DrugEx: deep learning for *de novo* drug design

– A case for A2B selective ligands –

Ninth Joint Sheffield Conference on Chemoinformatics

Universiteit Leiden
The Netherlands

LACDR

Sohvi Luukkonen, M. Sícho, L. Schoenmaker, H.W. van den Maagdenberg, O. Béquignon, J. Madern, D. Van der Es, G.J.P. Van Westen

de novo Drug Design

DrugEx

QSPRpred

Presentation by
H.W. van den Maagdenberg

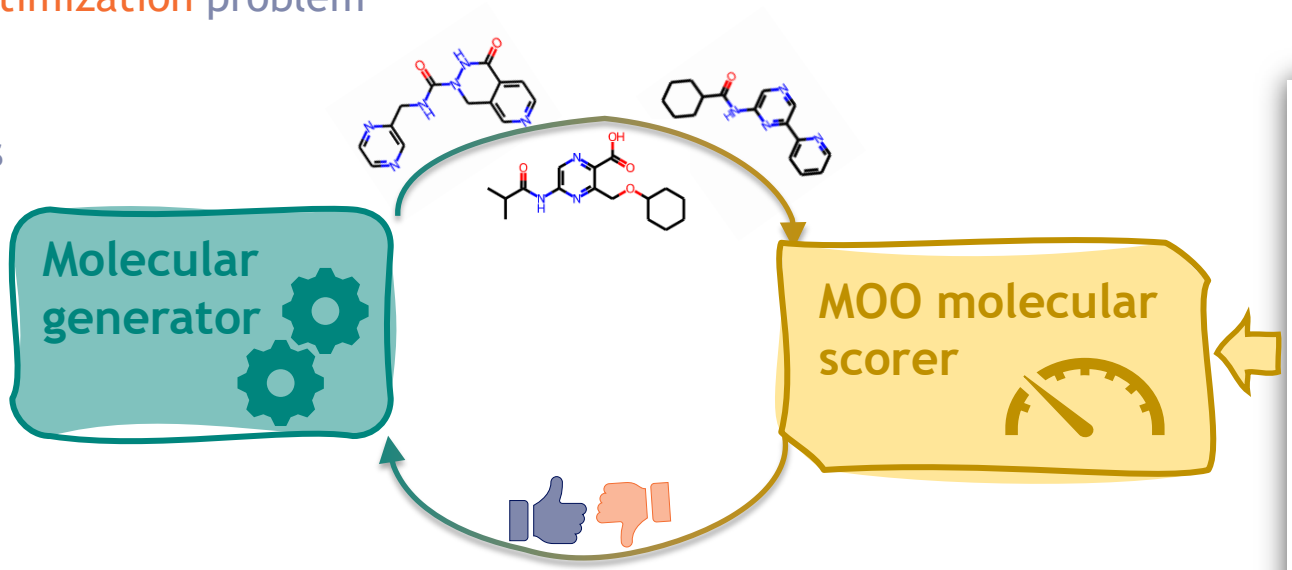
Chemical space of drug-like compounds
- ~10^8 synthesized molecule
- ~10^{33}-10^{60} estimated drug-like molecules

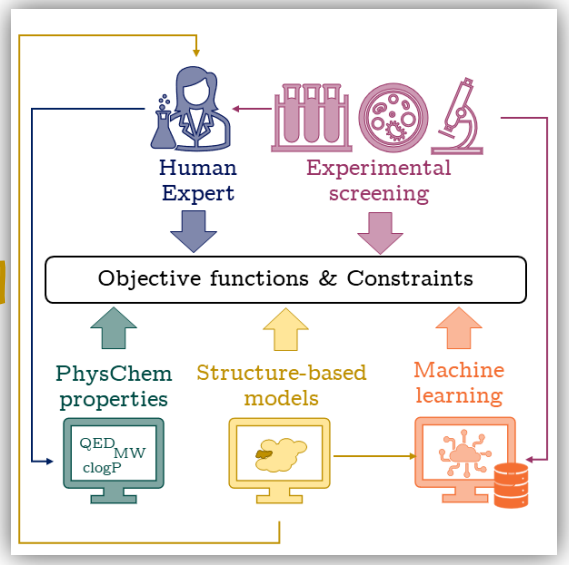
A good candidate fulfills multiple properties
- Maximize affinity, efficiency, synthezability, drug-likeness ...
- Minimize off-target effect, toxicity ...

⇒ a multi-objective optimization problem

Generators: genetic algorithms & deep learning (RNNs, VAEs, Transformers)

Training: iterative distribution, adversarial, reinforcement, *transfer* & *conditional* learning





Liu et al. (2021) *Methods Mol. Bio.*
Luukkonen et al. (2023) *Curr. Opin. Struct. Biol.*; Former and Coley (2023) *Patterns*

DrugEx – the Evolution of the Drug Explorer

DrugEx v1.0
RNN-based (GRU)
single-objective RL (SMILES)

*An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor, **2019**, J. Cheminf.*

By Xuhan Liu

DrugEx v2.0
RNN-based (LSTM)
multi-objective RL (SMILES)

*DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology, **2021**, J. Cheminf.*

By Xuhan Liu

DrugEx v3.0
Transformer-based
multi-objective RL (SMILES/Graph)

*DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning, **2023**, J. Cheminf.*

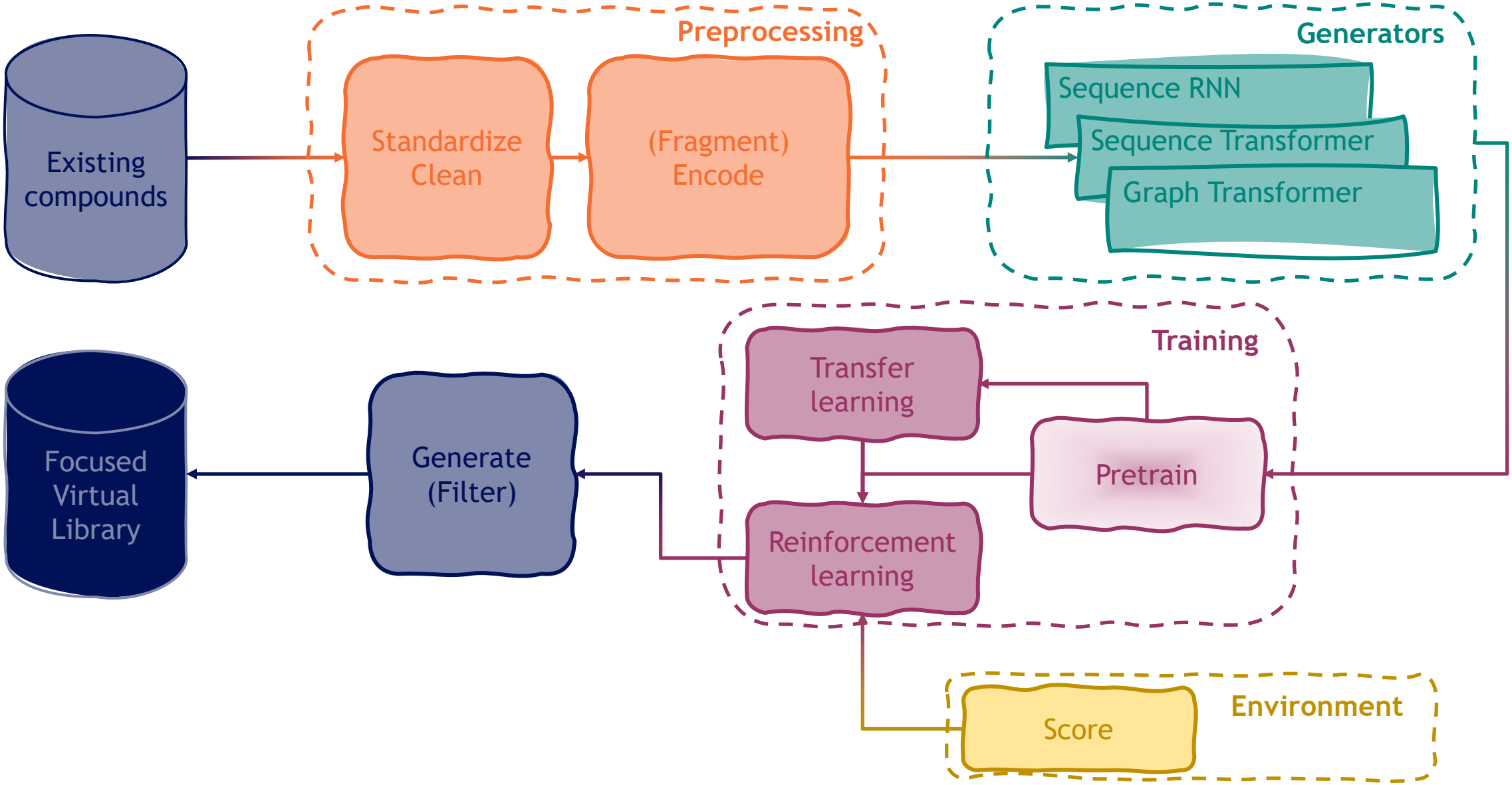
By Xuhan Liu

DrugEx (v3.4.4)
open-source software library for *de novo* design of small molecules
with deep learning generative models in a multi-objective reinforcement learning framework

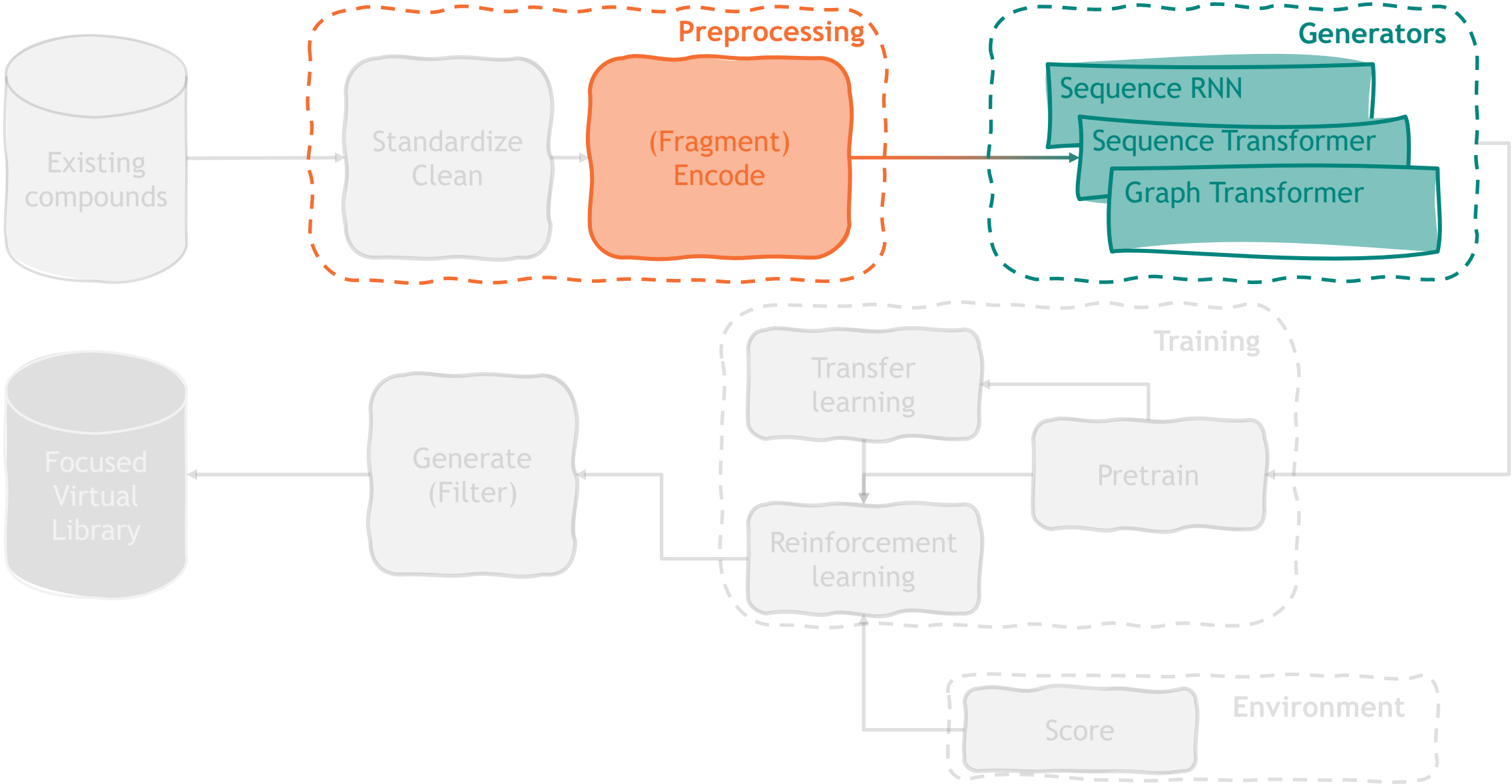
*DrugEx: Deep Learning Models and Tools for Exploration of Drug-like Chemical Space, **2023**, J. Chem. Inf. Model.*

By DrugEx dev team @ CDD Leiden

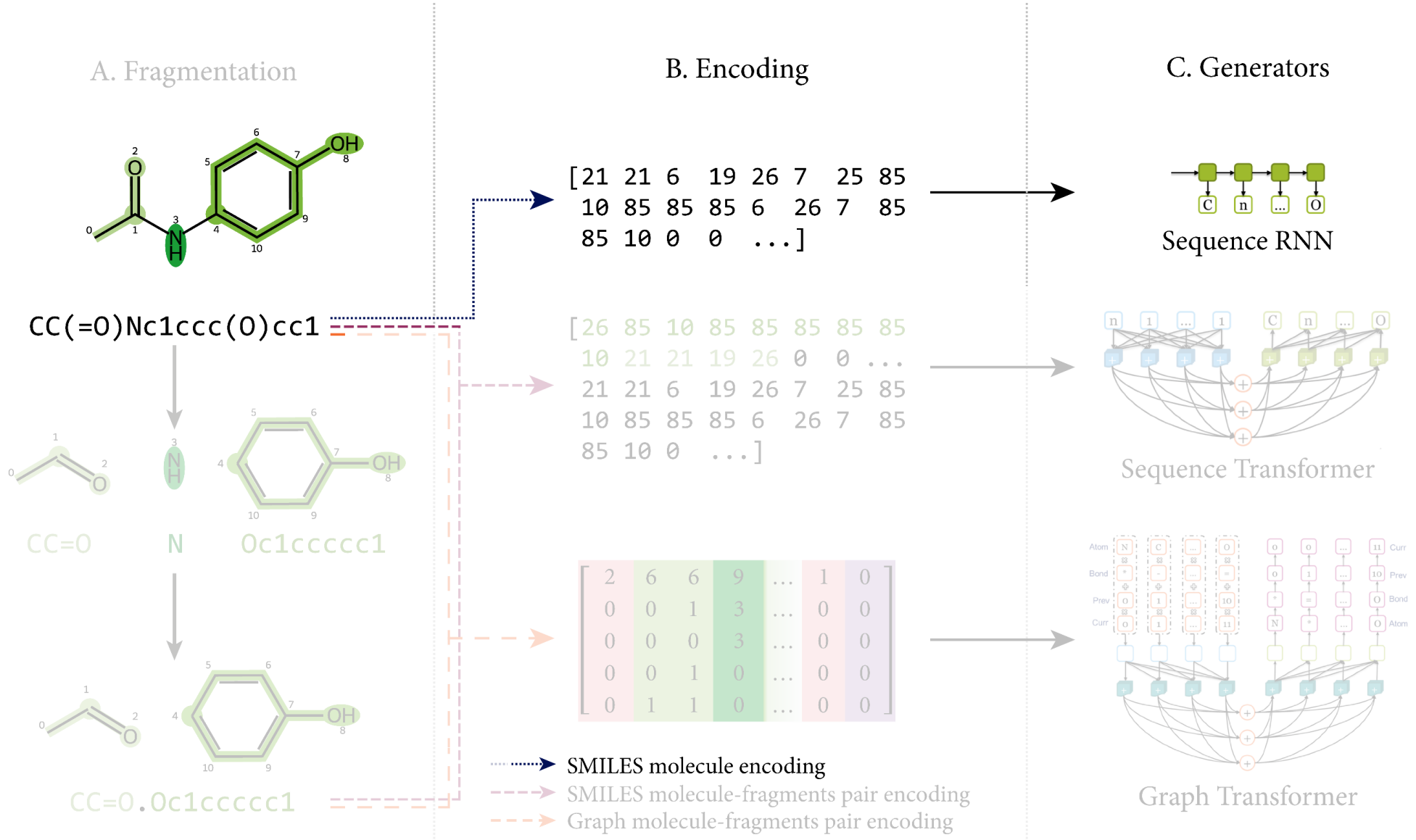
DrugEx – the Workflow



DrugEx – the Workflow

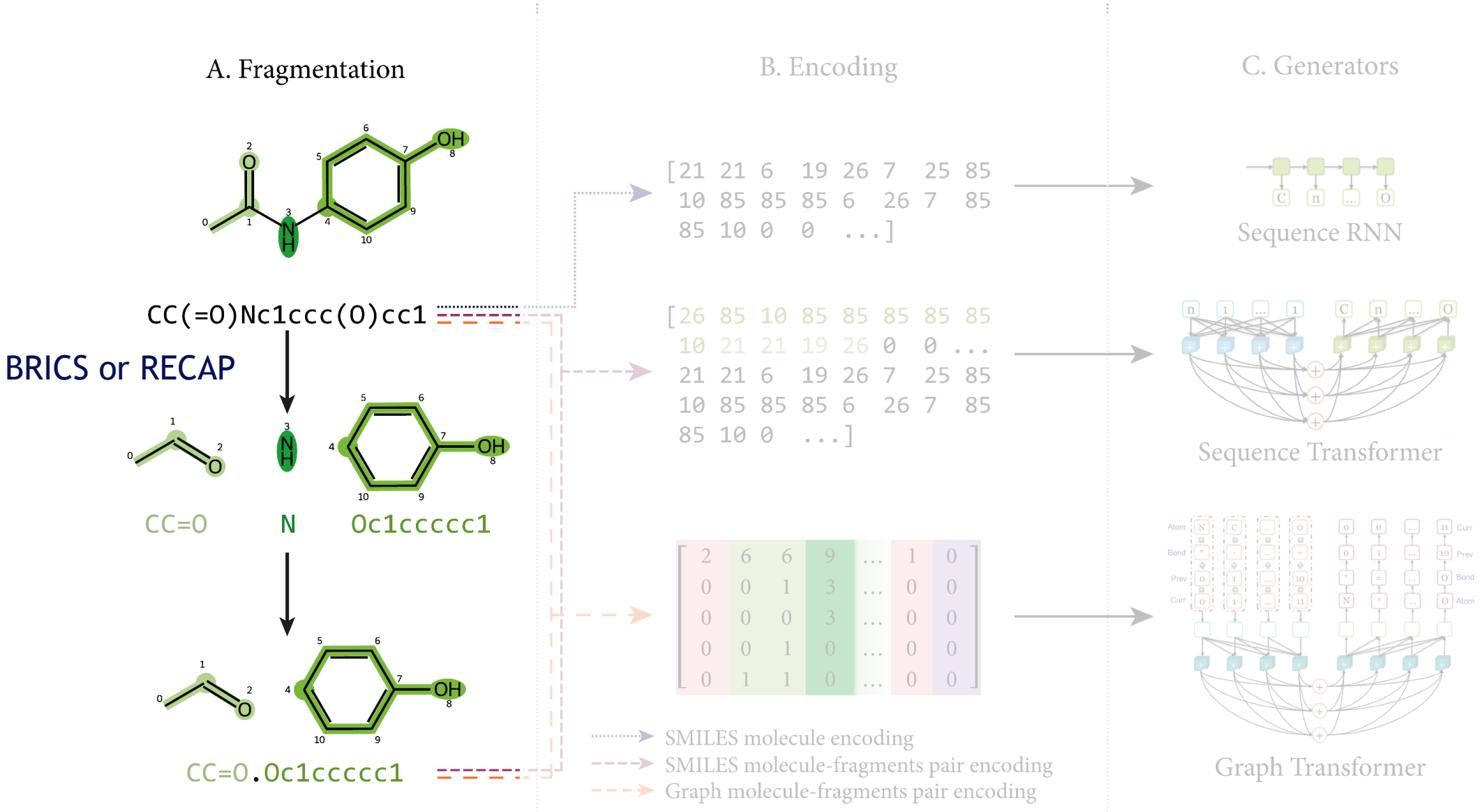


DrugEx - 3 Flavors: Sequence RNN



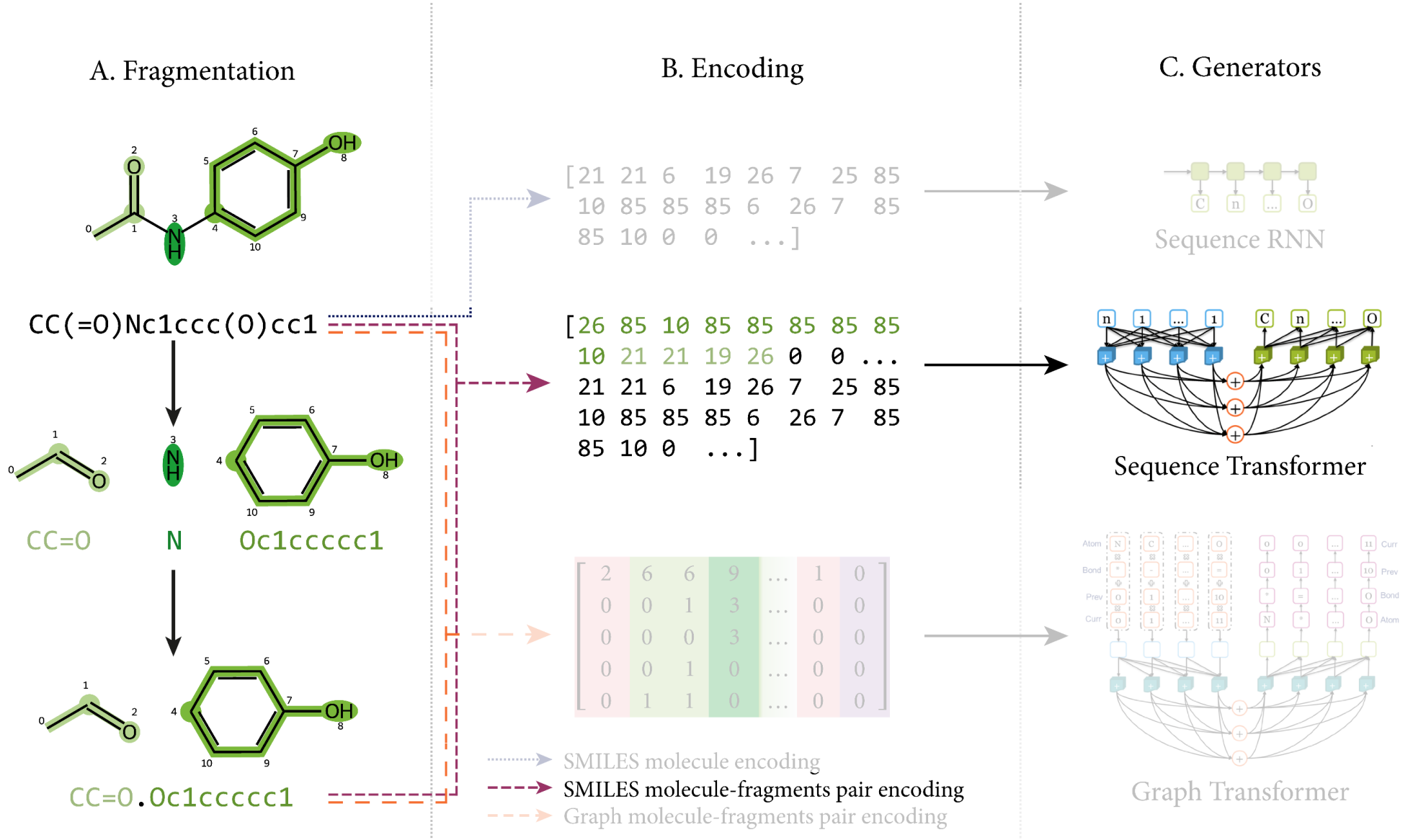
Liu et al. (2019) *J. Cheminf.*
Šícho et al. (2023) *J. Chem. Inf. Model.*

DrugEx - 3 Flavors: Transformers

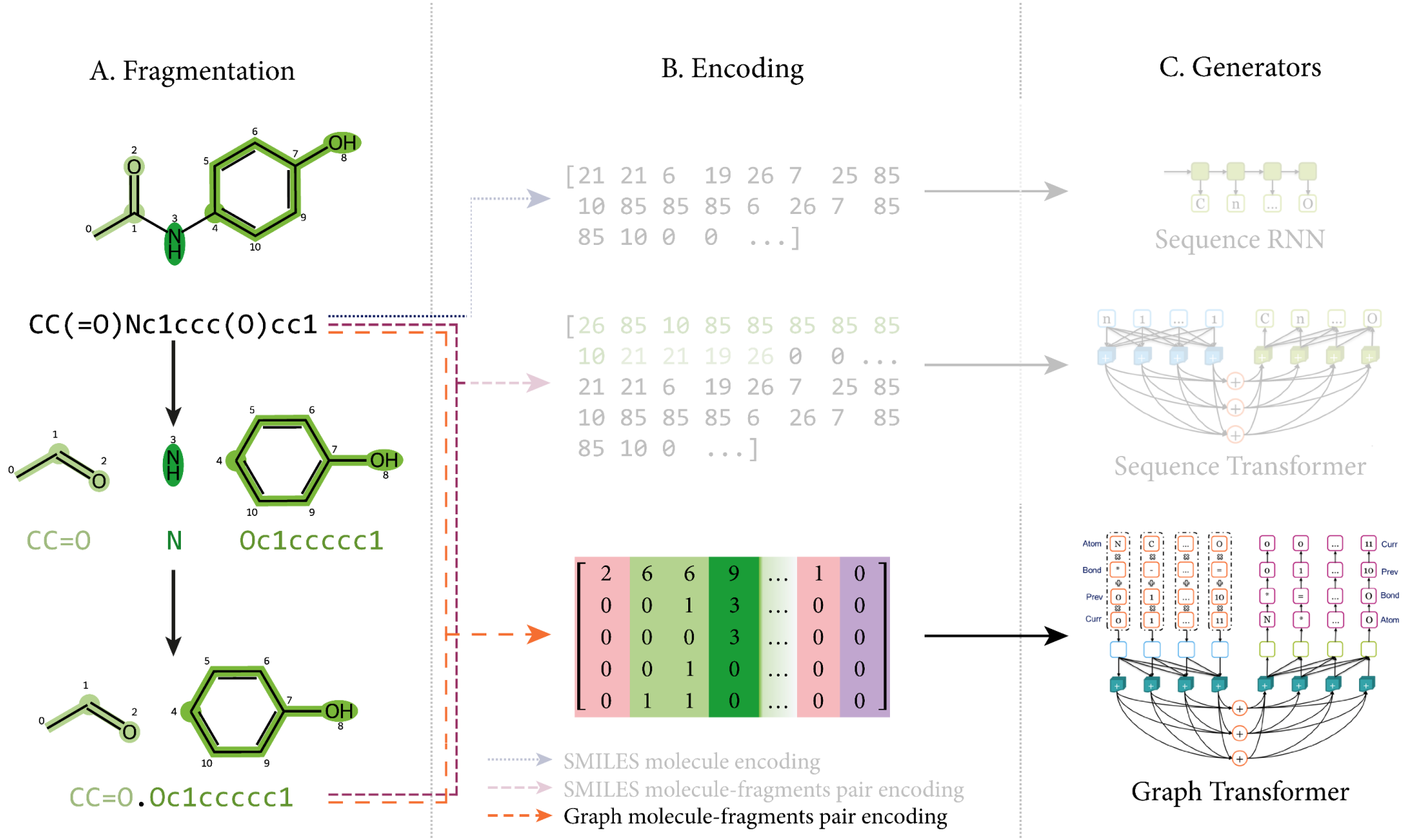


Liu et al. (2023) *J. Cheminf.*; Šícho et al. (2023) *J. Chem. Inf. Model.*; Degen et al. (2008) *ChemMedChem*; Lewell et al. (1998) *J. Chem. Inf. Comput. Sci.*

DrugEx - 3 Flavors: Sequence Transformer

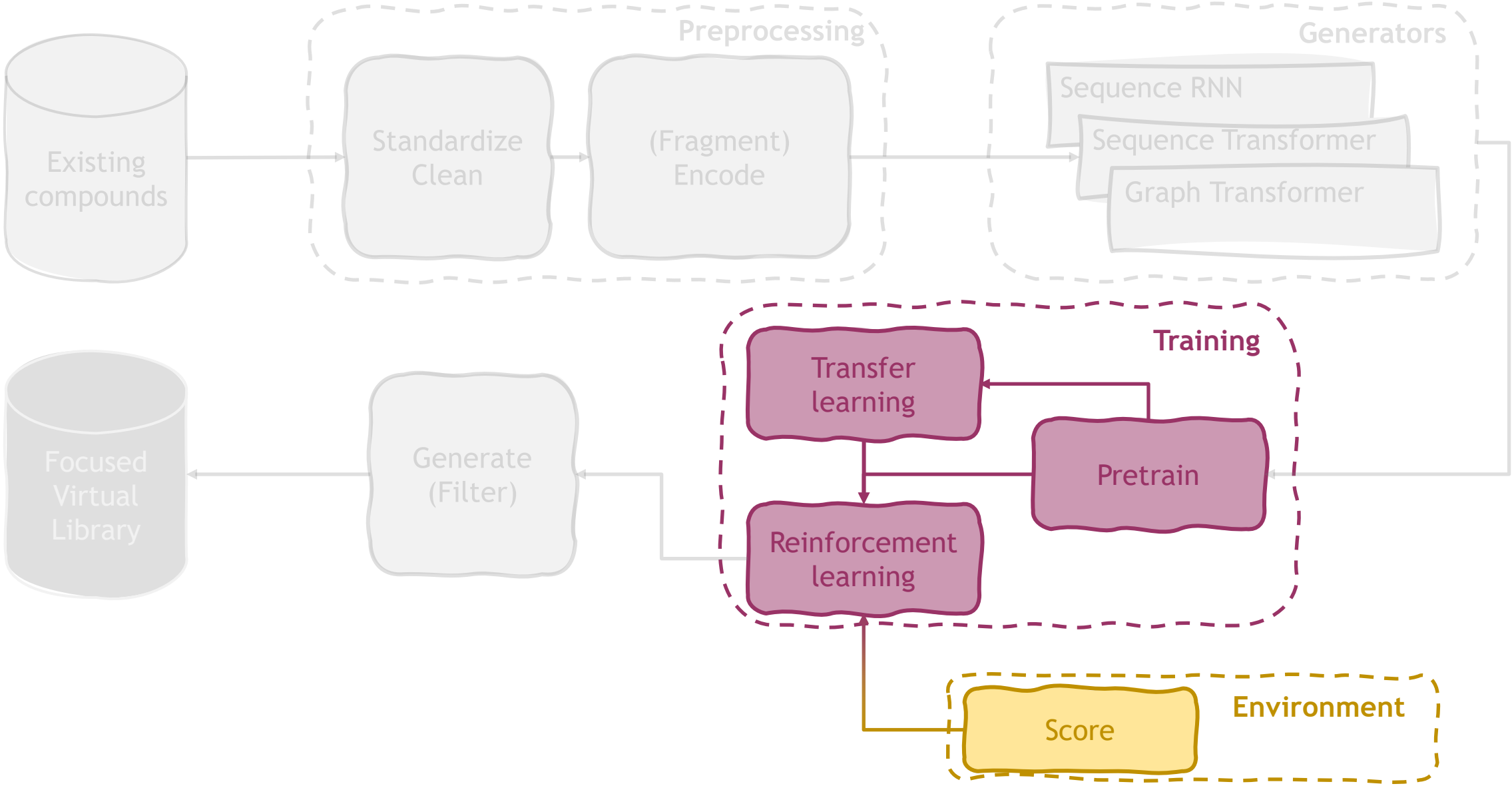


DrugEx - 3 Flavors: Graph Transformer



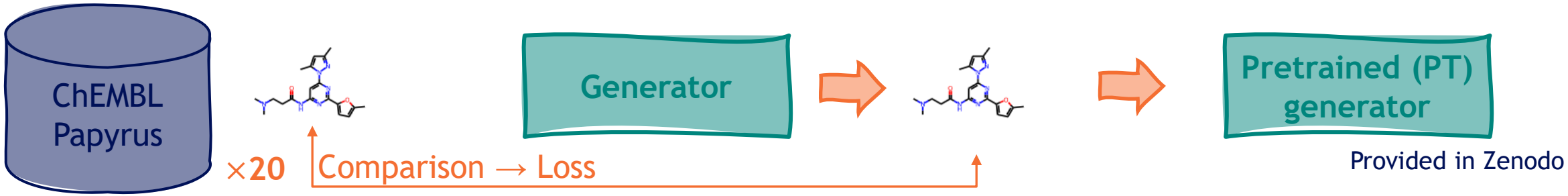
Liu et al. (2023) *J. Cheminf.*
Šícho et al. (2023) *J. Chem. Inf. Model.*

DrugEx – the Workflow



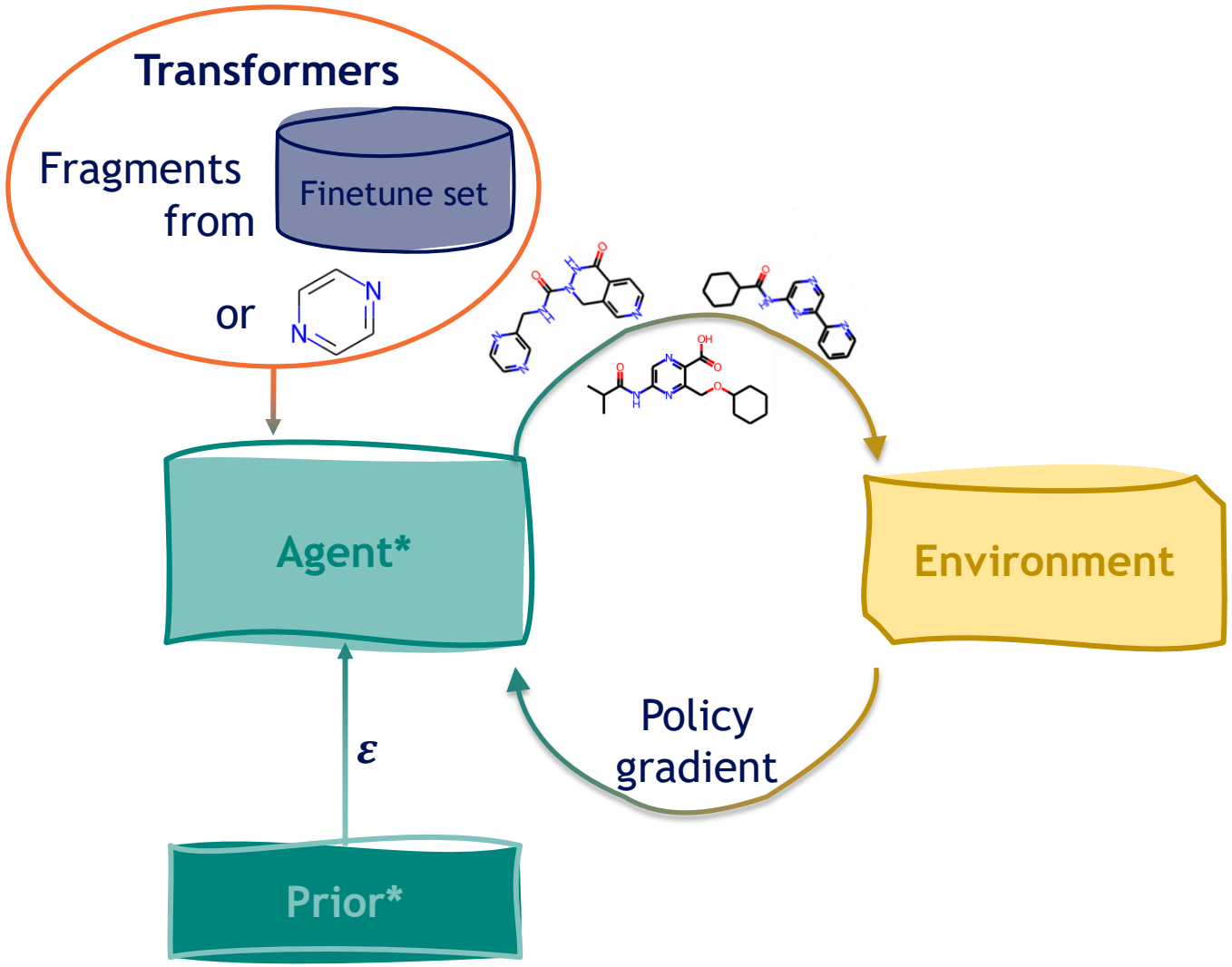
DrugEx – Pretraining & Finetuning

Pretraining

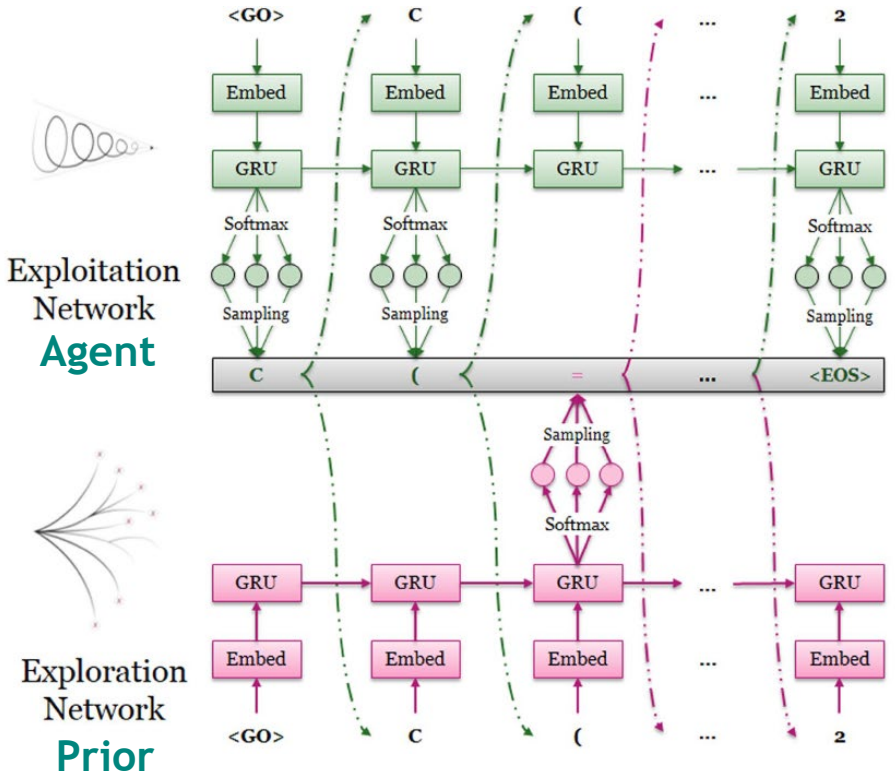


Finetuning / transfer learning

DrugEx – Reinforcement Learning



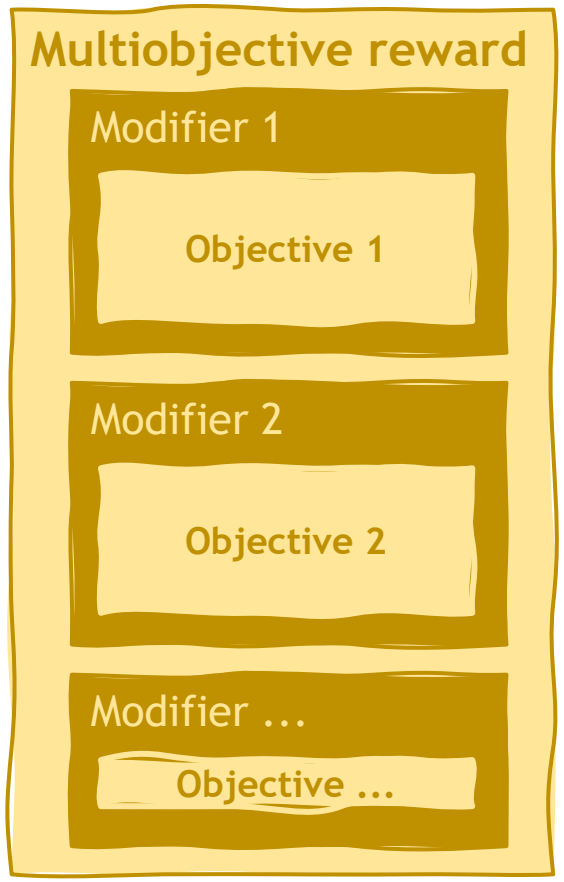
Exploitation-exploration strategy



From Liu et al. (2019) J. Cheminfo.

*PT or FT

DrugEx – the Environment



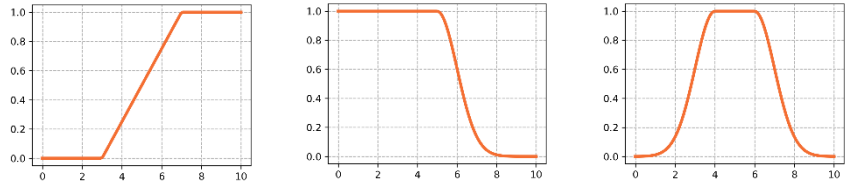
Objectives

- Over 20 predefined properties
  - PhysChem *MW, logP, QED, TPSA ...*
  - Similarity *Tversky, Fraggle & substructure*
  - Synthetic accessibility *SA, RA & LED3**
  - Efficiency *Ligand & Lipophilic*
- QSPR models from QSPRpred**
- Custom scorers *with the API*

Modifiers

All objectives need to be maximisation tasks and scaled between 0 and 1

- 10 predefined modifiers



- Custom modifiers *with the API*

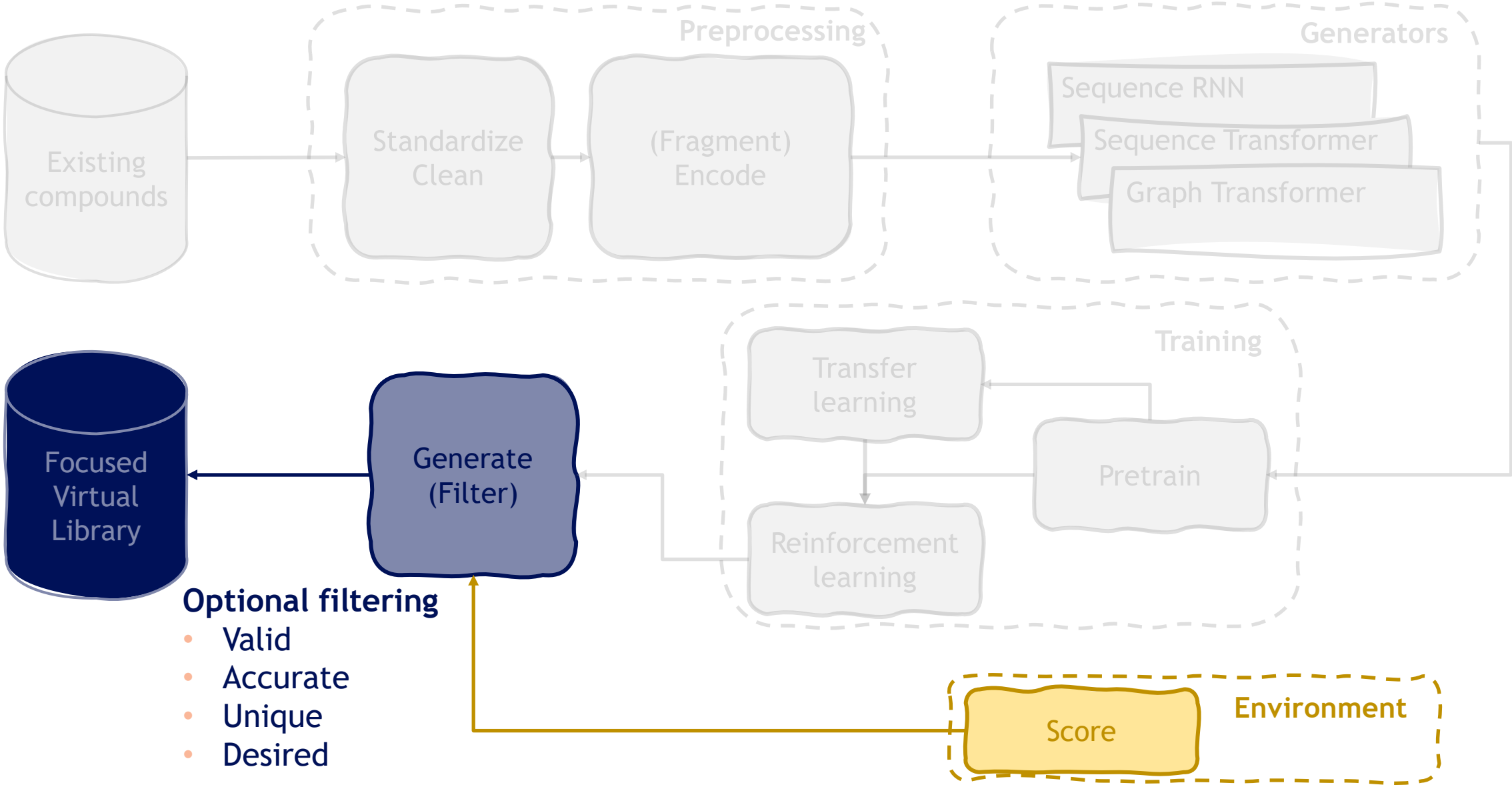
Multiobjective optimisation

- Scalarization
  - Dynamic/parametric weighted sum – WS
- Pareto ranking with
  - Crowding distance (NSGA-II) – PRCW
  - Tanimoto distance – PRTD
- Custom methods *with the API*

**Posters by M. Šícho and A.H. Kai*
***Presentation by H.W. van den Maagdenberg*

Deb et al. (2002) *IEEE Xplore*
Liu et al (2021) *J. Cheminf.*

DrugEx – the Workflow



DrugEx – Three interfaces

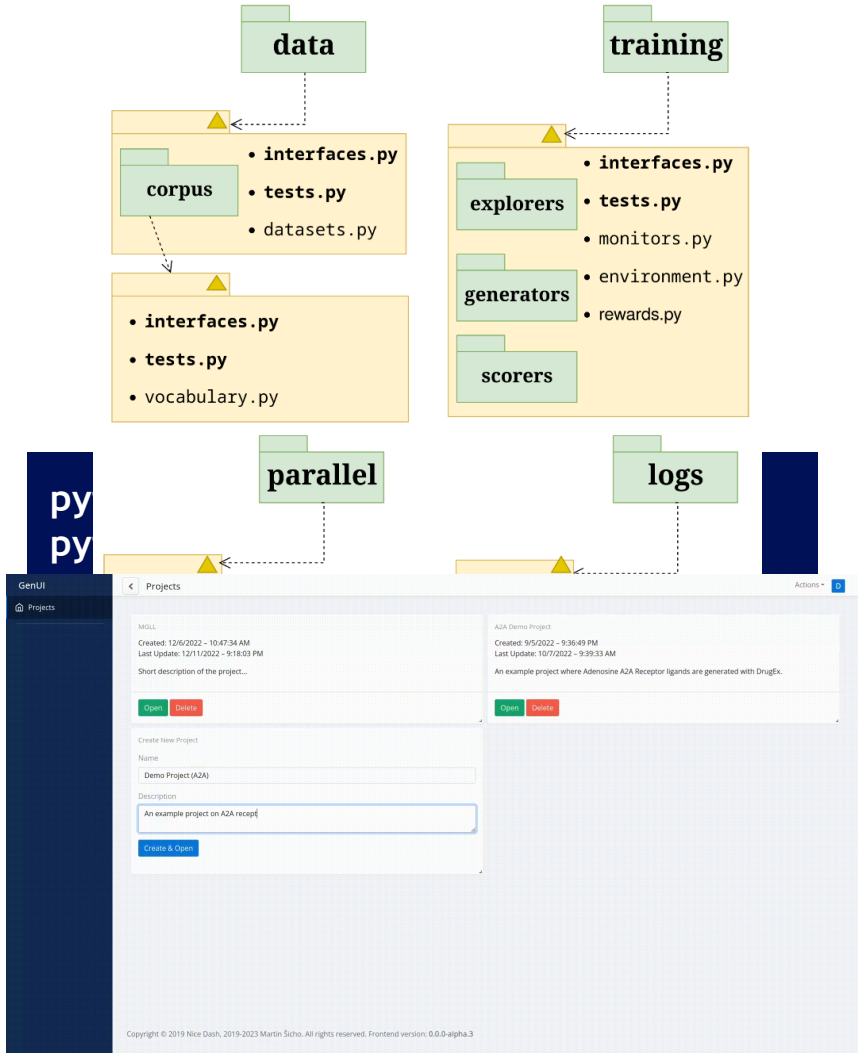
A Python package with an application programming interface (API)

@ github.com/CDDLeiden/DrugEx

Command line interface (CLI)

Graphical user interface (GenUI)

@ github.com/martin-sicho/genui-gui



Šícho et al. (2023) *J. Chem. Info. Model.*
Šícho et al. (2021) *J. Cheminfo.*

DrugEx - Getting Started

Application note

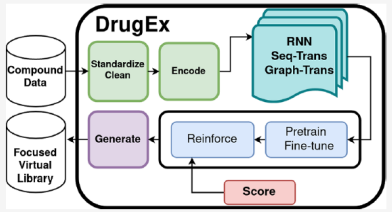
DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space

Martin Šícho, Sohvi Luukkonen, Helle W. van den Maagdenberg, Linde Schoenmaker, Olivier J. M. Béquignon, and Gerard J. P. van Westen*

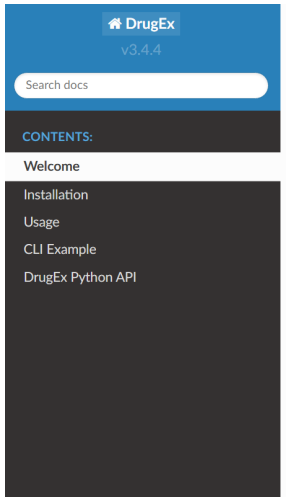
Cite This: https://doi.org/10.1021/acs.jcim.3c00434 Read Online

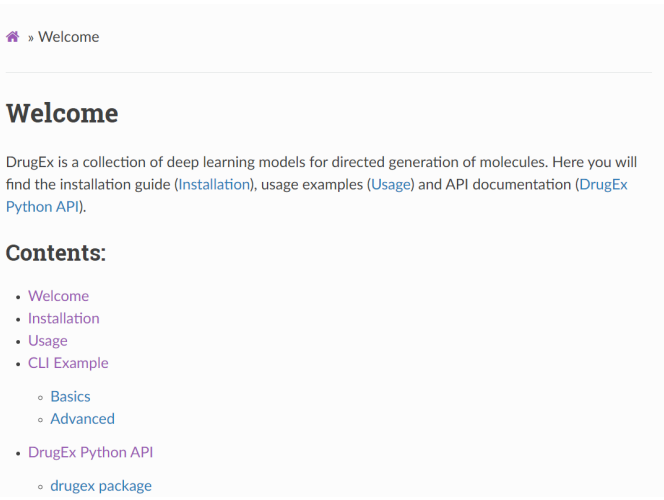
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The discovery of novel molecules with desirable properties is a classic challenge in medicinal chemistry. With the recent advancements of machine learning, there has been a surge of *de novo* drug design tools. However, few resources exist that are user-friendly as well as easily customizable. In this application note, we present the new versatile open-source software package DrugEx for multiobjective reinforcement learning. This package contains the consolidated and redesigned scripts from the prior DrugEx papers including multiple generator architectures and a variety of scoring tools and multiobjective optimization methods. It has a flexible application programming interface and can readily be used via the command line interface or the graphical user interface GenUI. The DrugEx package is publicly available at https://github.com/CDDLeiden/DrugEx.



Documentation





Tutorials (API & CLI)

Data Preparation

In this tutorial, we assume you already extracted the required data and models with the download utility as described in the README file. They should be located in the `data` directory in the current folder:

```
import os
os.listdir('data')
```

```
['models', 'logs', 'download.json', 'datasets', 'download.log']
```

We will only be preparing a fine-tuned model in this tutorial so we just need one data set that closely relates to our target of interest, which is the adenosine A2A receptor (A2AR) in this case. Data about ligands extracted from the Papyrus dataset is saved in the following folder:

```
DATASETS_PATH = 'data/data'
os.listdir(DATASETS_PATH)
```

```
['encoded', 'qsar', '.Papyrus', 'A2AR_LIGANDS.tsv']
```

Lets take a look at the data set file itself:

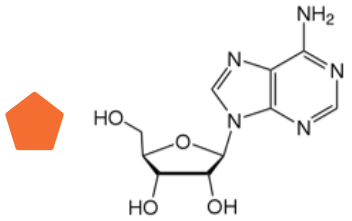
Recurrent neural network

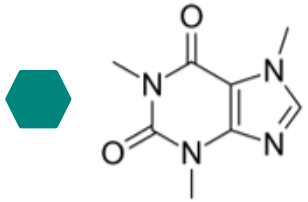
The most simple model is the RNN-based generator. This model gets the 'go' token as input and from there generates SMILES strings. Therefore, this model does not use input fragments for training or sampling. To preprocess the data for training an RNN-based generator the molecules are standardized and encoded based on the vocabulary of the pretrained model `-vf Papyrus05.5_smiles_voc.txt`, but no fragmentation is done `-nof`. To fine-tune an RNN-based generator on the A2AR set, the algorithm needs to be specified `-a rnn`. Here the generator is fine-tuned on the A2AR set and then used to generate new compounds.

```
python -m drugex.dataset -b ${BASE_DIR} -i A2AR_LIGANDS.tsv -mc SMILES -o rnn-example -nof -vf Papyrus05
python -m drugex.train -tm FT -b ${BASE_DIR} -i rnn-example -ag ${BASE_DIR}/models/pretrained/smiles-rnn
python -m drugex.generate -b ${BASE_DIR} -g rnn-example_smiles_rnn_FT -vfs Papyrus05.5_smiles_voc.txt -
```

A case for Adenosine A2B receptor selective ligands

Adenosine A2B Receptor





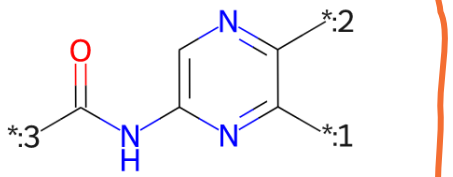
4 Adenosine receptors (ARs) – A1, A2A, A2B and A3

- Class A GPCRs
- Endogenous ligand: adenosine
- Known antagonists: xanthine-derivatives (caffeine)
- Conserved binding sites
- Situated in different organs – varied functions

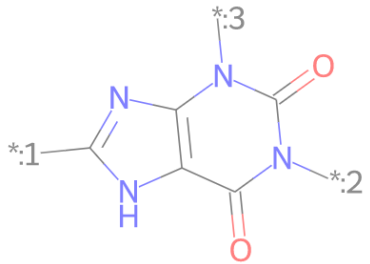
Adenosine A2B receptor ($A_{2B}R$)

- Lower affinity to adenosine than other ARs
- Activation linked to hallmarks of cancer ➢ Interesting to selectively inhibit

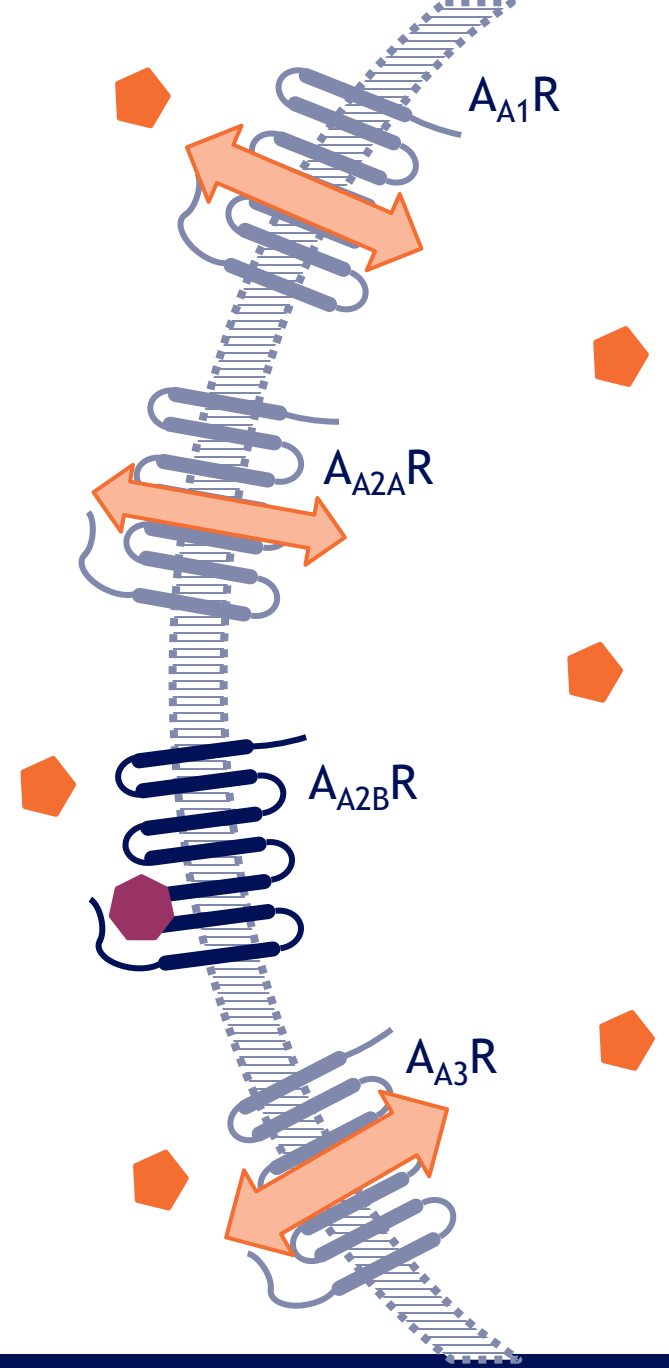
- Known antagonist scaffolds with some selectivity



Pyrazineamide



Xanthine



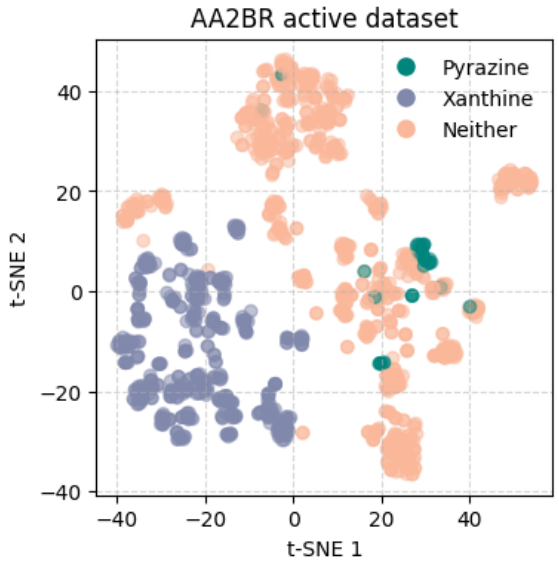
Sun and Hugan (2016) *Front. Chem.*
Eastwood et al. (2023) *ACS Med. Chem. Lett.*; Basu et al. (2017) *Eur. J. Med. Chem.*

AR data

AR dataset

5507 compounds

QSAR models



$A_{A2B}R$ active dataset

1204 compounds

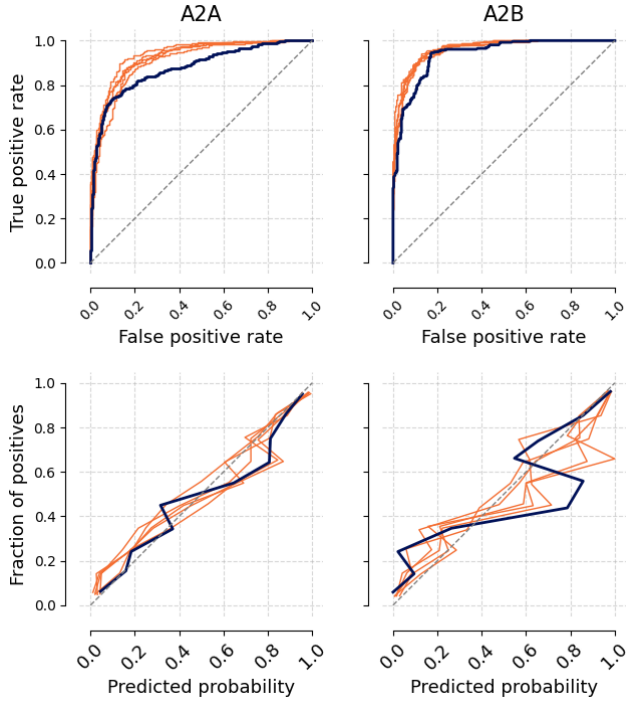
Finetuning

QSAR Models for ARs with QSPRpred

Random forest classifiers

⇒ Objectives during RL

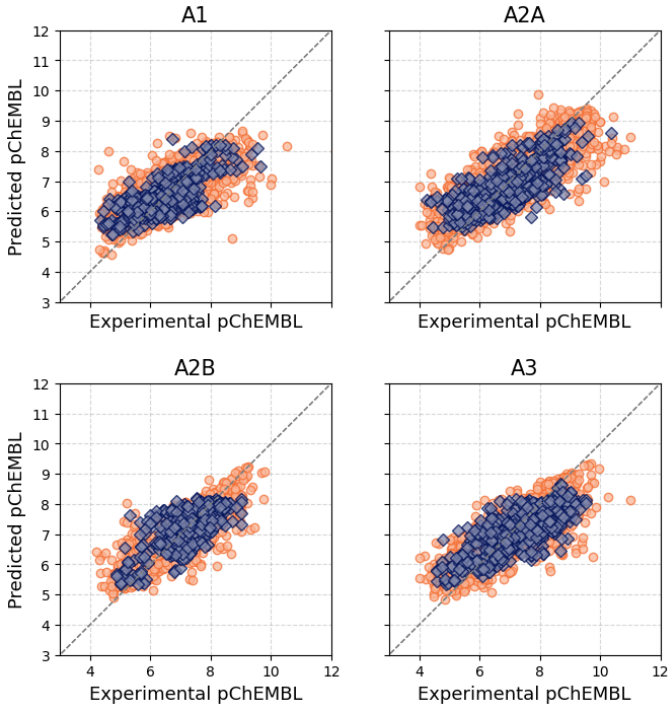
5-fold random cross-validation
Scaffold-based test set



	$A_{A2A}R$	$A_{A2B}R$
MCC	.70 (.03) .70	.78 (.01) .73
ROCAUC	.92 (.01) .88	.96 (.01) .94
MCA	.06 (.01) .06	.07 (.03) .12

Random forest regressors

⇒ Filtering of *de novo* compounds

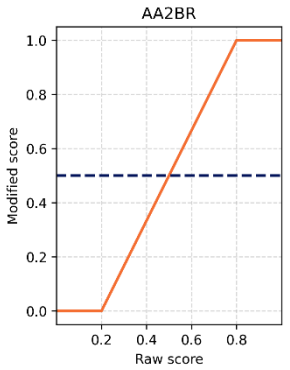


	$A_{A1}R$	$A_{A2A}R$	$A_{A2B}R$	$A_{A3}R$
RMSE	.68 (.04) .64	.68 (.02) .67	.60 (.03) .60	.67 (.04) .74
R^2	.55 (.04) .52	.69 (.01) .54	.69 (.03) .52	.66 (.03) .54
ρ	.76 (.02) .69	.84 (.01) .73	.84 (.02) .70	.82 (.02) .76

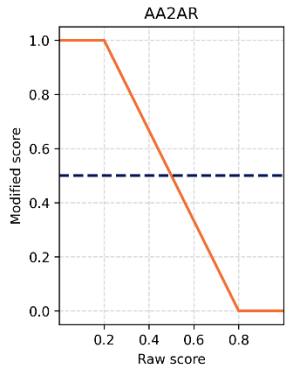
Creating the Environment

Dynamic weighted sum ⇒ Environment

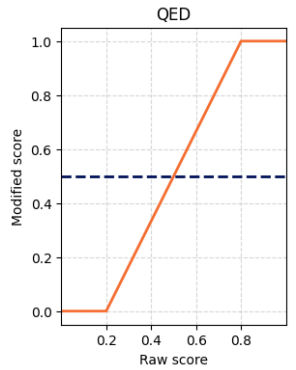
AA2BR classifier (target)



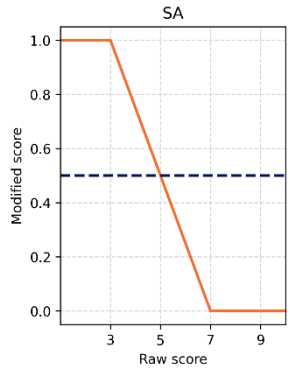
AA2AR classifier (anti-target)



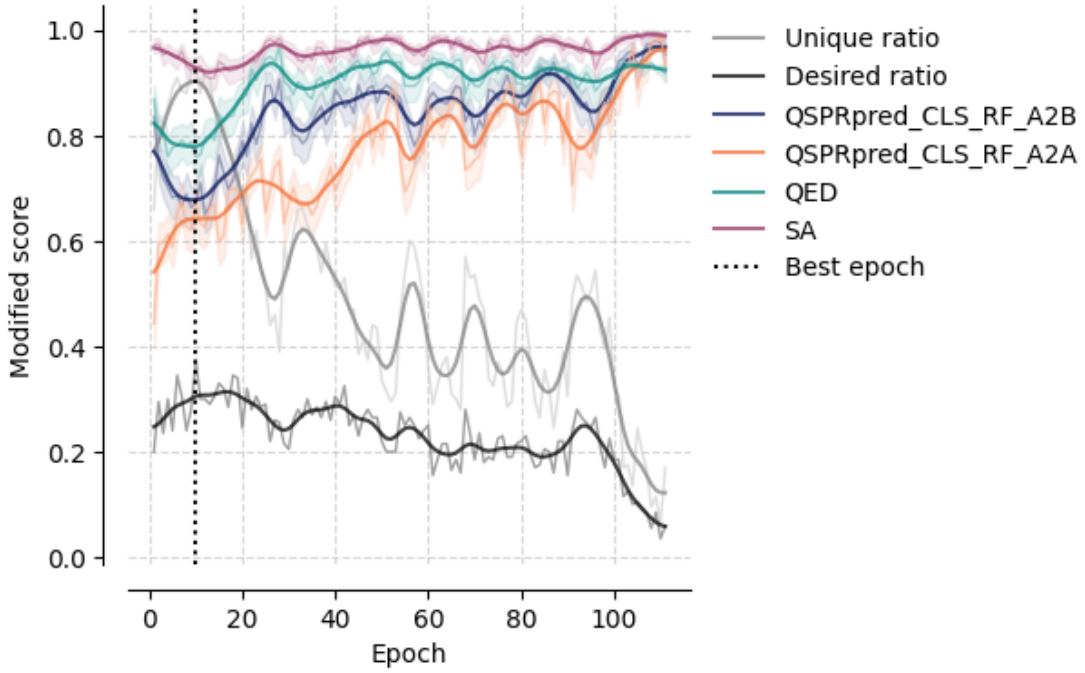
QED (drug-likeness)

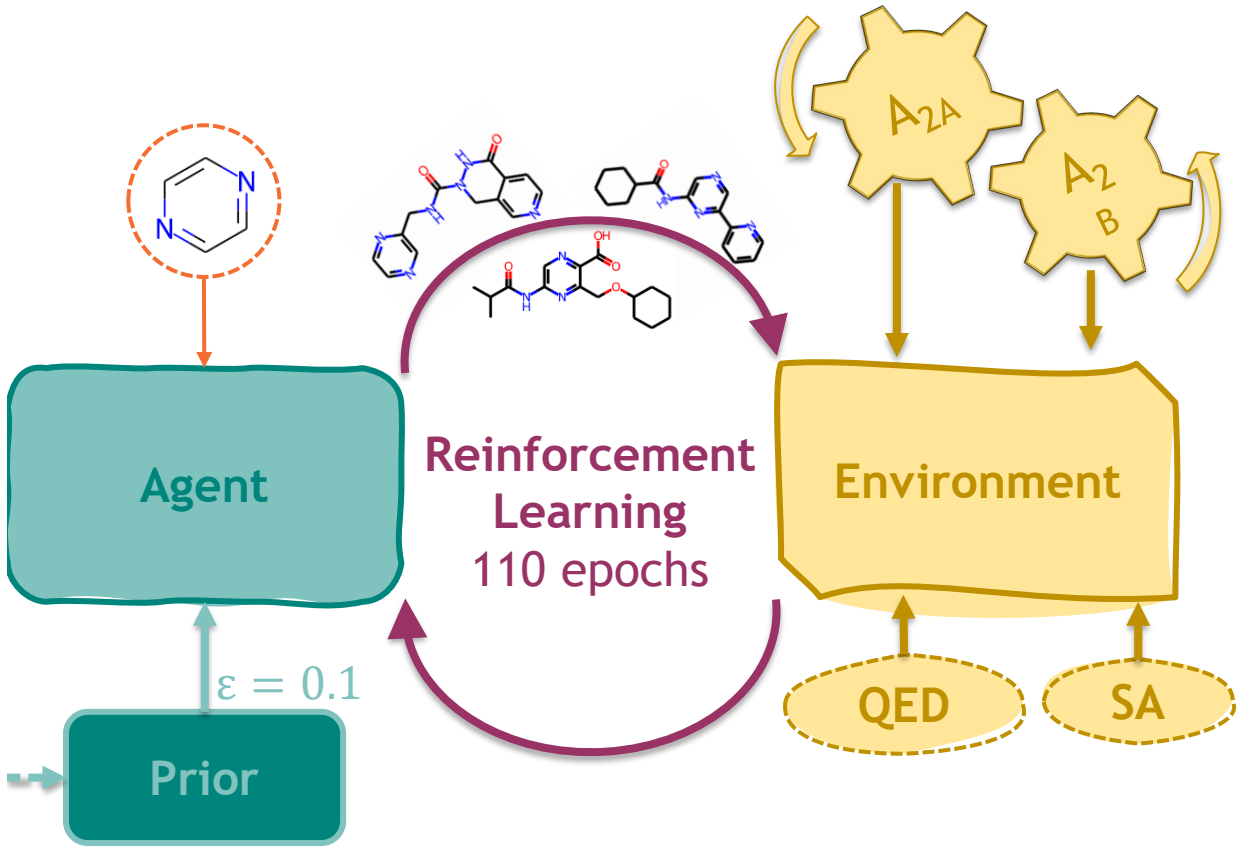


SA (complexity)



Training the Generator

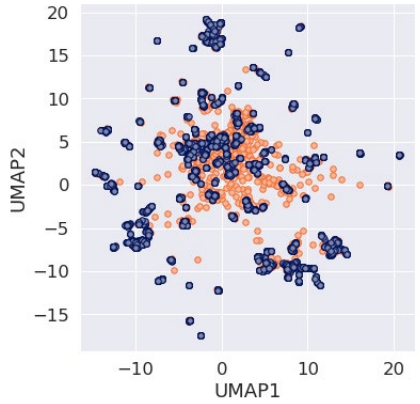




1st try* – Generation & Experimental Results

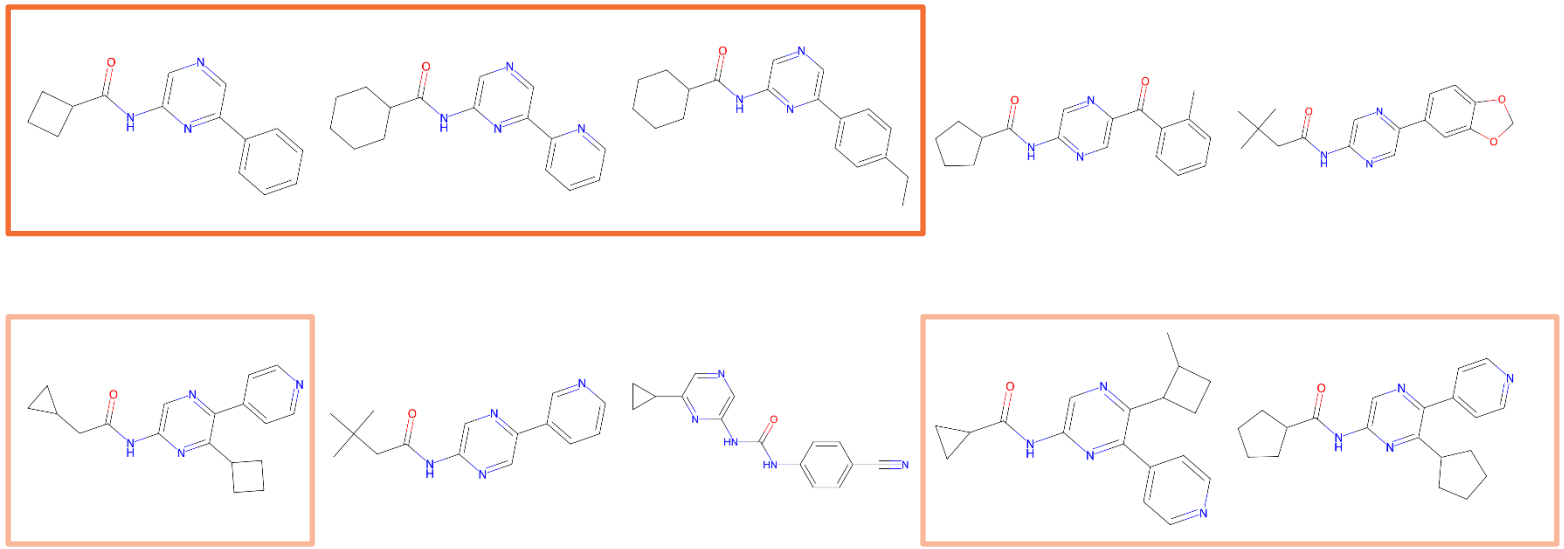
10 000 generated molecules

⇒ 3 791 novel unique desired compounds



Top 10 compounds (average score)

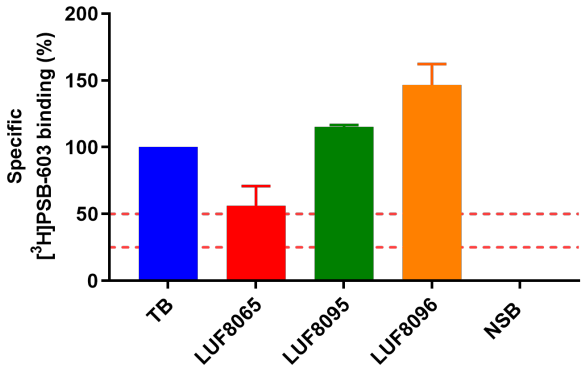
LUF8065 LUF8096 LUF8095



Top 3 compounds overall

Top 3 tri-substituted compounds

CHO-hA_{2B} at 10 µg/25µl, 1.5 nM [3H]PSB-603, 25°C ligands at 10 µM



Differences in workflow

Finetuning: AR active compounds

QSAR: less well optimised model

Extra objective: MW (200-500 Da)

Early stopping: mean score

Ligand-based Filtering

22,200 molecules saved during the RL process

2,357 novel unique desired & unsilly compounds

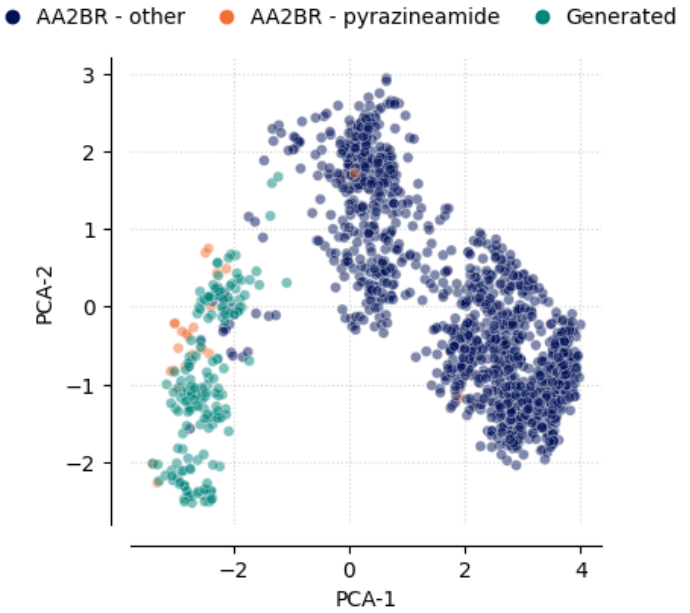
898 pyrazineamide scaffold

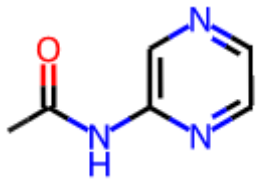
547 without macrocycles, with logP $\in [-2,3]$ & MW $\in [0,500]$

429 achiral

340 selective compounds (Δ>0.5 log units)

184 solved by AiZynthFinder



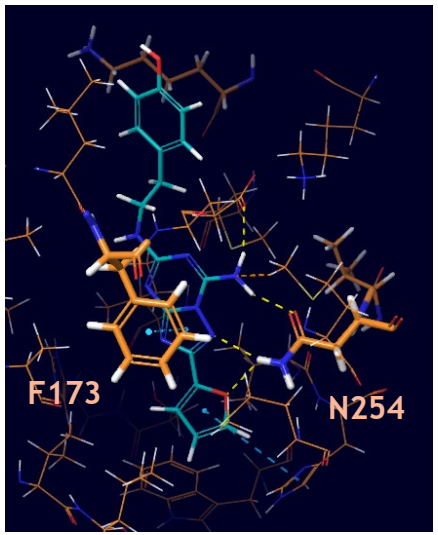


Ranking: a geometric mean between AA2BR activity and selectivity

Structure-based Filtering

Docking

- 184 compounds
- $A_{A2A}R$ – 4EIY
- $A_{A2B}R$ – AlphaFold inactive model from GPCRdb

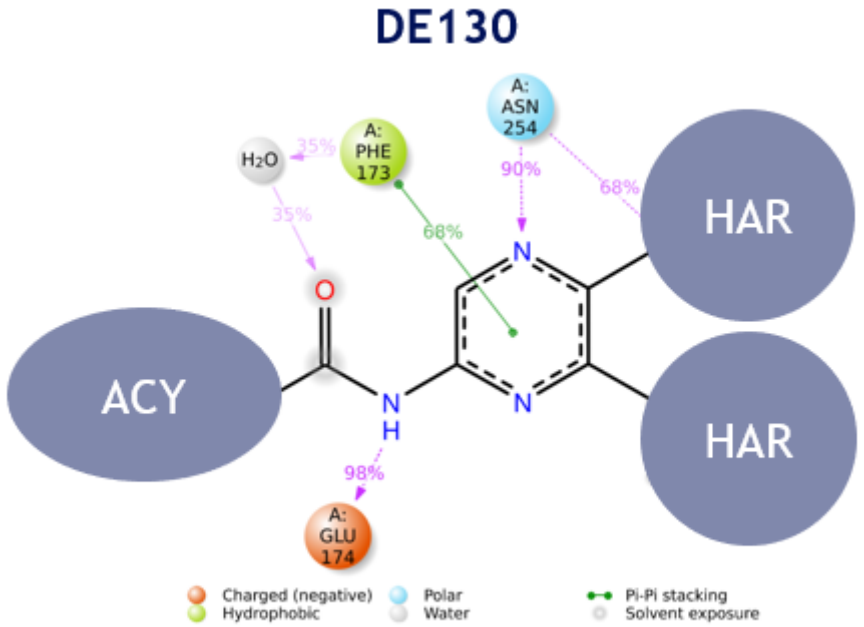


Reference compound: ZMA241385

π-stacking with Phe173 + 2 H-bonds with Asn254
π-stacking with Phe173 + 1 H-bond with Asn254
Missing π-stacking with Phe173 / H-bond with Asn254 / both

Molecular Dynamics

- 21 $A_{A2B}R$ complexes ($A_{A2A}R$: 5/11/4)



ACY – acyclic
HAR - heteroaryl

Jespers et al. (2018) *Trends Pharmacol. Sci.*

Conclusions & Perspectives

DrugEx – a production-ready open-source software library for *de novo* design of small molecules with deep learning generative models in a multi-objective reinforcement learning framework

Perspectives
- MOO with uncertainty quantification
- Many-objective optimization
- Scorers based on docking/pharmacophores

Design of A2B selective ligands
- First results were disappointing → to be smarter!
- Second try + ligand- and structure-based → better selection of compounds

Next step: continue synthesis and validation of pyrazineamine (xanthine and scaffoldless) series

Aknowledgements

DrugEx dev team

Martin Šícho, Helle van den Maadenberg,
Linde Schoenmaker, Olivier Béquignon & Xuhan Liu

CDD Leiden

Rosalie Drinkwaard, Willem Jespers
& Gerard van Westen

DDS4 team

Jerre Madern, Rongfang Lie & Daan van der Es

DrugEx: deep learning for *de novo* drug design

– A case for A2B selective ligands –

Ninth Joint Sheffield Conference on Chemoinformatics

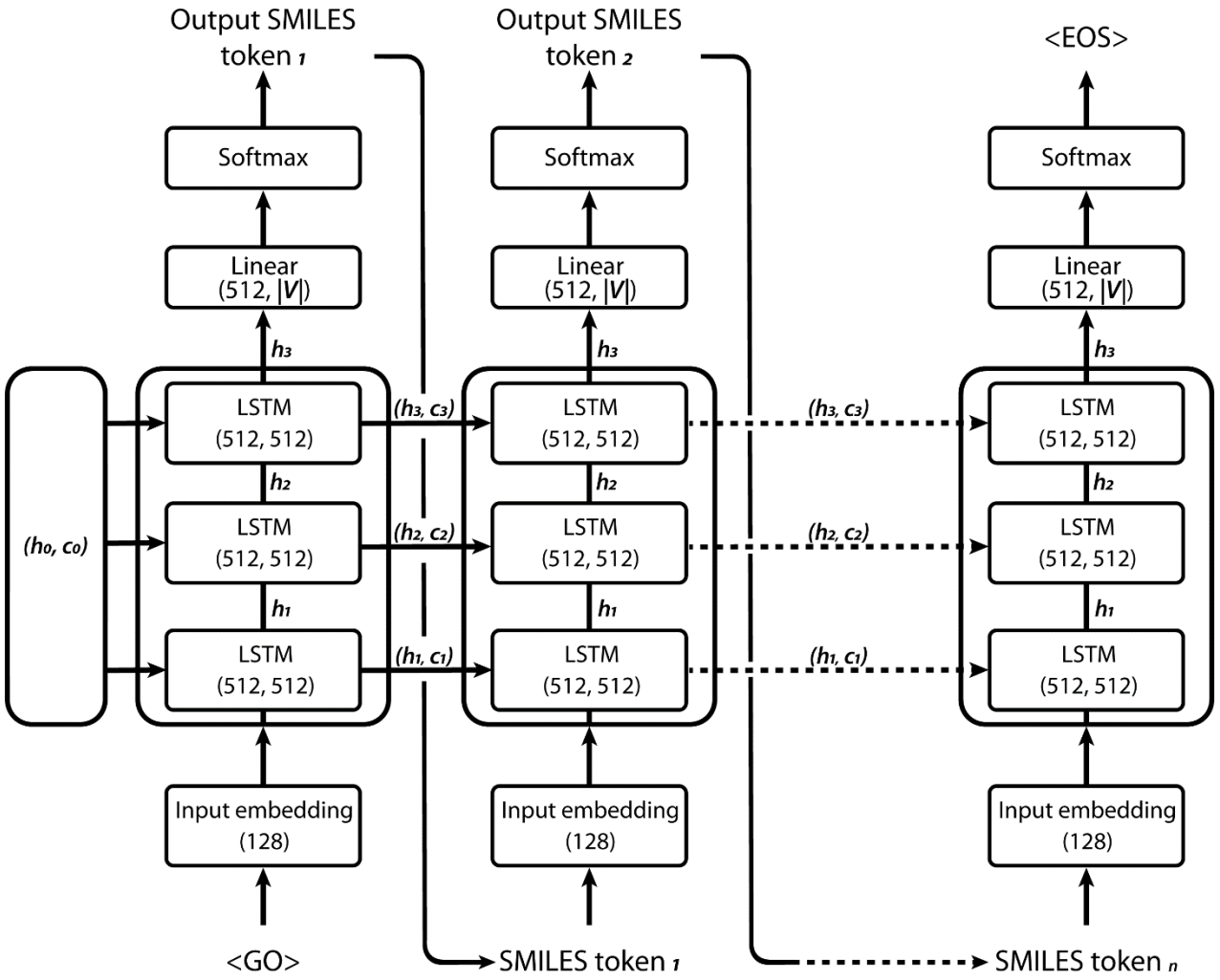
Universiteit Leiden
The Netherlands

LACDR

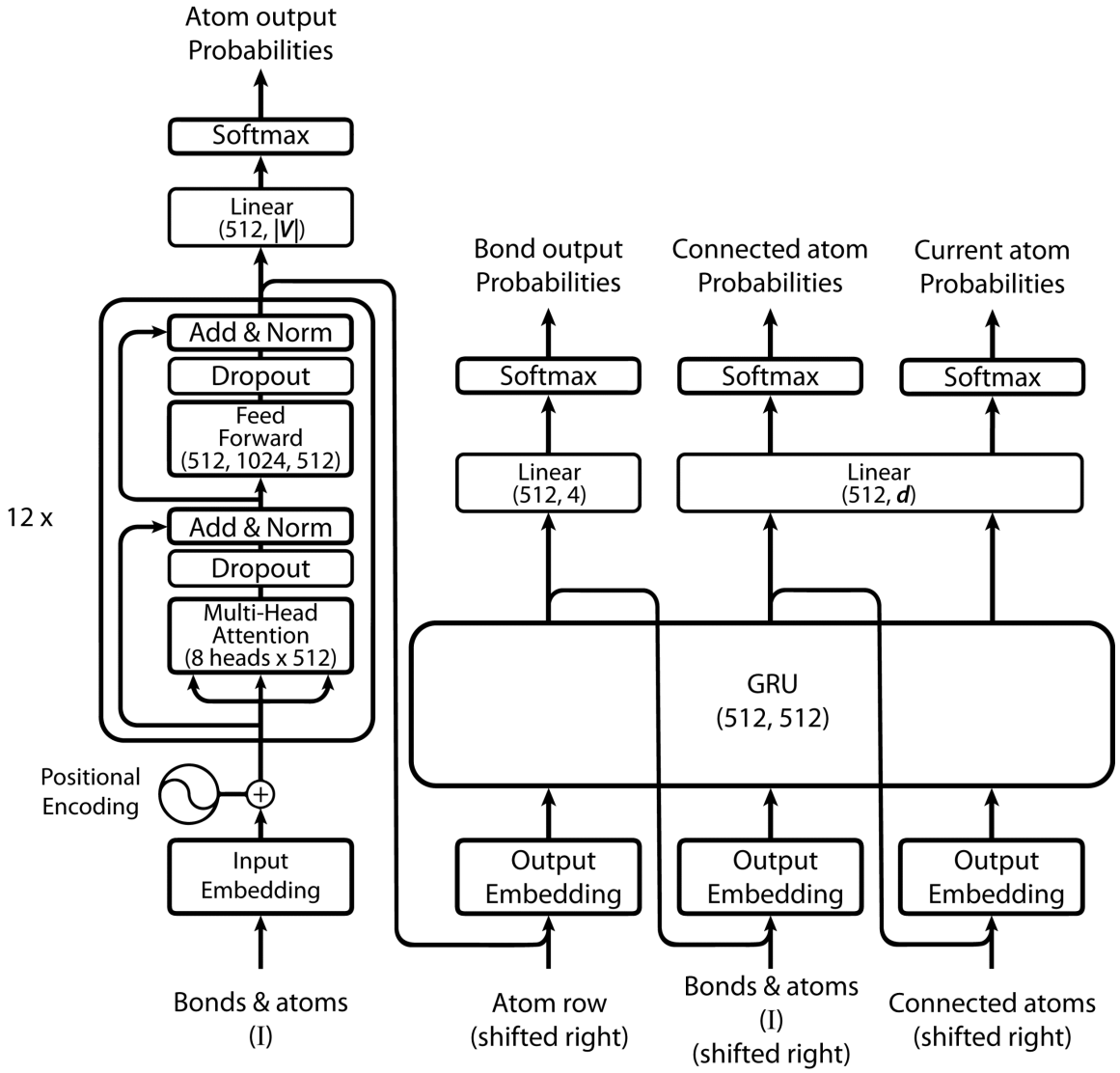
DrugEx

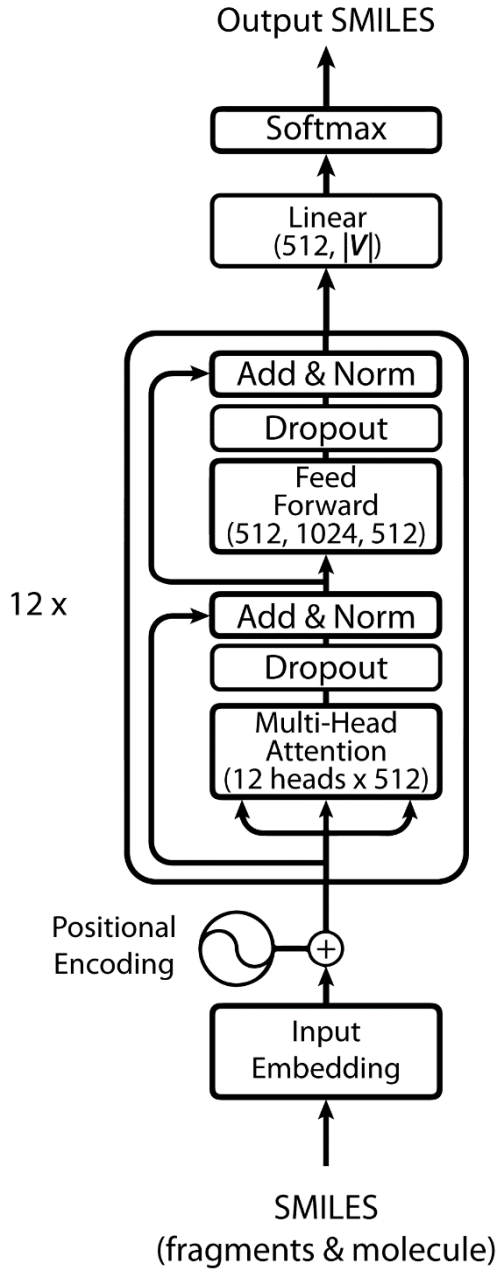
Discover the world at Leiden University

DrugEx RNN architecture



DrugEx Transformer architectures





DrugEx GraphTransformer statistics (exploration)

Table 2: the performance of the Graph Transformer with different exploration rates in the RL framework.

ε	Accuracy	Desirability	Uniqueness	Diversity
0.0	99.7%	74.6%	60.7%	0.879
0.1	99.7%	66.8%	75.0%	0.842
0.2	99.8%	61.6%	80.2%	0.879
0.3	99.7%	56.8%	89.8%	0.874
0.4	99.7%	54.8%	88.8%	0.859
0.5	99.7%	46.8%	88.5%	0.875

DrugEx Pretrained model statistics

Model Type	Training set	Fragmentation method	Validity	Accuracy	Uniqueness	Novelty	Relative sampling time	Ref.
SMILES GRU RNN	ChEMBL (v31)	-	1.000	-	0.996	0.999	0.705±0.049	7
SMILES GRU RNN	Papyrus (v05.5)	-	1.000	-	0.992	0.999	0.706±0.052	8
SMILES LSTM RNN	ChEMBL (v27)	-	0.999	-	0.600	0.865	1.000±0.000	9
SMILES LSTM RNN	ChEMBL (v31)	-	1.000	-	0.994	0.999	0.470±0.038	10
SMILES LSTM RNN	Papyrus (v05.5)	-	1.000	-	0.988	0.998	0.474±0.050	11
SMILES transformer	Papyrus (v05.5)	BRICS	0.947	0.622	0.591	0.995	86.628±50.843	12
SMILES transformer	Papyrus (v05.5)	RECAP	0.963	0.675	0.649	0.996	86.376±50.629	13
Graph transformer	ChEMBL (v27)	BRICS	1.000	0.796	0.791	1.000	23.292±10.249	14
Graph transformer	ChEMBL (v31)	BRICS	1.000	0.786	0.775	1.000	25.253±10.373	15
Graph transformer	Papyrus (v05.5)	BRICS	1.000	0.762	0.751	1.000	24.694±10.270	16
Graph transformer	Papyrus (v05.5)	RECAP	1.000	0.814	0.810	1.000	24.843±10.378	17

DrugEx Pretrained model distributions