RetroBridge

Modeling Retrosynthesis with Markov Bridges

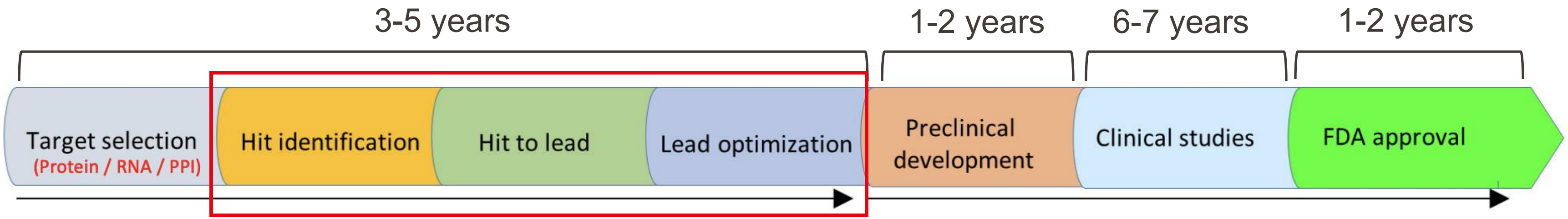
Arne Schneuing, EPFL

Swiss Joint Research Centre, Spring Workshop
April 10, 2024

@rneschneuing

arne.schneuing@epfl.ch

Computational Drug Discovery



3-5 years

1-2 years

6-7 years

1-2 years

Target selection
(Protein / RNA / PPI)

Hit identification

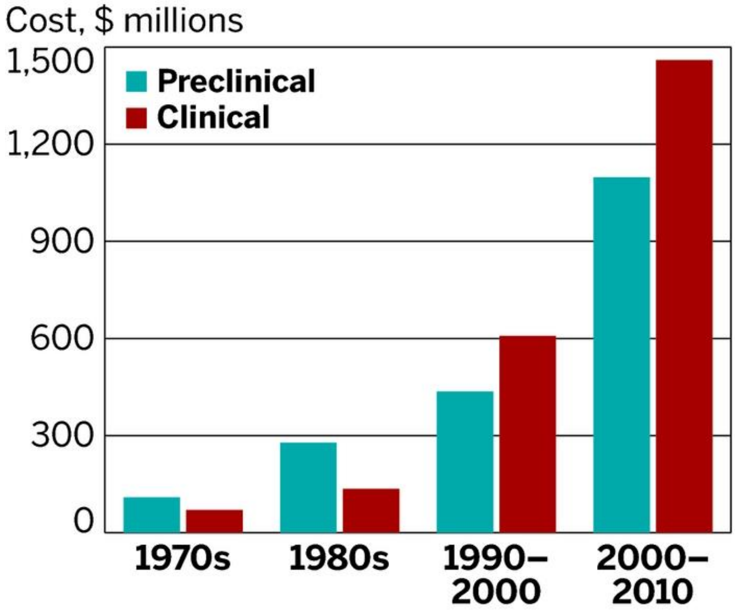
Hit to lead

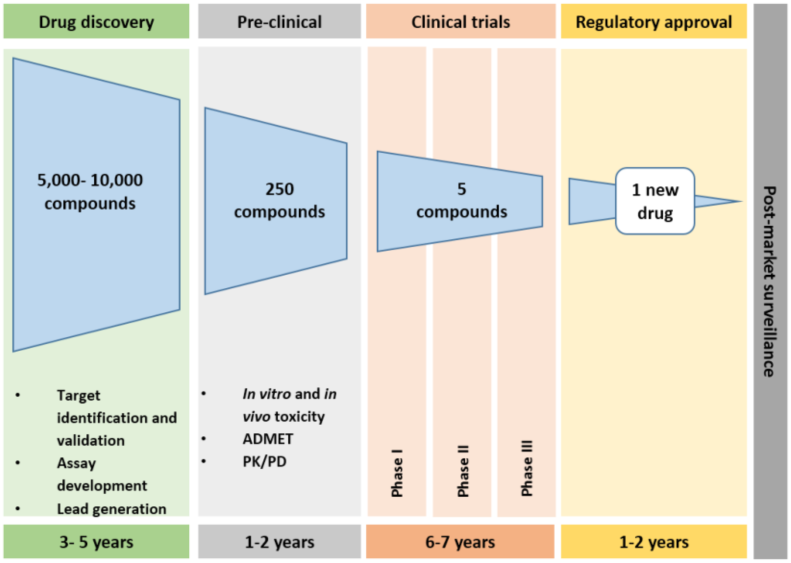
Lead optimization

Preclinical development

Clinical studies

FDA approval





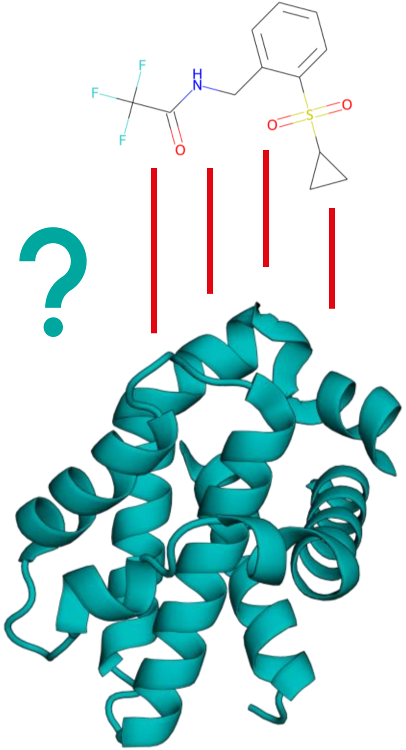
Drug discovery	Pre-clinical	Clinical trials	Regulatory approval
5,000- 10,000 compounds	250 compounds	5 compounds	1 new drug
- Target identification and validation - Assay development - Lead generation	- *In vitro* and *in vivo* toxicity - ADMET - PK/PD	Phase I Phase II Phase III	
3- 5 years	1-2 years	6-7 years	1-2 years

Post-market surveillance

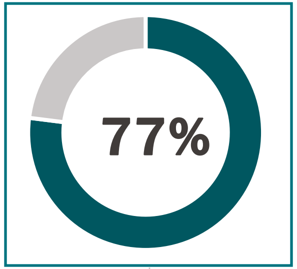
https://cen.acs.org/articles/92/web/2014/11/Tufts-Study-Finds-Big-Rise.html (Accessed 2023/10/23)

Matthews, Holly, James Hanison, and Niroshini Nirmalan. ""Omics"-informed drug and biomarker discovery: opportunities, challenges and future perspectives." *Proteomes* 4.3 (2016): 28.

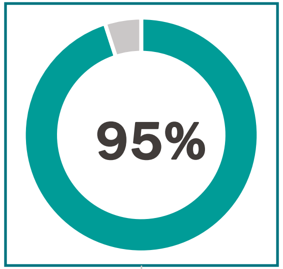
Targeting Proteins with Small Molecule Drugs



>10000 approved drugs



Drugs are small molecules

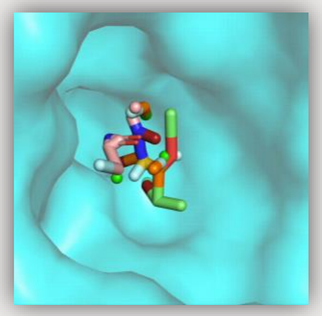


Drug targets are proteins

Ghadermarzi, Sina, et al. "Sequence-derived markers of drug targets and potentially druggable human proteins." Frontiers in Genetics 10 (2019): 475494.

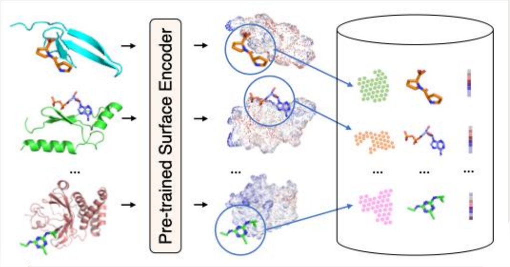
Structure-based Small Molecule Design Tools

DiffSBDD

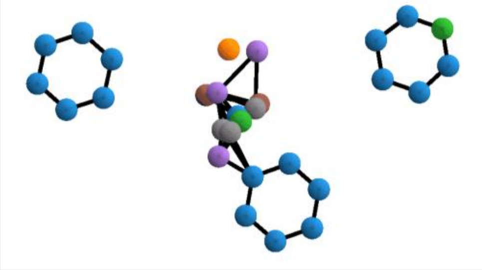


- ✨ 3D-conditional diffusion model
- ✨ Generates atomic point clouds given protein structures
- 📖 *Structure-based drug design with equivariant diffusion models* (preprint)

FragmentScope

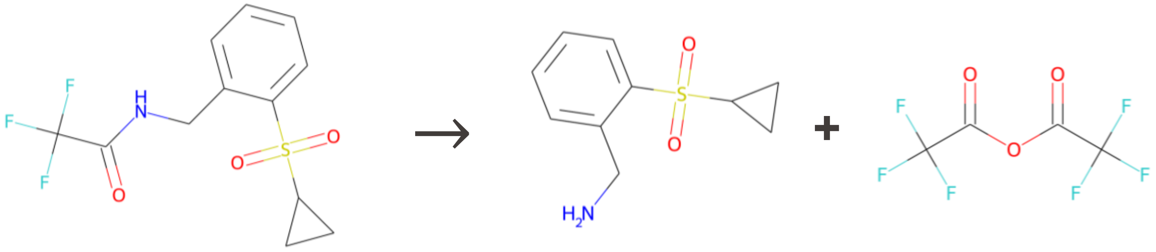


DiffLinker



- ✨ Select and place molecule fragments based on protein surface patterns
- ✨ Generate linker atoms with a diffusion model conditioned on the fragments
- 📖 *Equivariant 3d-conditional diffusion models for molecular linker design* (accepted at Nature Machine Intelligence)

Retrosynthesis Modeling



- Recursively decompose target molecule into simpler building blocks until available starting molecules are reached
- Single-step retrosynthesis prediction: predict possible reactants given a target molecule
- Multi-step retrosynthesis prediction: plan optimal reaction sequence that minimizes number of synthesis steps, cost of starting molecules etc.

Retrosynthesis as a Distribution Learning Task

- Each product molecule can be obtained starting from several valid sets of reactants

  → we want to sample from $p(\text{reactants}|\text{product})$
  → confidence estimates should reflect this probability

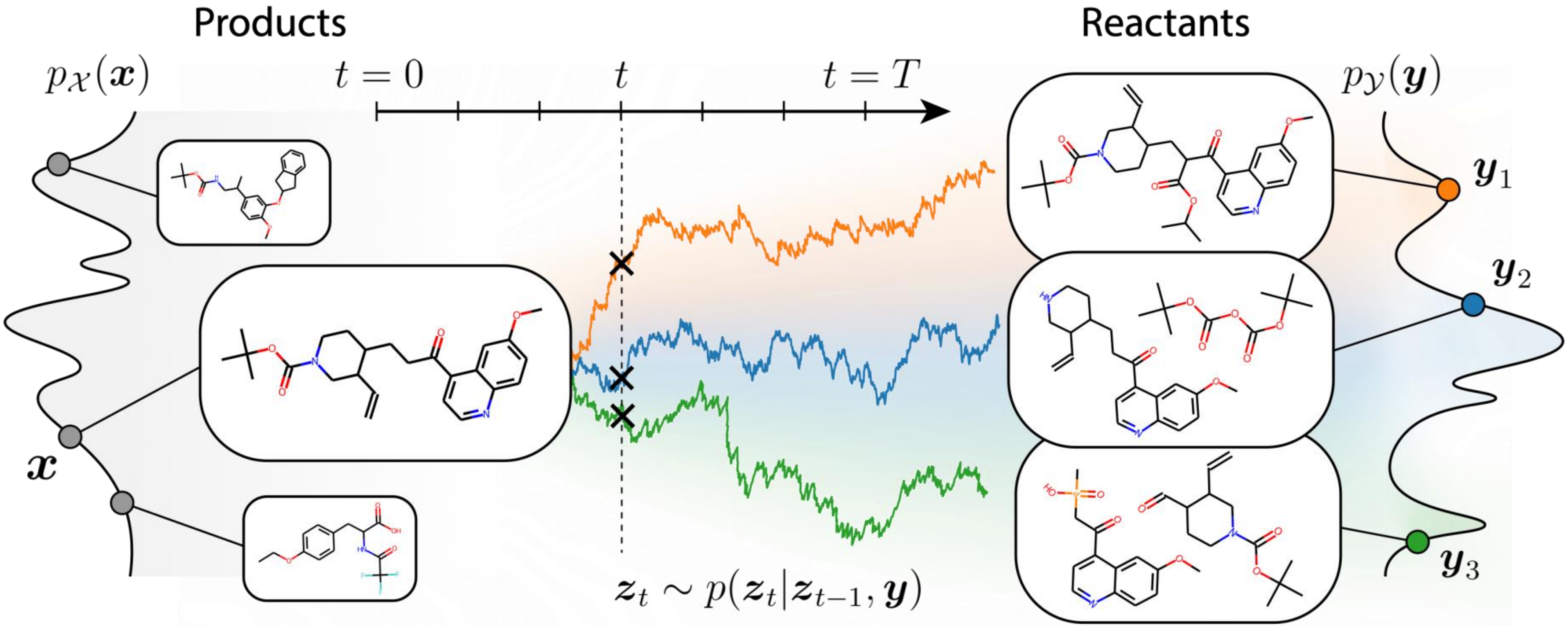
- Most existing retrosynthesis models optimize…
  - …the likelihood of single actions greedily
    ⚠ calibration and global consistency issues
  - …the likelihood of a sequence (SMILES, graph edits)
    ⚠ different sequences can lead to the same outcome

$$p(S) = \prod_{i=0}^{n} p(S_{i+1}|S_0, \ldots, S_i)$$

- We want to be able to sample from the data distribution instead

Our goal

⚠ **Discrete data**



⚠ **Intractable prior**

⚠ **Finite set of coupled samples**

⚠ **Looking for p(y|x)**

Summary

Framework	Intractable prior	Coupled samples	Likelihood optimisation	Discrete data
Requirement	✅	✅	✅	✅
Discr. Diffusion	🚫	🚫	✅	✅
CFM	✅	✅	🚫	🚫
Aligned SB	✅	✅	🚫	🚫
Diffusion SB	✅	🚫	🚫	🚫

Markov Bridge Model

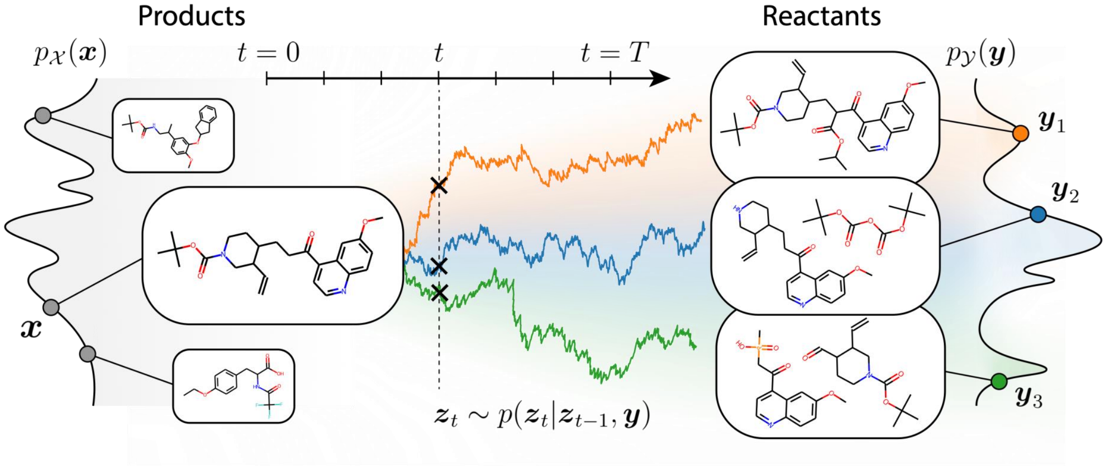
Coupled data points:

$$\boldsymbol{x}, \boldsymbol{y} \sim p_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y})$$

Intractable distributions:

$$p_{\mathcal{X}}(\boldsymbol{x}) = \int p_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y}$$

$$p_{\mathcal{Y}}(\boldsymbol{y}) = \int p_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x}$$

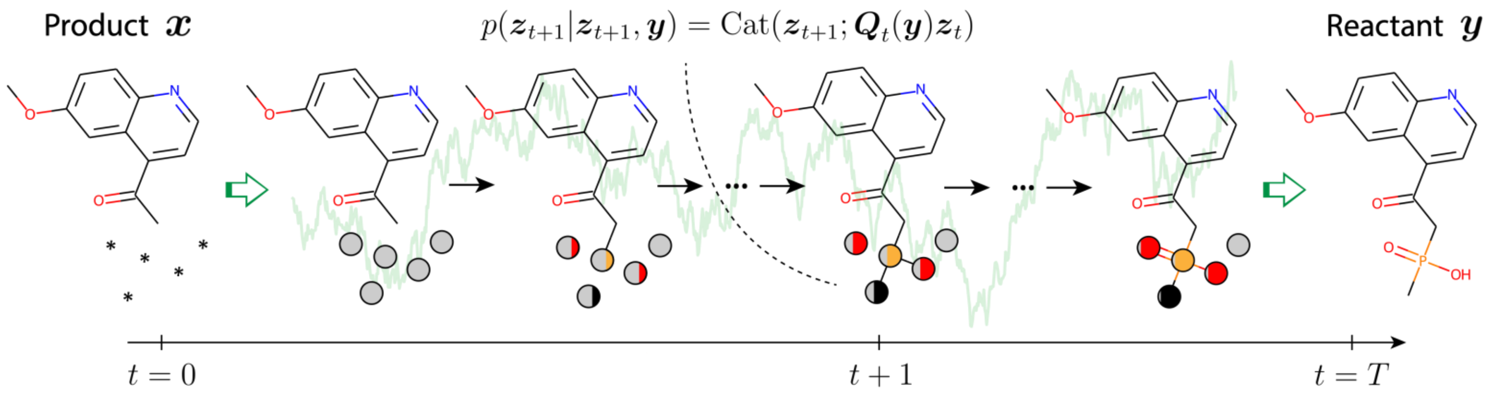


Markov bridge:

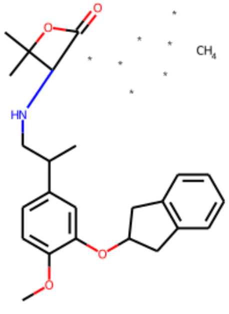
- A sequence of time steps and corresponding r.v.: $t = 0, \ldots, T, \quad (\boldsymbol{z}_t)_{t=0}^{T} \sim p$
- Trajectory starts at product: $\boldsymbol{z}_0 = \boldsymbol{x}$
- Markov property: $p(\boldsymbol{z}_t | \boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}, \boldsymbol{y}) = p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}, \boldsymbol{y})$
- Process is pinned in the end: $p(\boldsymbol{z}_T = \boldsymbol{y} | \boldsymbol{z}_{T-1}, \boldsymbol{y}) = 1$

Fitzsimmons, Pat, Jim Pitman, and Marc Yor. "Markovian bridges: construction, Palm interpretation, and splicing." Seminar on Stochastic Processes, 1992. Boston, MA: Birkhäuser Boston, 1992.

Markov Bridge Model



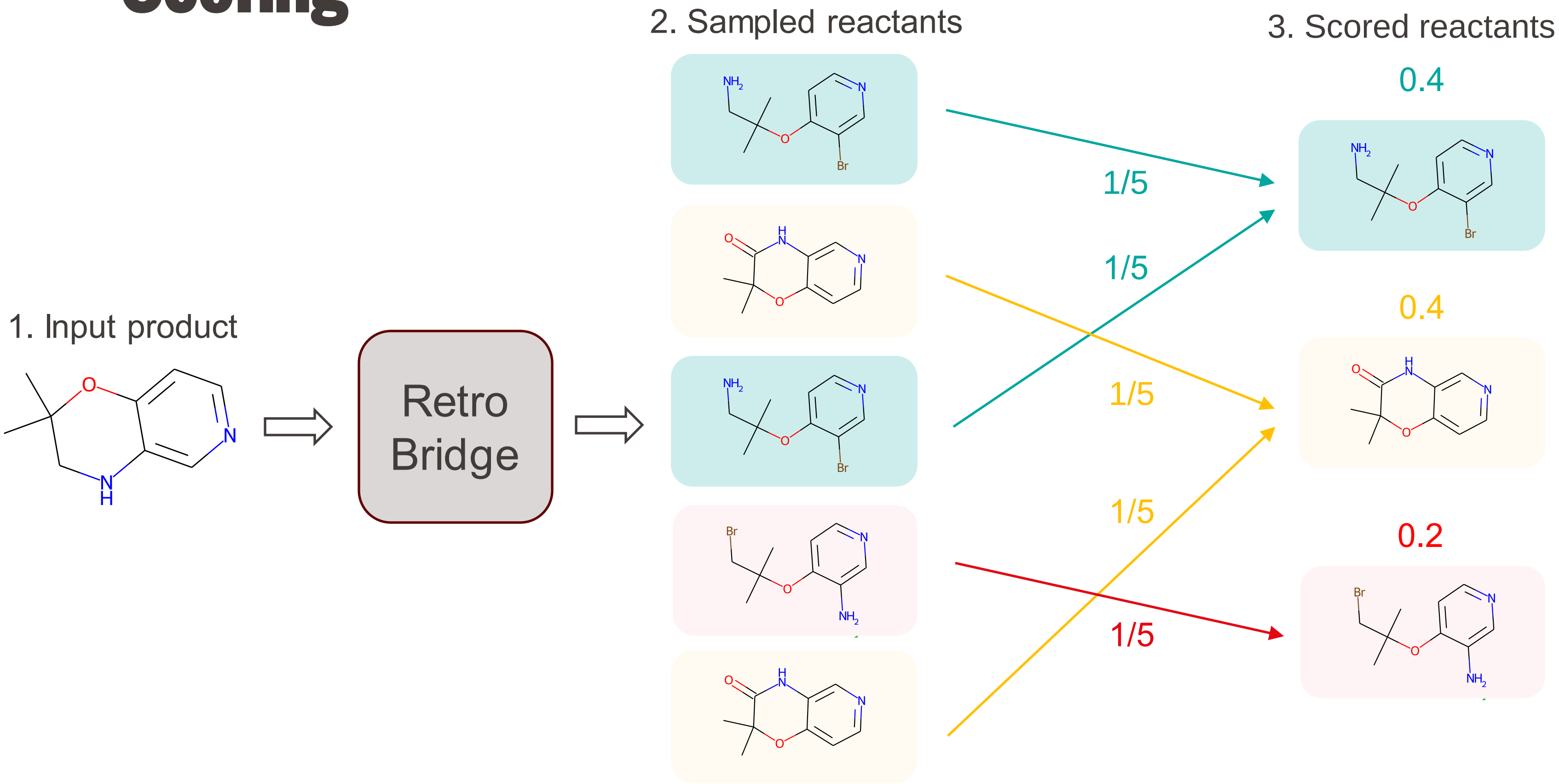
- Transition probabilities: $p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{y}) = \mathrm{Cat}\left(\boldsymbol{z}_{t+1}; \boldsymbol{Q}_t\boldsymbol{z}_t\right)$
- Transition matrices: $\boldsymbol{Q}_t \equiv \boldsymbol{Q}_t\boxed{(\boldsymbol{y})} = \alpha_t \boldsymbol{I}_K + (1 - \alpha_t\boxed{)\boldsymbol{y}\mathbf{1}}_K^\top$
- Neural network approximation: $\hat{\boldsymbol{y}} = \varphi_\theta(\boldsymbol{z}_t, t)$
- New transition kernel: $q_\theta(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) = \mathrm{Cat}\left(\boldsymbol{z}_{t+1}; \boldsymbol{Q}_t(\hat{\boldsymbol{y}})\boldsymbol{z}_t\right)$
- Train by maximizing log-likelihood: $\log q_\theta(\boldsymbol{y}|\boldsymbol{x})$



Frame 0

$$\log q_\theta(\boldsymbol{y}|\boldsymbol{x}) \geq -T \cdot \mathbb{E}_{t\sim\mathcal{U}(0,\dots,T-1)}\mathbb{E}_{\boldsymbol{z}_t\sim p(\boldsymbol{z}_t|\boldsymbol{x},\boldsymbol{y})} D_{\mathrm{KL}}(p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{y})\|q_\theta(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t))$$

Scoring



1. Input product

Retro Bridge

2. Sampled reactants

3. Scored reactants

0.4

1/5

1/5

0.4

1/5

1/5

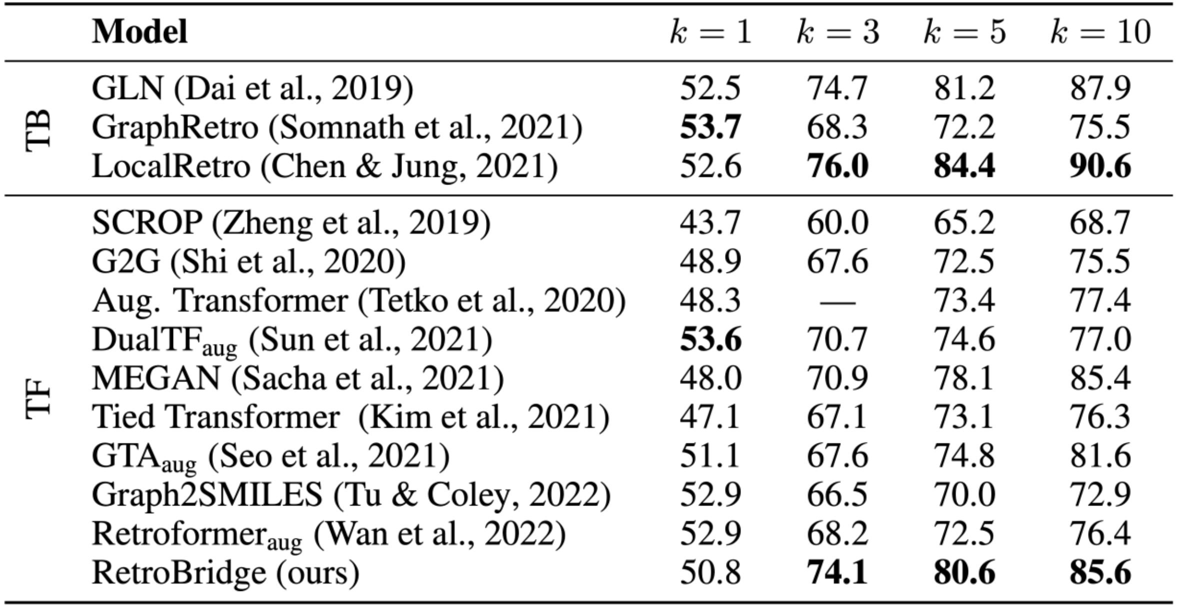
0.2

1/5

$$q_\theta(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}'\sim q_\theta(\cdot|\boldsymbol{x})}\mathbb{I}\{\boldsymbol{y}' = \boldsymbol{y}\}$$

Results

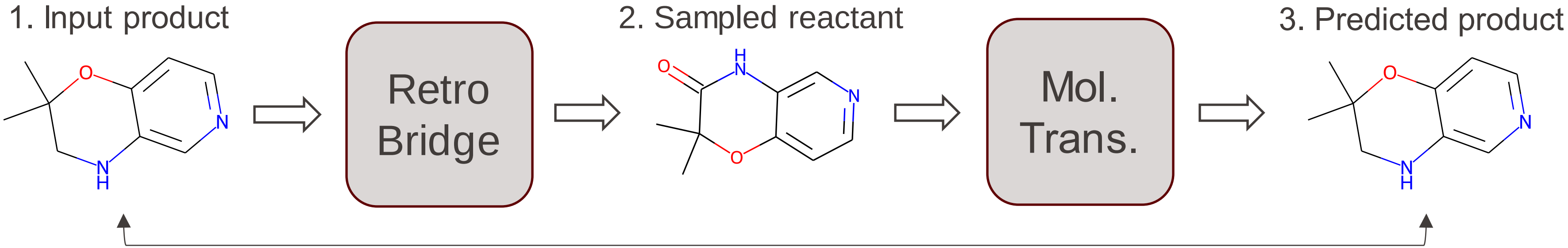
Top-k exact match: ground-truth set of reactants found among top-k samples



	Model	$k = 1$	$k = 3$	$k = 5$	$k = 10$
TB	GLN (Dai et al., 2019)	52.5	74.7	81.2	87.9
	GraphRetro (Somnath et al., 2021)	**53.7**	68.3	72.2	75.5
	LocalRetro (Chen & Jung, 2021)	52.6	**76.0**	**84.4**	**90.6**
TF	SCROP (Zheng et al., 2019)	43.7	60.0	65.2	68.7
	G2G (Shi et al., 2020)	48.9	67.6	72.5	75.5
	Aug. Transformer (Tetko et al., 2020)	48.3	—	73.4	77.4
	DualTF$_{\text{aug}}$ (Sun et al., 2021)	**53.6**	70.7	74.6	77.0
	MEGAN (Sacha et al., 2021)	48.0	70.9	78.1	85.4
	Tied Transformer (Kim et al., 2021)	47.1	67.1	73.1	76.3
	GTA$_{\text{aug}}$ (Seo et al., 2021)	51.1	67.6	74.8	81.6
	Graph2SMILES (Tu & Coley, 2022)	52.9	66.5	70.0	72.9
	Retroformer$_{\text{aug}}$ (Wan et al., 2022)	52.9	68.2	72.5	76.4
	RetroBridge (ours)	50.8	**74.1**	**80.6**	**85.6**

Limitation: does not account for other possible valid reactants

Results



1. Input product

Retro Bridge

2. Sampled reactant

Mol. Trans.

3. Predicted product

	Model	Coverage			Accuracy		
		$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
TB	GLN (Dai et al., 2019)	**82.5**	92.0	94.0	**82.5**	**71.0**	66.2
	LocalRetro (Chen & Jung, 2021)	82.1	**92.3**	**94.7**	82.1	**71.0**	**66.7**
TF	MEGAN (Sacha et al., 2021)	78.1	88.6	91.3	78.1	67.3	61.7
	Graph2SMILES (Tu & Coley, 2022)	—	—	—	76.7	56.0	46.4
	Retroformer$_{\text{aug}}$ (Wan et al., 2022)	—	—	—	78.6	71.8	67.1
	RetroBridge (ours)	**85.1**	**95.7**	**97.1**	**85.1**	**73.6**	**67.8**

Limitation: quality of the forward reaction prediction method

Schwaller, Philippe, et al. "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction." ACS central science 5.9 (2019): 1572-1583.

Project Summary

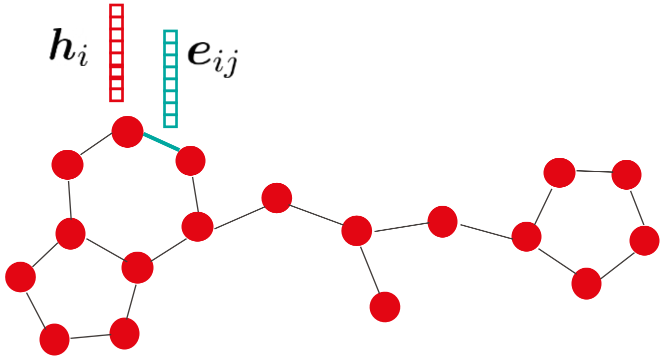
- 3D generative models for drug design
- Experimental validation requires chemical synthesis
- Graph generative model for synthesis planning
- Markov Bridge Model is an efficient framework to address the one-to-many mapping problem

Acknowledgements
Ilia Igashov
Marwin Segler
Max Welling
Michael Bronstein
Bruno Correia

Preprint

@igashov
@rneschneuing
ilia.igashov@epfl.ch
arne.schneuing@epfl.ch

Molecular graphs



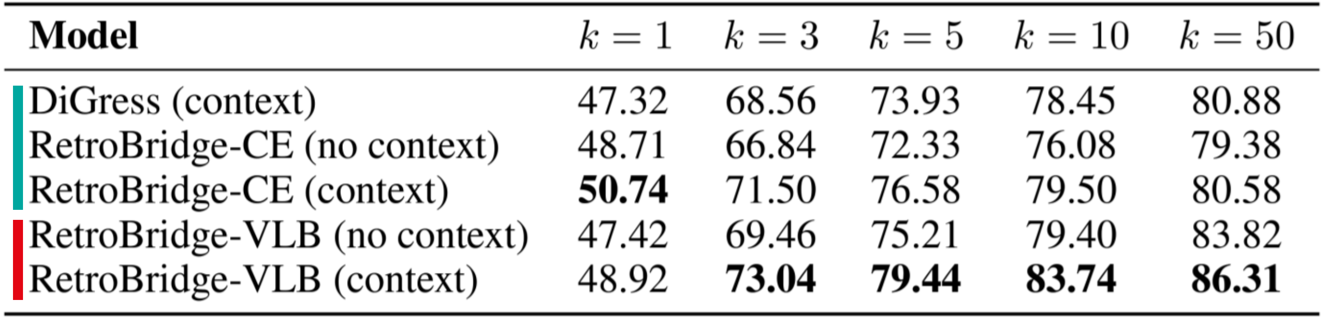
$$\begin{array}{rc} & \begin{matrix} \text{C} & \text{N} & \text{O} & \text{S} \end{matrix} \\ \boldsymbol{H} = & \begin{pmatrix} 1 & 0 & 0 & 0 \\ \multicolumn{4}{c}{\boxed{1 \quad 0 \quad 0 \quad 0} \; \boldsymbol{h}_i} \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Data representation:

- Graph of N nodes $\boldsymbol{x} = (\boldsymbol{H}, \boldsymbol{E})$
- Node features $\boldsymbol{H} \in \mathbb{R}^{N \times d_h}$
- Edge features $\boldsymbol{E} \in \mathbb{R}^{N \times N \times d_e}$
- All variables are categorical, represented as one-hot vectors

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_{11} & \boldsymbol{e}_{12} & \dots \\ \boldsymbol{e}_{21} & \boldsymbol{e}_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

Ablations



Model	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 50$
DiGress (context)	47.32	68.56	73.93	78.45	80.88
RetroBridge-CE (no context)	48.71	66.84	72.33	76.08	79.38
RetroBridge-CE (context)	**50.74**	71.50	76.58	79.50	80.58
RetroBridge-VLB (no context)	47.42	69.46	75.21	79.40	83.82
RetroBridge-VLB (context)	48.92	**73.04**	**79.44**	**83.74**	**86.31**

no context: $\hat{\boldsymbol{y}} = \varphi_\theta(\boldsymbol{z}_t, t)$

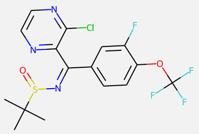
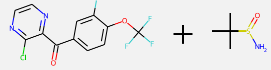
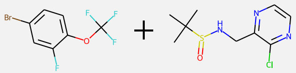
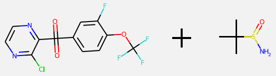
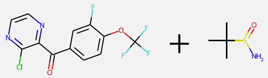
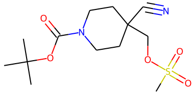
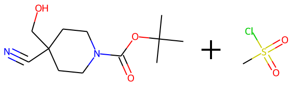
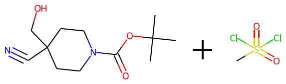
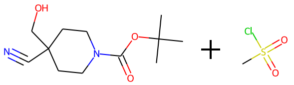
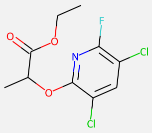
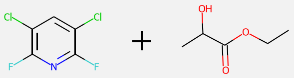
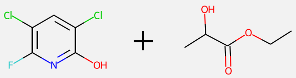
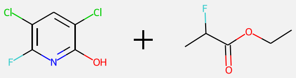
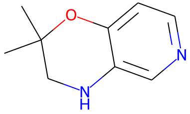
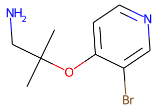
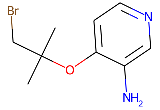
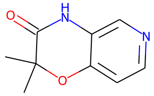
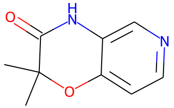
context: $\hat{\boldsymbol{y}} = \varphi_\theta(\boldsymbol{z}_t, \boldsymbol{x}, t)$

VLB: $\mathcal{L}_{\mathrm{VLB}}(\theta) = -T \cdot \mathbb{E}_{t\sim\mathcal{U}(0,\dots,T-1)} \mathbb{E}_{\boldsymbol{z}_t\sim p(\boldsymbol{z}_t|\boldsymbol{x},\boldsymbol{y})} D_{\mathrm{KL}}(p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{y}) \| q_\theta(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t))$

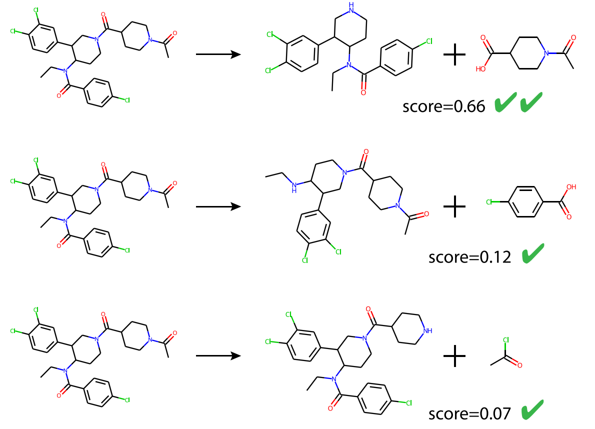
CE: $\mathcal{L}_{\mathrm{CE}}(\theta) = -T \cdot \mathbb{E}_{t\sim\mathcal{U}(0,\dots,T-1)} \mathbb{E}_{\boldsymbol{z}_t\sim p(\boldsymbol{z}_t|\boldsymbol{x},\boldsymbol{y})} \mathrm{CrossEntropy}(\boldsymbol{y}, \varphi_\theta(\boldsymbol{z}_t, t))$

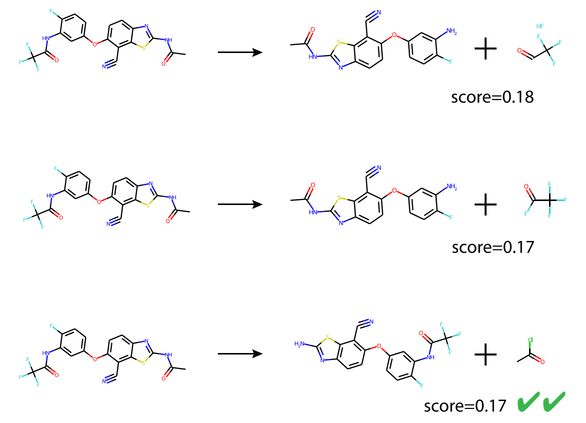
Vignac, Clement, et al. "Digress: Discrete denoising diffusion for graph generation." arXiv preprint arXiv:2209.14734 (2022).

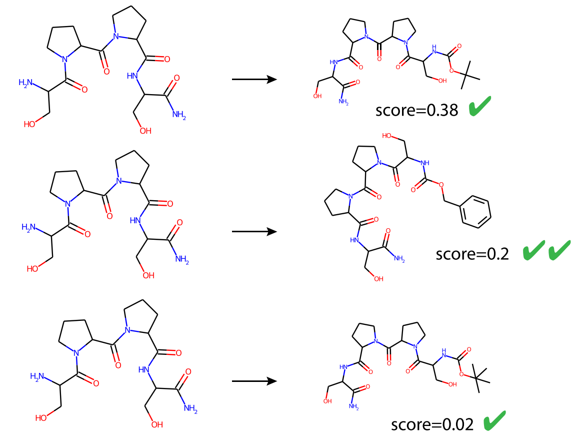
Examples

Input product	Sample 1	Sample 2	Sample 3	True reactants
	 score=0.97 ✓✓	 score=0.01	 score=0.01	
	 score=0.99 ✓✓	 score=0.01 ✓	—	
	 score=0.43 ✓	 score=0.4 ✓	 score=0.1	
	 score=0.49 ✓	 score=0.23 ✓	 score=0.07 ✓✓	

Examples


score=0.66 ✓✓
score=0.12 ✓
score=0.07 ✓


score=0.18
score=0.17
score=0.17 ✓✓

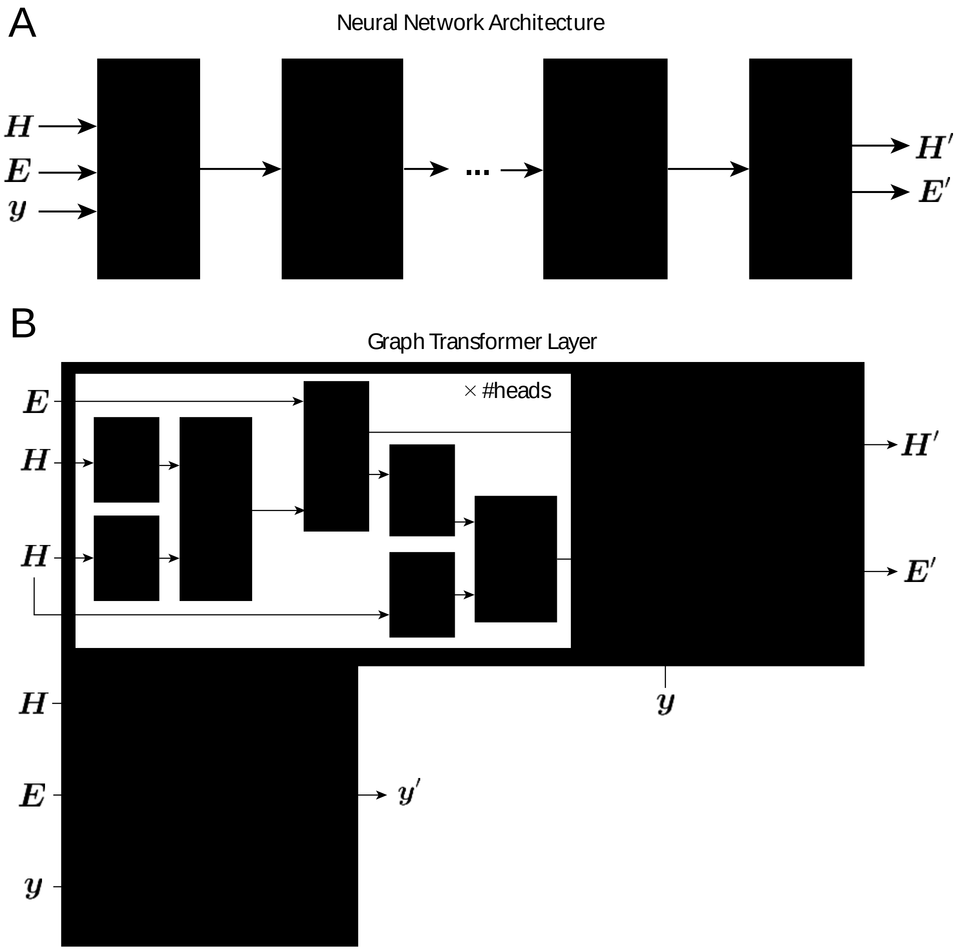

score=0.38 ✓
score=0.2 ✓✓
score=0.02 ✓

Neural network

- Fully-connected graph
- Graph transformer network

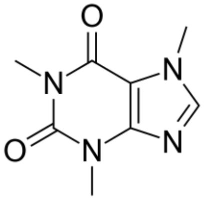
- Node (atom) features:
    - 16 atom types + 1 dummy
    - Additional spectral features
    - Number of cycles

- Edge (bond) features:
    - 3 bond types + “none” type



Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. 2020

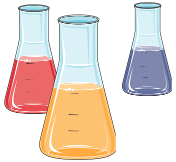
ML for Retrosynthesis Prediction

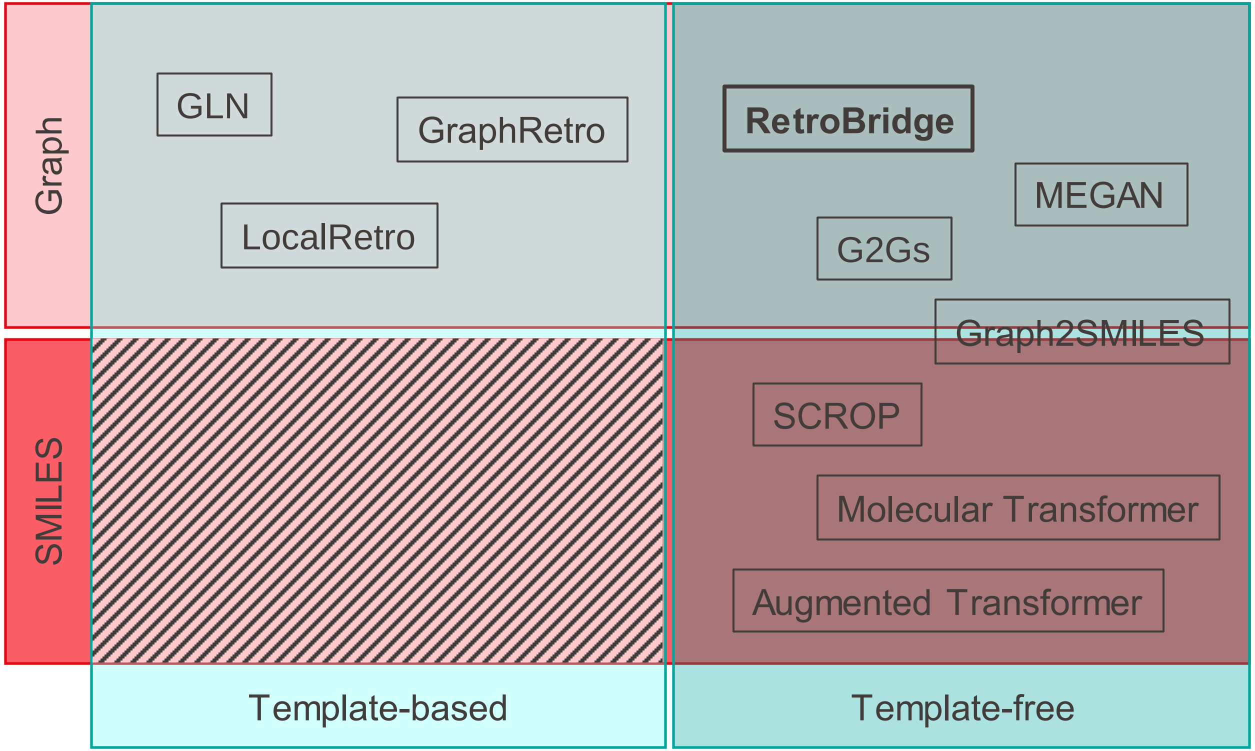


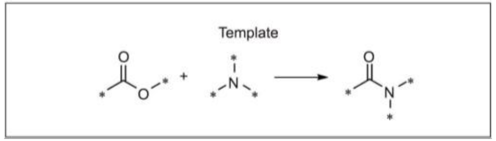
⚠ Typically requires atom mapping!

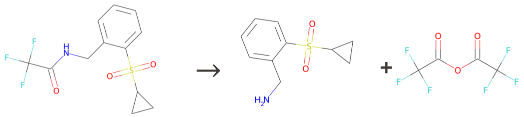
`CN1C=NC2=C1C(=O)N(C(=O)N2C)C`

⚠ Not permutation equivariant!





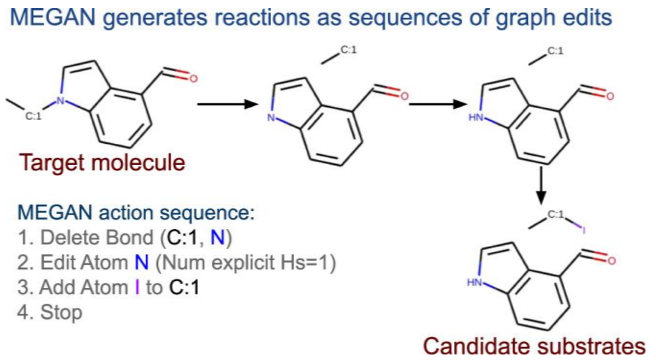




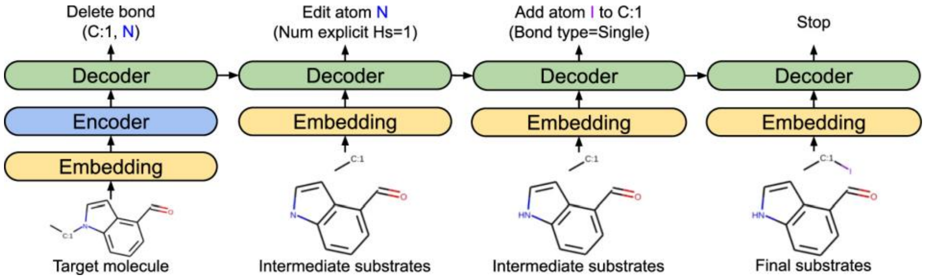
Jiang, Yinjie, et al. "Artificial intelligence for retrosynthesis prediction." *Engineering* (2022).

Molecule Edit Graph Attention Network (MEGAN)

- Predicts actions on atoms and bonds
- Optimizes likelihood of sequence of actions
- Imposes order on actions for training
- Teacher forcing


MEGAN generates reactions as sequences of graph edits

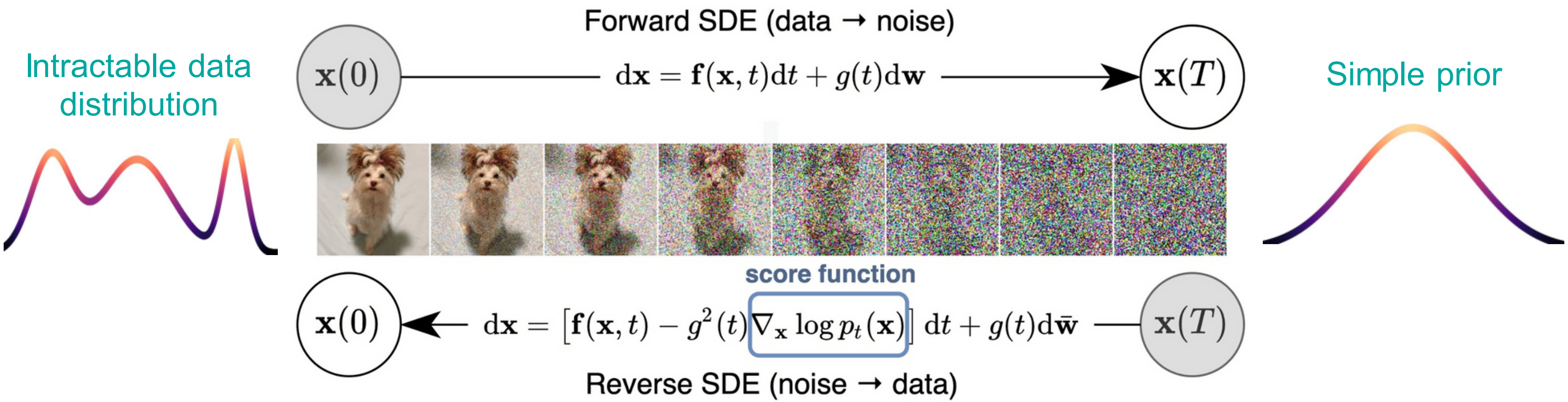
MEGAN action sequence:
1. Delete Bond (C:1, N)
2. Edit Atom N (Num explicit Hs=1)
3. Add Atom I to C:1
4. Stop



$$p(S) = \prod_{i=0}^{n} p(S_{i+1} | S_0, \ldots, S_i)$$

Sacha, Mikołaj, et al. "Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits." *Journal of Chemical Information and Modeling* 61.7 (2021): 3273-3284.

Diffusion models



Intractable data distribution

Forward SDE (data → noise)

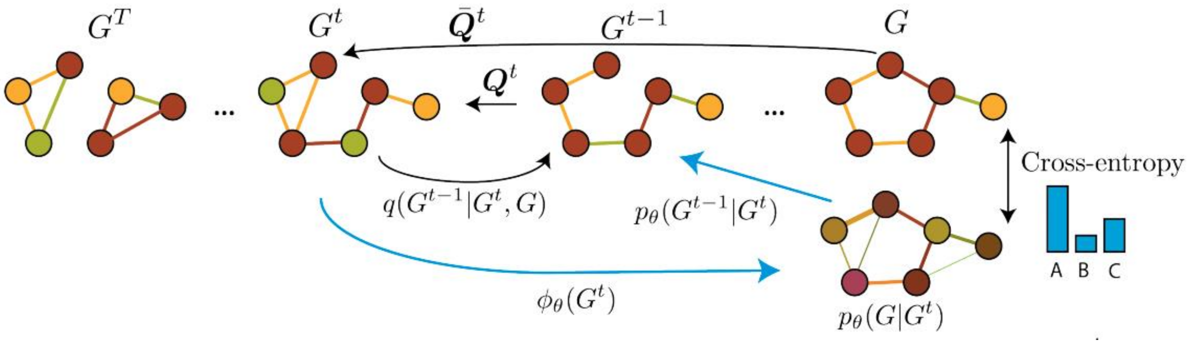
$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

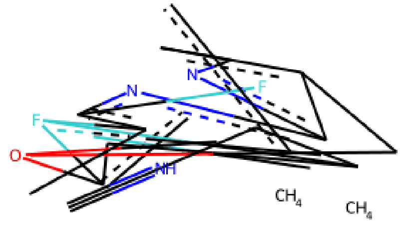
Simple prior

score function

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

Reverse SDE (noise → data)





Frame 0

Song, Yang, et al. "Score-based generative modeling through stochastic differential equations". 2020
Vignac, Clement, et al. "Digress: Discrete denoising diffusion for graph generation." 2022

Flow matching

ODE: $dx = u_t(x)\,dt,$

Loss: $\mathbb{E}_{t,q(z),p_t(x|z)}\|v_\theta(t,x) - u_t(x|z)\|^2$

Generating vector fields:

Diffusion / Variance Exploding:

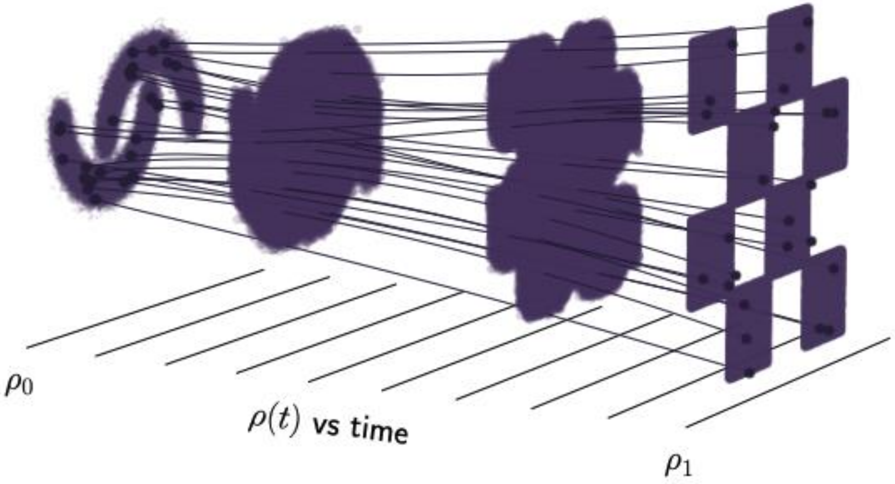
$$p_t(x|z) = \mathcal{N}(x \mid x_1, \sigma^2_{1-t}),$$
$$u_t(x|z) = -\frac{\sigma'_{1-t}}{\sigma_{1-t}}(x - x_1),$$

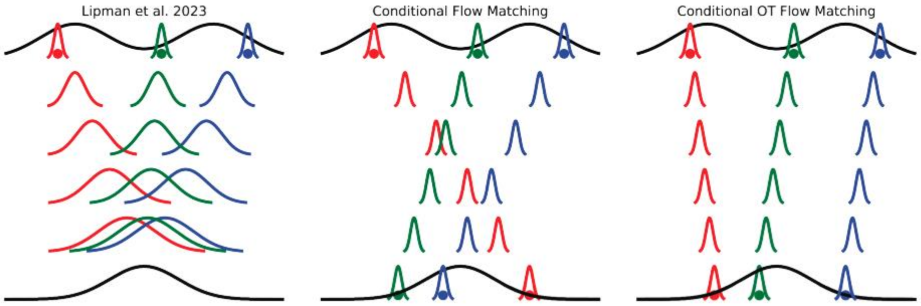
Optimal transport:

$$p_t(x|z) = \mathcal{N}(x \mid tx_1, (t\sigma - t + 1)^2),$$
$$u_t(x|z) = \frac{x_1 - (1-\sigma)x}{1 - (1-\sigma)t},$$

Independent coupling / stochastic interpolants:

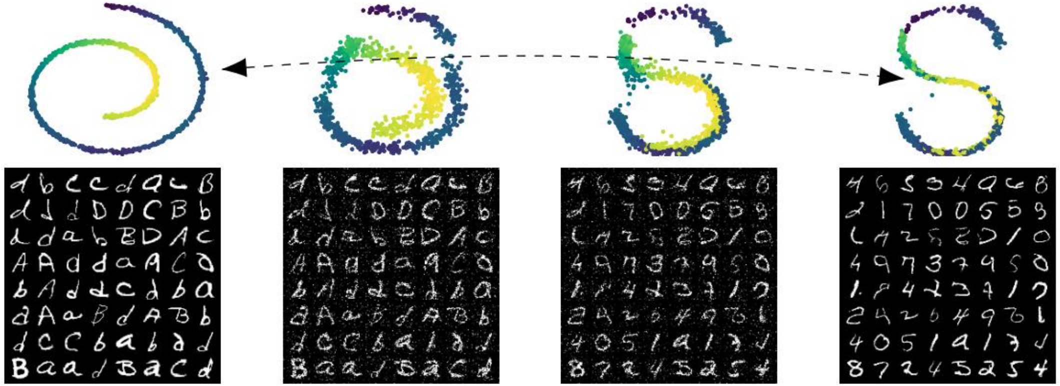
$$p_t(x|z) = \mathcal{N}(x \mid tx_1 + (1-t)x_0, \sigma^2),$$
$$u_t(x|z) = x_1 - x_0.$$





Lipman, Yaron, et al. Flow matching for generative modeling. 2022
Albergo, Michael S. et al. Stochastic interpolants: A unifying framework for flows and diffusions. 2023
Tong, Alexander, et al. Improving and generalizing flow-based generative models with minibatch optimal transport. 2023

Schrödinger bridge

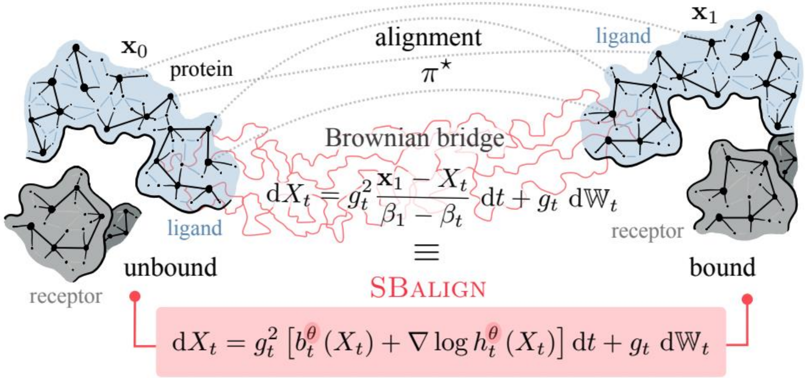


Schrödinger bridge Problem:

$$\min_{\mathbb{P}_0=\hat{\mathbb{P}}_0,\ \mathbb{P}_1=\hat{\mathbb{P}}_1} D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{Q}_t).$$

a.k.a. entropy-regularised OT:

$$\inf_{\pi} \int_{x_0}\int_{x_1} \frac{\|x_0 - x_1\|^2}{2} d\pi(x_0, x_1) - \gamma H(\pi)$$



Aligned SB: **samples are coupled**

Loss: $L(\theta) := \mathbb{E}\left[\int_0^1 \left\| \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + \nabla \log h_t^\theta(X_t)\right)\right\|^2 \mathrm{d}t\right]$

Schrödinger, Erwin. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. 1932
De Bortoli, Valentin, et al. Diffusion Schrödinger bridge with applications to score-based generative modelling. 2021
Somnath, Vignesh Ram, et al. Aligned Diffusion Schrödinger Bridges. 2023

Chen, Tianrong, et al. Likelihood training of Schrödinger bridge using forward-backward sdes theory. 2021
Vargas, Francisco, et al. Solving Schrödinger bridges via maximum likelihood. 2021
Liu, Guan-Horng, et al. Deep generalized Schrödinger bridge. 2022