

# De novo drug design

Guzel Minibaeva

Ph.D. student

Institute of Molecular and Translational Medicine  
Faculty of Medicine and Dentistry  
Palacky University

# Size of explored and enumerated chemical space

## real datasets



~ 160 M compounds



~ 105 M compounds

Commercial



~ 102 M compounds

Free

## ZINC

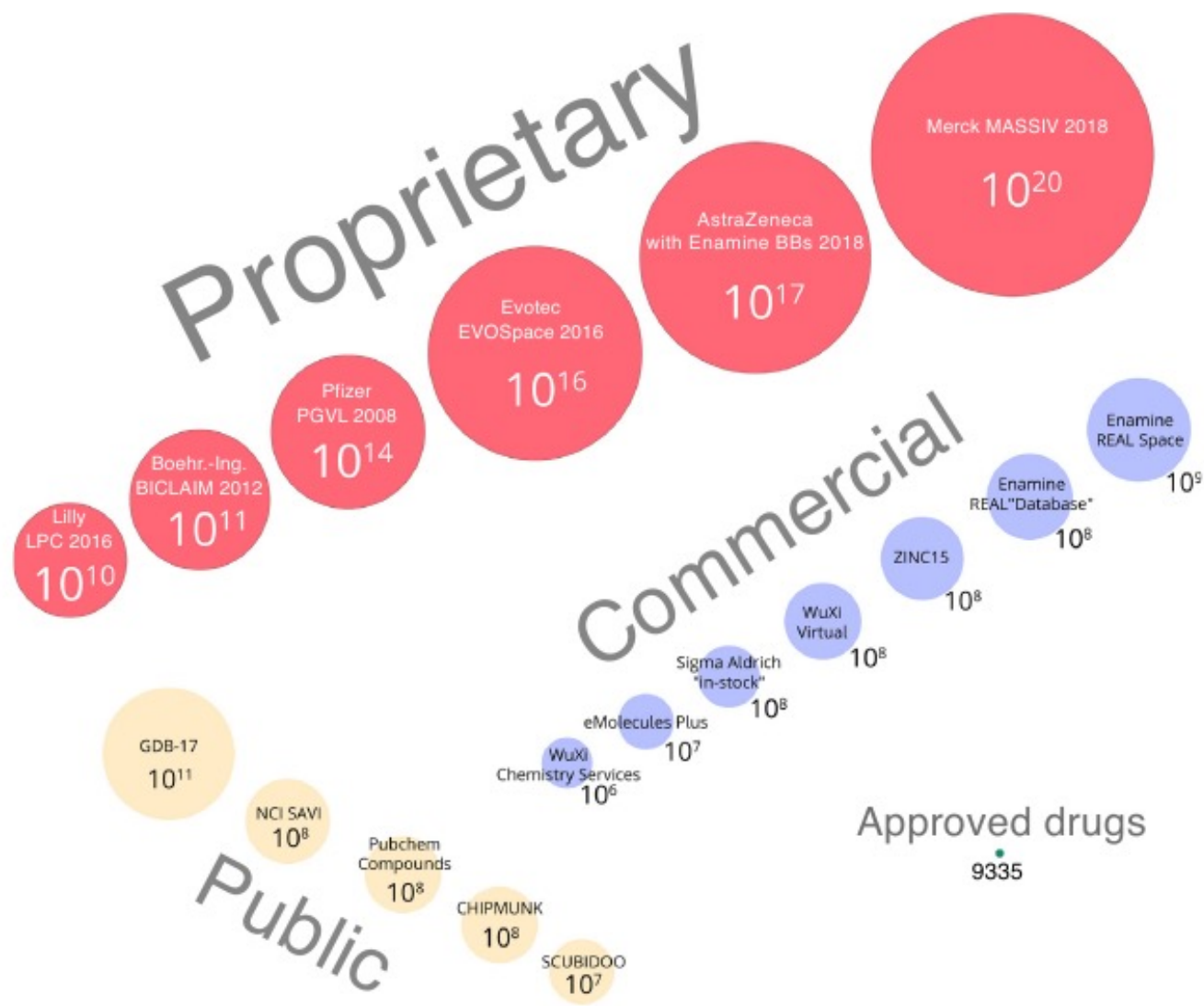
up to 1 B commercially available compounds

## virtually enumerated dataset

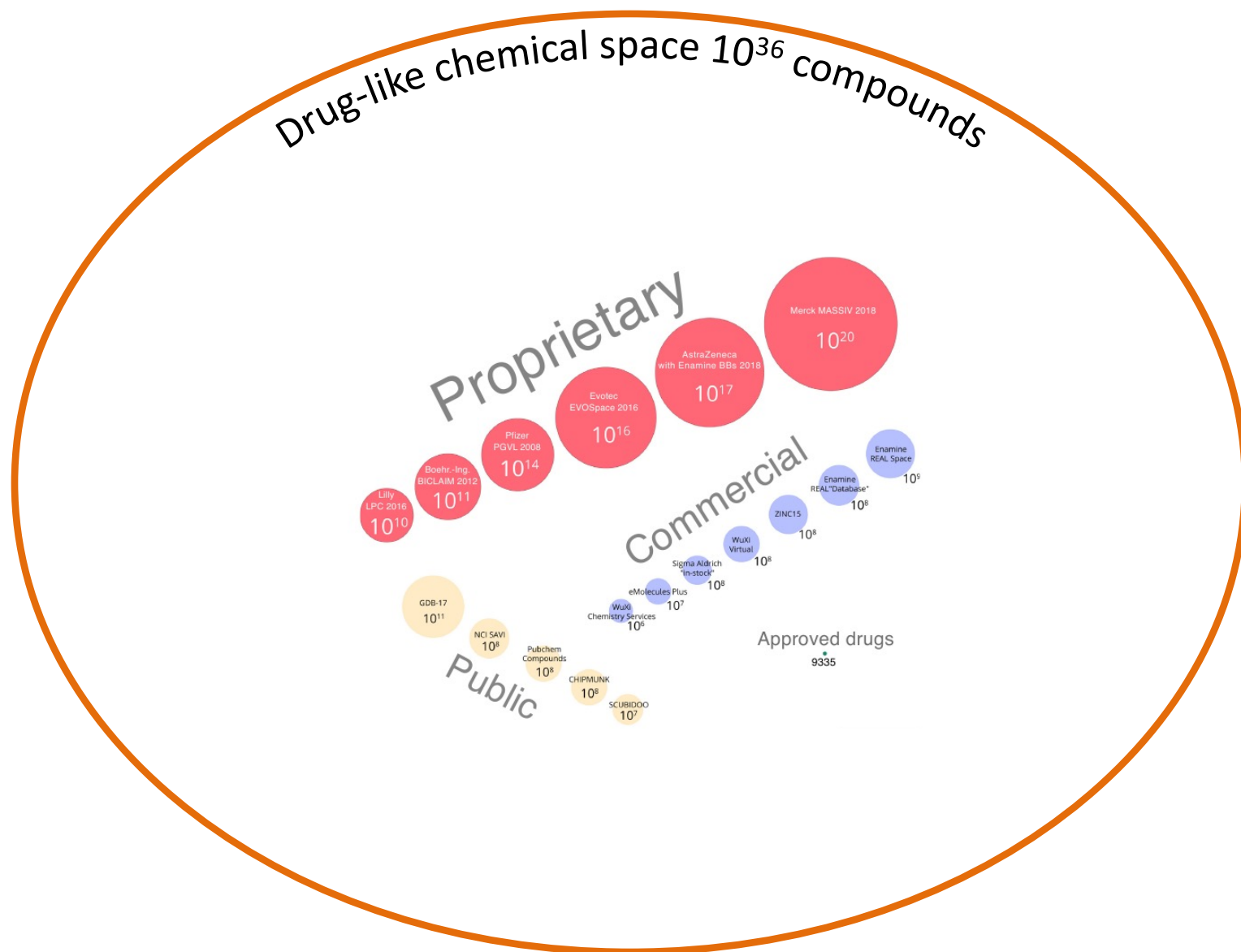
## GDB-17

166 B compounds =  $1.66 \times 10^{11}$

# Size of explored and enumerated chemical space



# Size of explored and enumerated chemical space

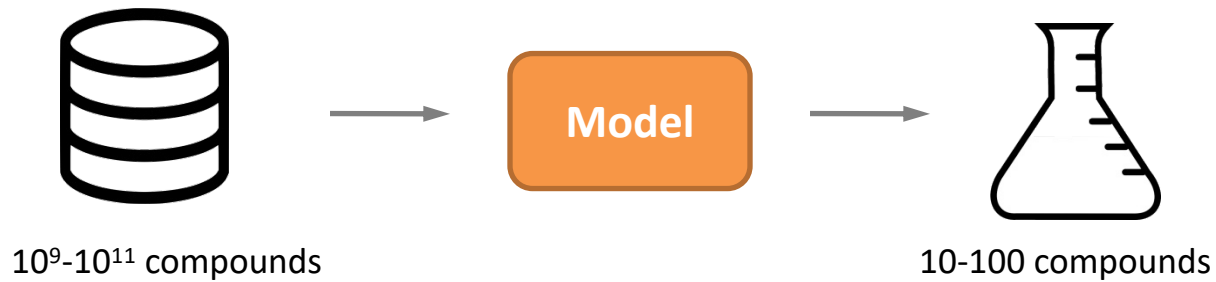


Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., Estimation of the size of drug-like chemical space based on GDB-17 data.

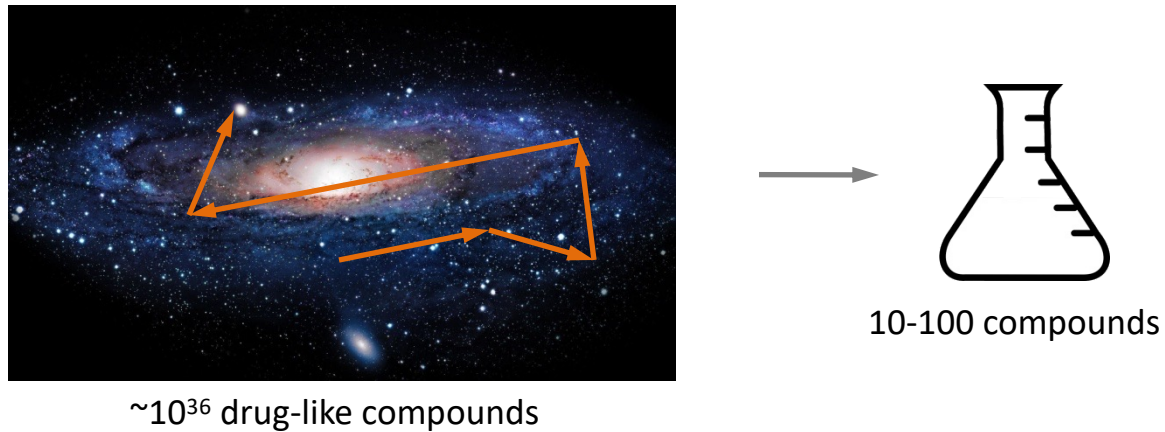
*Journal of Computer-Aided Molecular Design* **2013**, 27, 675-679. (<http://dx.doi.org/10.1007/s10822-013-9672-4>)

# Virtual screening vs. de novo design

## Virtual screening



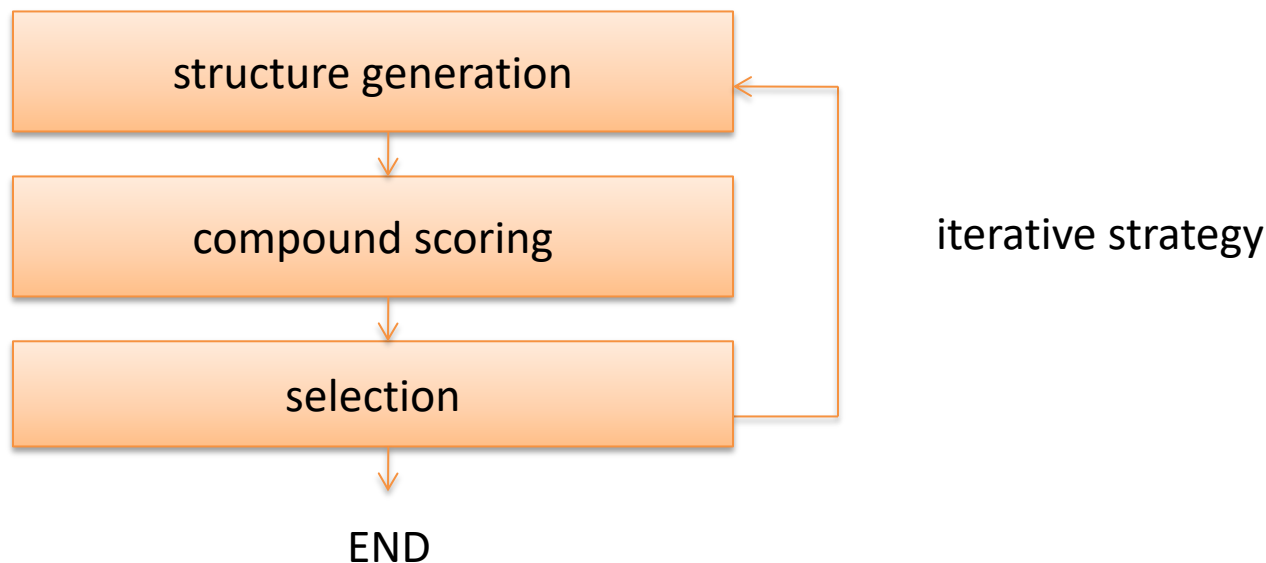
## De novo design



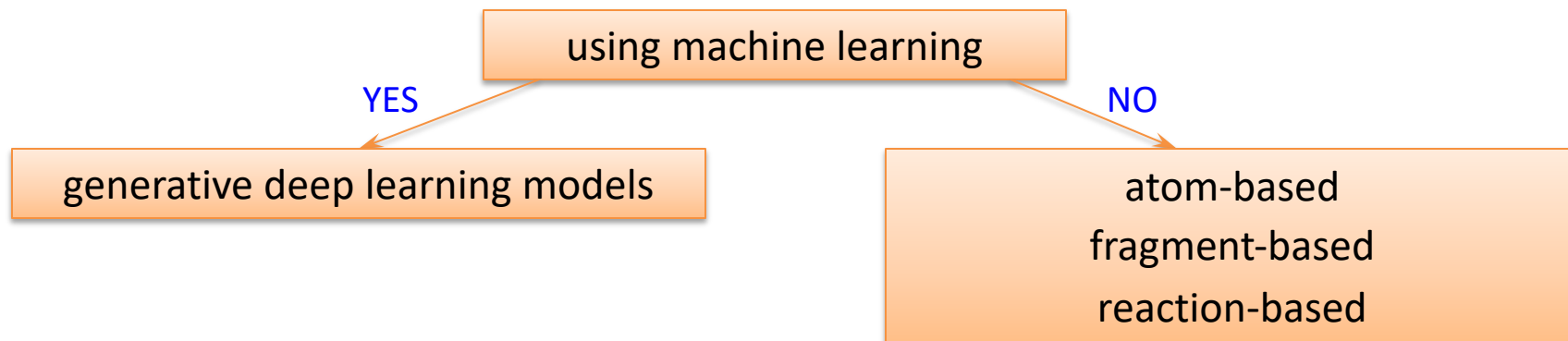
Model

# Iterative workflow of de novo design

1. **Structure generation** - how to create/assembly new structures
2. **Compound scoring** - how to estimate/predict a property of a compound
3. **Search strategy** - how to find compounds with optimal properties



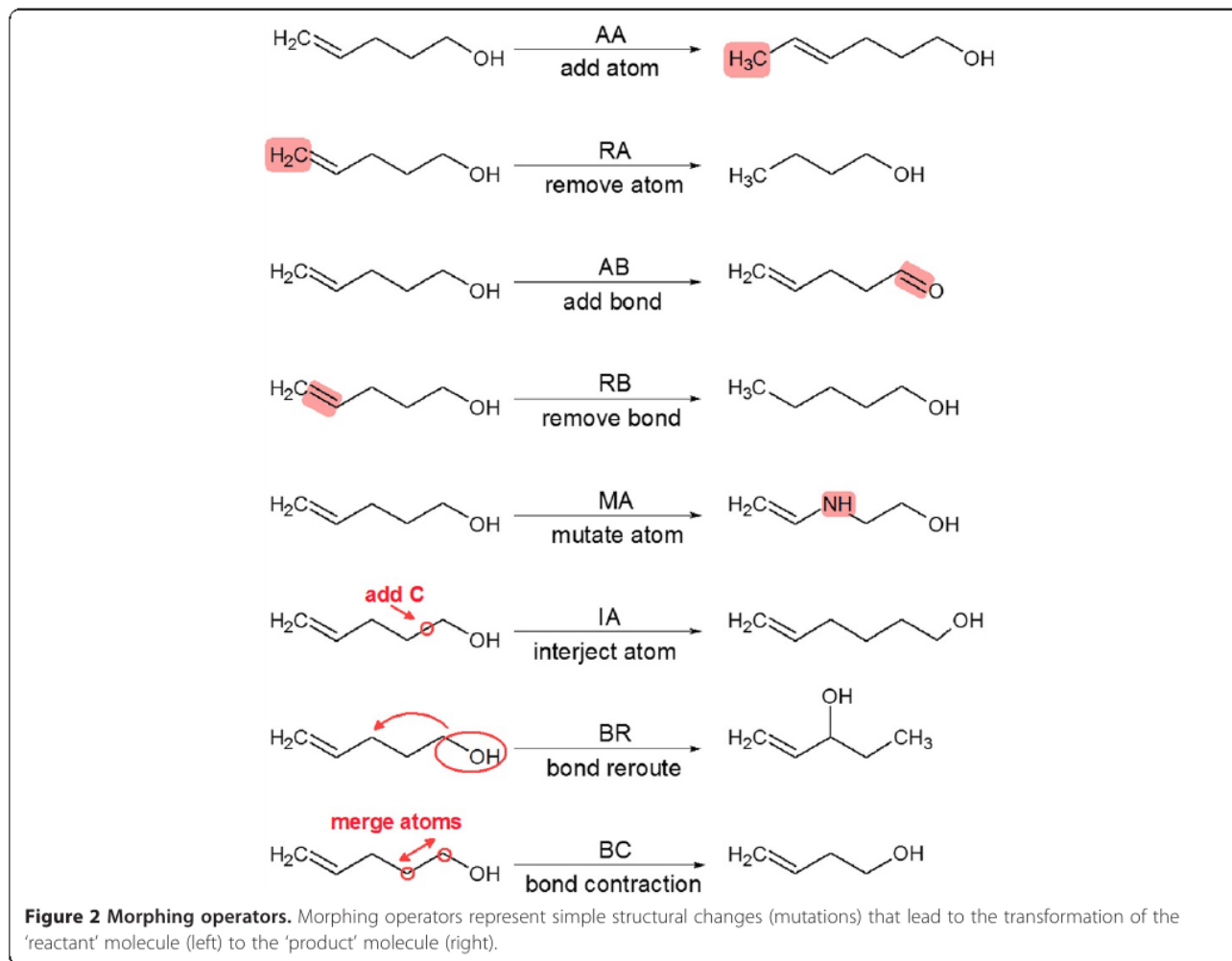
# De novo structure generation



- **atom-based** - uses simple rules like add/change/remove atom/bond to perturb structures
- **fragment-based** - uses fragment library to create structures
- **reaction-based** - uses a set of reaction rules and a library of reactants

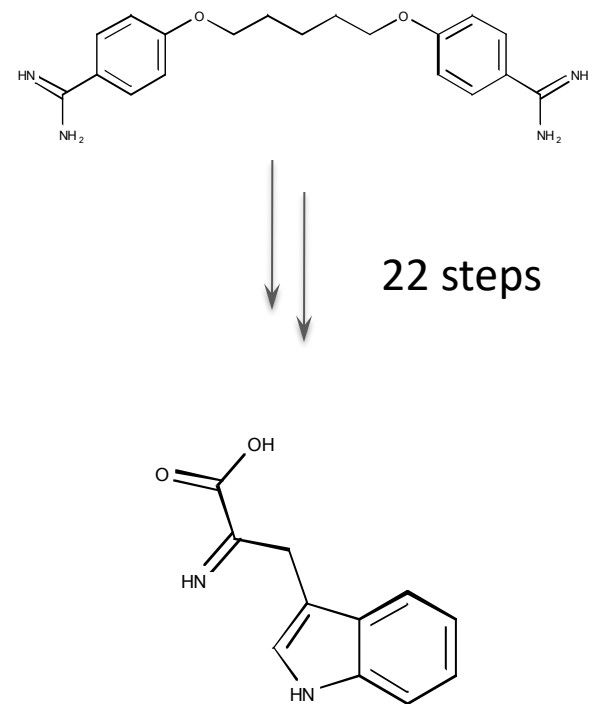
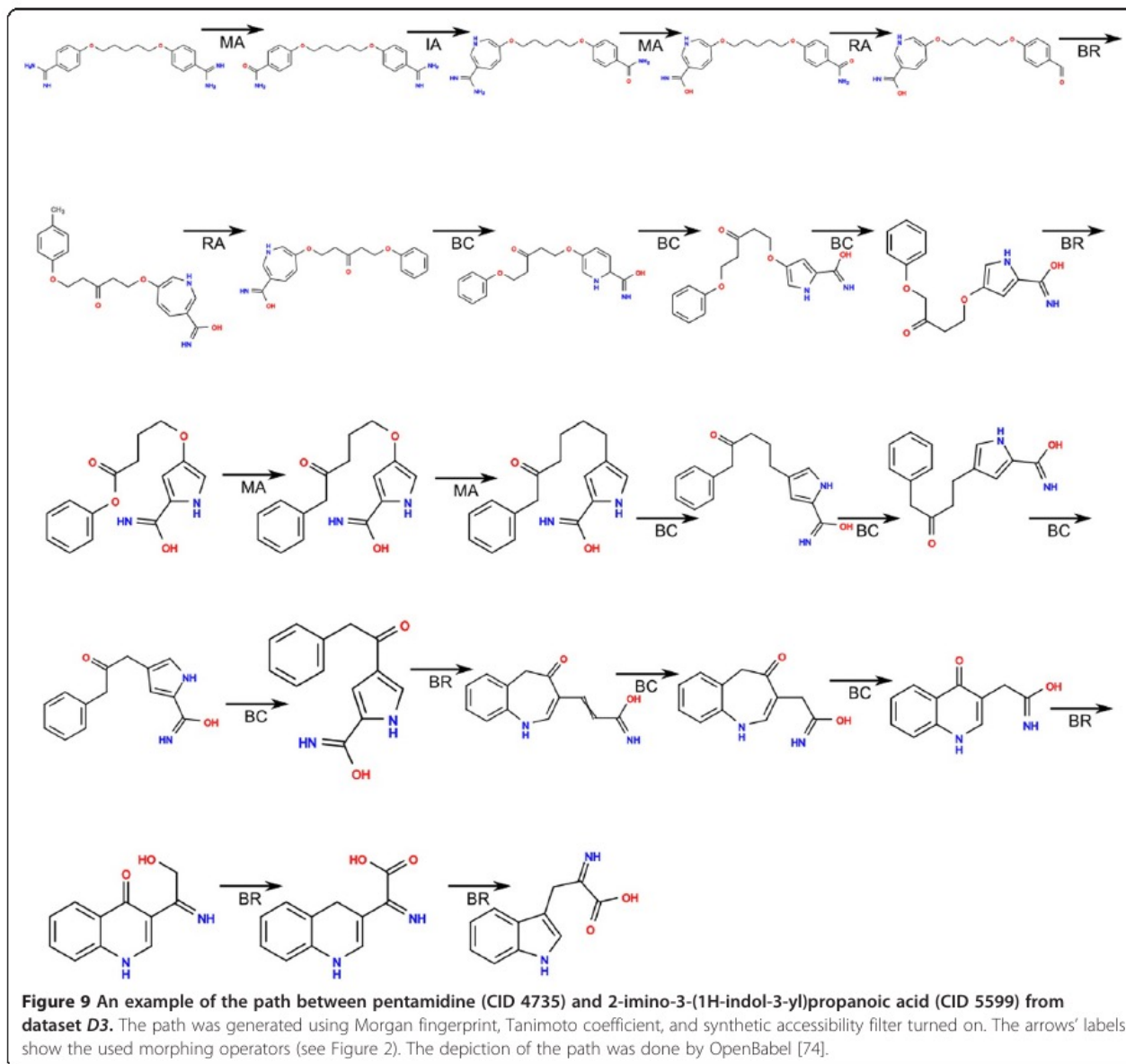
# Atom-based structure generation

## Molpher





# Atom-based structure generation



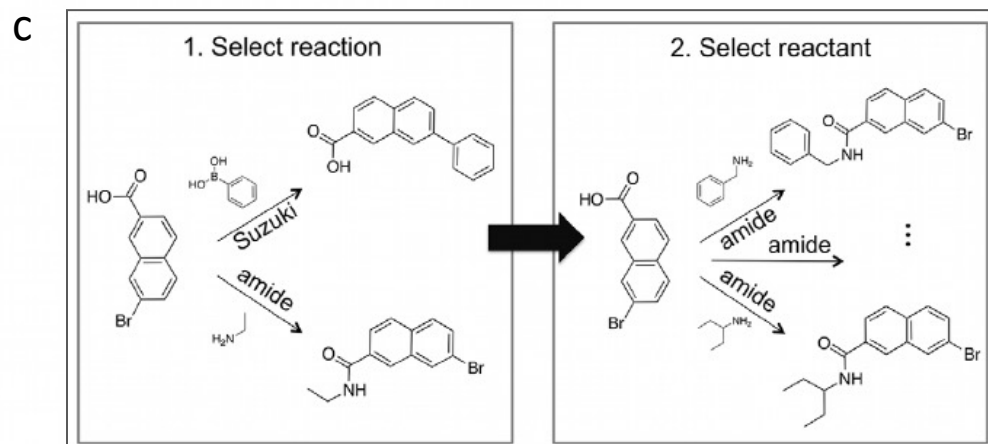
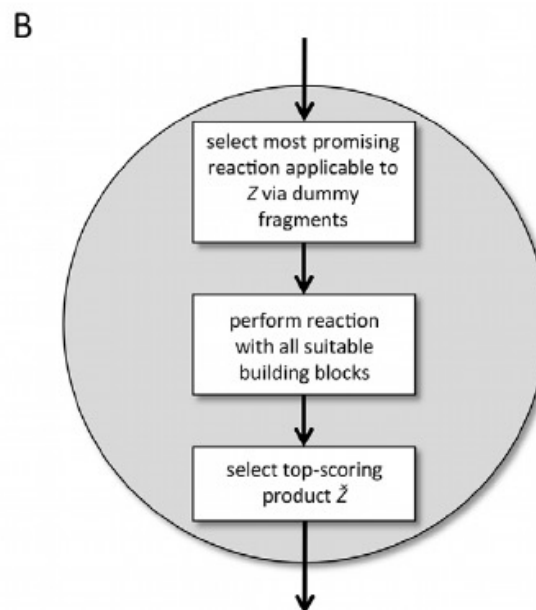
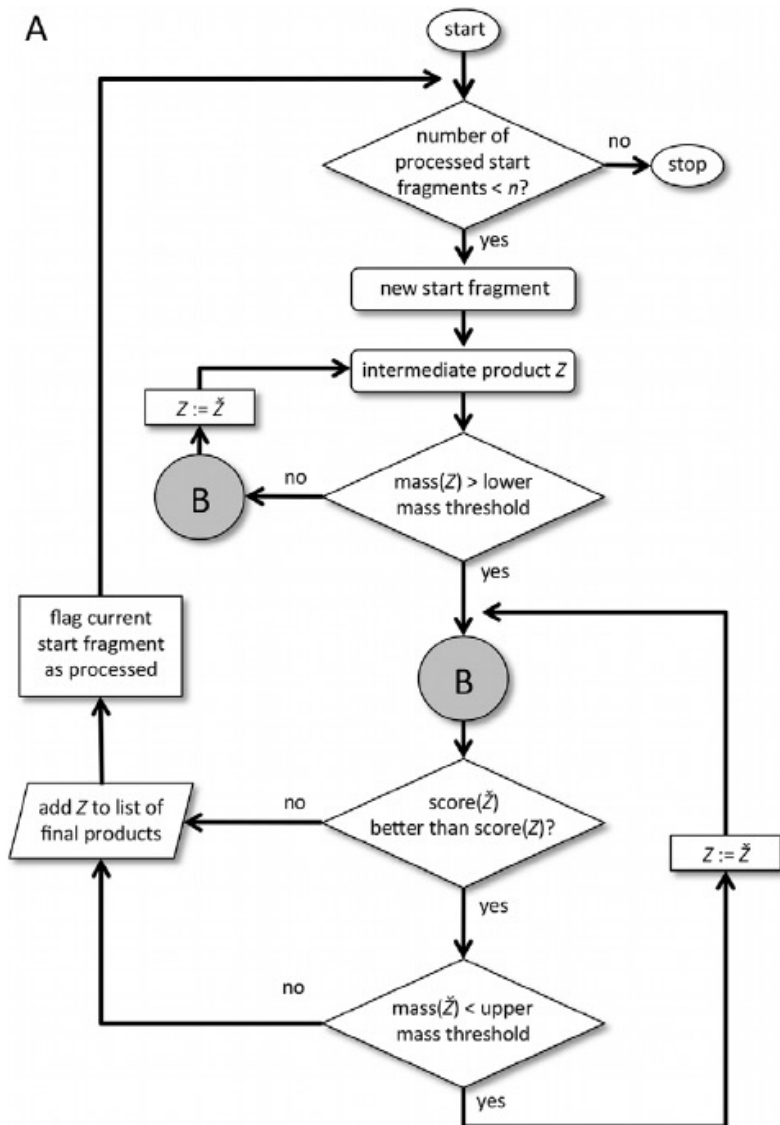
# Atom-based structure generation

parameters	atom-based
exhaustiveness of chemical space search	++++ very small steps; more suitable for systematic exploration of local chemical space
structure novelty	+++*
structure diversity	+++*
chemically valid structures	-
synthetically feasible	---
combinatorial explosion / time consuming	---

atom-based  $\approx$  *ab initio*

# Reaction-based structure generation

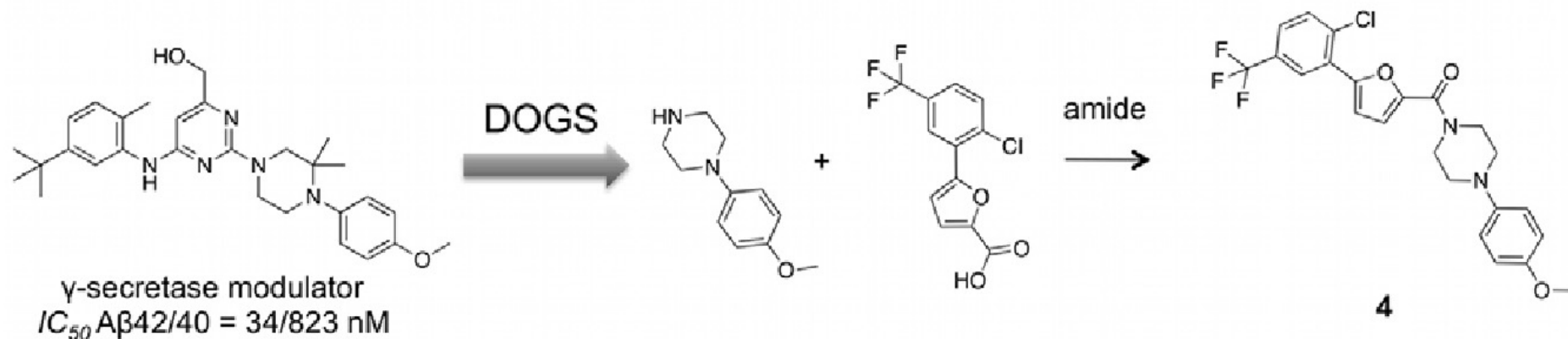
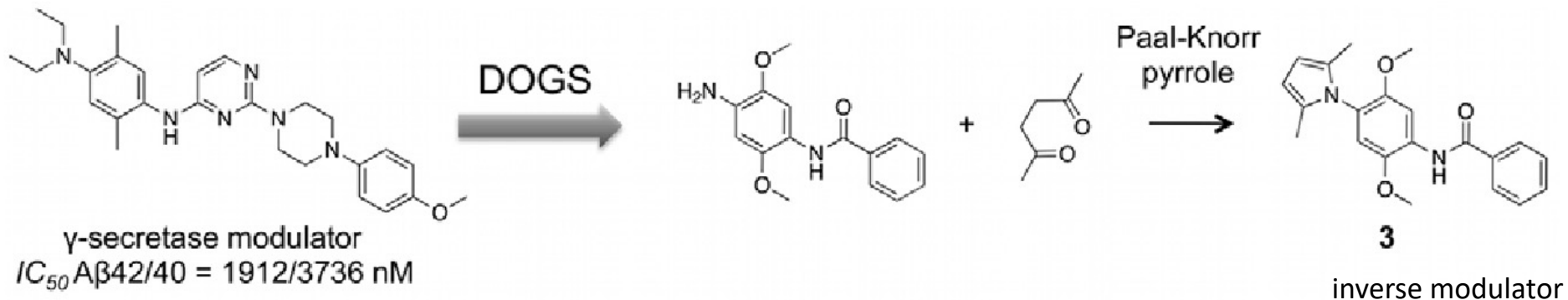
## DOGS



# Reaction-based structure generation

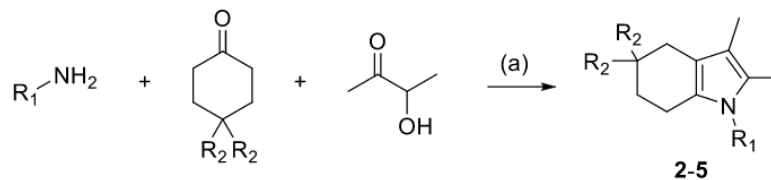
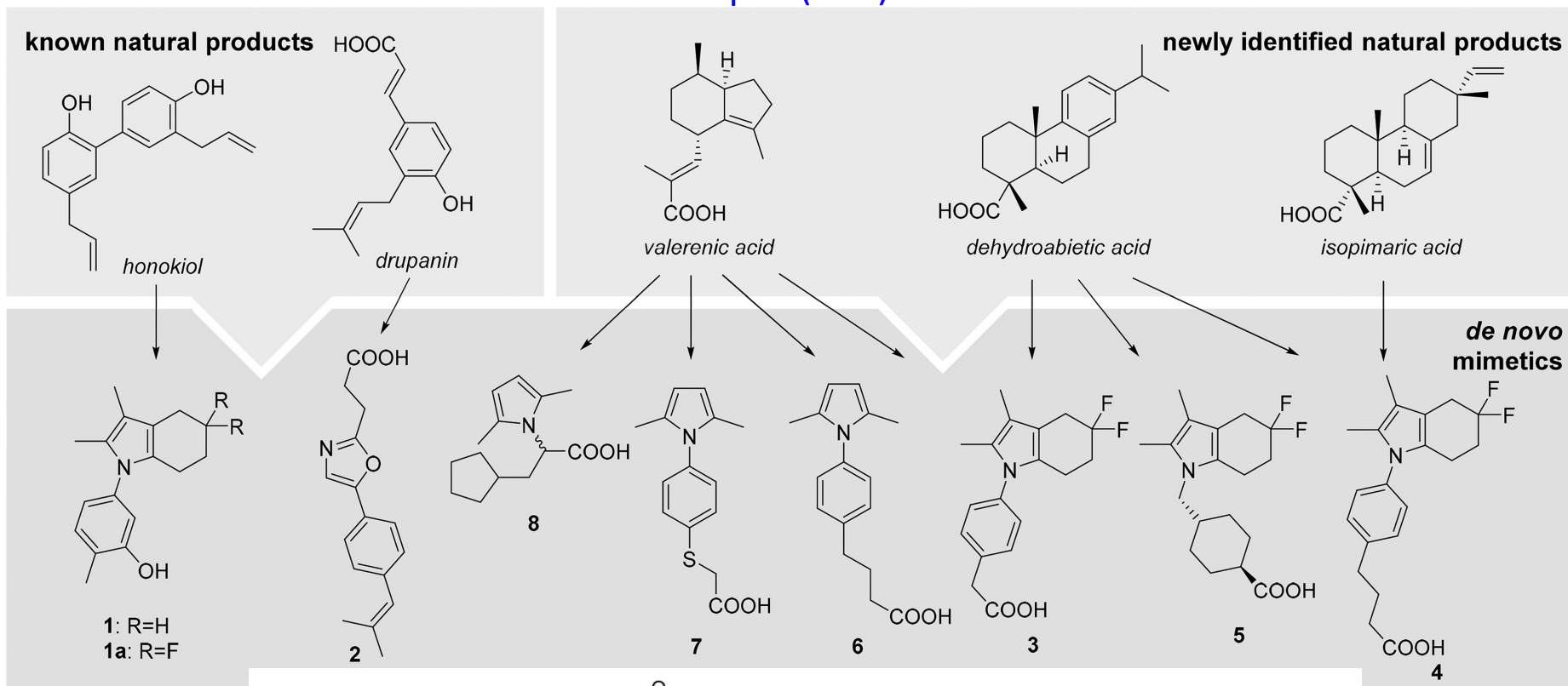
## DOGS

### $\gamma$ -secretase modulators



# Reaction-based structure generation

## Retinoid X Receptor(RXR) Modulators



isopimaric acid  
dehydroabietic acid  
valerenic acid  
sclareol  
conocarpan

**Supporting figure 5:** Synthesis of de novo mimetics **1a** and **3-8**. Reagents and conditions: (a) EtOH, HOAc,  $\mu w$ , 100°C, 3-6 h, 43-78%; (b) montmorillonite K10,  $\mu w$ , 90°C, 30 min, 41-85%.

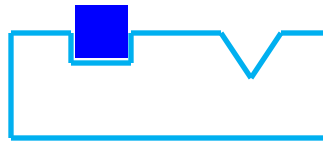
# Reaction-based structure generation

	reaction-based
exhaustiveness of chemical space search	+ depends on reactant library and reaction rules; only grow molecules
structure novelty	+
structure diversity	+
chemically valid structures	+++
synthetically feasible	+++
combinatorial explosion / time consuming	+++

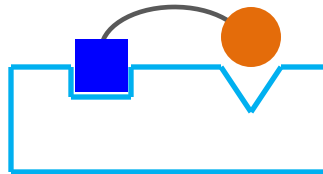
reaction-based  $\approx$  empirical

# Fragment-based structure generation

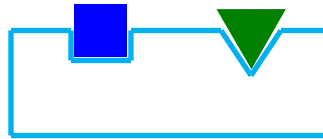
GROW



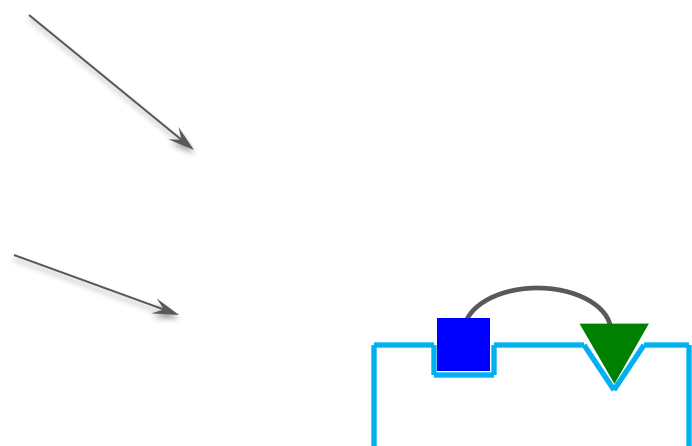
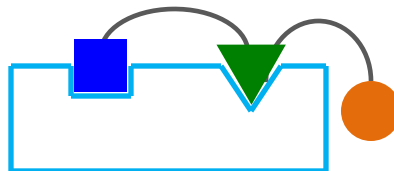
MUTATE



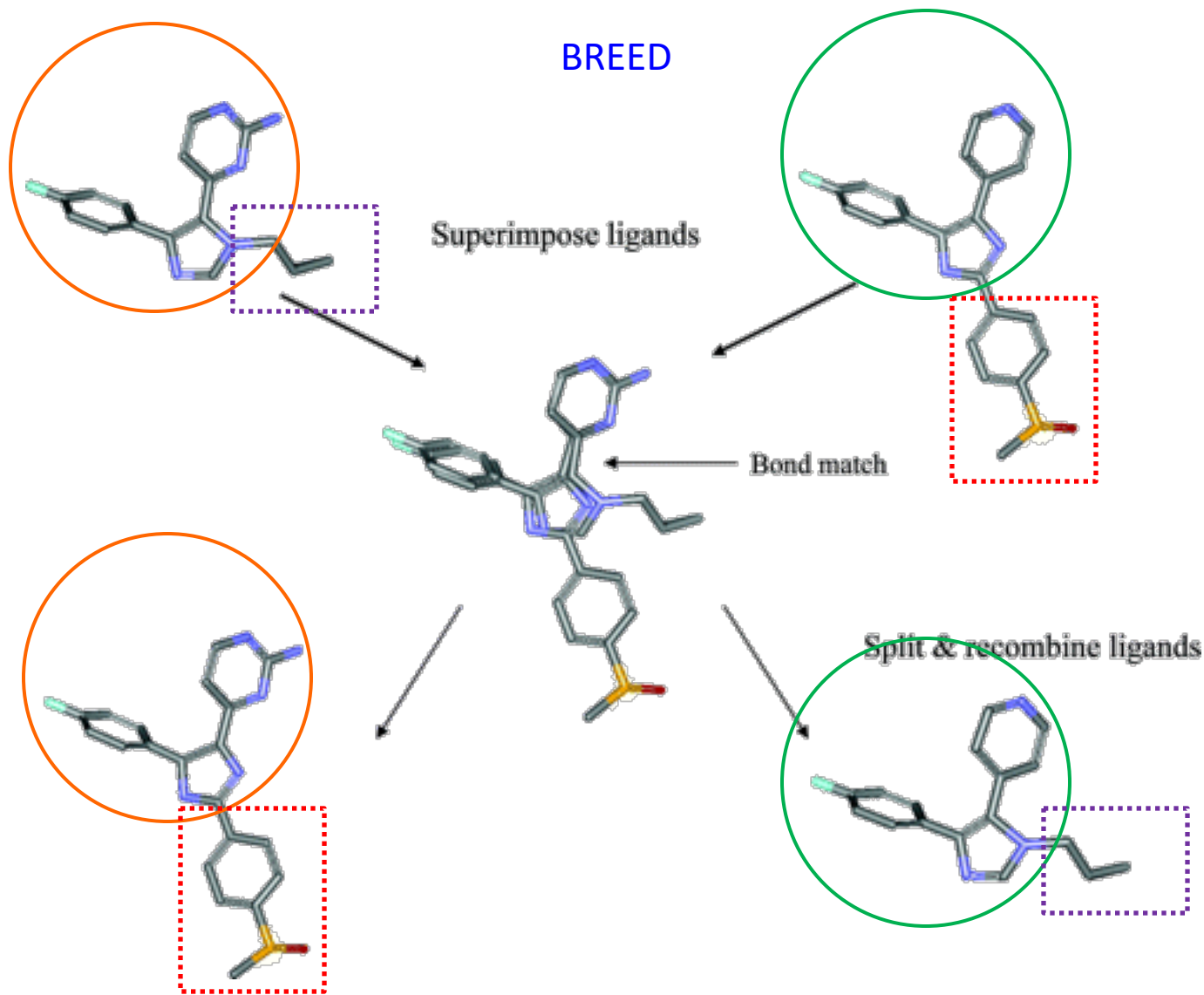
LINK



REDUCE



# Fragment-based structure generation

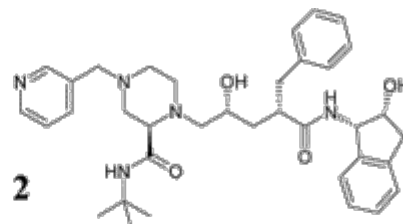
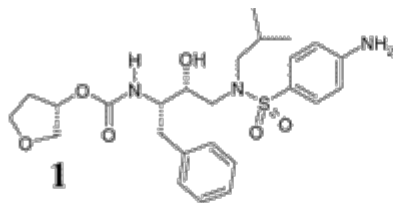




# Fragment-based structure generation

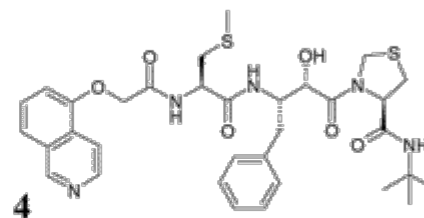
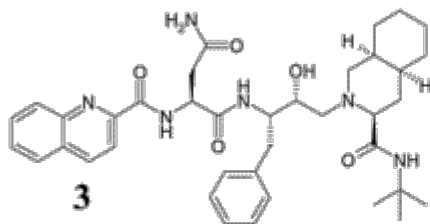
## BREED: HIV-1 protease inhibitors

$K_i = 0.4\text{-}0.6\text{ nM}$



$K_d = 1.1\text{ nM}$

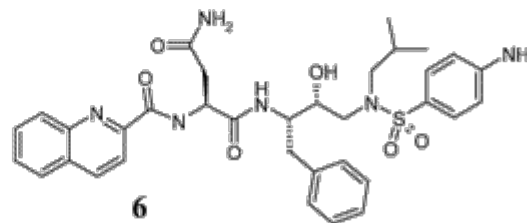
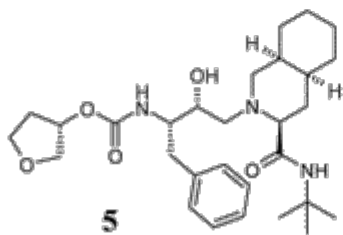
$K_i = 1.7\text{ nM}$



$K_d = 0.3\text{ nM}$

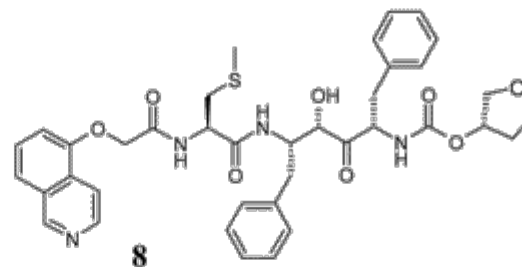
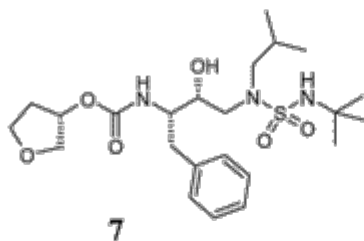
known  
designed

$IC_{50} = 160\text{ nM}$



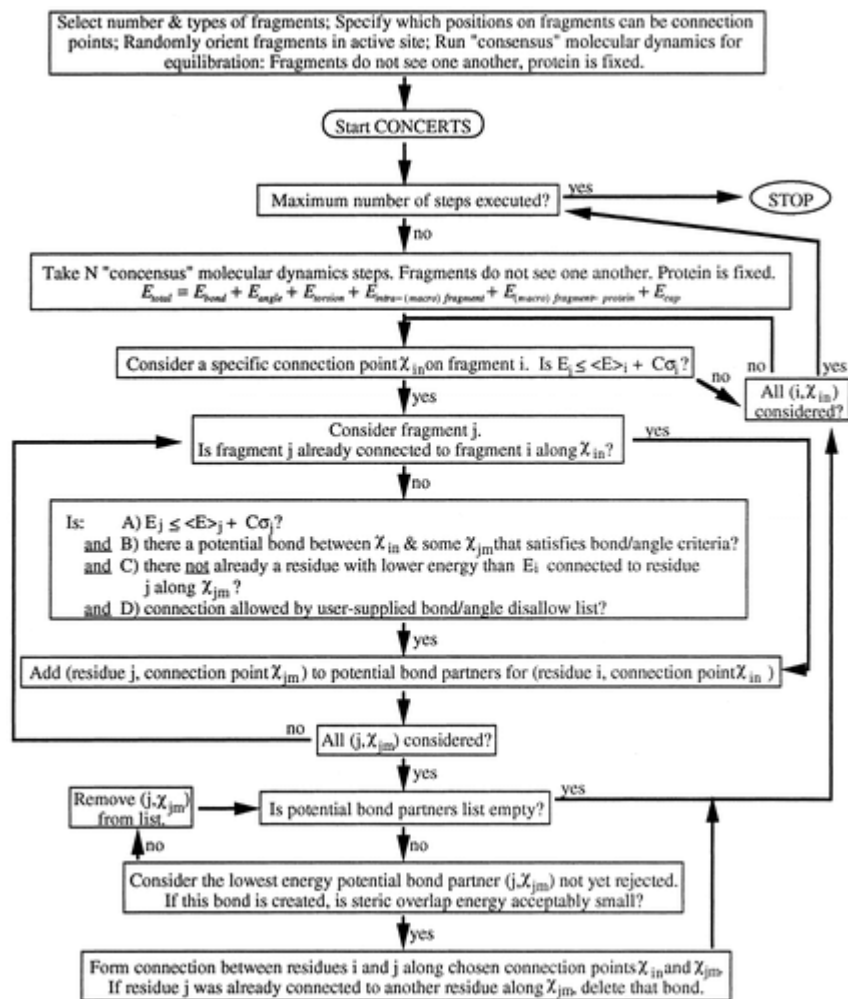
$K_i = 0.1\text{ nM}$

$K_i = 42\text{ nM}$



# Fragment-based structure generation

## CONCEPTS

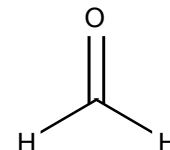
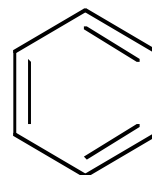
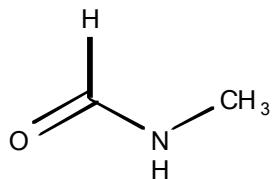


MD of fragments which are linking or breaking during the simulation in order to create more favorable structures

formation of certain bonds was forbidden:  
O-O, N-N, N-O, S-O, O-C-O, O-N-O, N-C-N,  
C $_{\alpha}$ -C $_{\alpha}$ , C-C $_{\alpha}$ -C

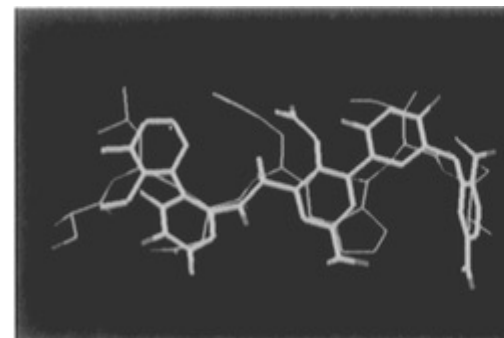
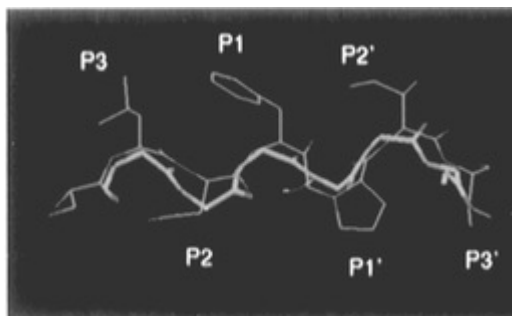
# Fragment-based structure generation

## CONCEPTS: HIV-1 protease inhibitors

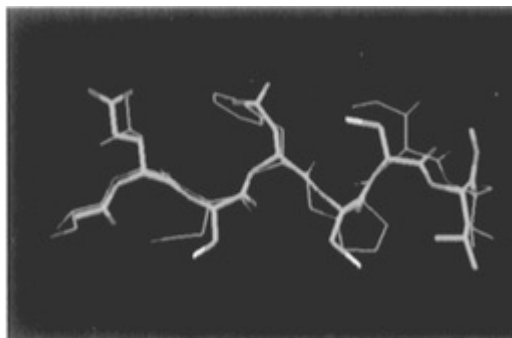


CH<sub>4</sub> H<sub>2</sub>O

NH<sub>3</sub>

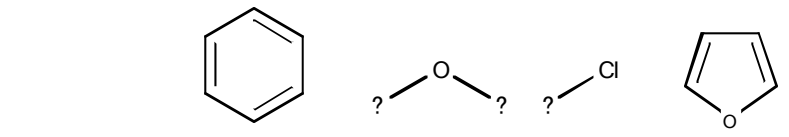


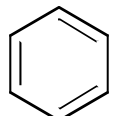
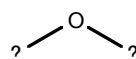
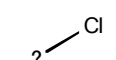
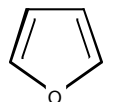
+ 19 side chains

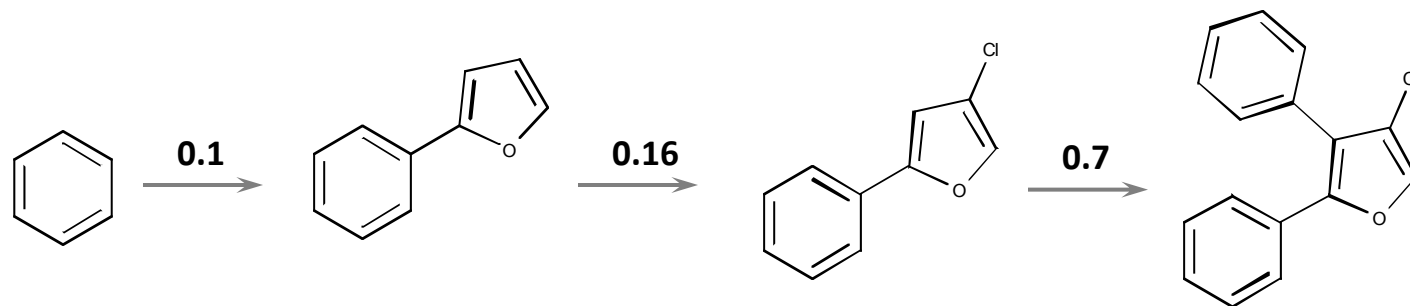
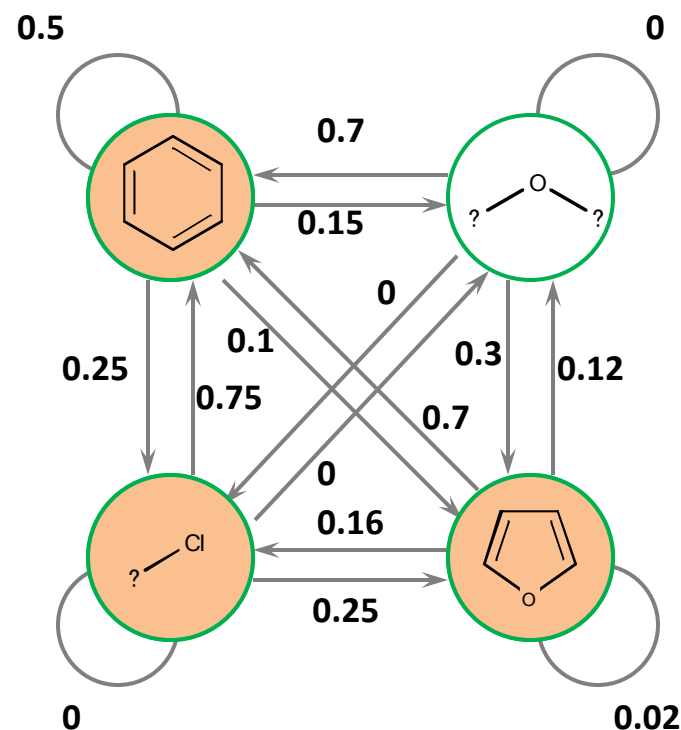


# Fragment-based structure generation

FOG

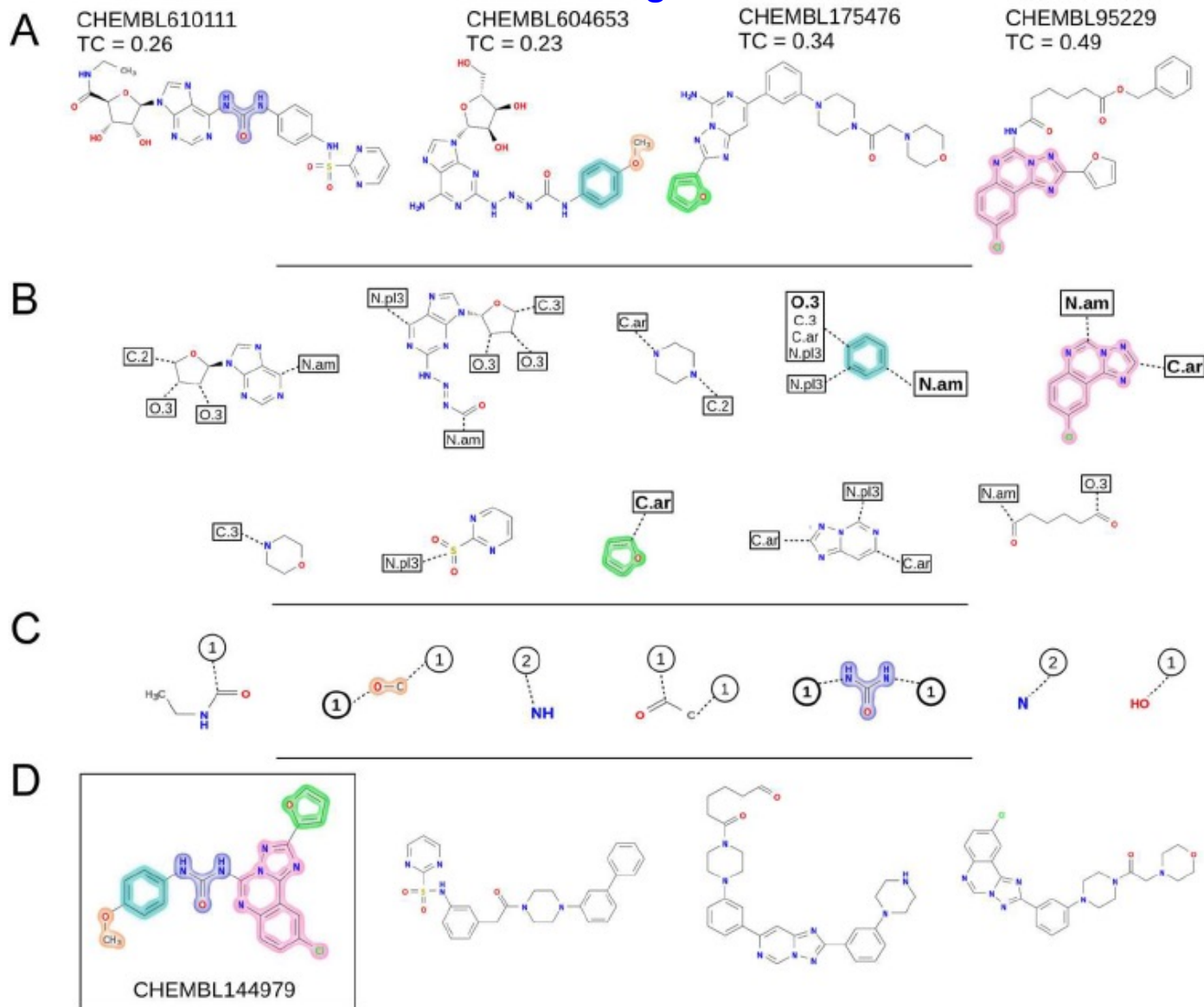


	0.5	0.15	0.25	0.1
	0.7	0	0	0.3
	0.75	0	0	0.25
	0.7	0.12	0.16	0.02



# Fragment-based structure generation

## eMolFrag

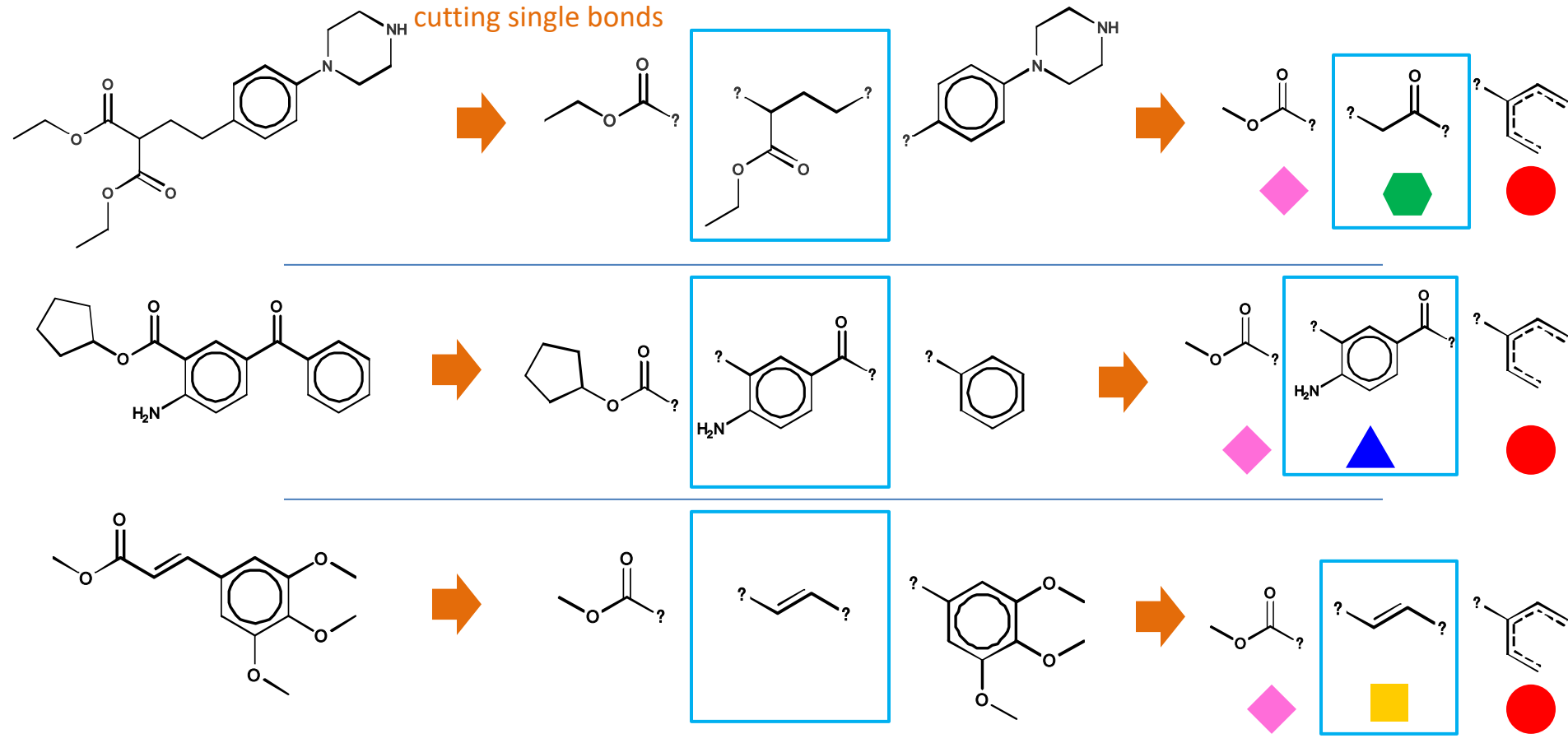


# Fragment-based structure generation

## CReM: chemically reasonable mutations

exhaustive fragmentation  
cutting single bonds

taking context of radius R (here R = 3)



DB of replacements

environment (radius = 3)

fragments



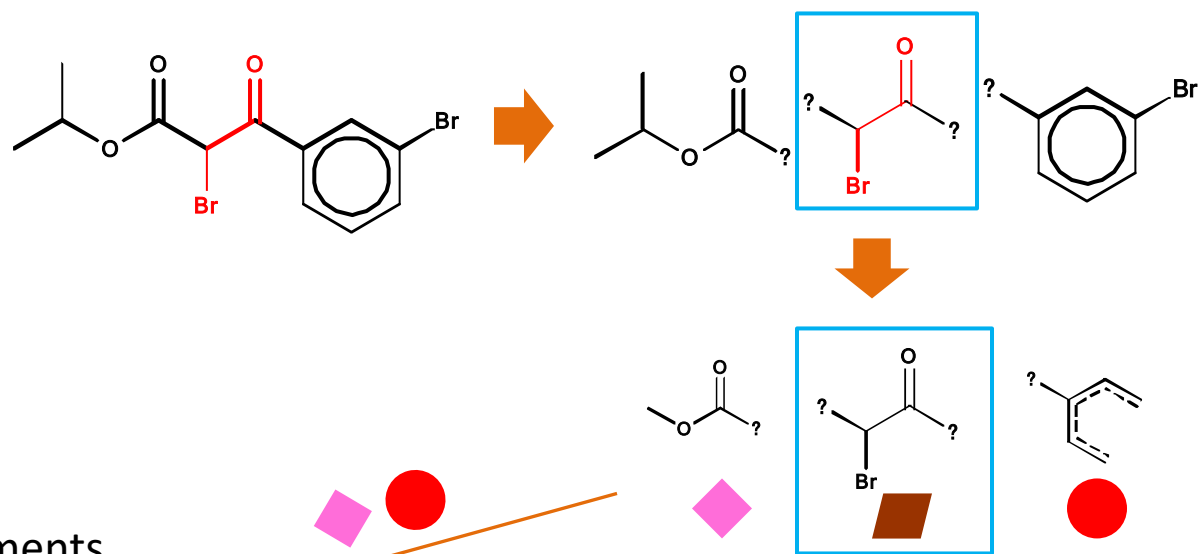
mutually exchangeable  
fragments

...

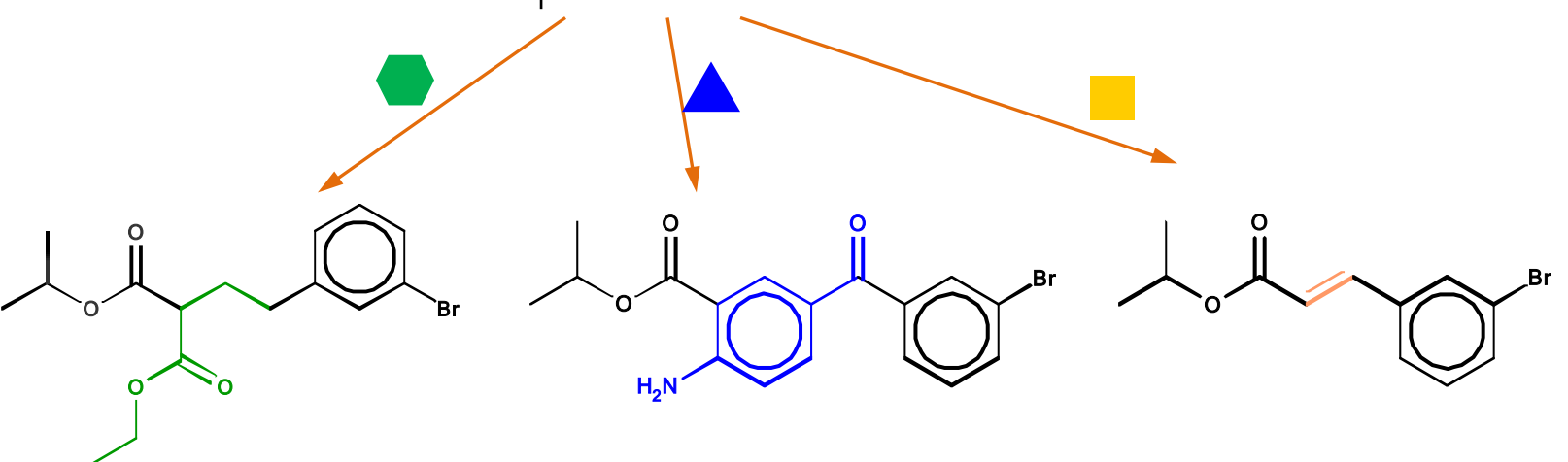
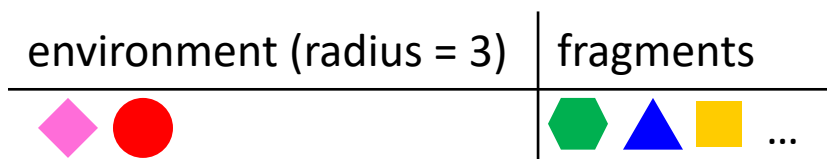
...

# Fragment-based structure generation

CReM: chemically reasonable mutations



DB of replacements



**Generated structures are always chemically valid!**

# Fragment-based structure generation

	fragment-based
exhaustiveness of chemical space search	++ variable, controlled by the size of fragments to replace
structure novelty	++
structure diversity	++
chemically valid structures	-/+ (+++)
synthetically feasible	-/+ (++)
combinatorial explosion / time consuming	++

fragment-based  $\approx$  semi-empirical



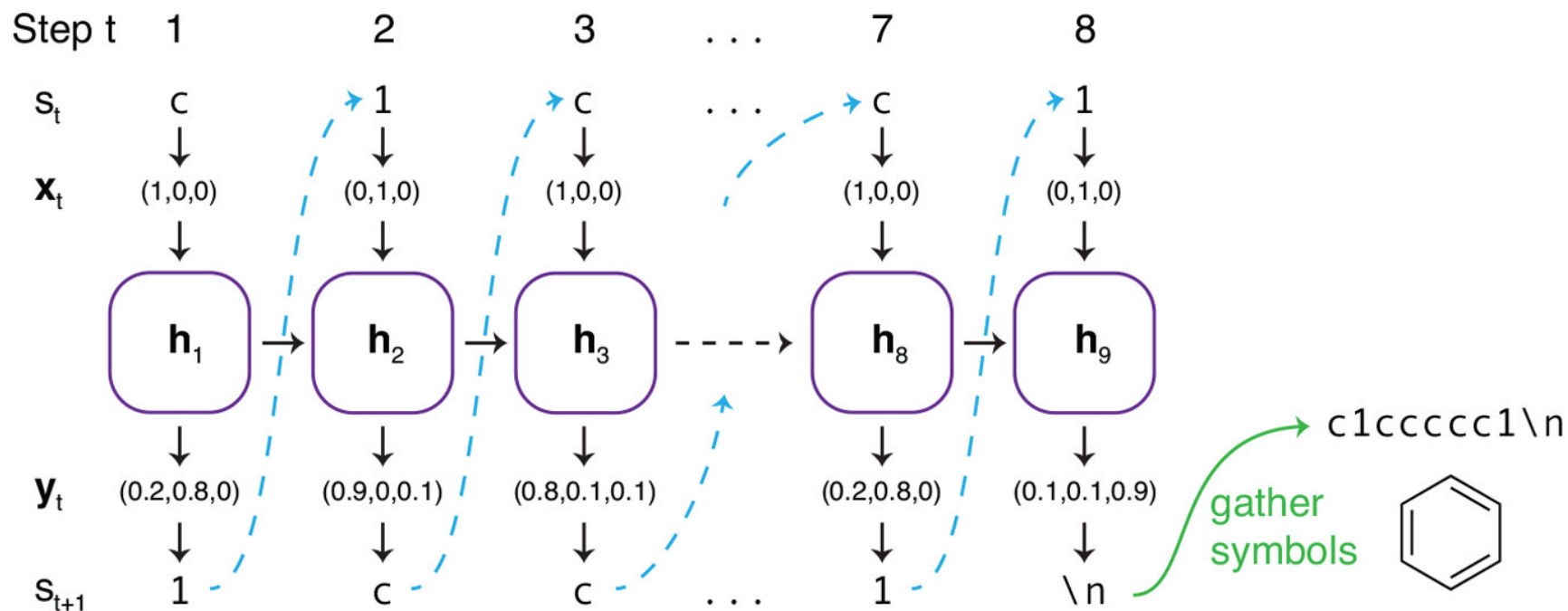
# De novo structure generation

## Summary

	atom-based	fragment-based	reaction-based
exhaustiveness of chemical space search	++++ very small steps; more suitable for systematic exploration of local chemical space	++ variable, controlled by the size of fragments to replace	+ depends on reactant library and reaction rules; only grow molecules
structure novelty	+++*	++	++
structure diversity	+++*	++	++
chemically valid structures	-	(+++)	+++
synthetically feasible	---	(++)	+++
combinatorial explosion / time consuming	---	++	+++

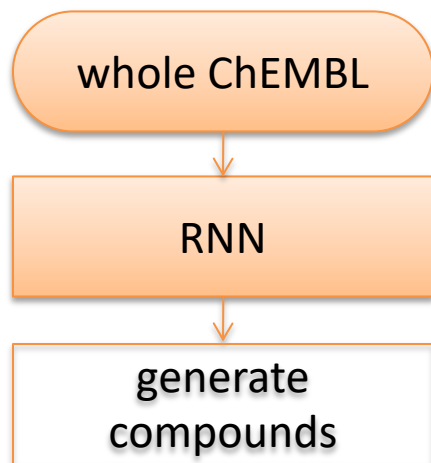
# Deep learning models for structure generation

## Recurrent neural network (RNN)

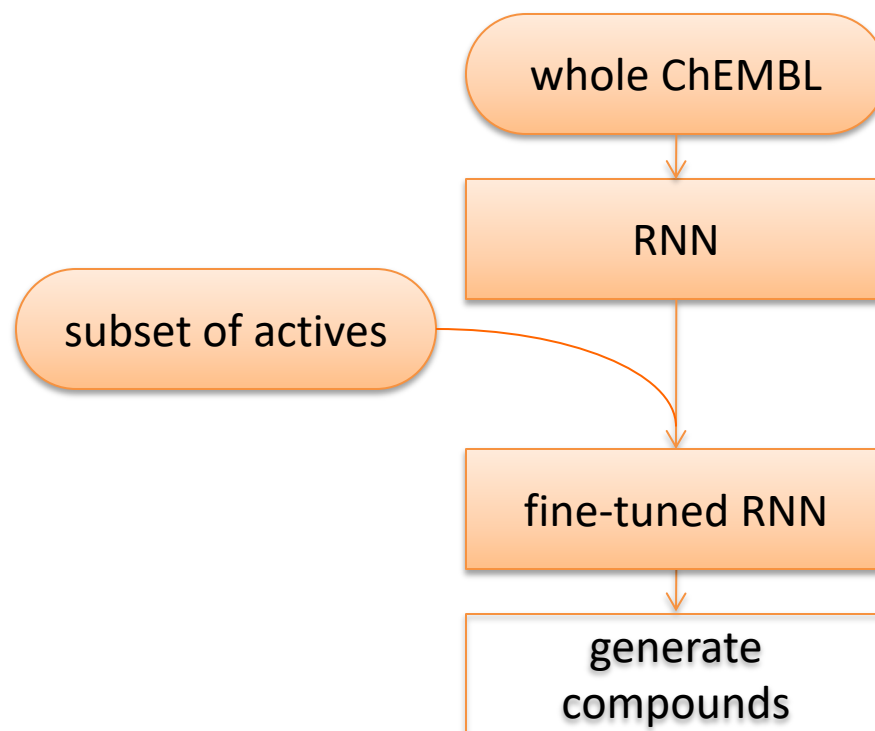


# Deep learning models for structure generation

## unsupervised generation

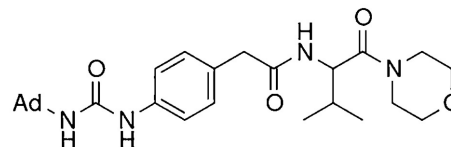
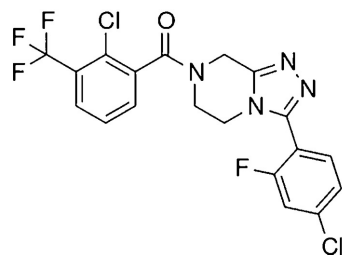
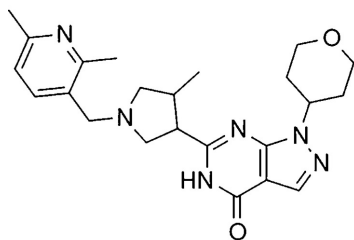
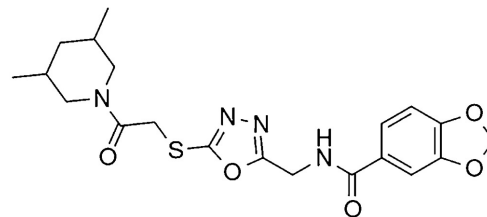
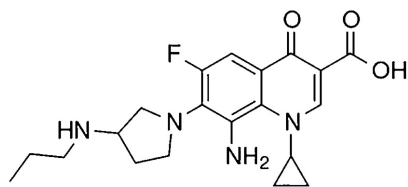
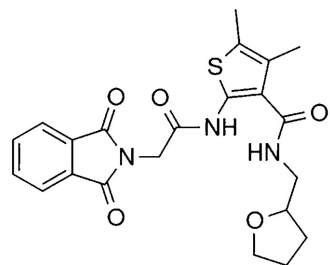


## transfer learning



# Deep learning models for structure generation

## unsupervised generation



976 327 compounds

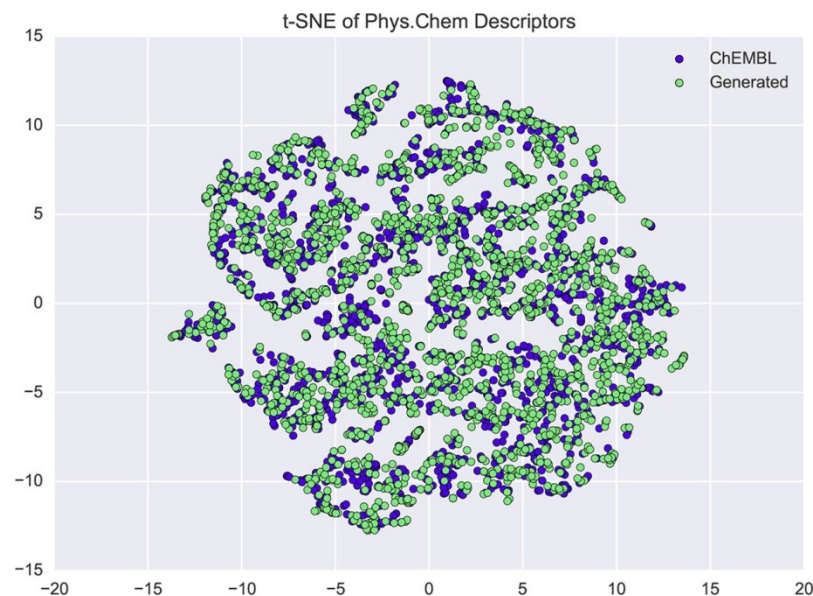
97.7% chemically valid

11.5% were duplicated with ChEMBL

1.7% of duplicates

75% passed AZ filters (similar to ChEMBL)

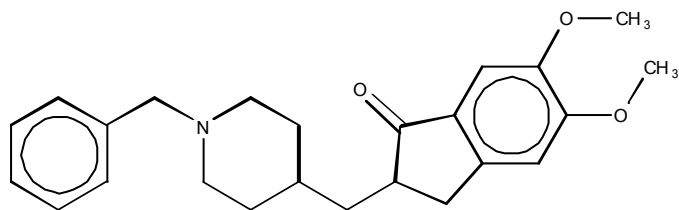
12% of scaffolds were common with ChEMBL



# Deep learning models for structure generation

	deep learning
exhaustiveness of chemical space search	++
structure novelty	++
structure diversity	++
chemically valid structures	++
synthetically feasible	?
combinatorial explosion / time consuming	+++

Issue of SMILES based representation -  
the same structure can be represented by different SMILES



```
COc1cc2CC(CC3CCN(Cc4ccccc4)CC3)C(=O)c2cc1OC  
COc1cc2c(cc1OC)C(=O)C(CC1CCN(Cc3ccccc3)CC1)C2
```

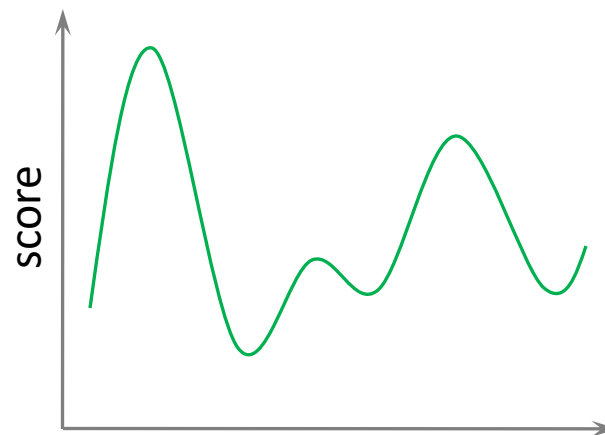
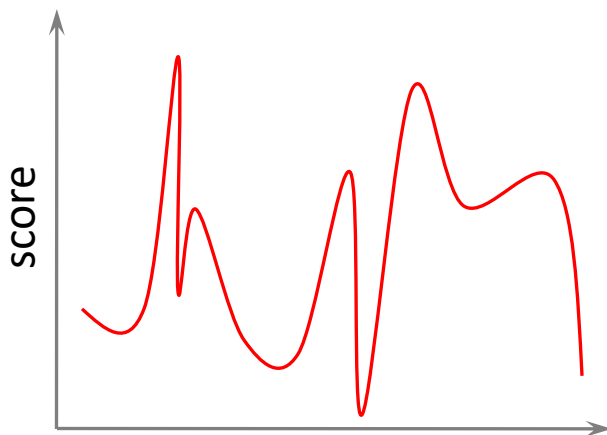
# Scoring functions

Can be any but preferably smooth to follow the chemical similarity principle:

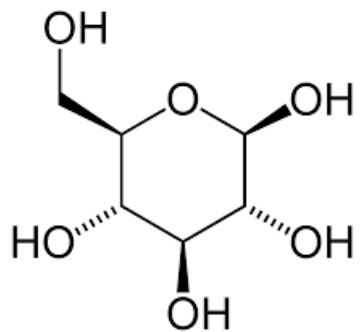
- similarity measures
  - QSAR model prediction
  - pharmacophore fit
  - docking score
  - molecular dynamics
- ...

ligand-based scoring functions

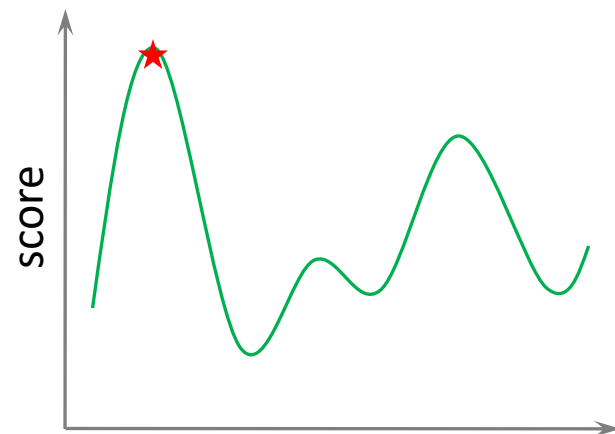
structure-based scoring functions



# Inverse QSAR



$D_1$	$D_2$	$D_3$	...	$D_N$
1	0	9	...	1
4	0	1	...	1
0	2	3	...	3
...	...	...	...	...
4	0	0	...	1



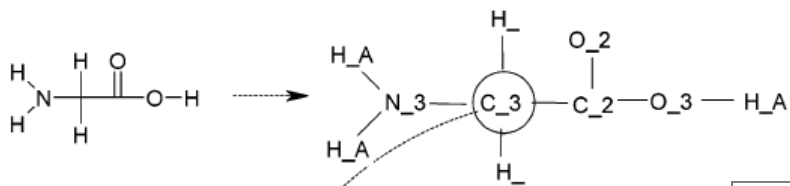
$D_1$	$D_2$	$D_3$	...	$D_N$
11	3	1	...	15



**STRUCTURE ?**

# Inverse QSAR

## Atom signatures



$\sigma^0$   
 $\sigma^1$   
 $\sigma^2$   
 $\sigma^3$

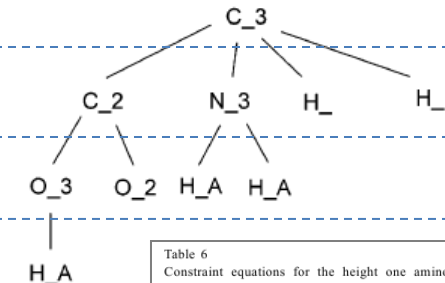
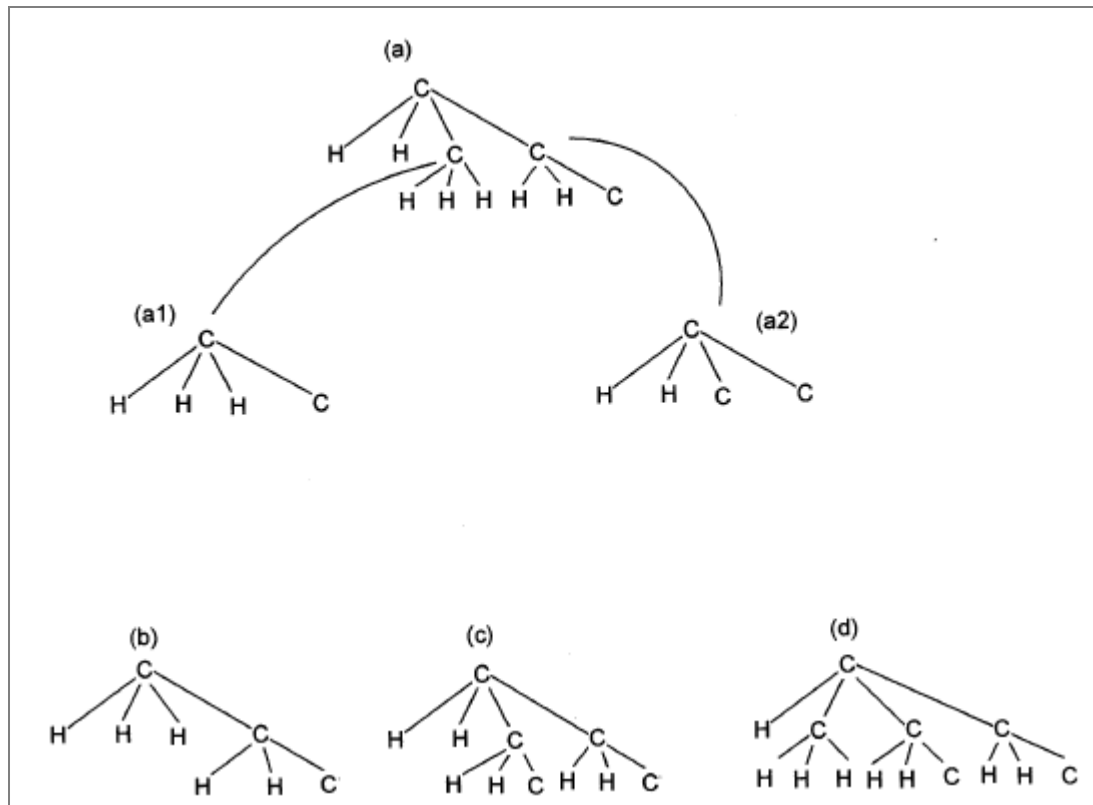


Table 6  
Constraint equations for the height one amino acid signatures in the training set

- (1)  $-x_{44} + x_{46} = 0$
- (2)  $-x_{38} + x_{47} = 0$
- (3)  $-x_{22} - x_{27} + x_{45} + x_{47} = 0$
- (4)  $-x_{10} + x_{45} + x_{46} = 0$
- (5)  $-x_{34} - x_{37} + x_{41} + x_{42} + x_{43} + x_{44} = 0$
- (6)  $-x_{21} + x_{43} = 0$
- (7)  $-x_{16} + x_{40} = 0$
- (8)  $-x_{13} + x_{39} + x_{42} = 0$
- (9)  $-x_2 - x_5 + x_{39} + x_{40} + x_{41} = 0$
- (10)  $-x_{28} - x_{30} - 2x_{31} + x_{33} + x_{35} + x_{36} + x_{37} + x_{38} = 0$
- (11)  $-x_{18} - x_{24} - x_{26} - x_{27} + x_{32} + x_{36} = 0$
- (12)  $-x_{14} + x_{35} = 0$
- (13)  $-x_3 - x_4 - 2x_6 + x_{32} + x_{33} + x_{34} = 0$
- (14)  $-x_{15} - x_{16} + 2x_{29} + x_{30} = 0$
- (15)  $-x_5 + x_{28} = 0$
- (16)  $(x_{20} + x_{25} + x_{26}) \% 2 = 0$
- (17)  $-x_{15} + x_{23} + x_{25} = 0$
- (18)  $-x_{12} - x_{14} + x_{19} + x_{23} + x_{24} = 0$
- (19)  $-x_9 + x_{17} + x_{19} + x_{20} + x_{21} + x_{22} = 0$
- (20)  $-x_1 - x_4 + x_{17} + x_{18} = 0$
- (21)  $-x_8 + x_{11} + x_{12} + x_{13} = 0$
- (22)  $-x_3 + x_{11} = 0$
- (23)  $(x_7 + x_8 + x_9 + x_{10}) \% 2 = 0$
- (24)  $-x_1 - x_2 + x_7 = 0$

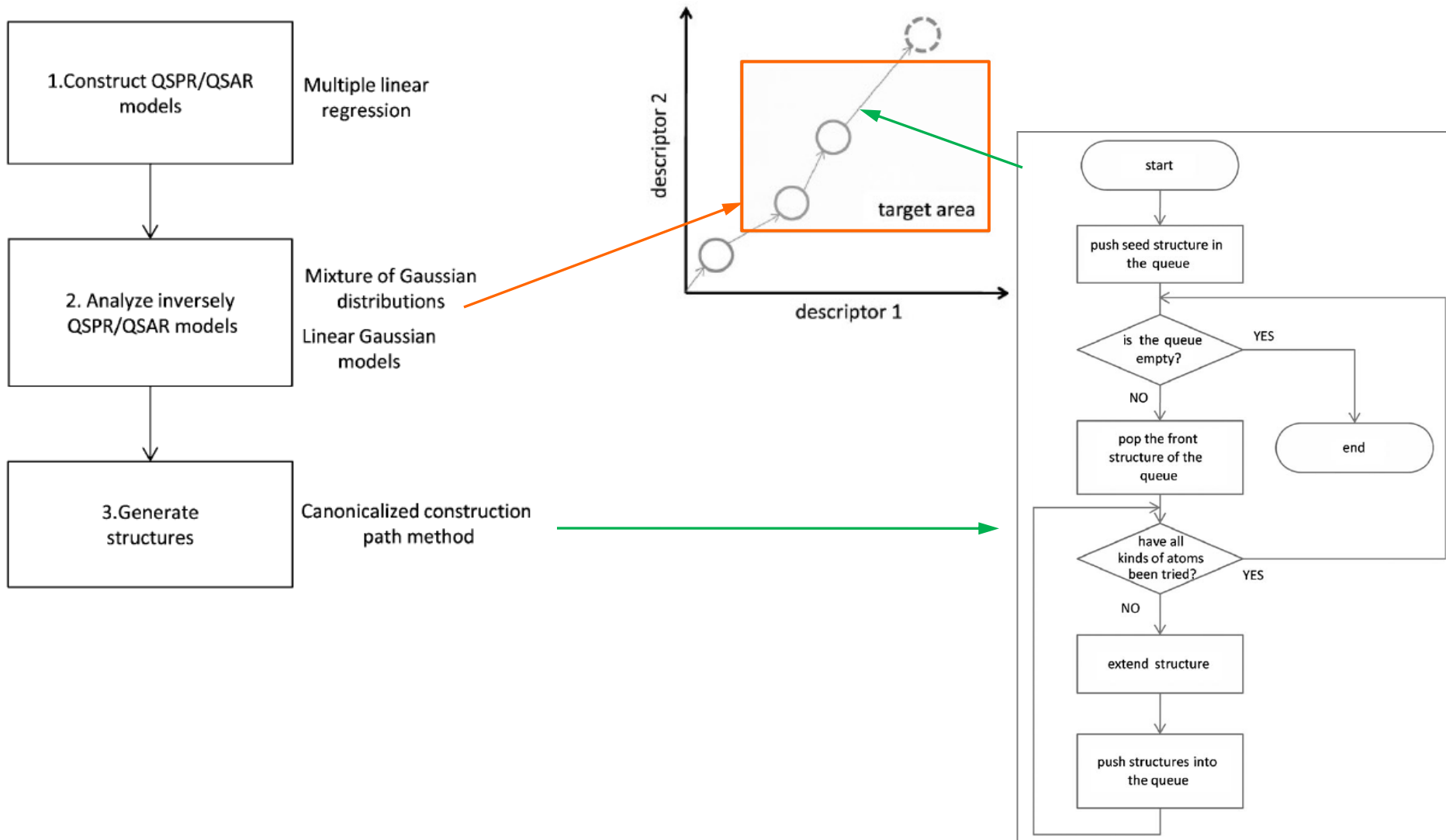
Eqs. (16) and (23) are modulus equations, which can be expressed as homogeneous equations by adding a dummy variable. For example Eq. (16) would read  $x_{20} + x_{25} + x_{26} - 2z_1 = 0$ . The % sign indicates the modulus is to be used.





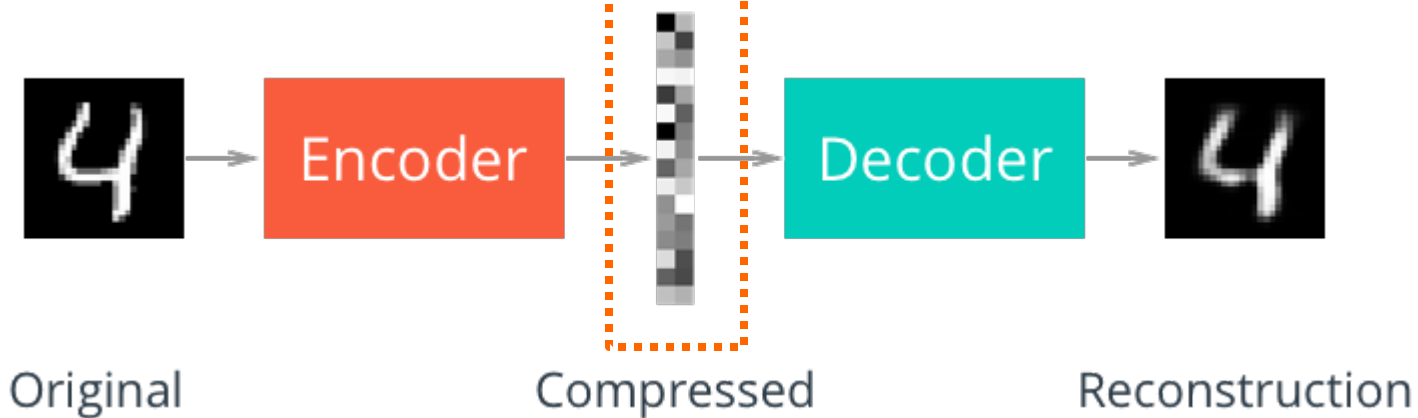
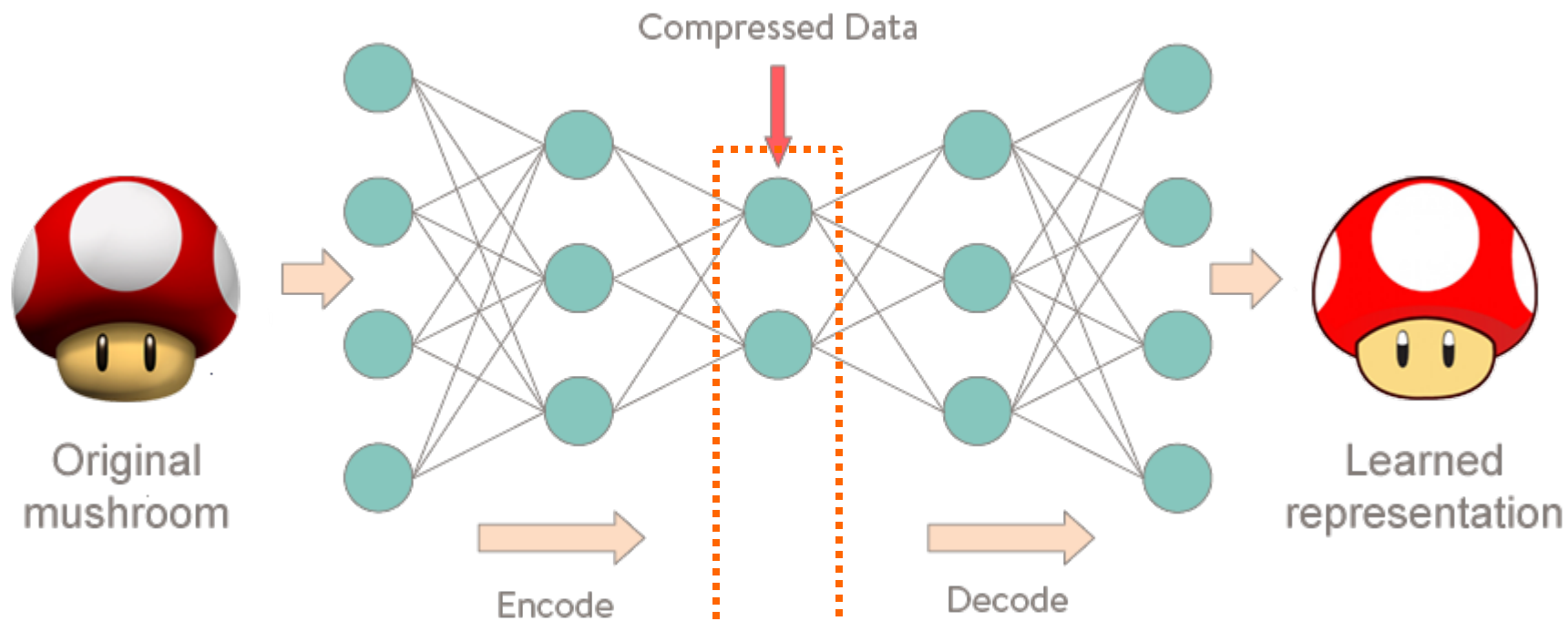
# Inverse QSAR

## Inverse QSAR with monotonically changed descriptors

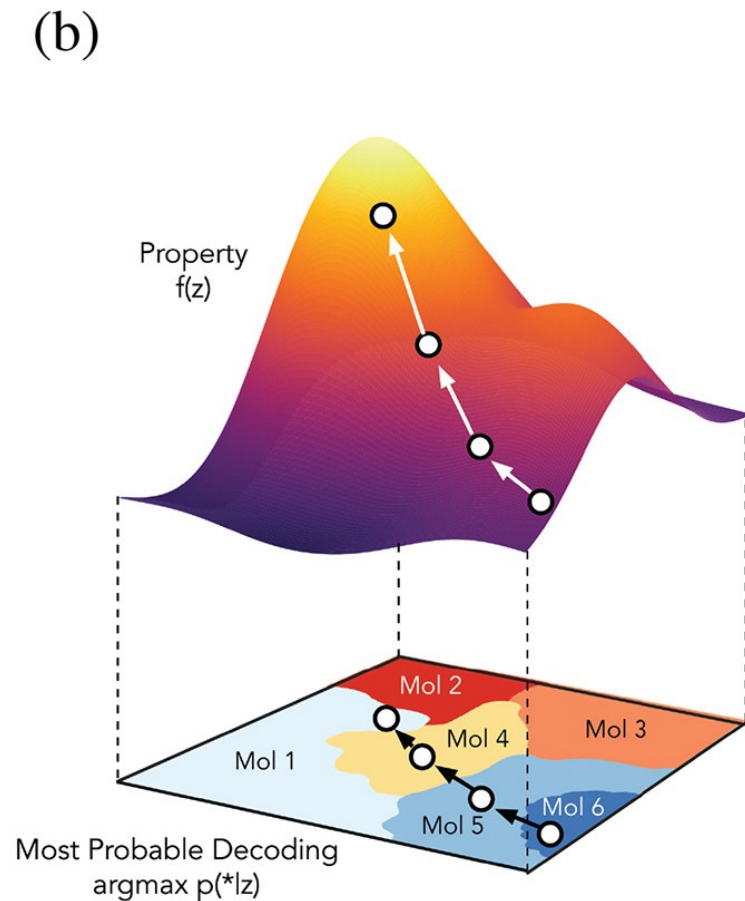
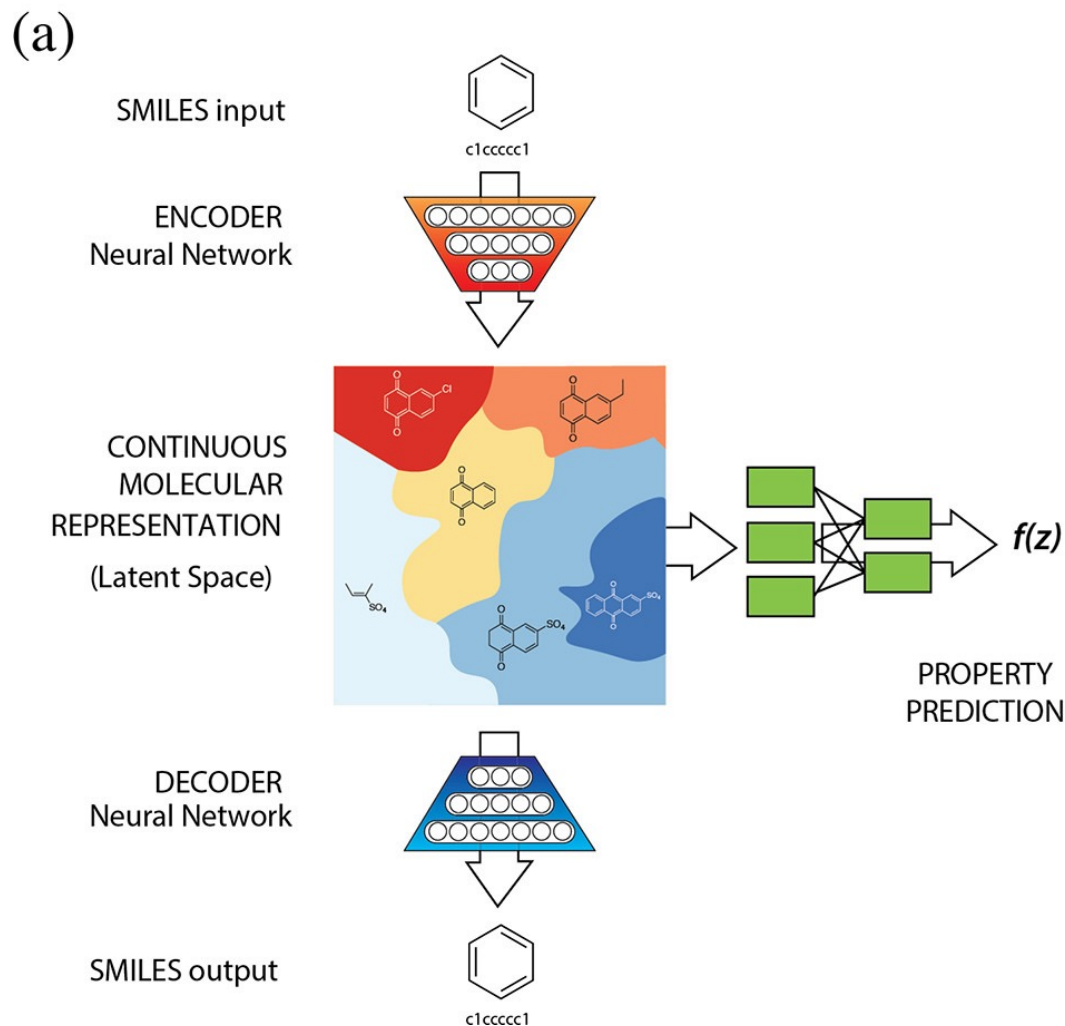


# Inverse QSAR: deep learning

## Autoencoder

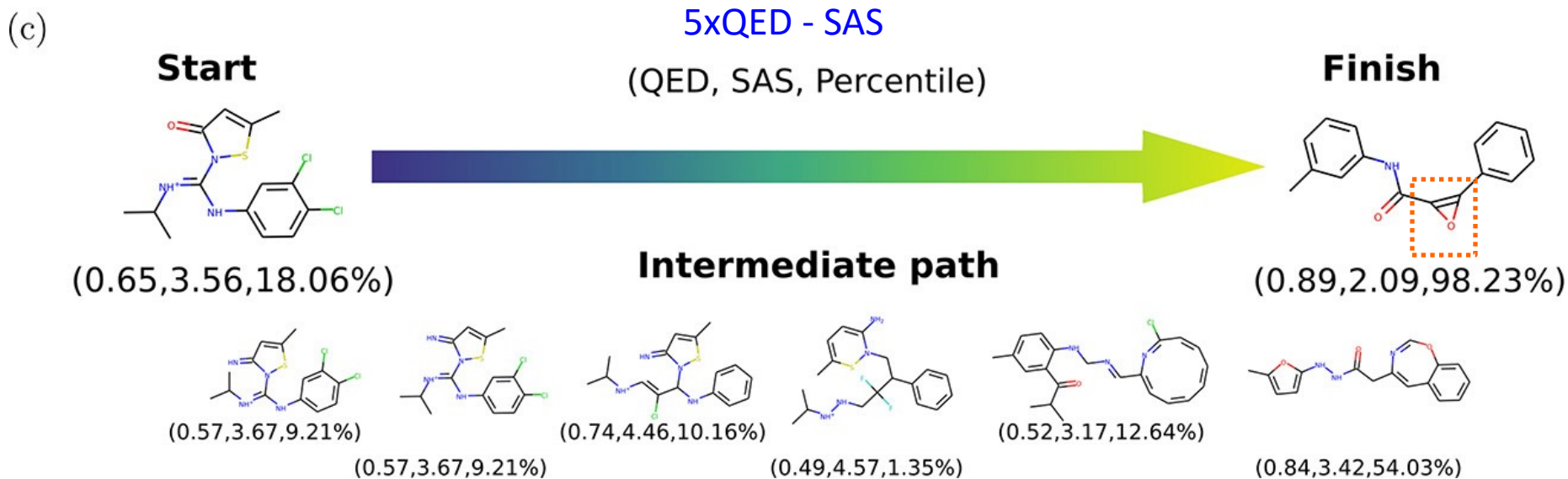
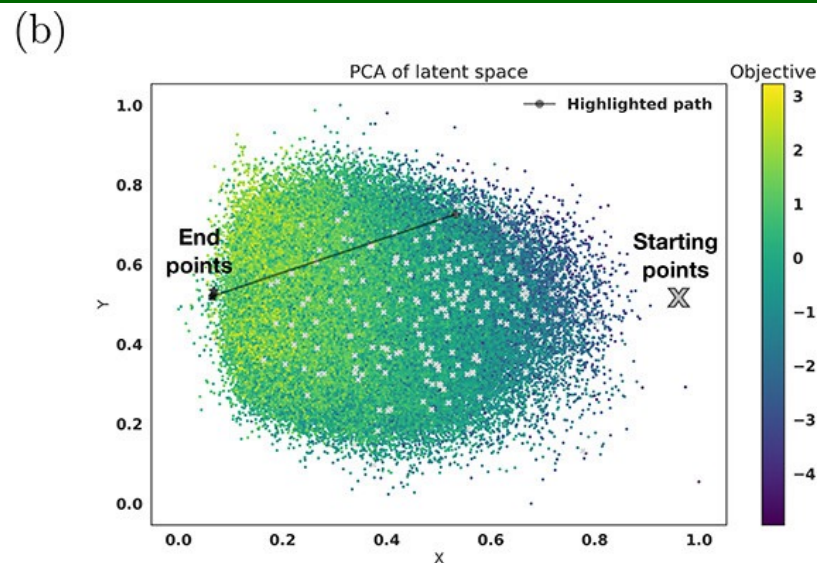
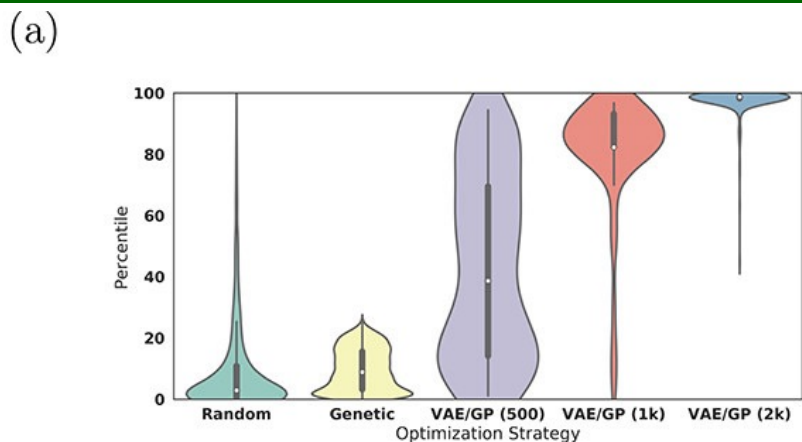


# Inverse QSAR: deep learning



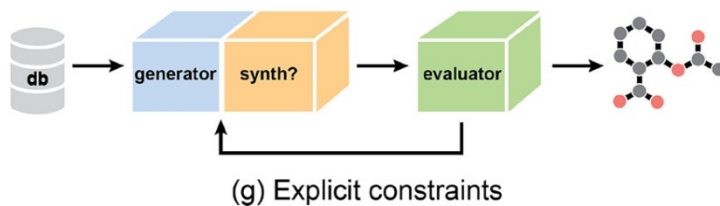
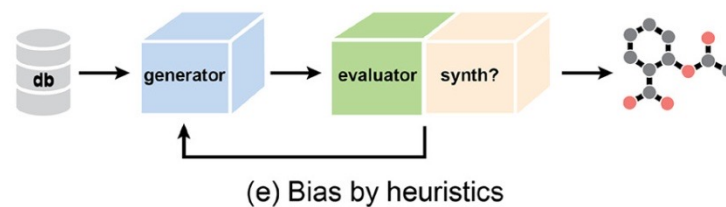
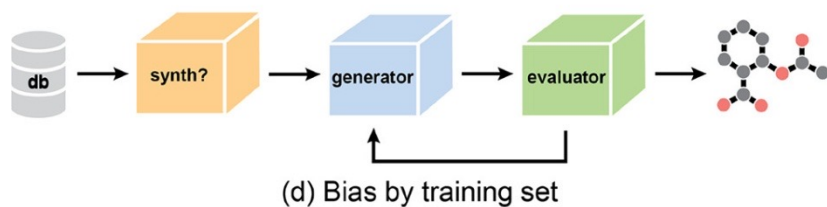
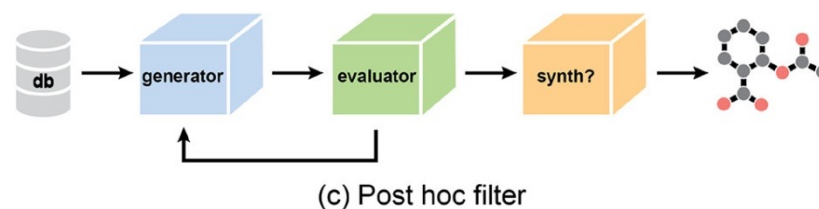
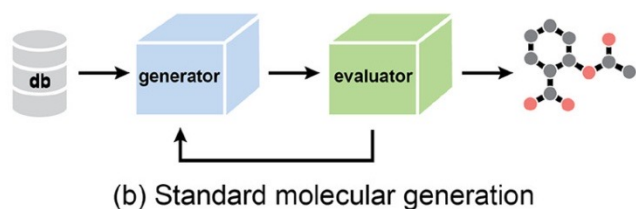
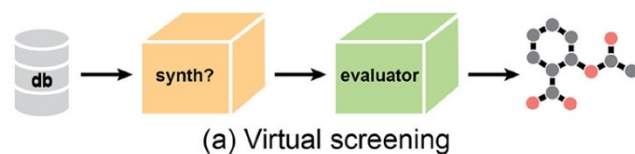
Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, 4, 268-276.

# Inverse QSAR: deep learning



Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, 4, 268-276.

# Control over synthetic feasibility



# Assessment of synthetic feasibility

Genheden et al. *J Cheminform* (2020) 12:70  
<https://doi.org/10.1186/s13321-020-00472-1>

Journal of Cheminformatics

## SOFTWARE

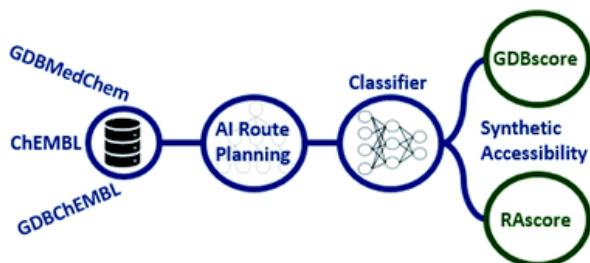
## Open Access

### AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning



Samuel Genheden<sup>1\*</sup>, Amol Thakkar<sup>1,2</sup>, Veronika Chadimová<sup>1</sup>, Jean-Louis Reymond<sup>2</sup>, Ola Engkvist<sup>1</sup> and Esben Bjerrum<sup>1\*</sup>

## Chemical Science



## EDGE ARTICLE

[View Article Online](#)  
[View Journal](#) | [View Issue](#)



Cite this: *Chem. Sci.*, 2021, 12, 3339

All publication charges for this article have been paid for by the Royal Society of Chemistry

### Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning†

Amol Thakkar, \*<sup>ab</sup> Veronika Chadimová, <sup>a</sup> Esben Jannik Bjerrum, <sup>a</sup> Ola Engkvist <sup>a</sup> and Jean-Louis Reymond \*<sup>b</sup>

Voršilák et al. *J Cheminform* (2020) 12:35  
<https://doi.org/10.1186/s13321-020-00439-2>

Journal of Cheminformatics

## RESEARCH ARTICLE

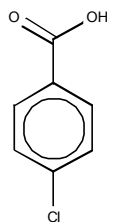
## Open Access

### SYBA: Bayesian estimation of synthetic accessibility of organic compounds



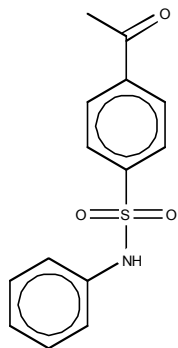
Milan Voršilák<sup>1,2</sup> , Michal Kolář<sup>3,4</sup> , Ivan Čmelo<sup>1</sup> and Daniel Svozil<sup>1,2\*</sup>

# Examples of SA scores (ChEMBL22)



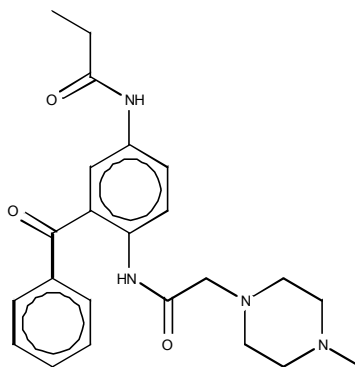
1.2

CHEMBL618



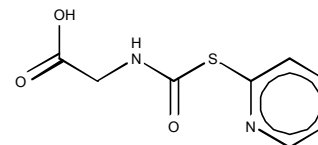
1.5

CHEMBL3310985



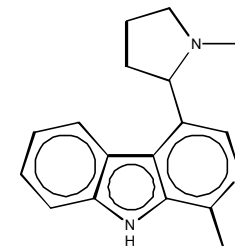
2.0

CHEMBL595820



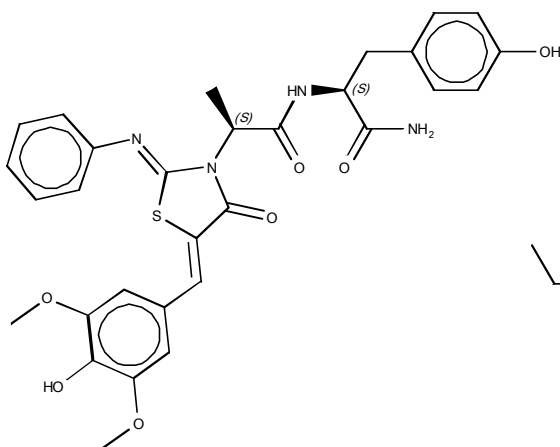
2.5

CHEMBL503660



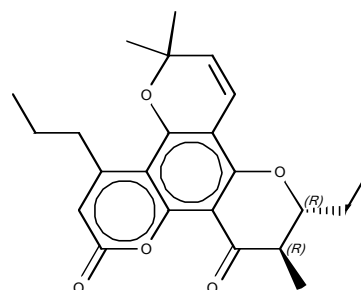
3.0

CHEMBL500286



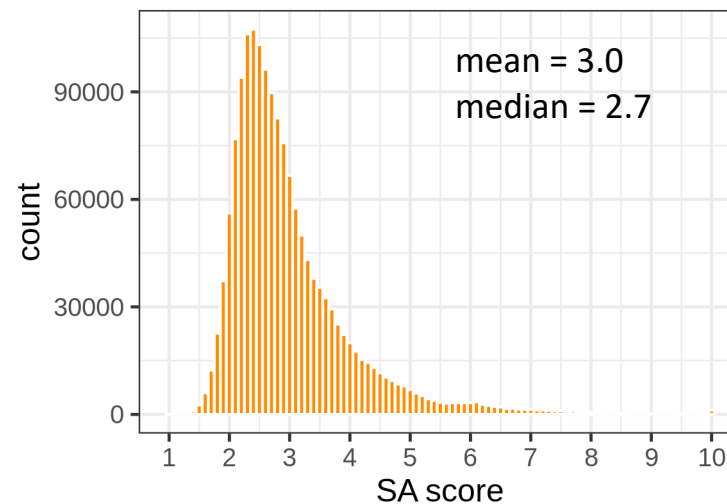
3.5

CHEMBL582554



4.0

CHEMBL7633





# Control of synthetic feasibility within CReM

## Content of fragmented library



all ChEMBL  
compounds  
(1 554 160)

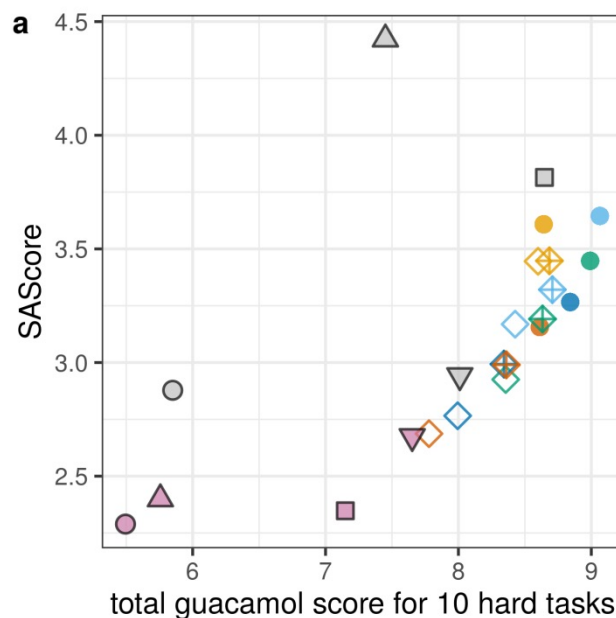
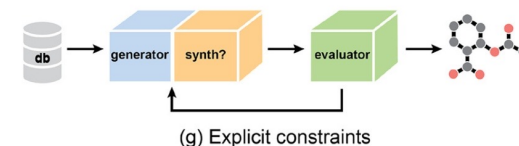
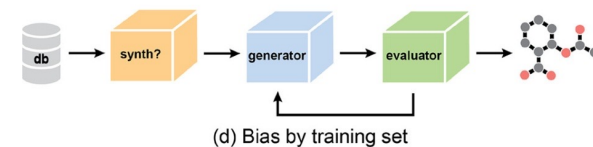
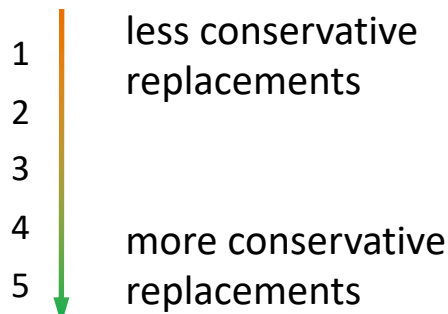


compounds with  
SA score  $\leq 2.5$   
(572 527)

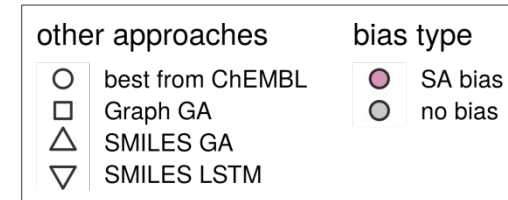
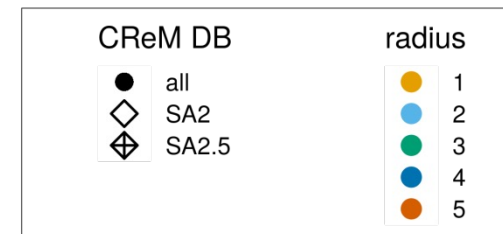


compounds with  
SA score  $\leq 2$   
(107 806)

## Context radius

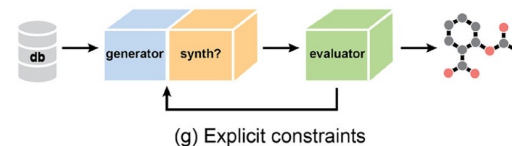
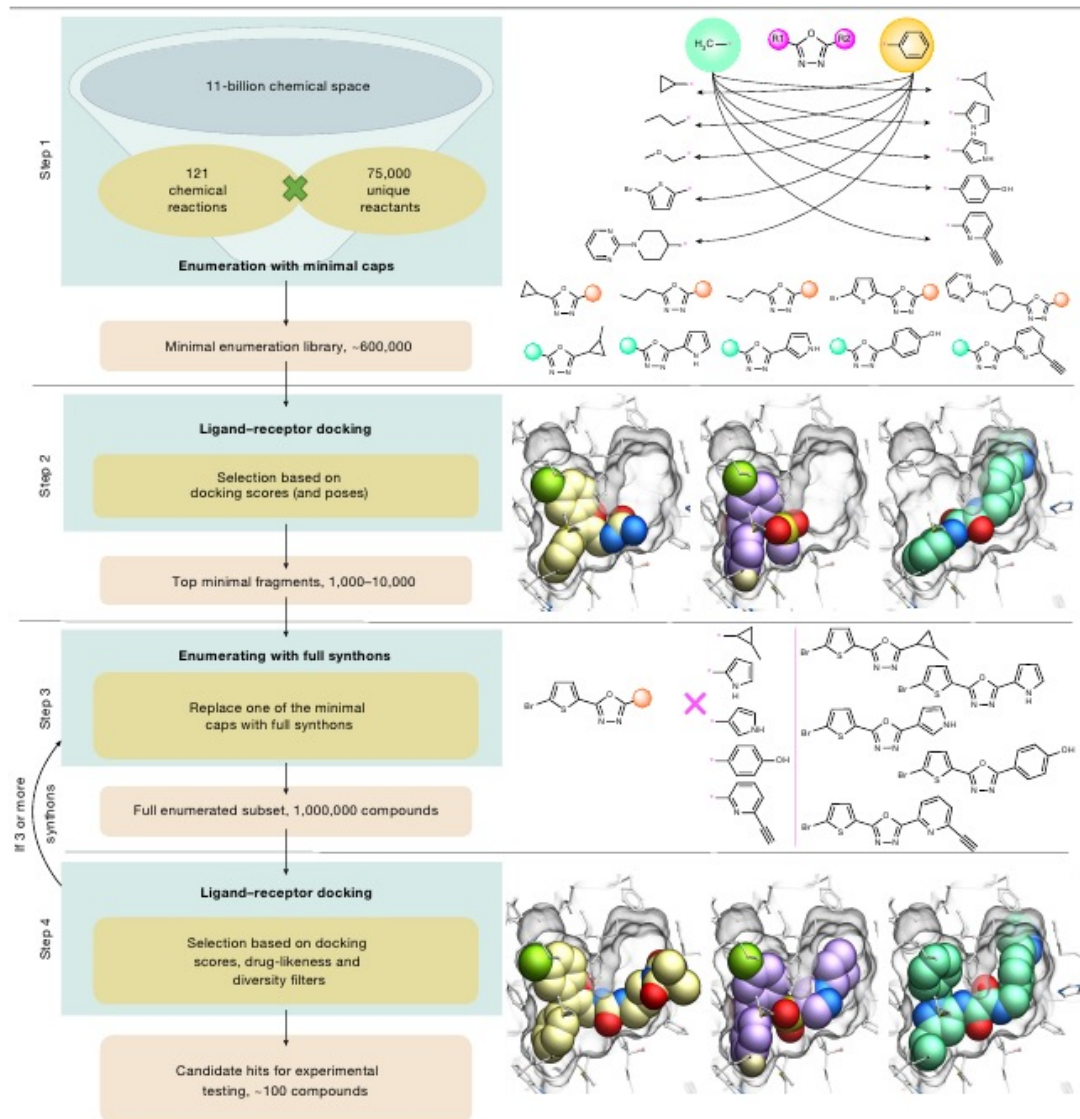


**b**



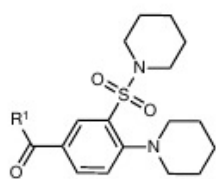


# V-SYNTHESIS



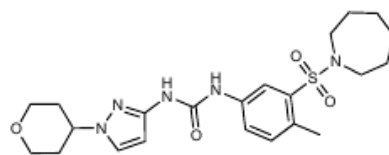
**Fig. 1 | V-SYNTHESIS approach to modular screening of Enamine REAL Space.** A general overview of the four-step algorithm (left) and examples for each step (right). Asterisks in step one show the attachment points of synthons; arrows show possible pairing of minimal synthons with real synthons.

# V-SYNTHES



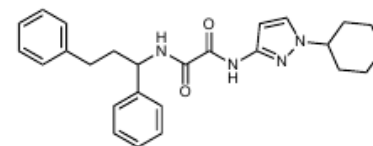
**523 scaffold**

**a**



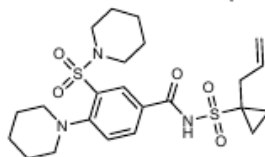
**505**

CB<sub>1</sub> K<sub>i</sub> 0.28 (0.22–0.36) μM  
CB<sub>2</sub> K<sub>i</sub> 0.54 (0.43–0.67) μM



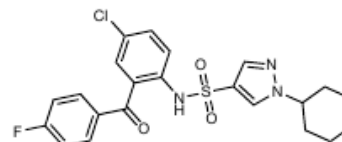
**610**

0.76 (0.62–0.93) μM  
4.17 (3.14–5.62) μM



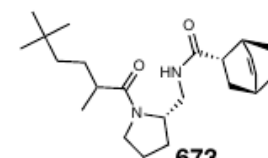
**523**

CB<sub>1</sub> K<sub>i</sub> 1.82 (1.46–2.28) μM  
CB<sub>2</sub> K<sub>i</sub> 1.59 (1.27–1.98) μM



**665**

0.30 (0.32–0.47) μM  
0.82 (0.71–0.95) μM

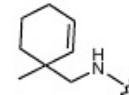
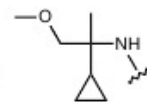
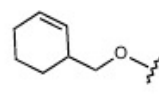
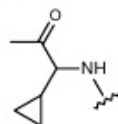
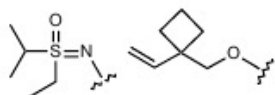


**673**

0.97 (0.84–1.14) μM  
3.66 (2.98–4.51) μM

**c**

R1



Compound

**733**

**736**

**738**

**742**

**747**

**749**

CB<sub>1</sub> functional  
potency

K<sub>i</sub> (nM)

871

1,185

856

2,340

455

209

CI 95% (nM)

(720–1,051)

(868–1,603)

(725–1,009)

(1,878–2,919)

(373–558)

(177–248)

CB<sub>2</sub> functional  
potency

K<sub>i</sub> (nM)

10.9

48.5

125

120

9.6

49.2

CI 95% (nM)

9.3–12.9

38.6–61.0

105–148

101–144

8.58–10.8

42.1–57.6

CB<sub>1</sub> binding  
affinity

K<sub>i</sub> (nM)

43.2

140

23.1

394

228

689

CI 95% (nM)

28.2–66.1

105–186

13.9–38.6

281–551

172–303

472–1,004

CB<sub>2</sub> binding  
affinity

K<sub>i</sub> (nM)

1.2

2.8

13.0

6.4

0.9

4.0

CI 95% (nM)

0.9–1.6

2.0–3.7

10.2–16.6

5.2–7.8

0.6–1.2

2.5–6.5

# Take home message

- De novo design can efficiently explore much larger chemical space than virtual screening
- There are multiple approaches to generate chemically valid structures, all of them have their pros and cons
- The main issue of de novo design is synthetic feasibility of generated compounds
- There are several ways how to control synthetic feasibility

**Thank you for your attention**