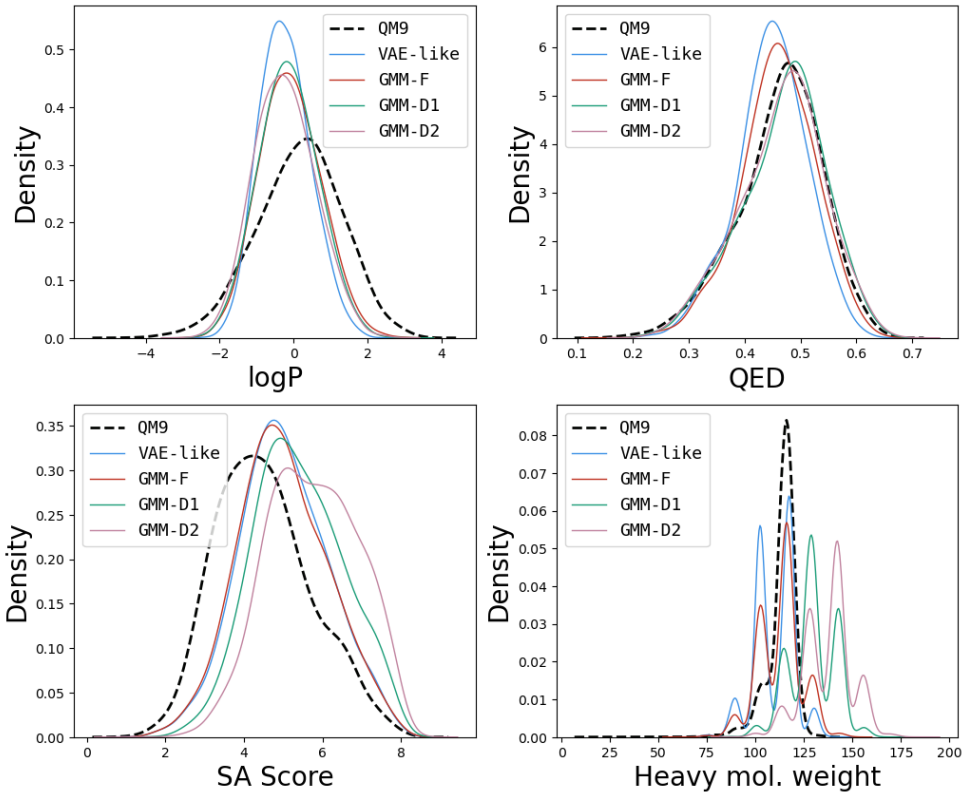
Can resampling cause trouble?



With more explorative priors, the model is able to keep a steady ratio of unique novel molecules

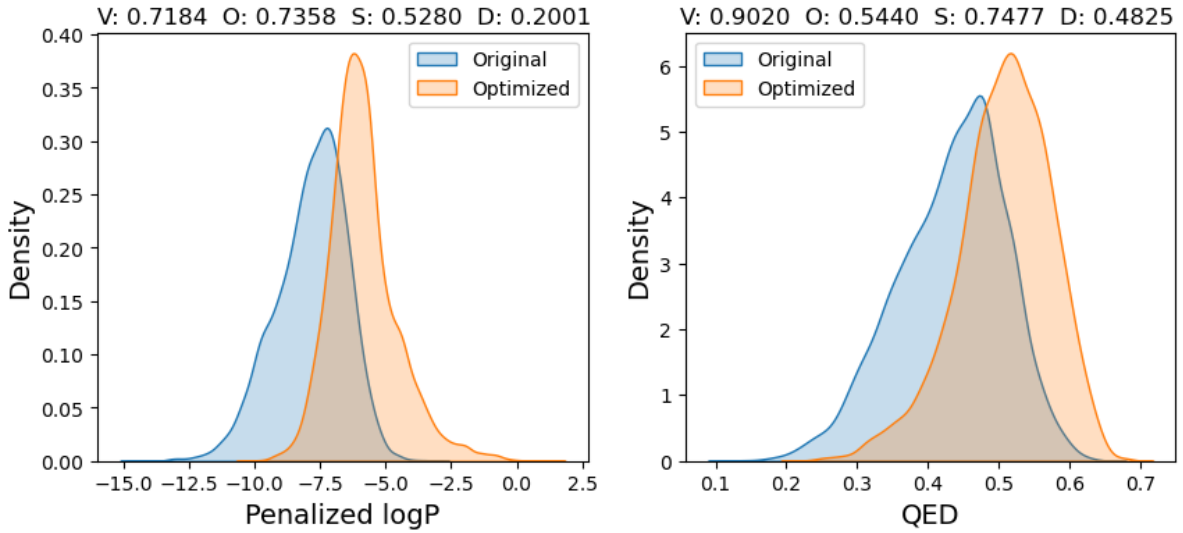
Can resampling cause trouble?

The generated molecules follow the original molecular property distributions



Property optimization

We can see a shift in the molecular property distributions



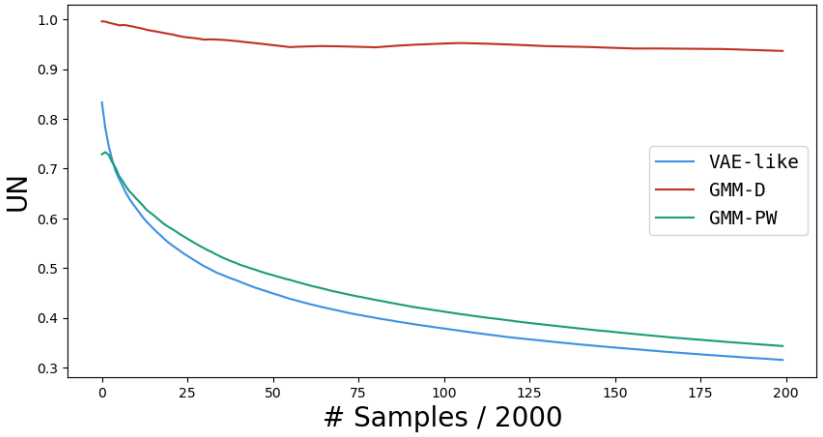
Learning ZINC dataset

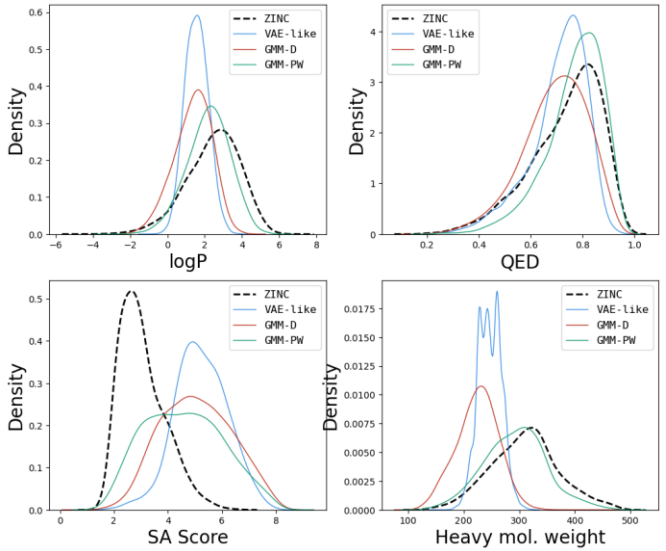
- ~250k molecules
- 9 atom types (C, O, N, F, P, S, Cl, Br, I)
- 58 atomic features
- 13 bond features
- 3 annotated properties

Learning ZINC dataset

- AMCG shows promising results, with a steady UN ratio and a competing VUN for the GMM-D prior
- However, the lower validity rates and the smaller molecular weight show room for improvement

Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
GraphNVP	-	0.426	0.948	1.000	0.4038
GRF	-	0.734	0.537	1.000	0.3942
GraphAF	1.000	0.68	0.991	1.000	0.9910
MoFlow	1.000	0.5030	0.9999	1.000	0.9999
GraphDF	1.000	0.8903	0.9916	1.000	0.9916
Ours - VAE	1.000	0.2323	0.0437	0.8902	0.0389
Ours - VAE-like	1.000	0.0262	0.7054	1.000	0.7054
Ours - GMM-D	1.000	0.0144	0.9900	1.000	0.9900
Ours - GMM-PW	1.000	0.2630	0.9190	0.7636	0.7017



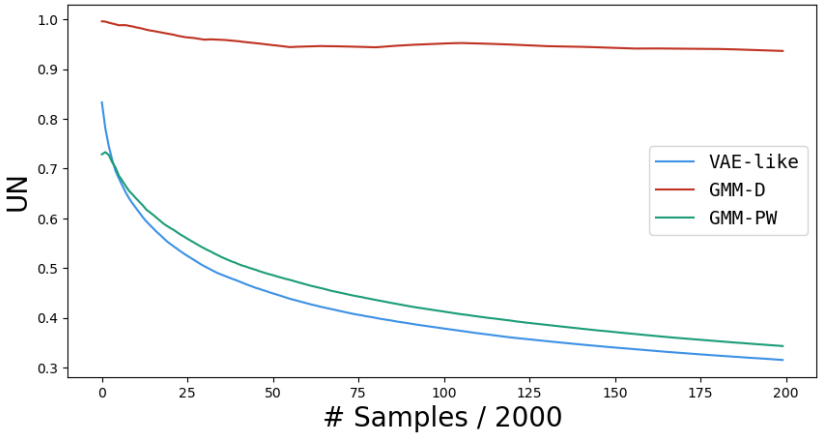


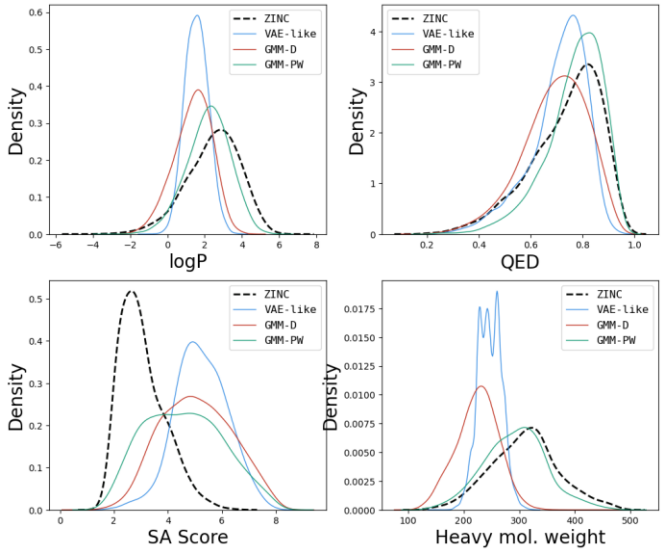
Learning ZINC dataset

Possible solutions

- More expressive encoders, requiring less information and smaller latent spaces → graph pooling techniques

Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
GraphNVP	-	0.426	0.948	1.000	0.4038
GRF	-	0.734	0.537	1.000	0.3942
GraphAF	1.000	0.68	0.991	1.000	0.9910
MoFlow	1.000	0.5030	0.9999	1.000	0.9999
GraphDF	1.000	0.8903	0.9916	1.000	0.9916
Ours - VAE	1.000	0.2323	0.0437	0.8902	0.0389
Ours - VAE-like	1.000	0.0262	0.7054	1.000	0.7054
Ours - GMM-D	1.000	0.0144	0.9900	1.000	0.9900
Ours - GMM-PW	1.000	0.2630	0.9190	0.7636	0.7017



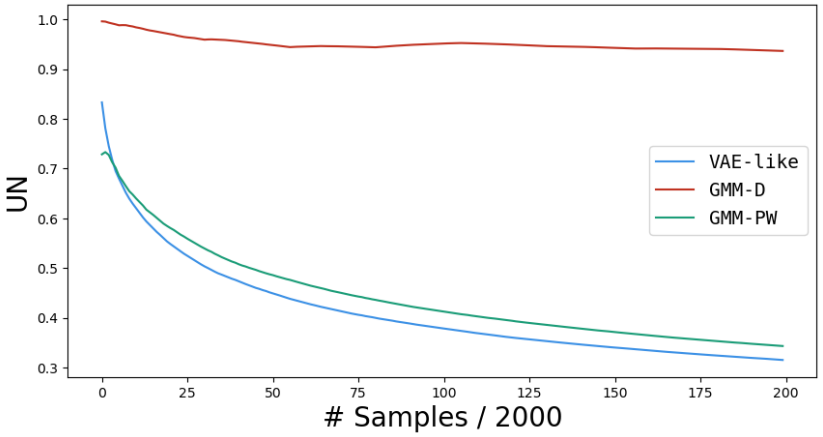


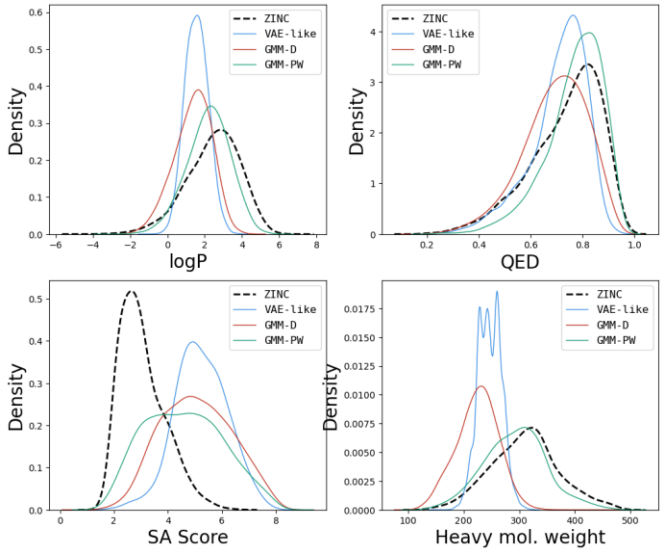
Learning ZINC dataset

Possible solutions

- More expressive encoders, requiring less information and smaller latent spaces → graph pooling techniques
- Different decoding strategies and training objectives

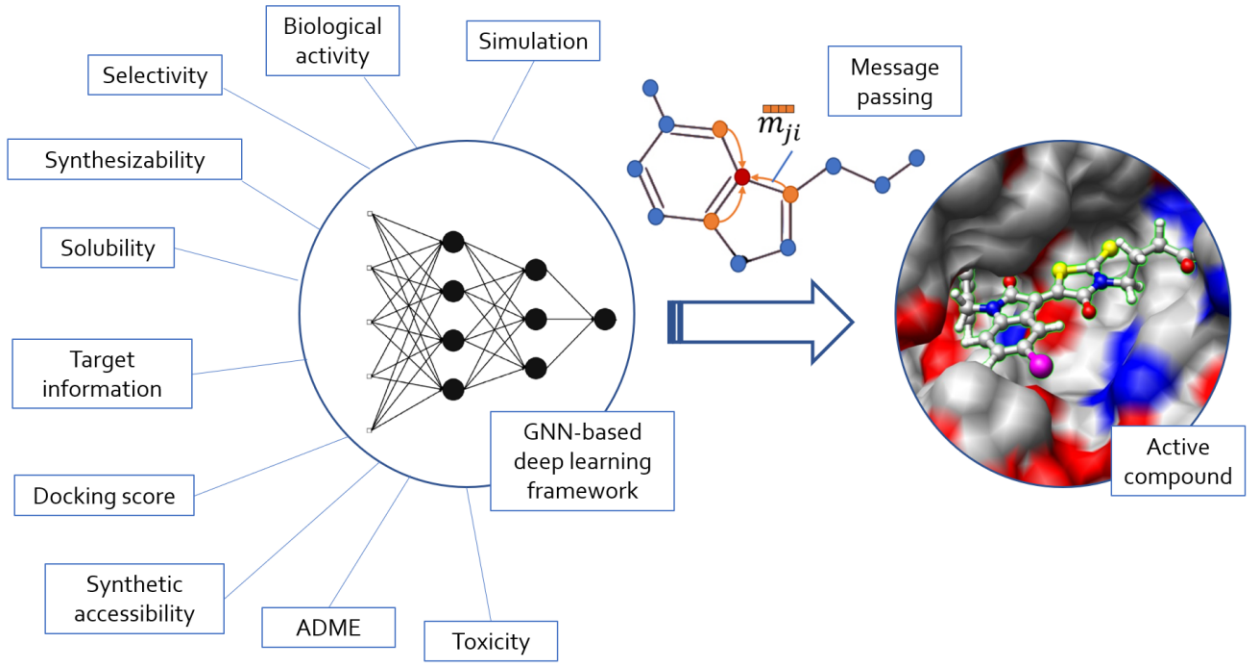
Model	Validity	Validity w/o check	Uniqueness	Novelty	VUN
GraphNVP	-	0.426	0.948	1.000	0.4038
GRF	-	0.734	0.537	1.000	0.3942
GraphAF	1.000	0.68	0.991	1.000	0.9910
MoFlow	1.000	0.5030	0.9999	1.000	0.9999
GraphDF	1.000	0.8903	0.9916	1.000	0.9916
Ours - VAE	1.000	0.2323	0.0437	0.8902	0.0389
Ours - VAE-like	1.000	0.0262	0.7054	1.000	0.7054
Ours - GMM-D	1.000	0.0144	0.9900	1.000	0.9900
Ours - GMM-PW	1.000	0.2630	0.9190	0.7636	0.7017





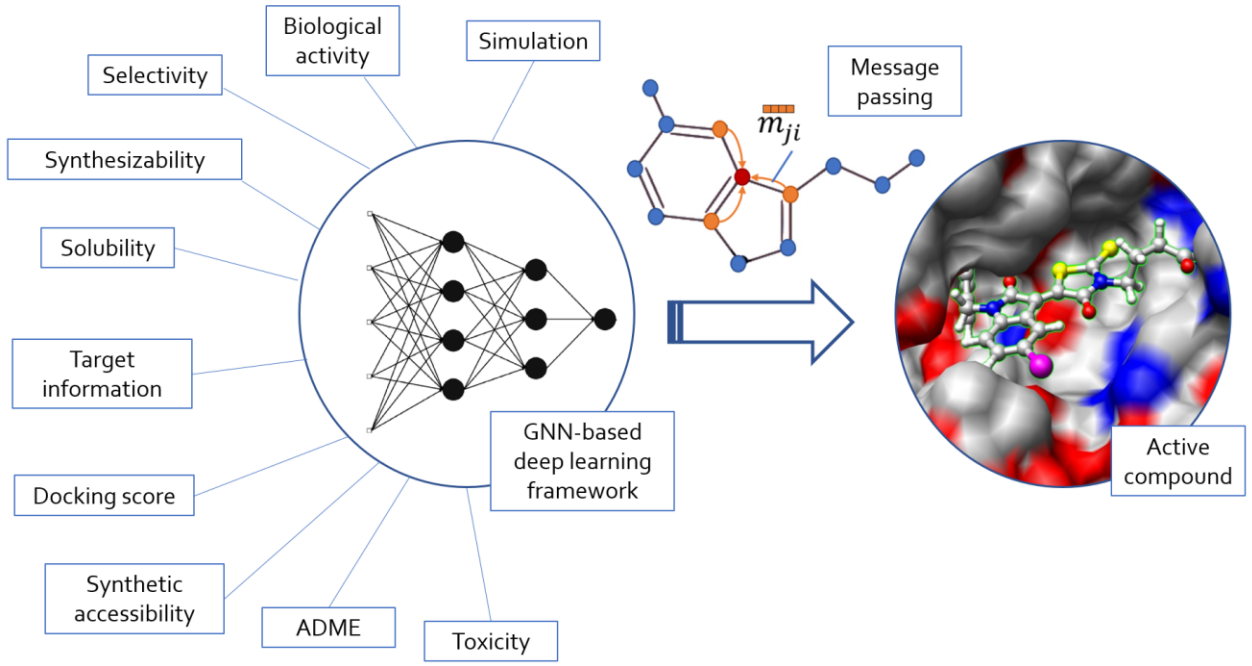
Summary

- Several approaches for molecular graph generation are available



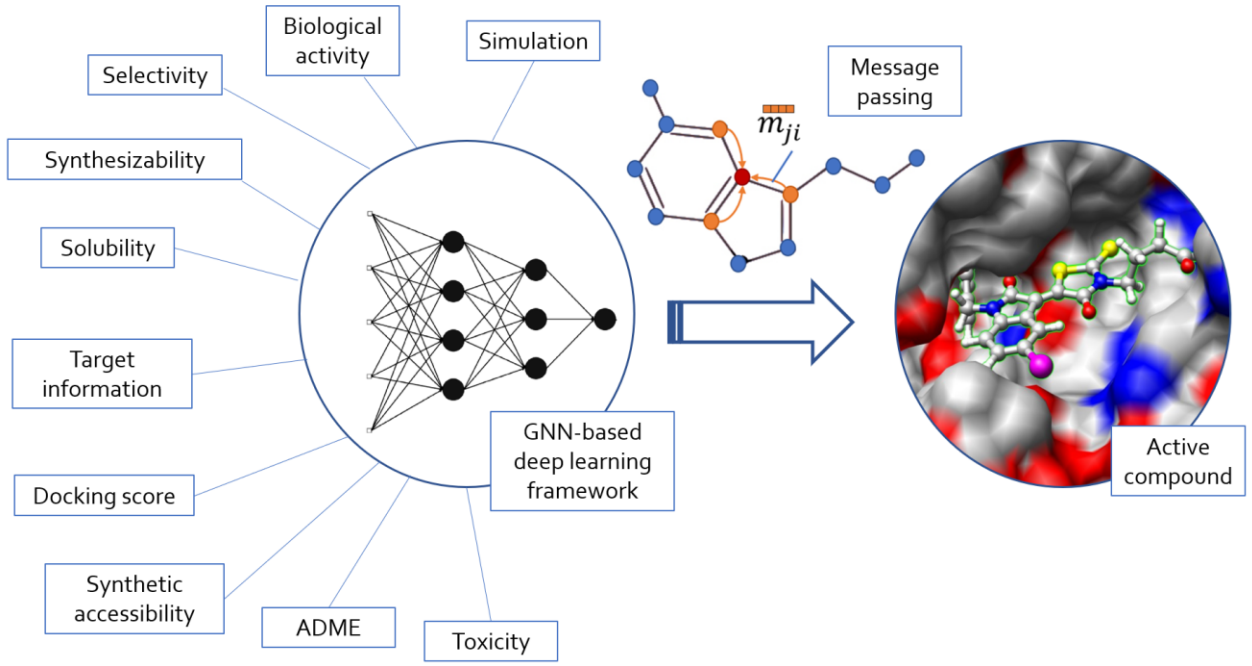
Summary

- Several approaches for molecular graph generation are available
- No free lunch theorem applies, but careful design choices enable the selection of the right trade-off for specific molecular design challenges

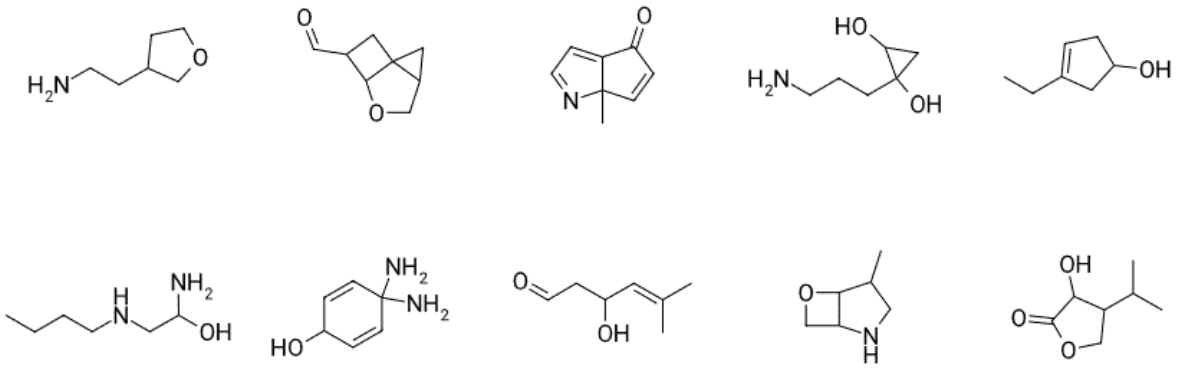


Summary

- Several approaches for molecular graph generation are available
- No free lunch theorem applies, but careful design choices enable the selection of the right trade-off for specific molecular design challenges
- As an example, we introduced AMCG model, deliberately trading-off *validity* for *speed* and *fantasy*



Abate, C., Decherchi, S., & Cavalli, A. (2024). AMCG: a graph dual atomic-molecular conditional molecular generator. *Machine Learning: Science and Technology*, **5**(3), 035004.



Thank you