Summer School Lab

Molecular Generation

June 14, 2024

Presented by

IVADO Valence Labs Mila

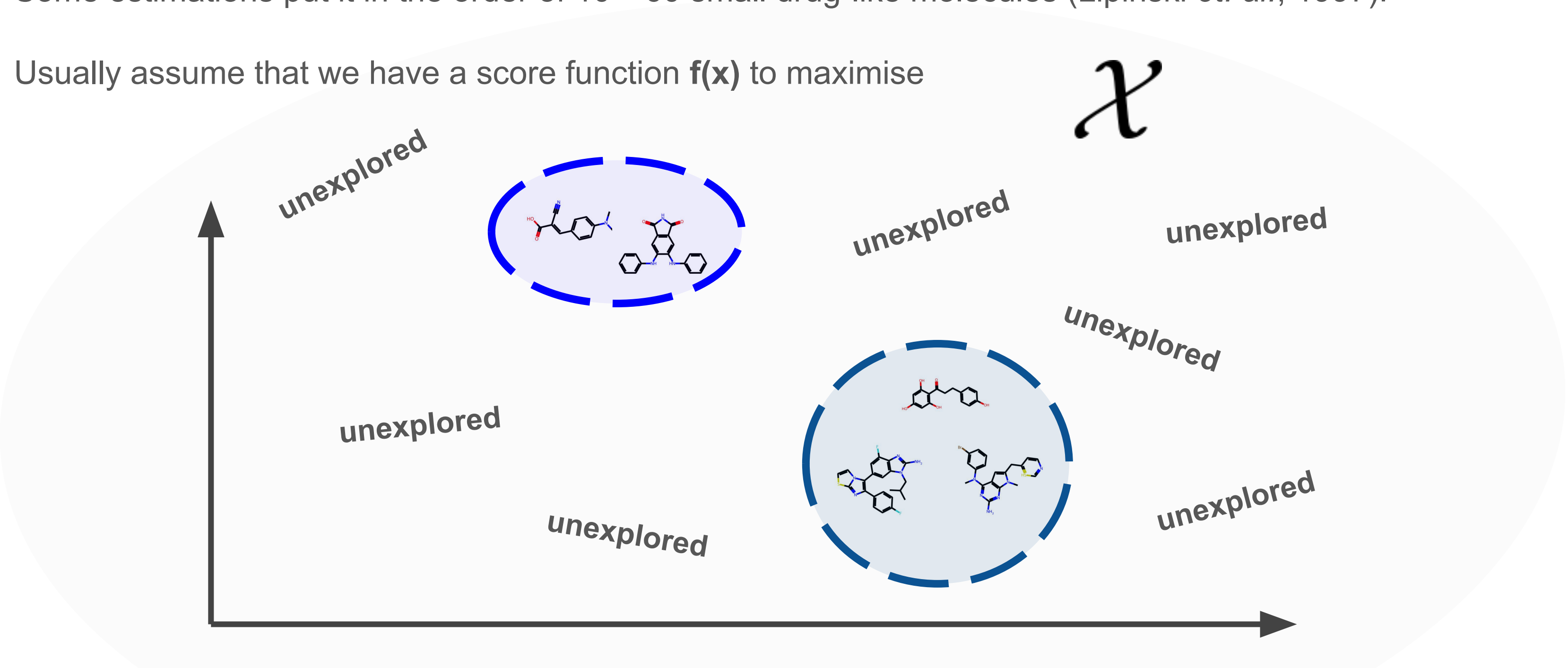
Agenda

15:00 - 15:30	**Graph GA**
15:30 - 16:00	Break
16:00 - 16:45	**Fragment GFlowNet**
16:45 - 17:15	**SAFE Scaffold Decoration**
17:15 - 17:30	Recap

Motivation

- **The Molecular Space is vast**
  It contains all the molecules that have ever been manufactured, and all the molecules that *could* exist !
  Some estimations put it in the order of 10 ^ 60 small drug-like molecules (Lipinski *et. al.*, 1997).

- Usually assume that we have a score function **f(x)** to maximise



Genetic Algorithms for molecular generation

- **Iterative application of selection and mutation steps**

  Selection: select the members of the population (molecule) with the highest fitness (score)

  Mutation: combine different elements of the population members to create the next generation

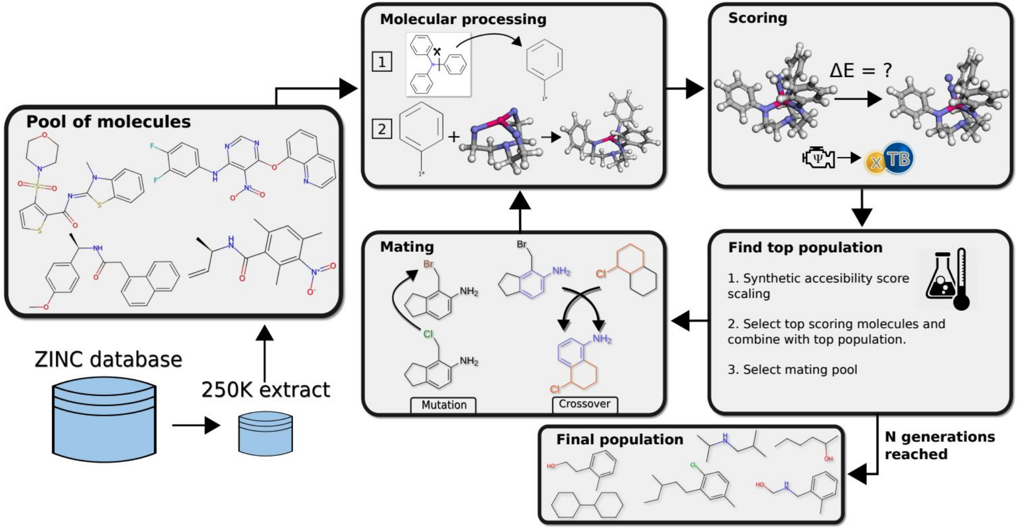
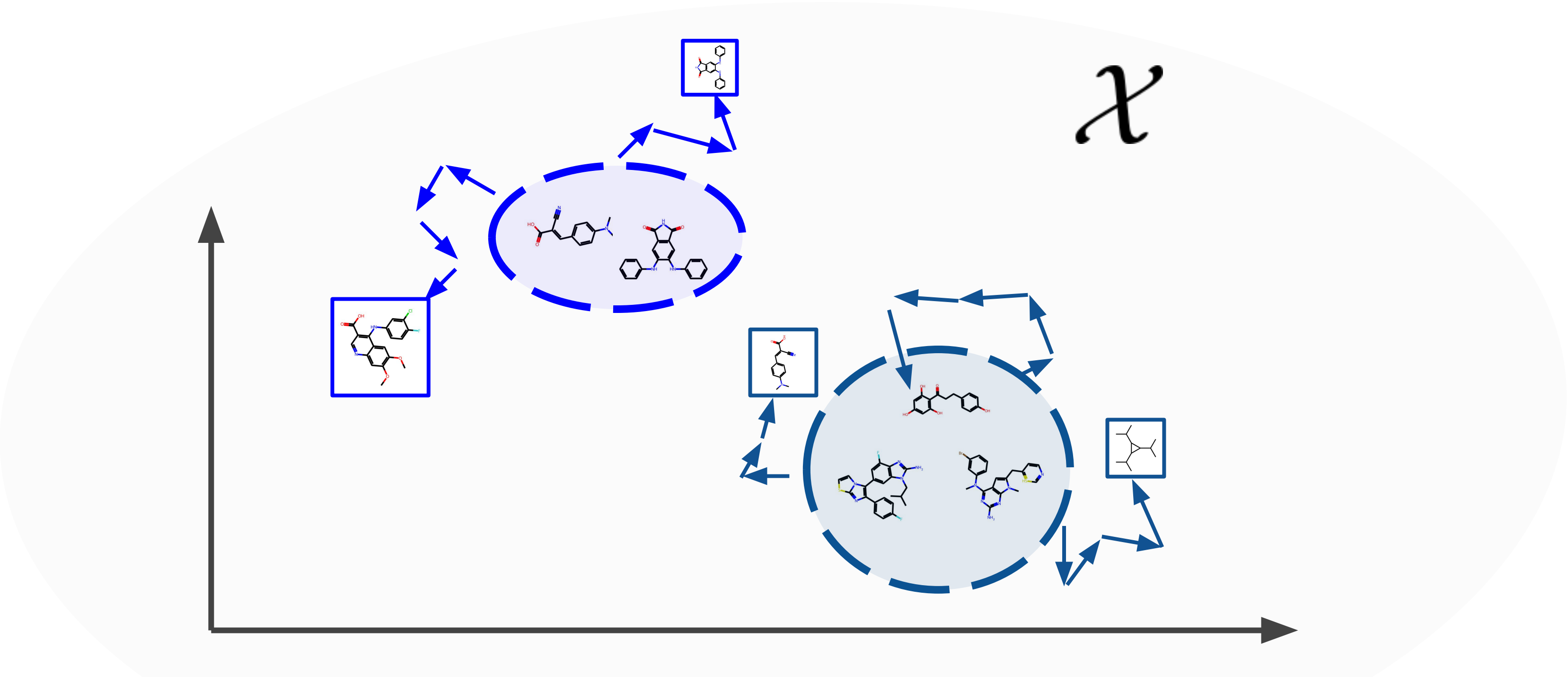


Figure from: Strandgaard, M., Seumer, J., Benediktsson, B., Bhowmik, A., Vegge, T., & Jensen, J. H. (2023). Genetic algorithm-based re-optimization of the Schrock catalyst for dinitrogen fixation. *PeerJ physical chemistry*, *5*, e30.

Genetic Algorithms for molecular generation

In our molecular space, GAs perform local search around existing clusters.



Part 1: Setting up and exploring `medchem` and `graph-ga`

- **MedChem**

  Offers several filtering options for medicinal chemistry

  For more info, go to: medchem-docs.datamol.io/stable/

- **Graph GA**

  Offers several filtering options for medicinal chemistry

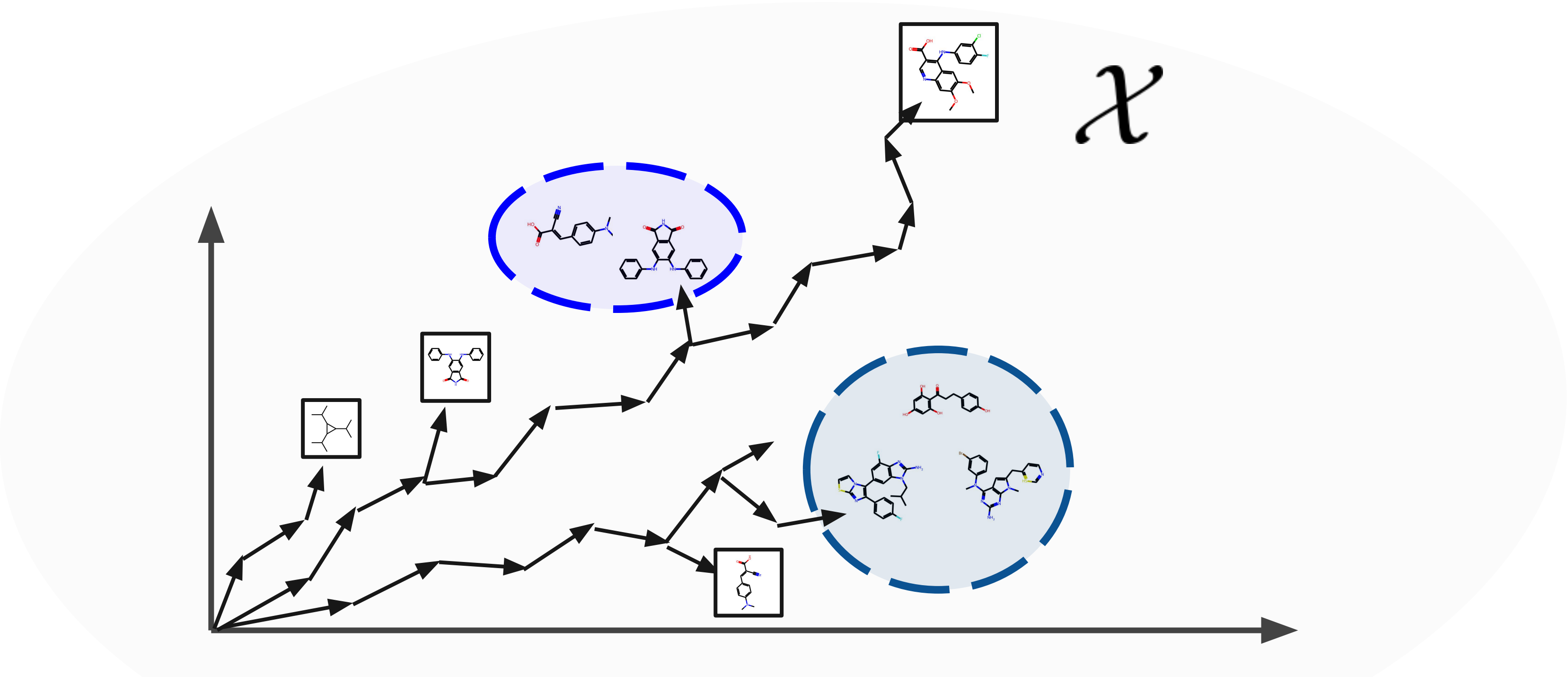
  For more info, go to: github.com/AustinT/mol_ga

Go through Part 1 in the Colab Notebook

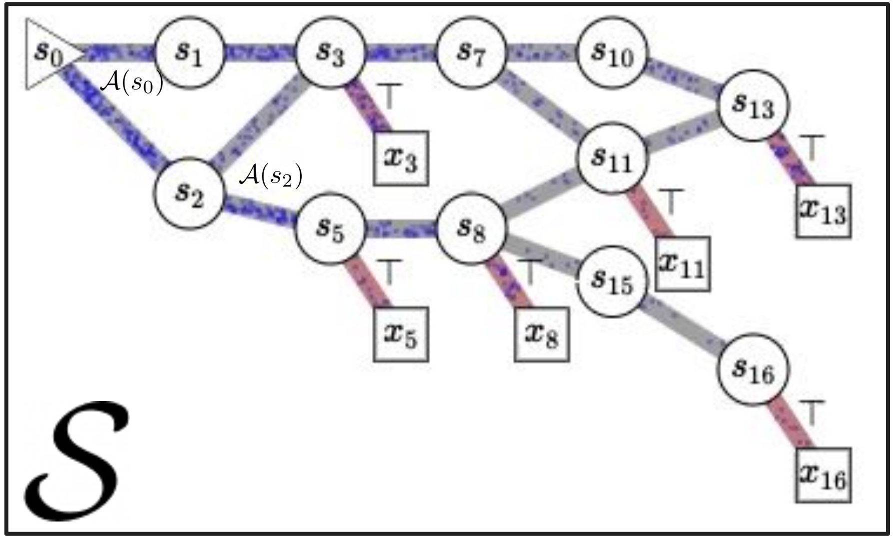
Ends at 15:30 !

Molecular Generation as Sequential Decision Making

In our molecular space, GFlowNets create new molecules by starting from the origin (empty set)



GFlowNets basics



A GFlowNet learns a flow of probability from state to state to sample objects proportionally to their reward:

$$p_\pi(x) = \frac{r(x)}{Z} \quad , \quad Z := \sum_{x' \in \mathcal{X}} x'$$

Fragment GFN

In a fragment GFN for molecular generation, states are molecules and actions are fragments

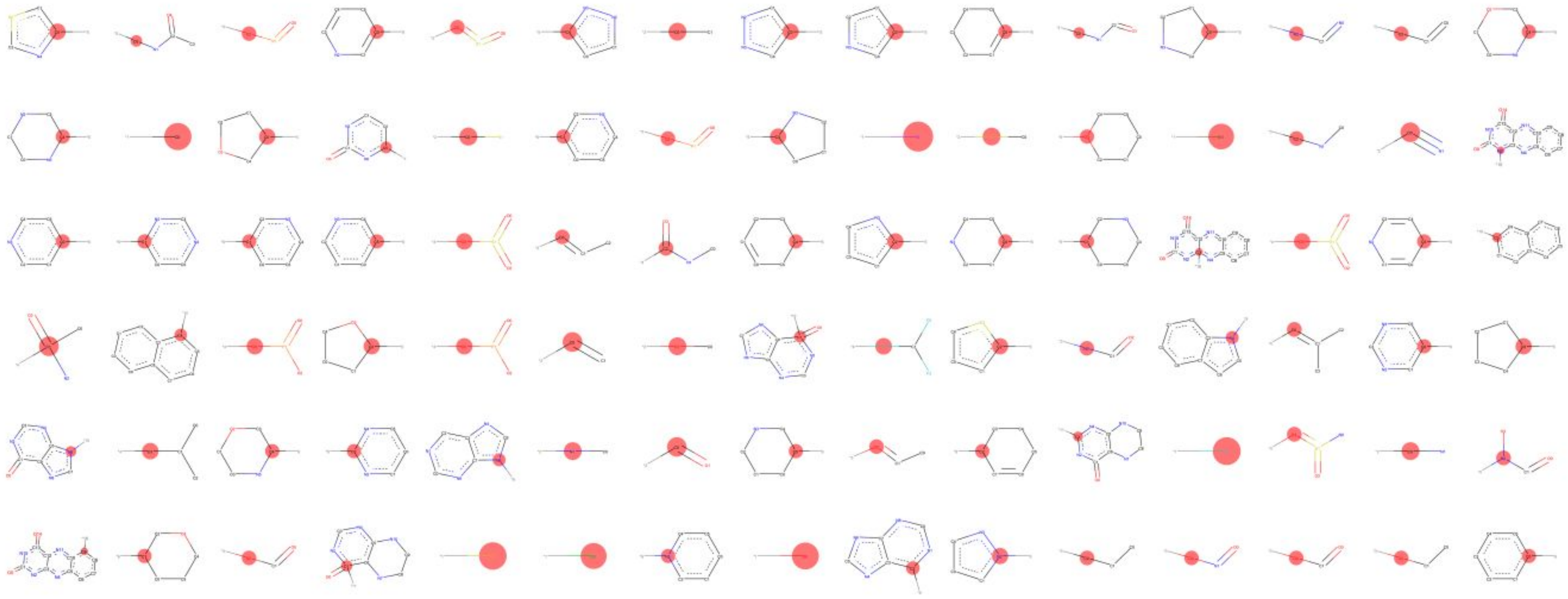
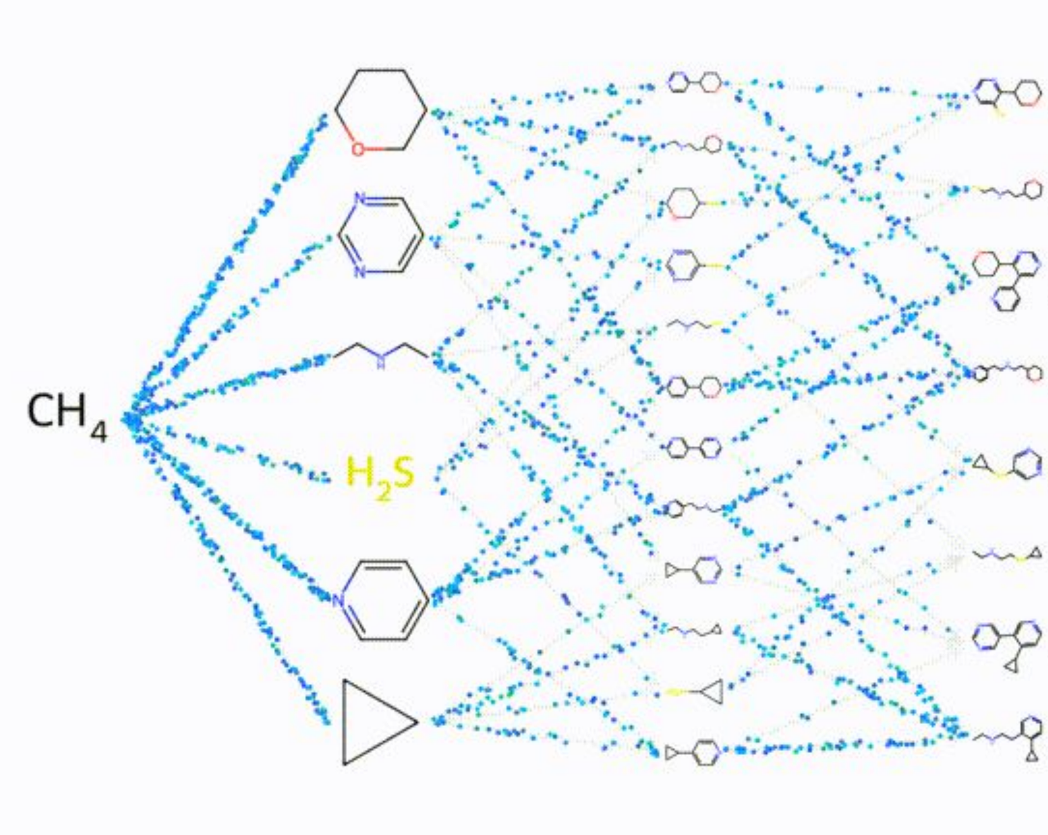


Figure from: Bengio, E., Jain, M., Korablyov, M., Precup, D., & Bengio, Y. (2021). Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, *34*, 27381-27394.

Fragment GFN

Given enough time to learn and explore, our GFN will learn to sample terminal states (finished molecules) according to their score.



Part 2: Exploring `gflownet` for fragment-based design

- **Gflownet codebase**

  Specialised for compositional object generation on graphs
  In our case, we are building graphs of molecular fragments!

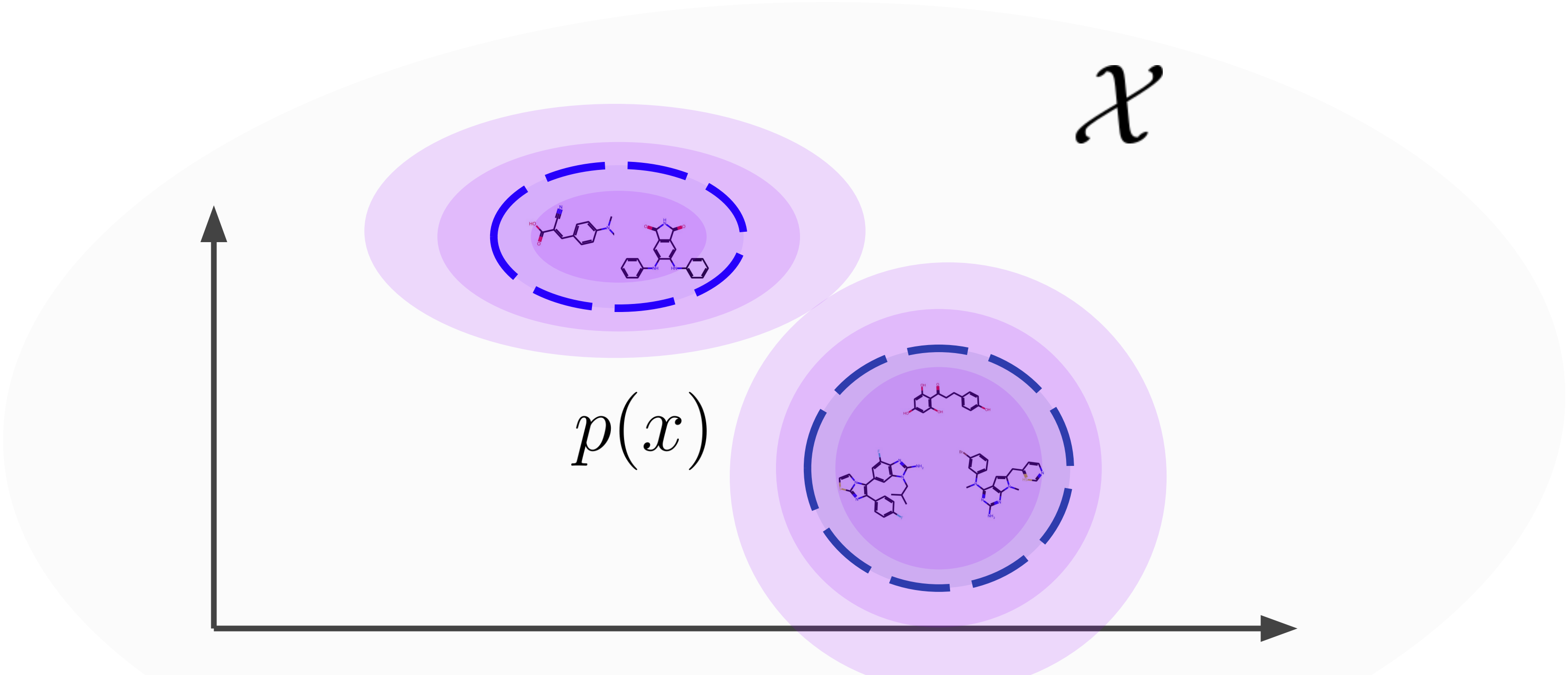
  For more info, go to: github.com/recursionpharma/gflownet

Go through Part 2 in the Colab Notebook

Ends at 16:45 !

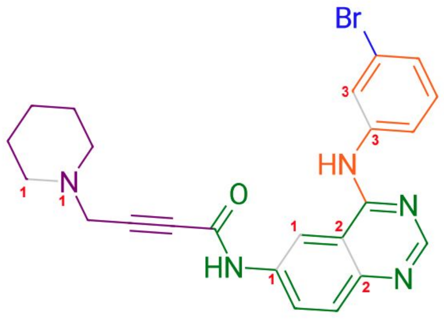
Molecular Generation from Likelihood Models

In our molecular space, language models learn a probability distribution around the known molecular clusters.



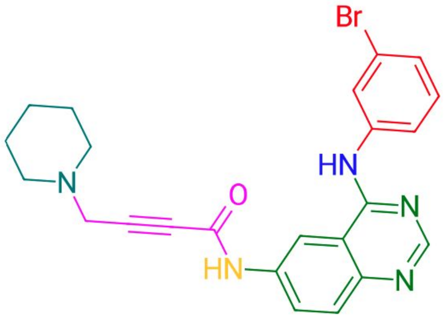
SAFE encoding

SMILES



O=C(C#CCN1CCCCC1)Nc1ccc2ncnc(Nc3cccc(Br)c3)c2c1

SAFE

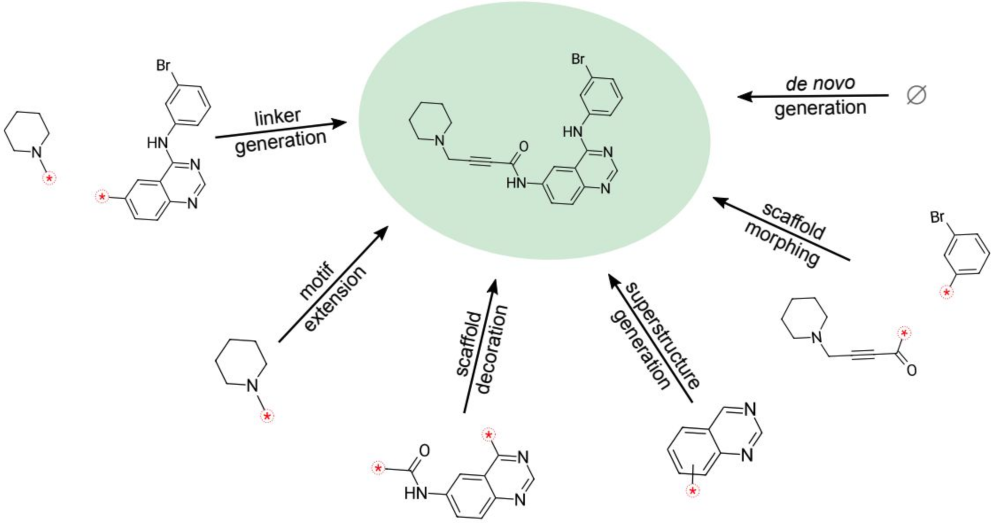


N18CCCCC1.O=C6C#CC8.N67.c17ccc2ncnc4c2c1.N45.c15cccc(Br)c1

SAFE Scaffold Decoration

Dataset

```
C12C3C14.CC2C.C3(C)C.C4(C)C
c14cc2ncc8c6c2cc15.c17ccc(F)c(Cl)c1.C8(=O)O.CO4.N67.O5C
O=C1NC(=O)c2cc4c6cc21.c15ccccc1.c17ccccc1.N45.N67
c12ccc(C=3)cc1.C=3(C#N)C(=O)O.CN2C
c13nc(N)nc2c1cc5n2C.c14cccc(Br)c1.C5C1=[SH]C=NC=C1.CN34
O=C1NC(=O)c2cc4c6cc21.c15ccccc1.c17ccccc1.N45.N67
c12c(O)cc(O)cc1O.c13ccc(O)cc1.O=C2CC3
n15c(N)nc2c(F)cc6cc21.c16c7nc2sccn12.c17ccc(F)cc1.CC(C)C5
```



Transformer

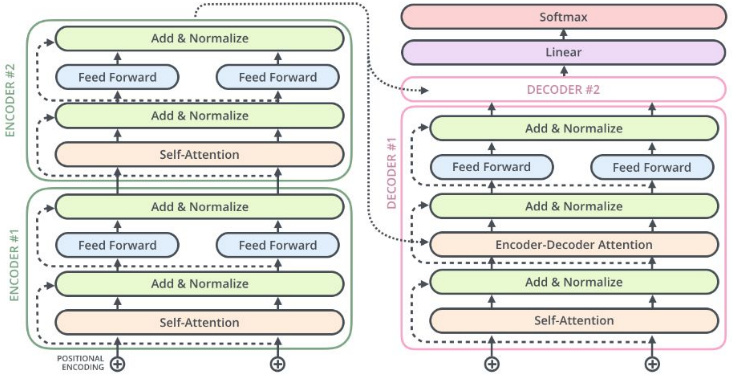


Figure from: jalammar.github.io/illustrated-transformer/

Part 3: Exploring `safe` for language-based design

- **SAFE**

  SAFEs are a newly proposed molecular string representation which allows for a variety of tasks from a single language model.

  In this lab, we use SAFE-GPT to refine a candidate molecules with scaffold decoration.

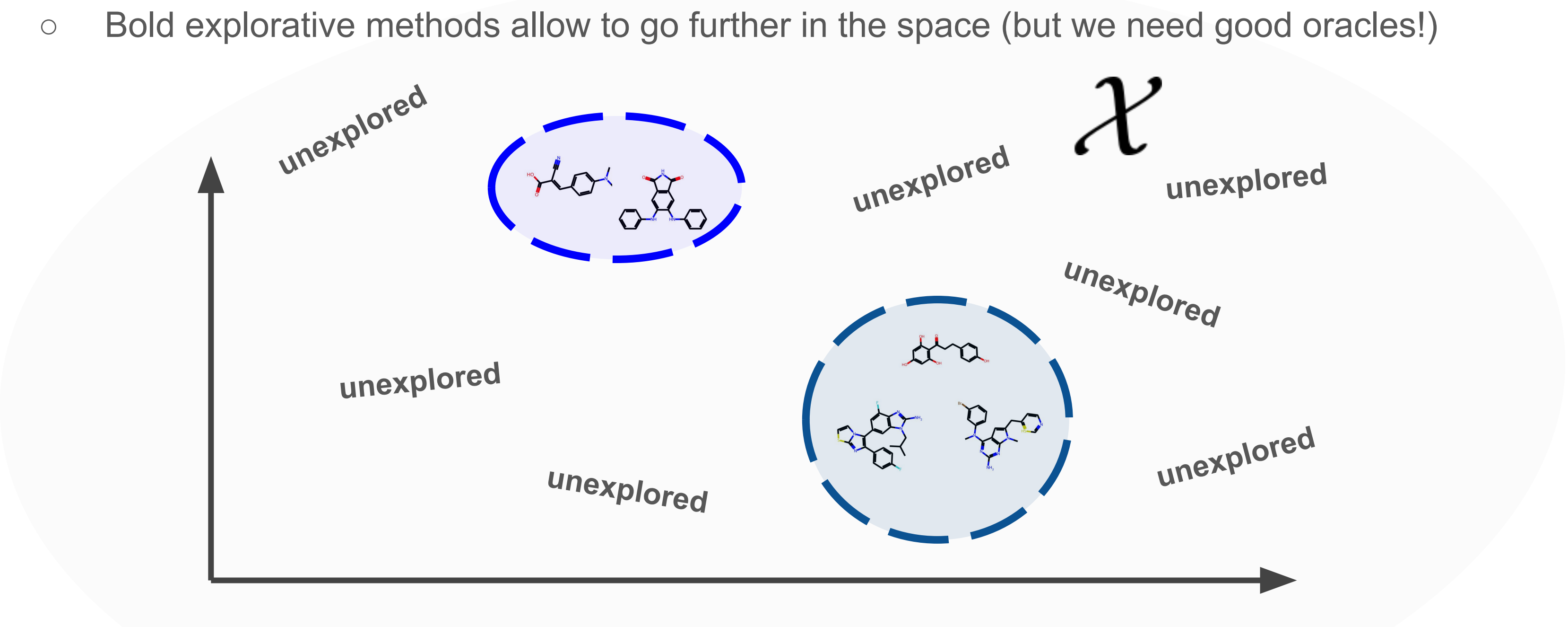
  For more info, go to: github.com/datamol-io/safe

Go through Part 3 in the Colab Notebook

Ends at 17:15 !

Recap: different approaches to molecular generation

- **The Molecular Space is vast, and we need different approaches for different tasks**
  - Fine-grained methods and local search are crucial for lead optimization
  - Likelihood-based methods represent inductive biases towards what we know works
  - Bold explorative methods allow to go further in the space (but we need good oracles!)



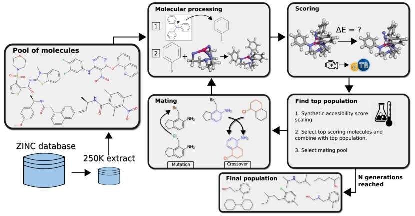
Recap: different approaches to molecular generation

- ## Genetic Algorithms

  Local search based on fitness function.

  Genetic algorithms are strong baselines for molecule generation

  A graph-based genetic algorithm [...] for the exploration of chemical space



- ## Generative Flow Networks (GFlowNets)

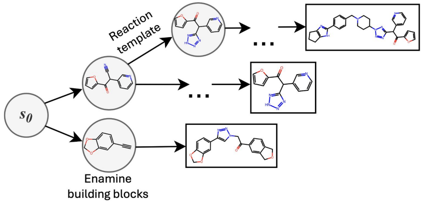
  Molecular generation as a sequential decision making process.

  Learns to samples new molecules proportionally to their score.

  Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation

  Goal-conditioned GFlowNets for Controllable Multi-Objective Molecular Design

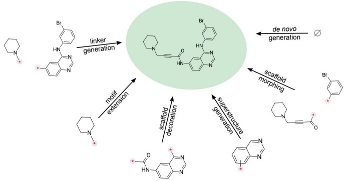
  SynFlowNet: Towards Molecule Design with Guaranteed Synthesis Pathways



- ## SAFE GPT

  Language-based molecular generation applying to a variety of generative tasks.

  Gotta be SAFE: A New Framework for Molecular Design



Summer School Lab

The End.

Connect with us

Any questions, ideas or other feedback?
We would love to hear from you!

Presented by

IVADO Valence Labs Mila