Proteochemometrics (PCM) de novo generation (DrugEx)

Willem Jespers

Sequence alignment

Alignment where match is cheap, mismatch is cheap and gap is costly

```
This is a hard example.------
||   ||||        |   |
That is another easy example.
**:. ****   .    *.  *
```

Algorithm wants to have maximal points! Aligning correct get's points, misaligning and gaps cost points..!

An * (asterisk) indicates positions which have a single, fully conserved residue.
A : (colon) indicates conservation between groups of strongly similar properties
A . (period) indicates conservation between groups of weakly similar properties

Adapted from Hala Iqbal & Ethan Karp, Online Computational Biology Textbook, Pairwise Sequence Alignment and Substitution Matrices, http://compbio.pbworks.com/w/page/16252891/FrontPage, retrieved August 20th 2015

BLOcks SUbstitution (BLOSUM) Matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Less frequently occurring or high impact residues aligned correctly increase score

Adapted from Hannes Röst

Aligning two sequences

- CALHP & CLLHL
- Gap = -10

	Cys	Ala	Leu	His	Pro
Cys	9	0	-1	-3	-3
Leu	-1	-1	4	-3	-3
Leu	-1	-1	4	-3	-3
His	-3	-2	-3	8	-2
Leu	-1	-1	4	-3	-3

Option 1 : Score -7

```
C-ALHP
|    |
CLL-HL
```

Option 2 : Score 0

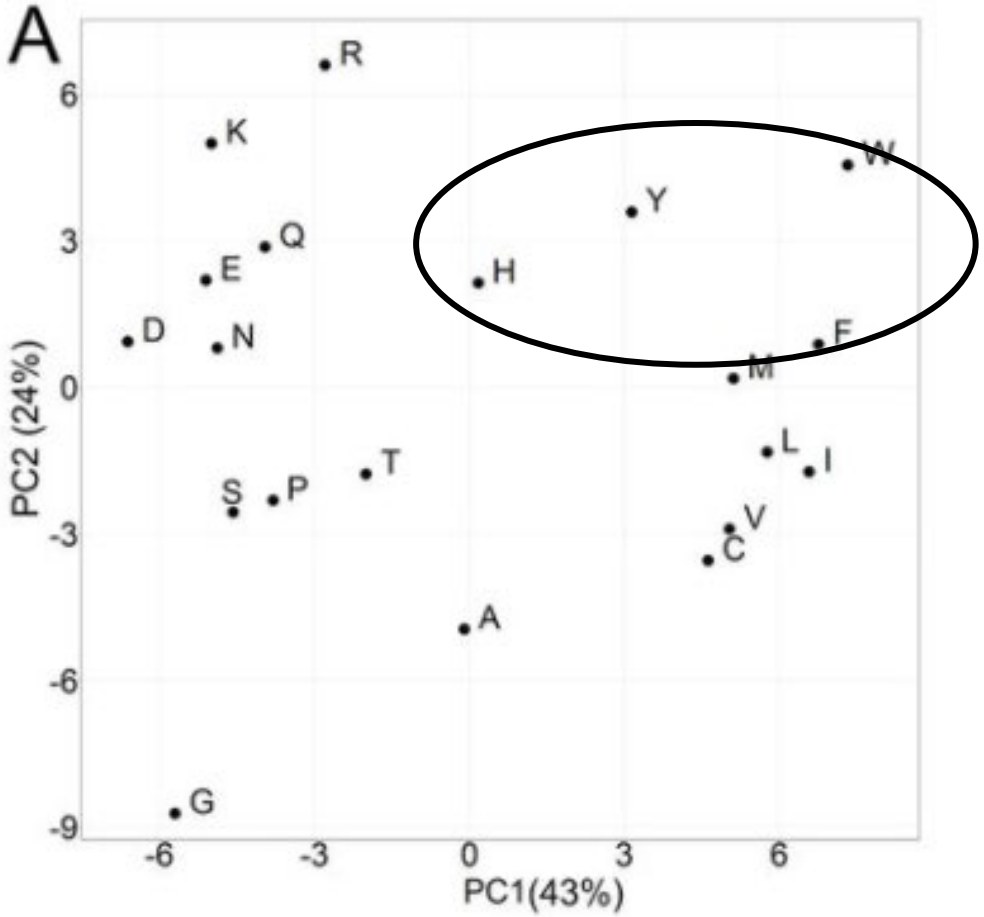
```
CALH-P
|  ||
CLLHL-
```

Option 3 : Score 17

```
CALHP
|  ||
CLLHL
```

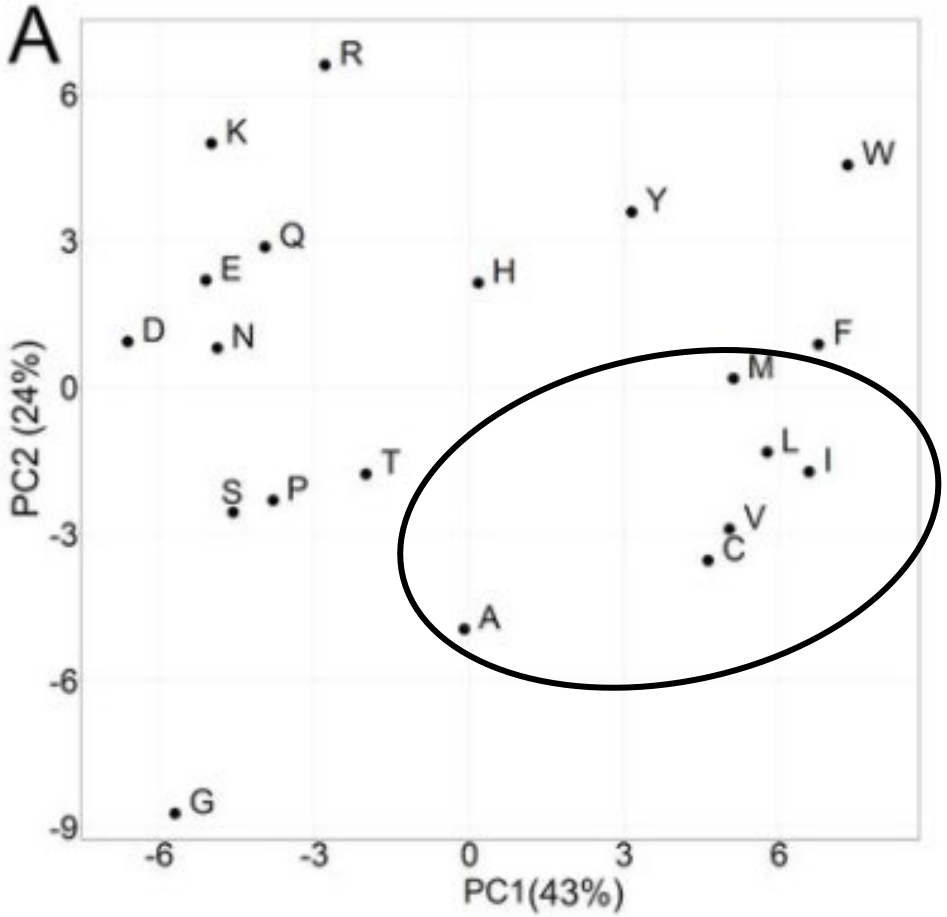
Peptide / binding site descriptors

- Similar to the fingerprints used for small molecules, we can convert amino acids to descriptors.



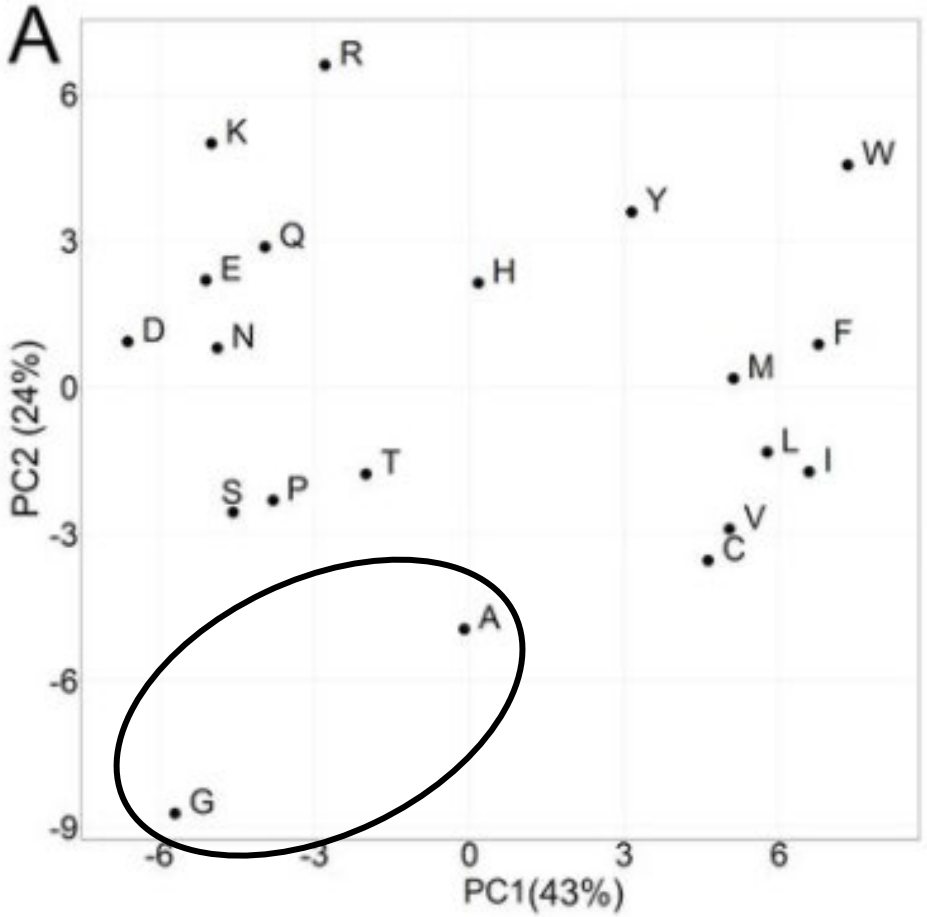
Peptide / binding site descriptors

- Similar to the fingerprints used for small molecules, we can convert amino acids to descriptors.



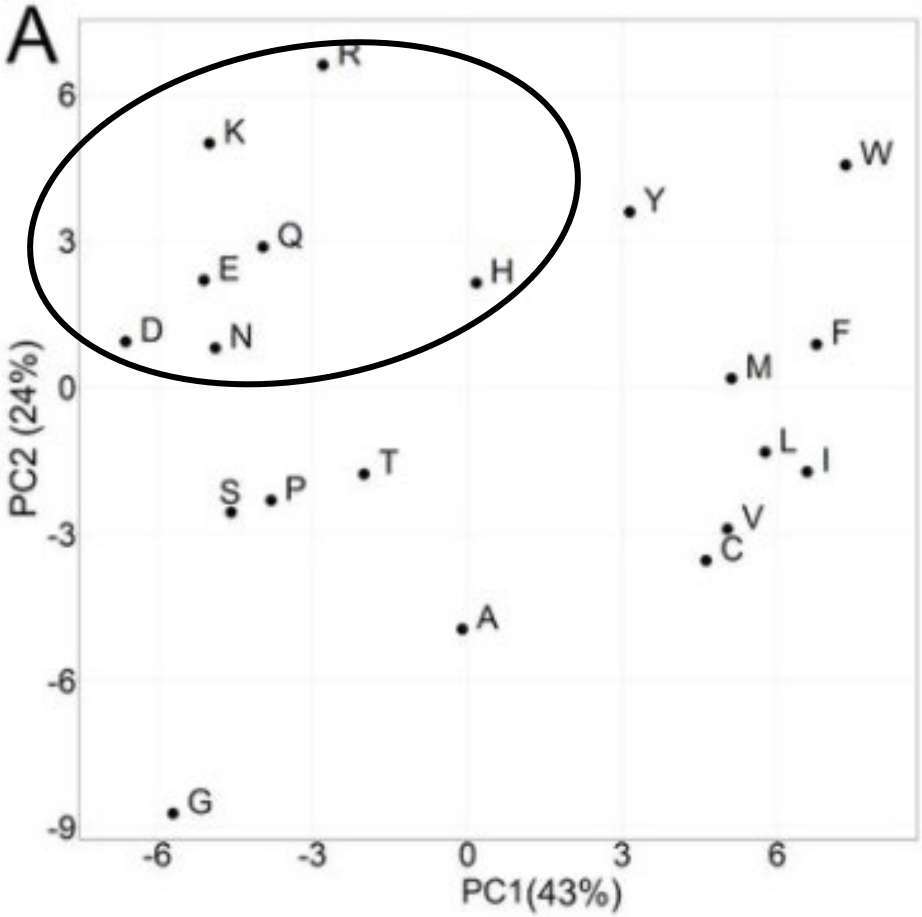
Peptide / binding site descriptors

- Similar to the fingerprints used for small molecules, we can convert amino acids to descriptors.



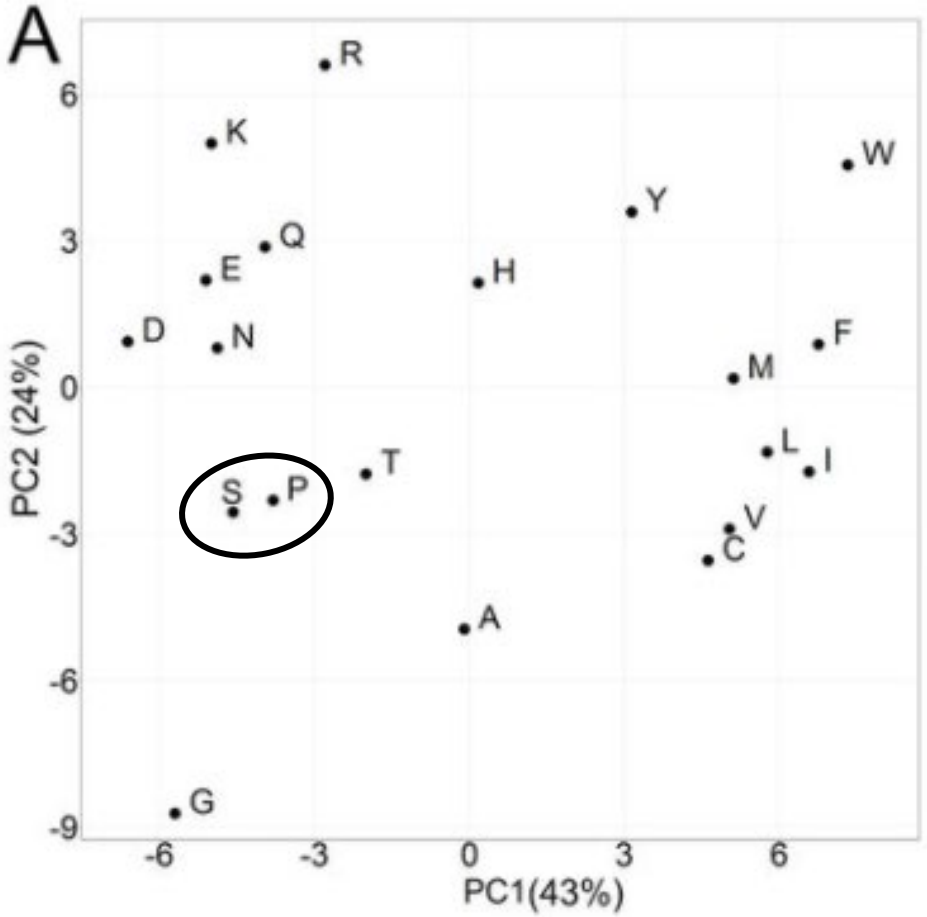
Peptide / binding site descriptors

- Similar to the fingerprints used for small molecules, we can convert amino acids to descriptors.



Peptide / binding site descriptors

- Similar to the fingerprints used for small molecules, we can convert amino acids to descriptors.



How to quantify sequence similarity

- Take the examples from just now and add 4 more sequences
  - CALHP
  - CLLHL
  - CLWHL
  - CLLYP
  - GLLWT
  - GLLYT

How to quantify sequence similarity

- Take the examples from just now and add 4 more sequences

```
CLUSTAL O(1.2.4) multiple sequence alignment


seq_E       GLLWT-
seq_F       GLLYT-
seq_D       -CLLYP
seq_A       -CALHP
seq_B       -CLLHL
seq_C       -CLWHL
```

Converting these to descriptors

Receptor	Amino_acid_sequence	Amino_acid_number	Z1	Z2	Z3
A	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	A	3	0.24	-2.32	0.6
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	P	6	-1.66	0.27	1.84
B	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	L	3	-4.28	-1.30	-1.49
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	L	6	-4.28	-1.30	-1.49

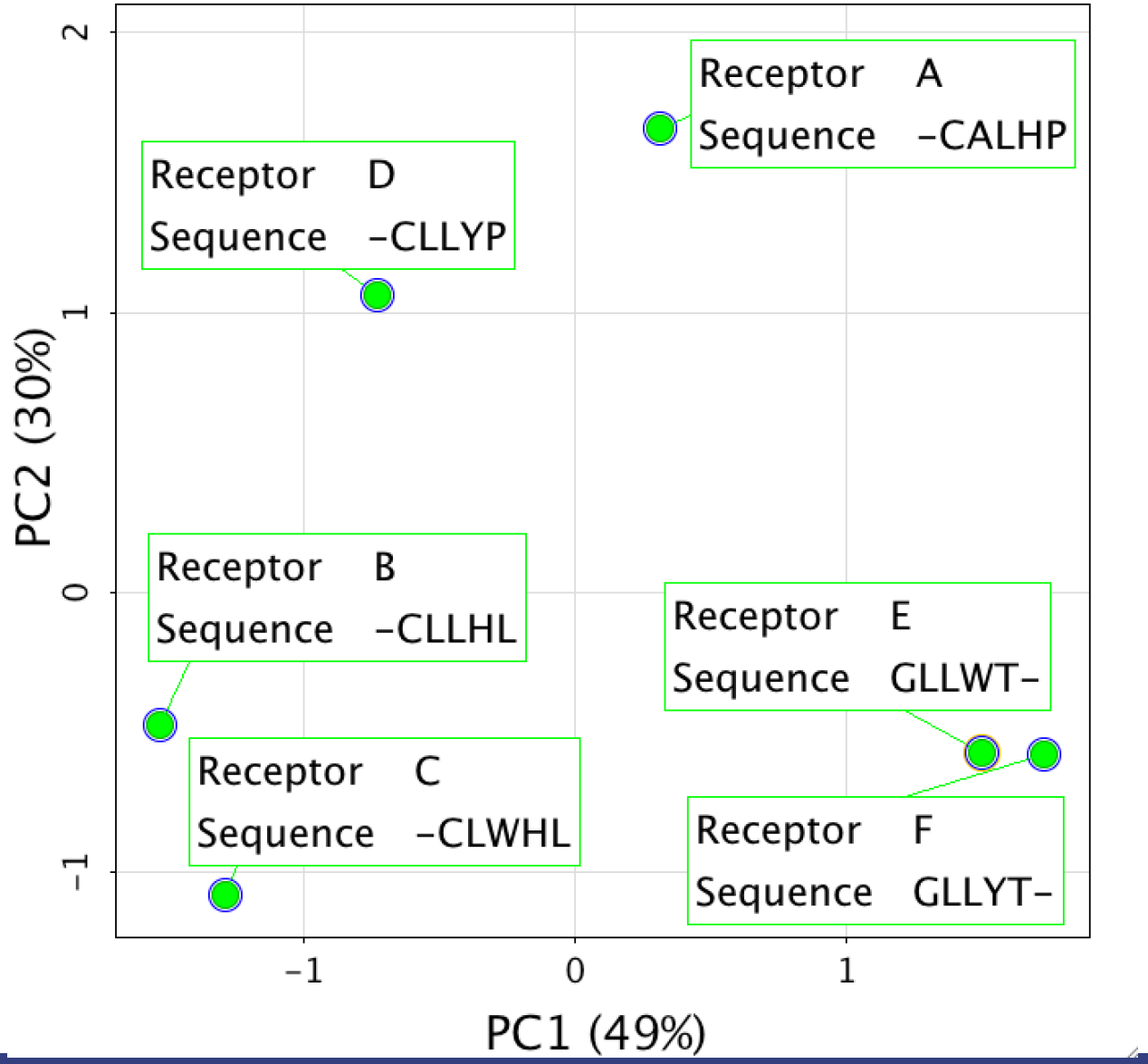
Converting these to descriptors

Receptor	Amino_acid_sequence	Amino_acid_number	Z1	Z2	Z3
A	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	A	3	0.24	-2.32	0.6
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	P	6	-1.66	0.27	1.84
B	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	L	3	-4.28	-1.30	-1.49
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	L	6	-4.28	-1.30	-1.49

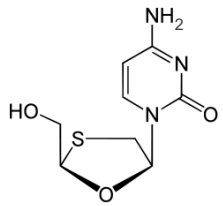
Converting these to descriptors

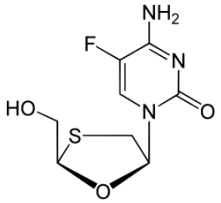
Receptor	Amino_acid_sequence	Amino_acid_number	Z1	Z2	Z3
A	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	A	3	0.24	-2.32	0.6
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	P	6	-1.66	0.27	1.84
B	-	1	0	0	0
	C	2	0.84	-1.67	3.71
	L	3	-4.28	-1.30	-1.49
	L	4	-4.28	-1.30	-1.49
	H	5	2.47	1.95	0.26
	L	6	-4.28	-1.30	-1.49

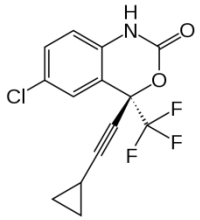
Quantify distance between sequences

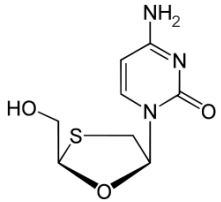


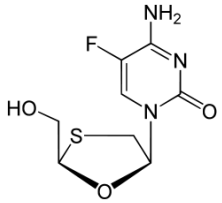
Molecular Similarity

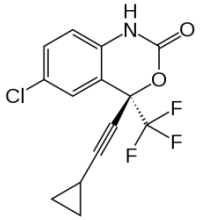






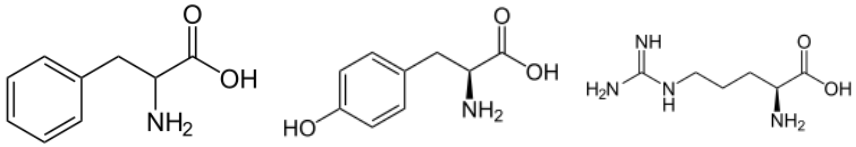


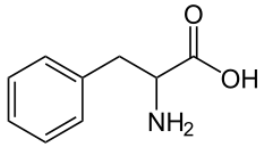
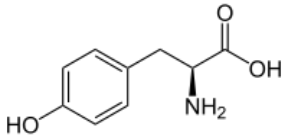
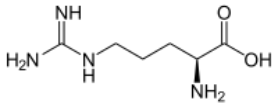




1.0	0.9	0.3
0.9	1.0	0.4
0.3	0.4	1.0

Sequence Similarity



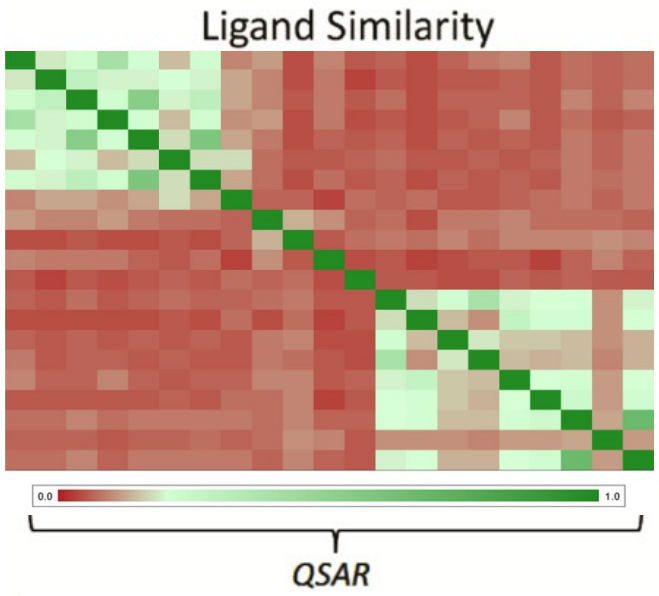
Phenylalanine 	1.0	0.9	0.3
Tyrosine 	0.9	1.0	0.4
Arginine 	0.3	0.4	1.0

Sequence Similarity

	FYI	IYF	WTF
FYI	1.0	0.9	0.3
IYF	0.9	1.0	0.4
WTF	0.3	0.4	1.0

The how… *what is PCM ?*

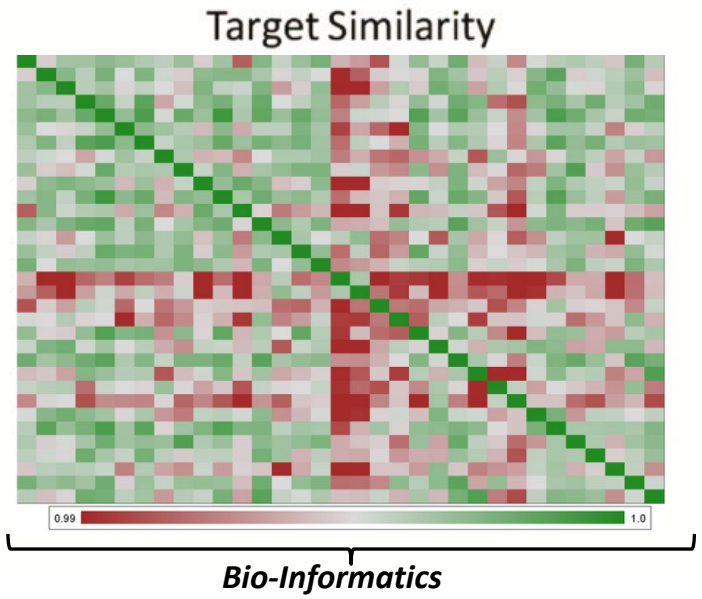
- Proteochemometric modeling combines both a ligand descriptor and target descriptor



Van Westen, Wegner *et al. MedChemComm* **(2011)**,16-30, 10.1039/C0MD00165A

What is PCM ?

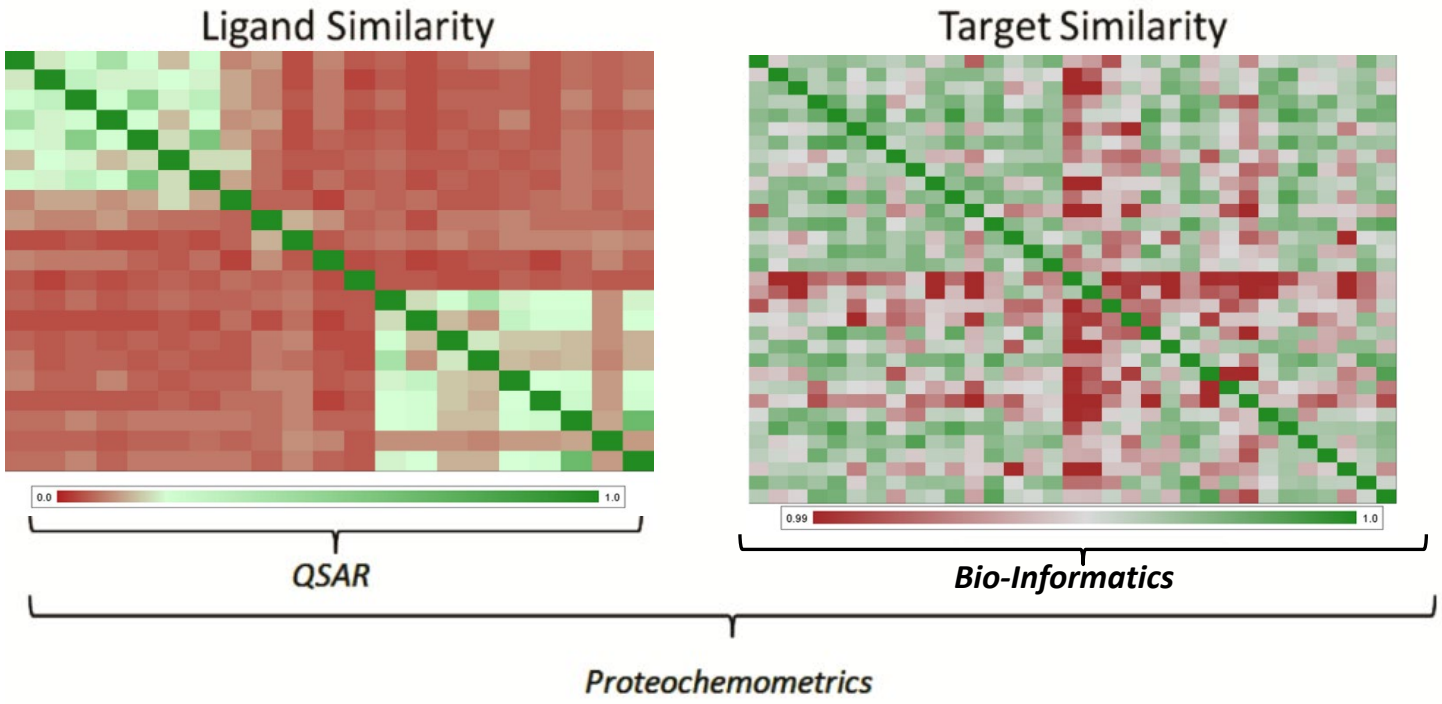
- Proteochemometric modeling combines both a ligand descriptor and target descriptor



Van Westen, Wegner *et al. MedChemComm* **(2011)**,16-30,
10.1039/C0MD00165A

What is PCM ?

- Proteochemometric modeling combines both a ligand descriptor and target descriptor



Van Westen, Wegner *et al. MedChemComm* **(2011)**,16-30, 10.1039/C0MD00165A

Describe protein properties

Describe protein properties

- Simple way to derive protein descriptors
  1. Align the relevant residues
  2. Convert to physicochemical properties

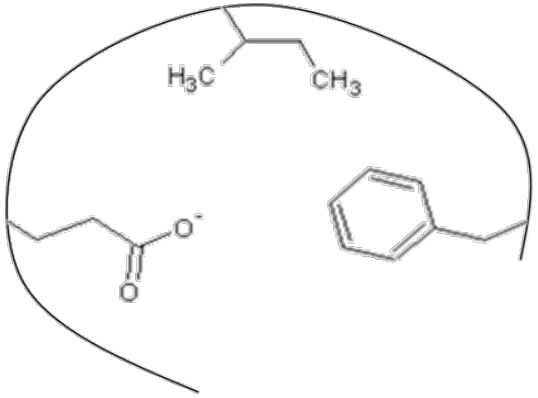
Describe protein properties

Iso-leucine:

- Lipophilicity : -3.89
- Volume : -1.73
- Polarity : -1.71
- … : …

Glutamic Acid:

- Lipophilicity : 3.11
- Volume : 0.26
- Polarity : -0.11
- … : …



Phenylalanine:

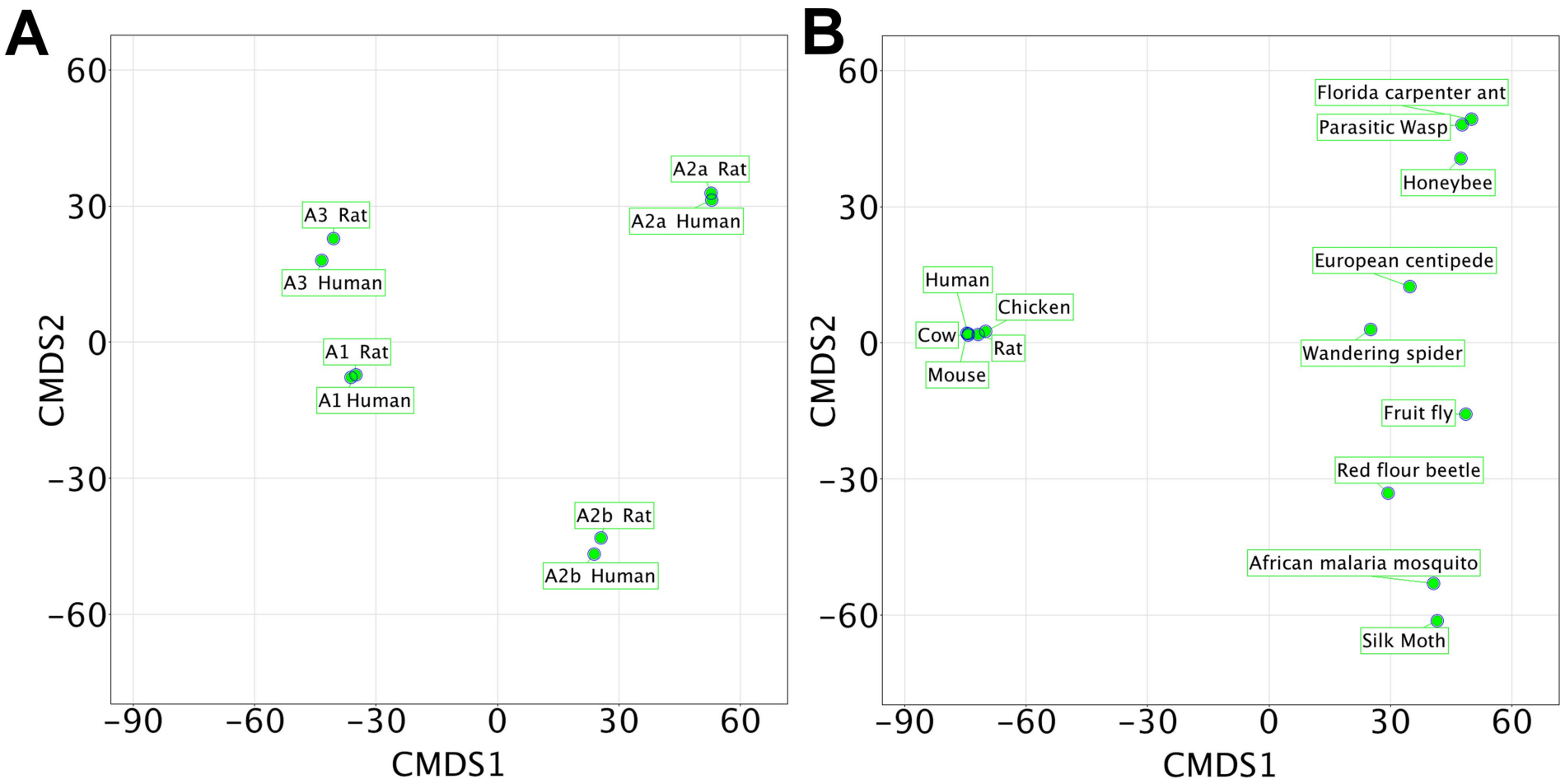
- Lipophilicity : -4.22
- Volume : 1.94
- Polarity : 1.06
- … : …

Sequence	Descriptor		
	Z1	Z2	Z3
E	3.11	0.26	-0.11
I	-3.89	-1.73	-1.71
F	-4.22	1.94	1.06
…	…	…	…

Van Westen, Swier *et al. J Cheminf* **(2013)**, 5:42, 10.1186/1758-2946-5-42

So what can we use this method for?

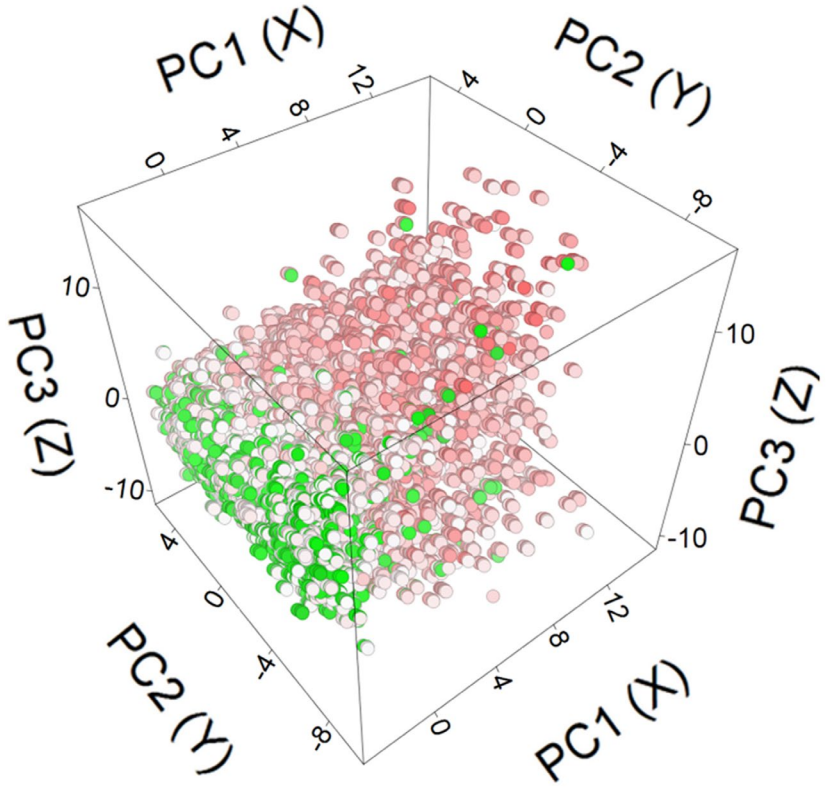
Target descriptors

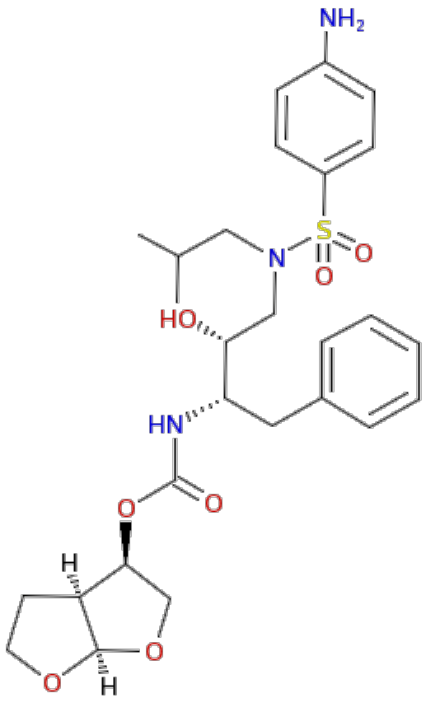


a Similarity between the different adenosine receptors in human and rat.
b Similarity between the GABA-A ligand-gated ion channels in mammals arthropods.

Van Westen, Bender, and Overington. *J Chem Biol* **(2014)**, 10.1007/s12154-014-0112-2

Prevent anti-HIV treatment failure





Sensitive **Resistant**

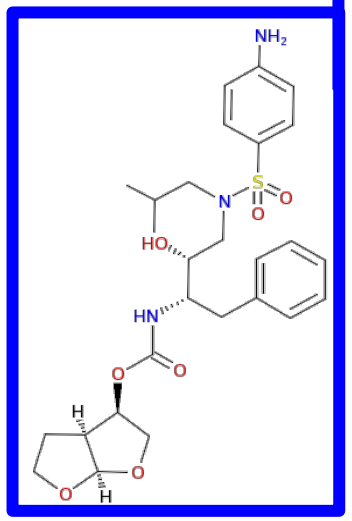
van Westen, et al., PLoS Comp Biol **(2013)**, e1002899

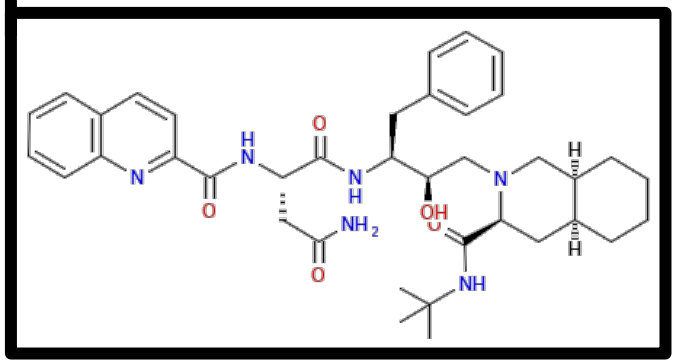
Prevent anti-HIV treatment failure

Patient
(Viral Genotype)

	APV	ATV	DRV	IDV	LPV	NFV	RTV	SQV	TPV
P81G	0.26	0.54	0.43	0.08	0.18	0.10	0.32	0.75	
V82T	0.23	0.19	-0.02	0.30	0.34	0.18	0.44	0.09	0.16
V82D	0.24	0.13	-0.29	0.53	0.79	0.18		-0.32	-0.05
V82A	0.30	0.21	0.05	0.34	0.45	0.24	0.56	0.09	0.09
I84C	0.81	0.37	0.36	0.62	0.14	1.00	0.74	1.45	0.71
R87K	-0.20	0.28	0.26	-0.16	-0.16	-0.14	-0.44	-0.05	1.15
N88S	0.22	0.54	-0.27	0.42	0.06	0.66	0.09	0.30	0.05
N88G	0.39	0.97	-0.27	0.77	0.18	1.14	0.10	0.49	0.22
L90M	0.33	0.27	0.11	0.29	0.22	0.42	0.46	0.40	0.06
I93L	0.02	0.01	0.00	0.06	-0.01	0.07	-0.02	0.00	0.01
G94N	-0.12	-0.08	0.60	-0.25	-0.02	-0.11	0.01	-0.20	-0.35

-0.94 1.45

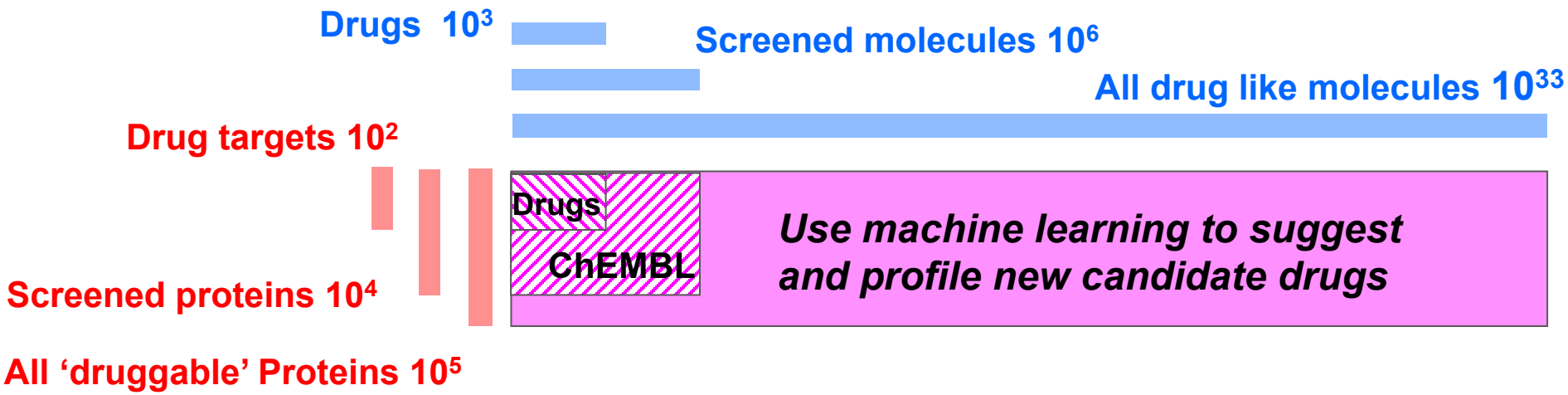




van Westen, et al., PLoS Comp Biol ***(2013)****, e1002899*

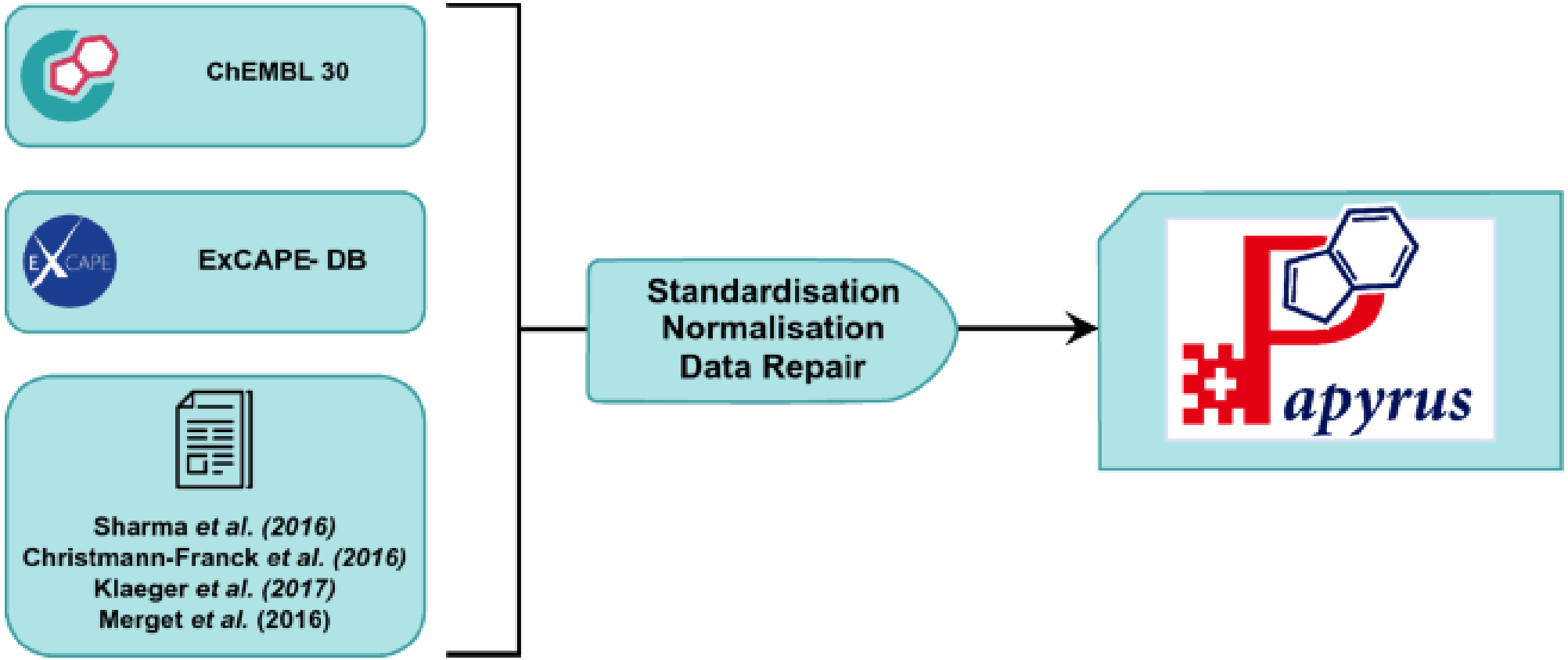
de novo generation

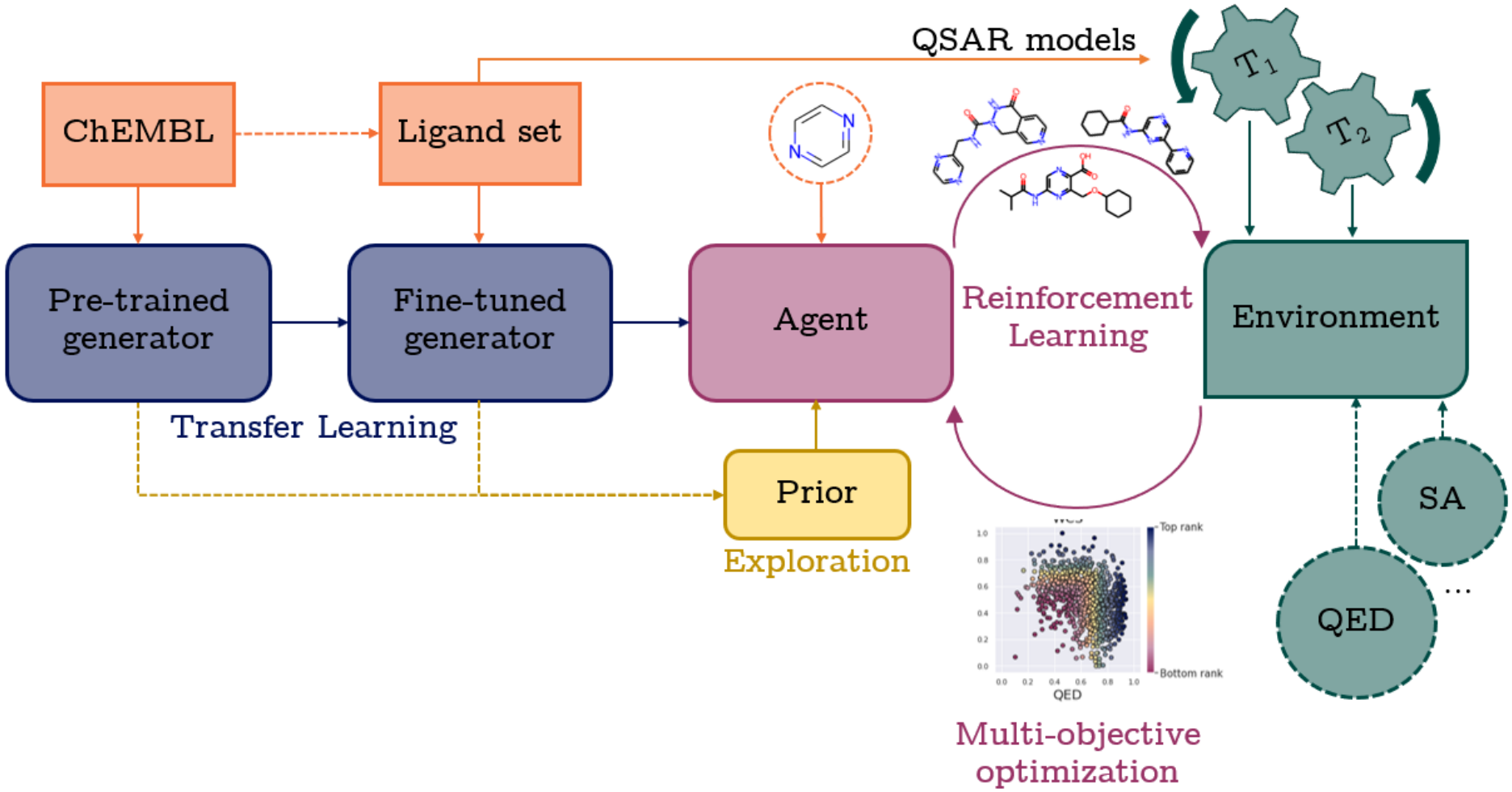
Learning a machine the grammar of molecules..

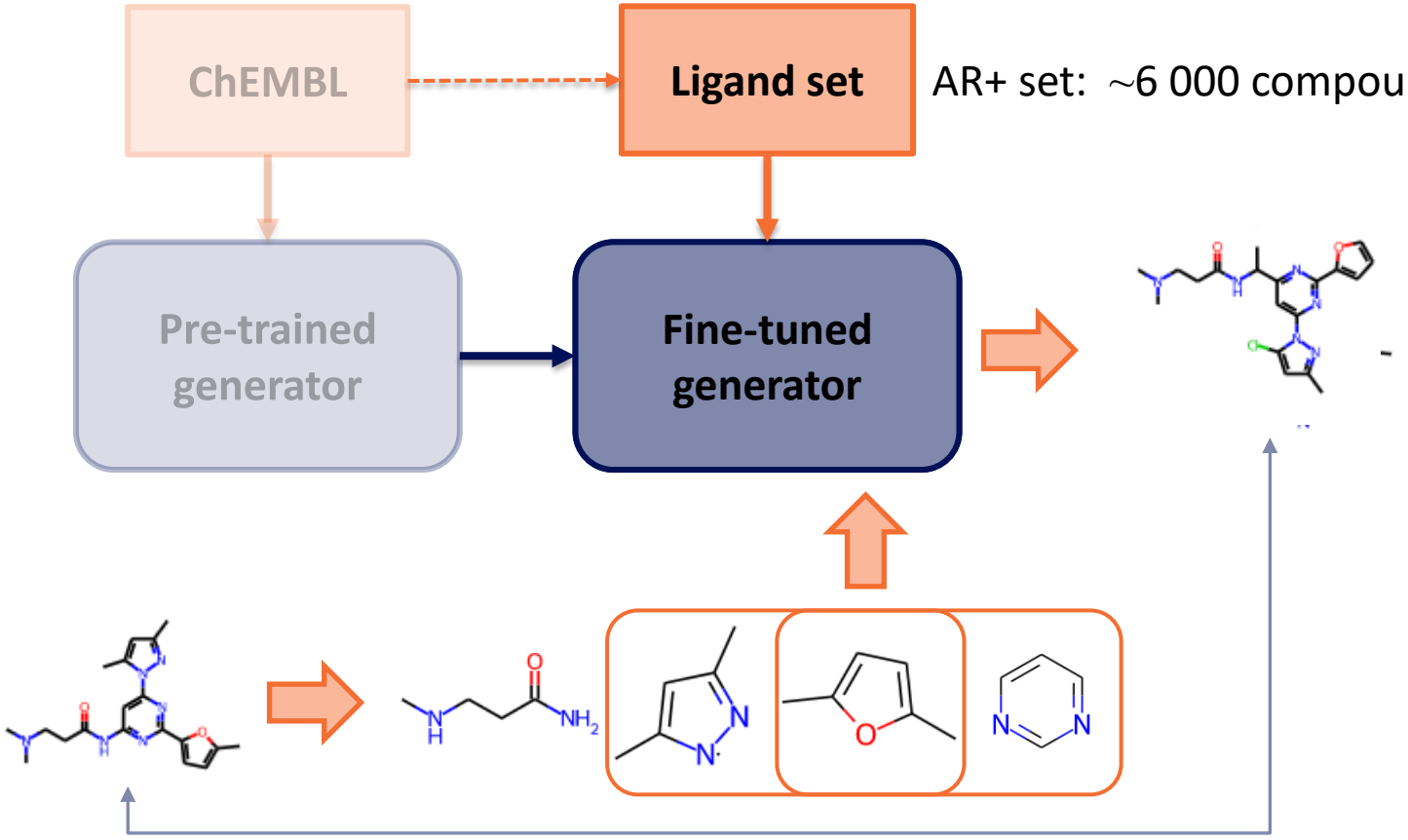


[1] Reymond, J.-L.; (**2015**) Acc. Chem. Res., 48, 722.
[2] Berman, H.M.; et al. (**2012**) Structure, 20, 391.
[3] Polishchuk, P.G.; (2013) J. Comput. Aided. Mol. Des., 27, 675 And J.P. Overington

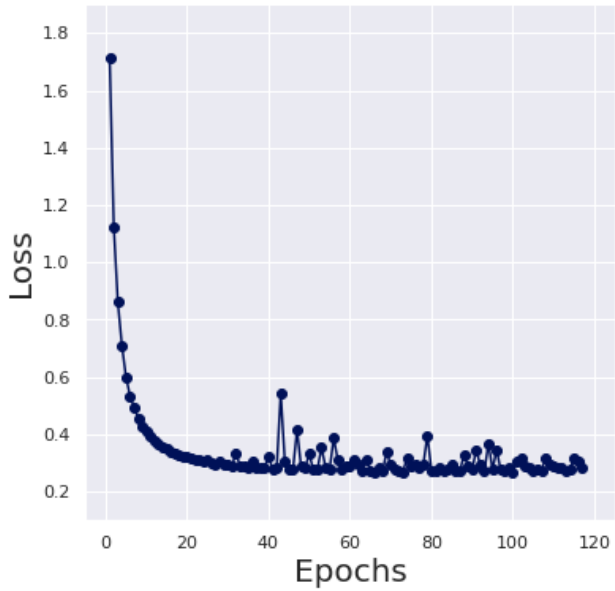
Input data: Papyrus



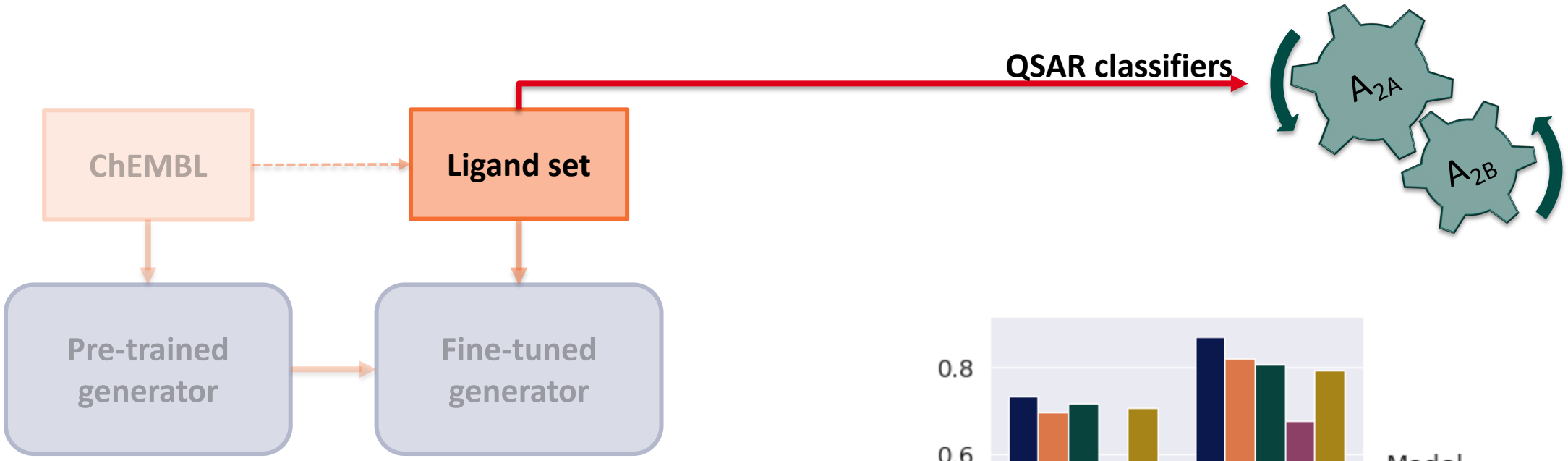

QSAR models
ChEMBL
Ligand set
Pre-trained generator
Fine-tuned generator
Agent
Reinforcement Learning
Environment
Transfer Learning
Prior
Exploration
SA
QED
...
Top rank
Bottom rank
QED
Multi-objective optimization



AR+ set: ~6 000 compounds active on at least one of the AR



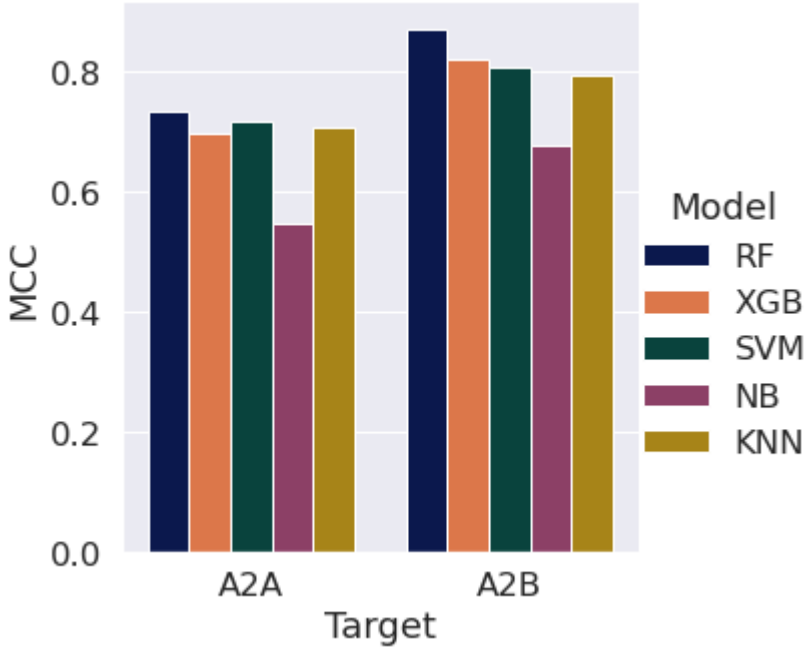
QSAR: quantitative structure-activity relationship

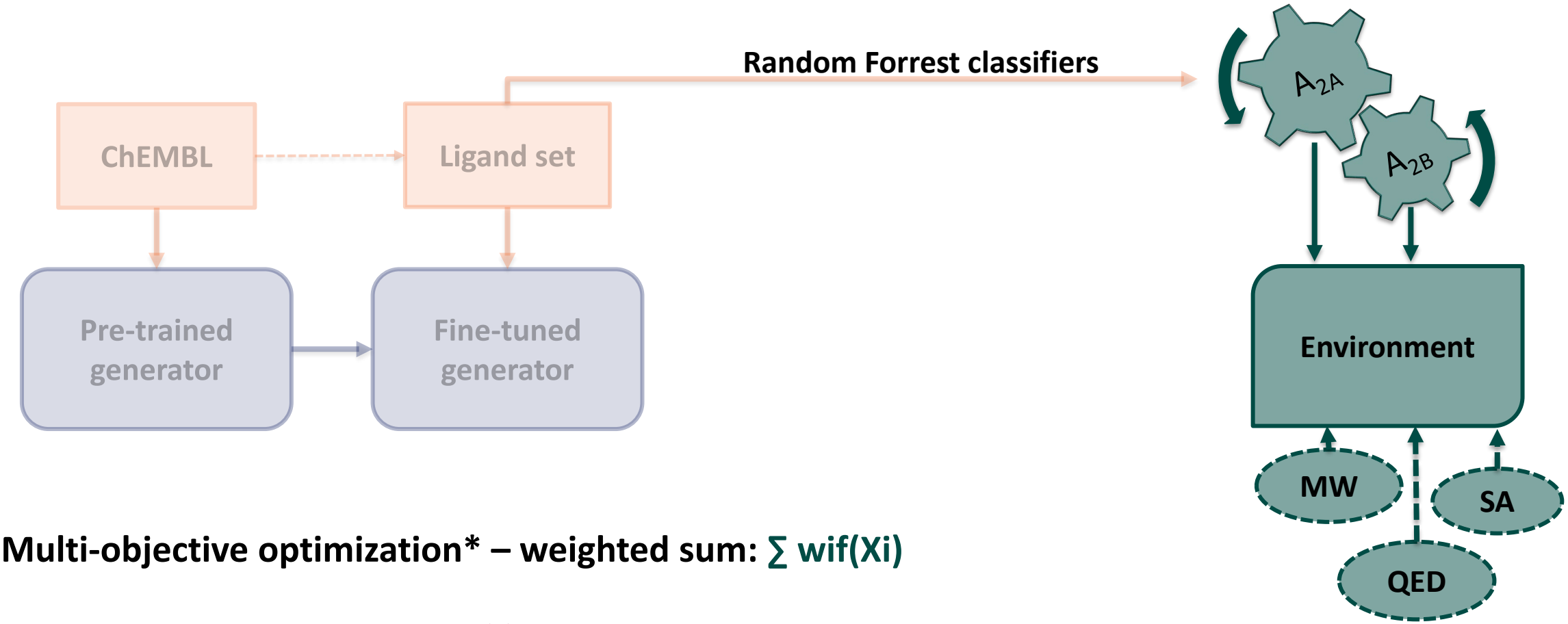


	Exp +	Exp −
Pred +	45	20
Pred −	5	30

+ active
− inactive

→ MCC ∈ [-1,1]

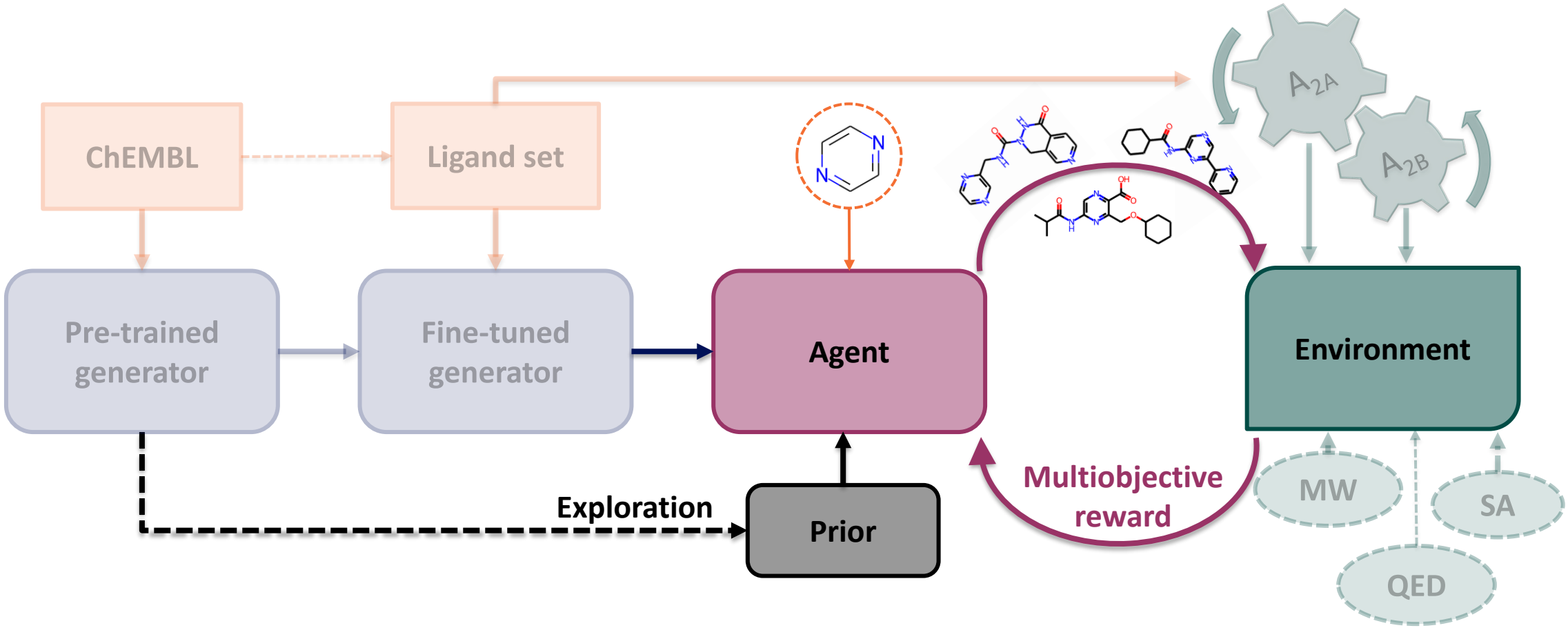




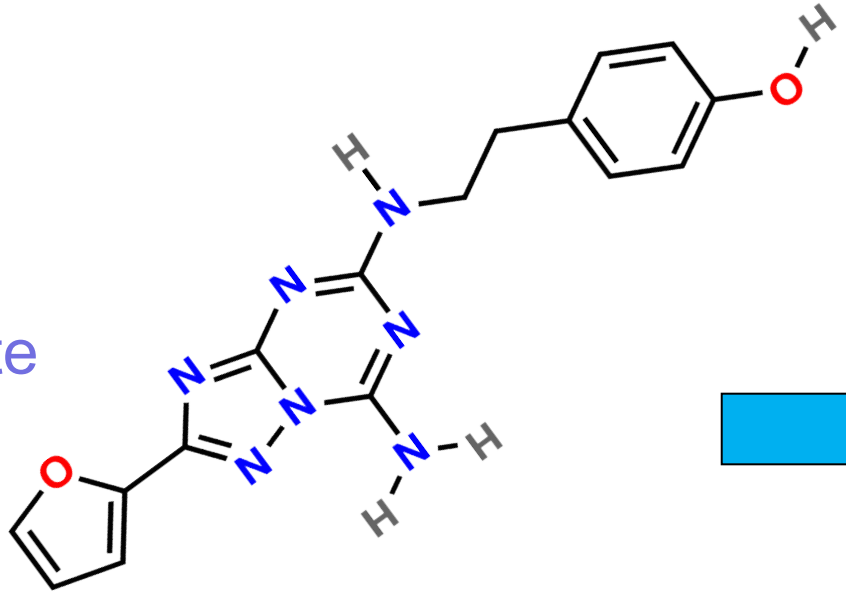
Multi-objective optimization* – weighted sum: ∑ wif(Xi)

f(X) – clips values between 0 (✗) and 1 (✓)

MW – Molecular weight : < 400 Da
SA – Synthetic accessibility
QED – Quantitative estimate of drug-likeness


ChEMBL
Ligand set
Pre-trained generator
Fine-tuned generator
Agent
Environment
Exploration
Prior
Multiobjective reward
A2A
A2B
MW
SA
QED

1 - Translate

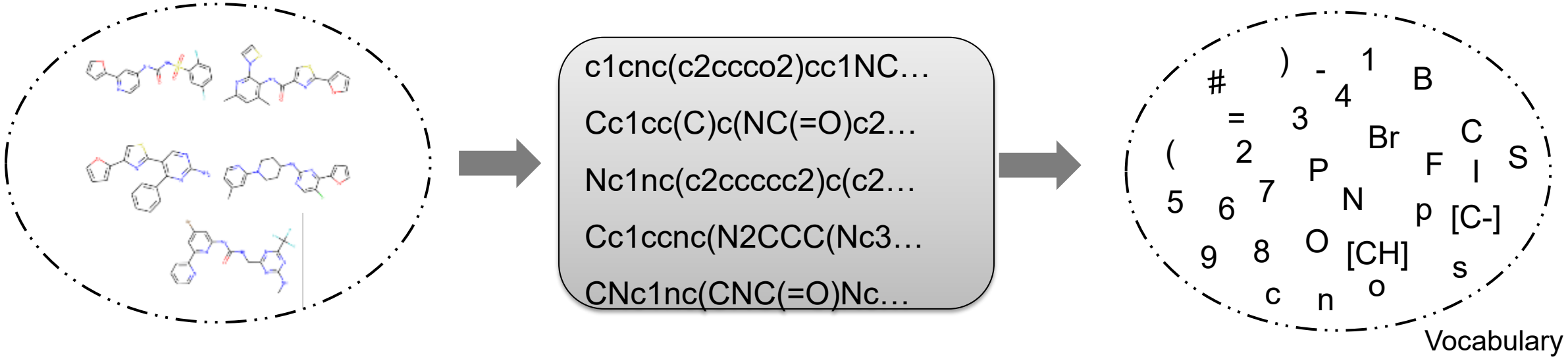


ZM241385

Nc1nc(NCCc2ccc(O)cc2)nc3nc(nn13)c4occc4

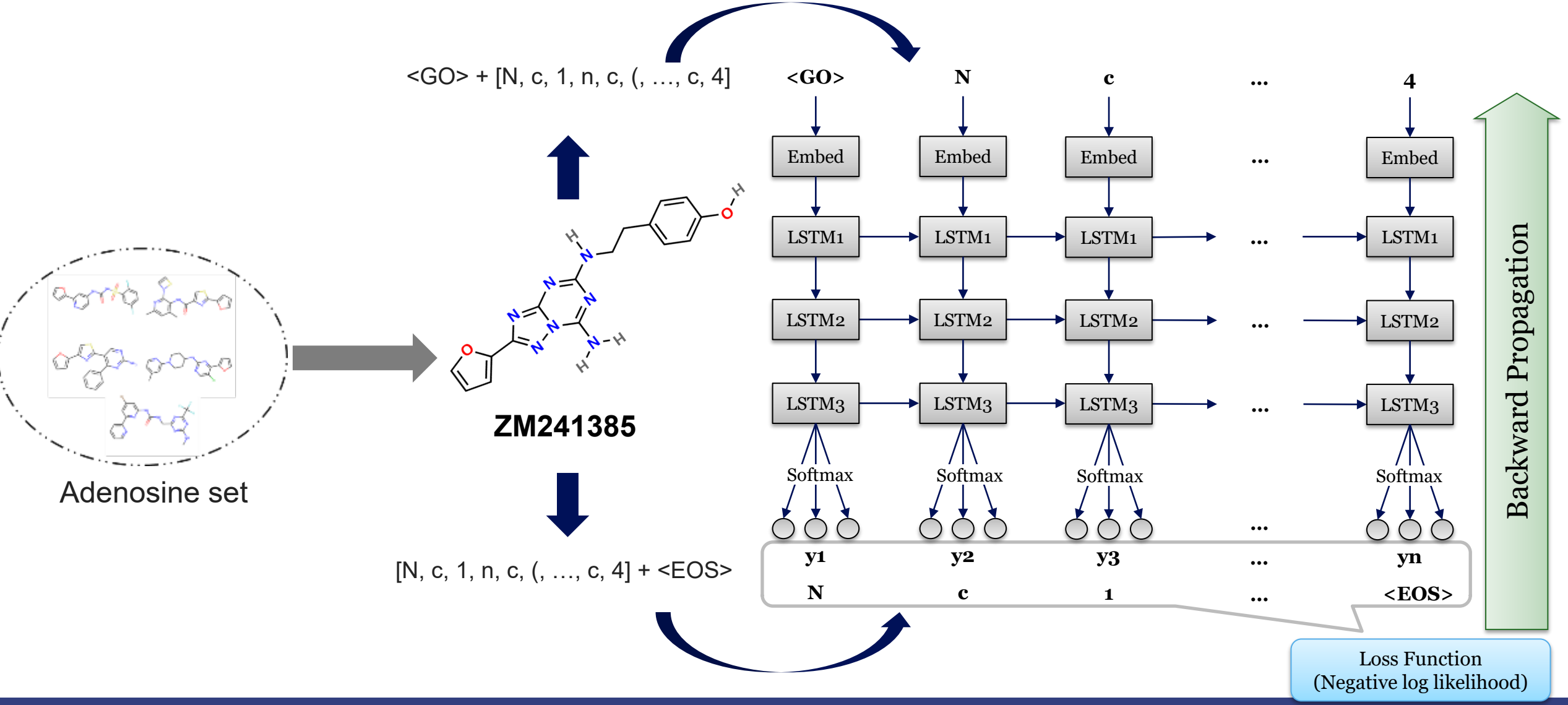
Adenosine dataset

- All public compounds tested on the adenosine receptors ChEMBL (v 24).



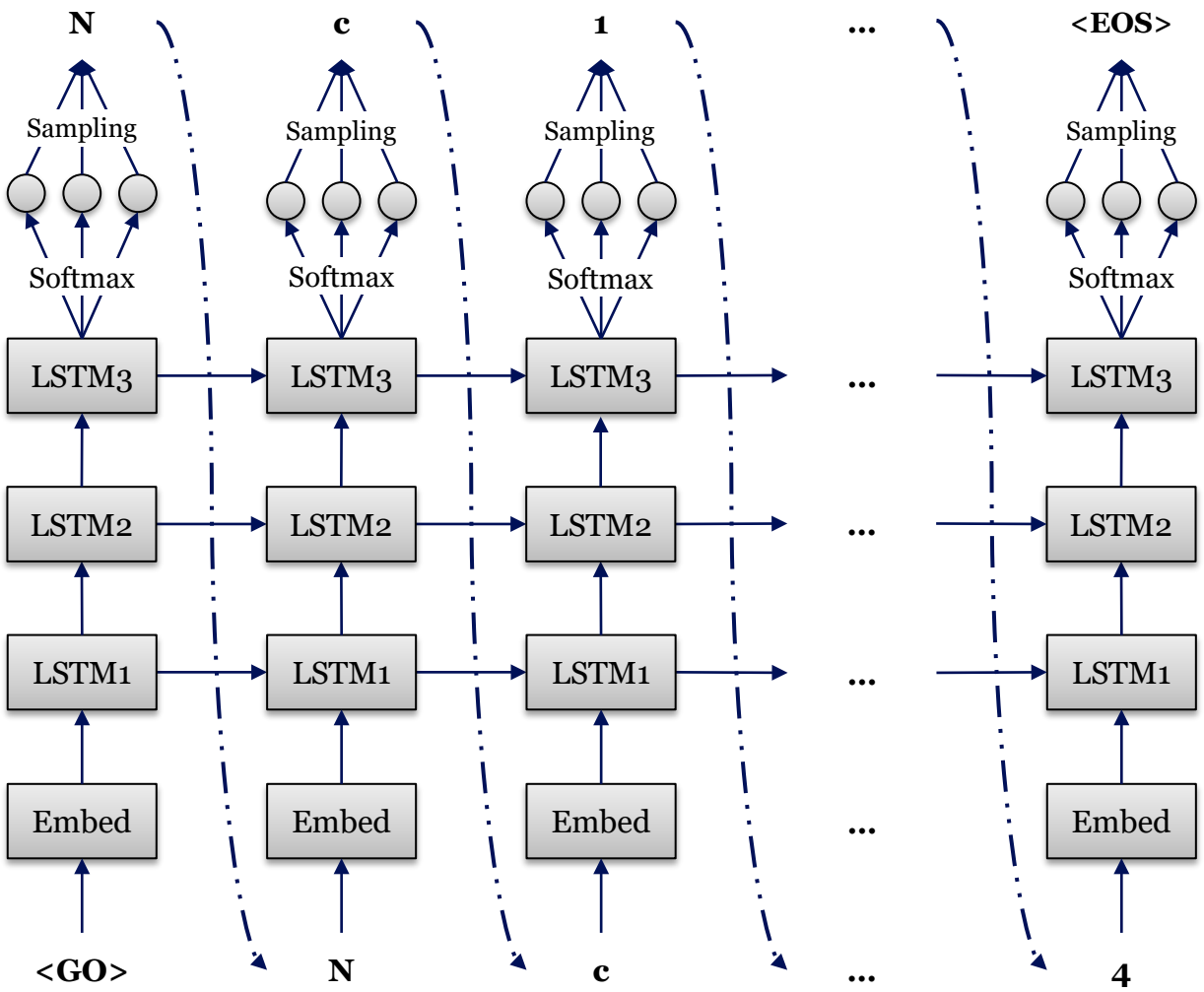
Liu, X. et al, (**2019**), J Cheminf, 10.1186/s13321-019-0355-6

RNN Training

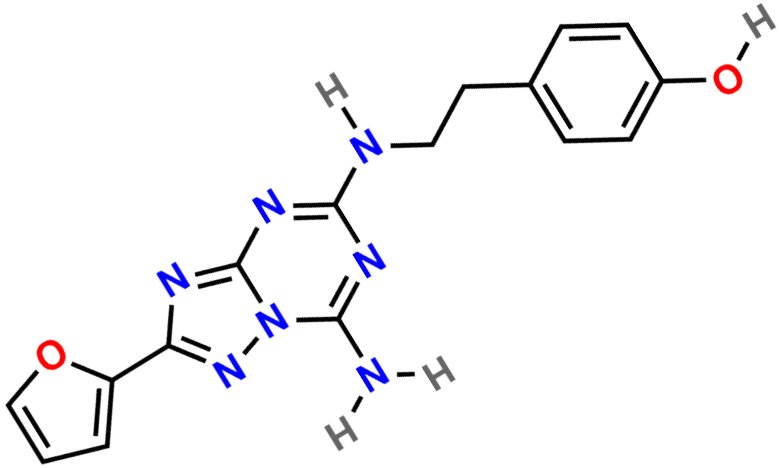


Liu, X. et al, (**2019**), J Cheminf, 10.1186/s13321-019-0355-6

Molecule Generation



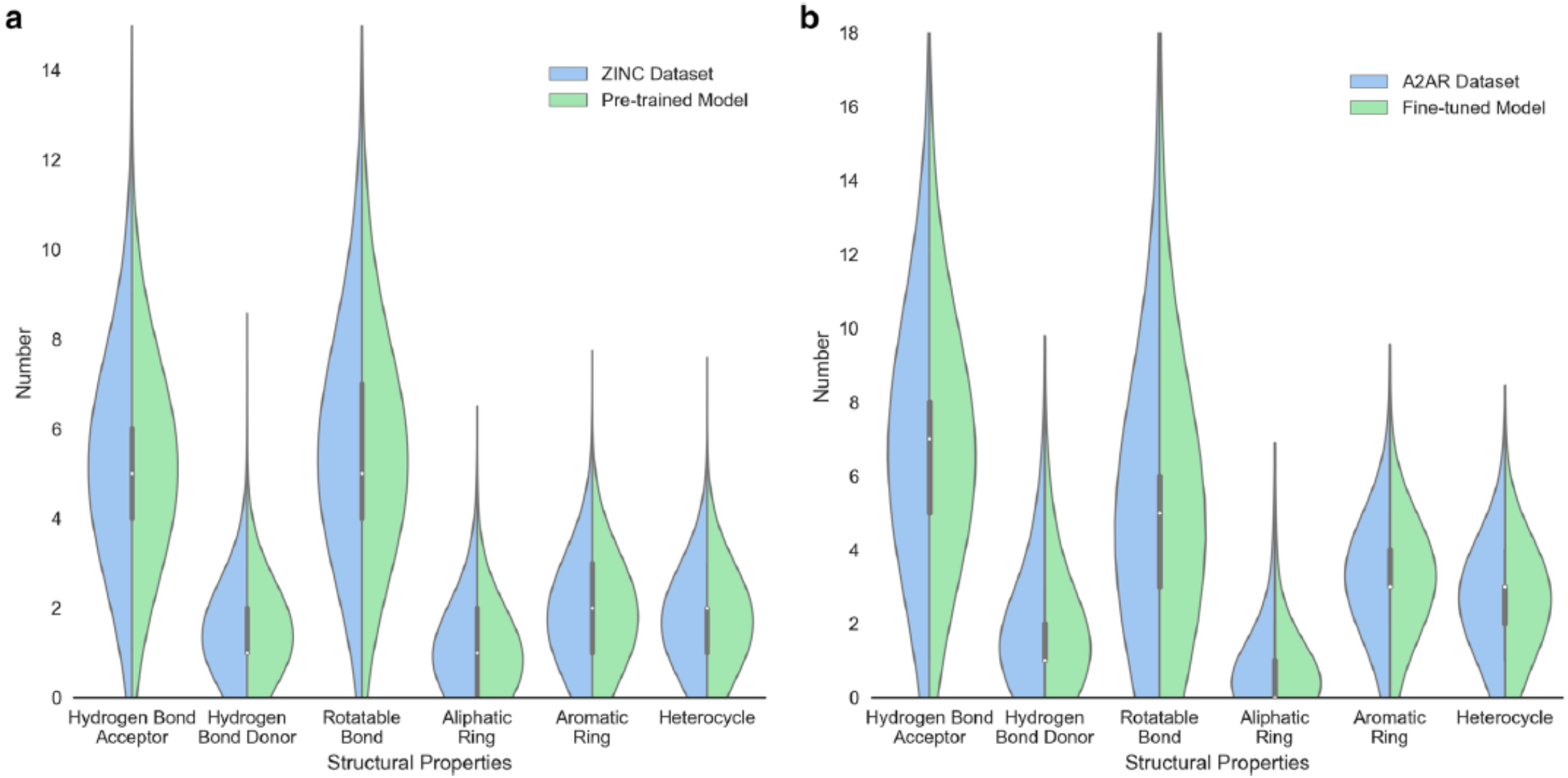
Nc1nc(NCCc2ccc(O)cc2)nc3nc(nn13)c4occc4



ZM241385 ($C_{13}H_{16}N_2O_2$)

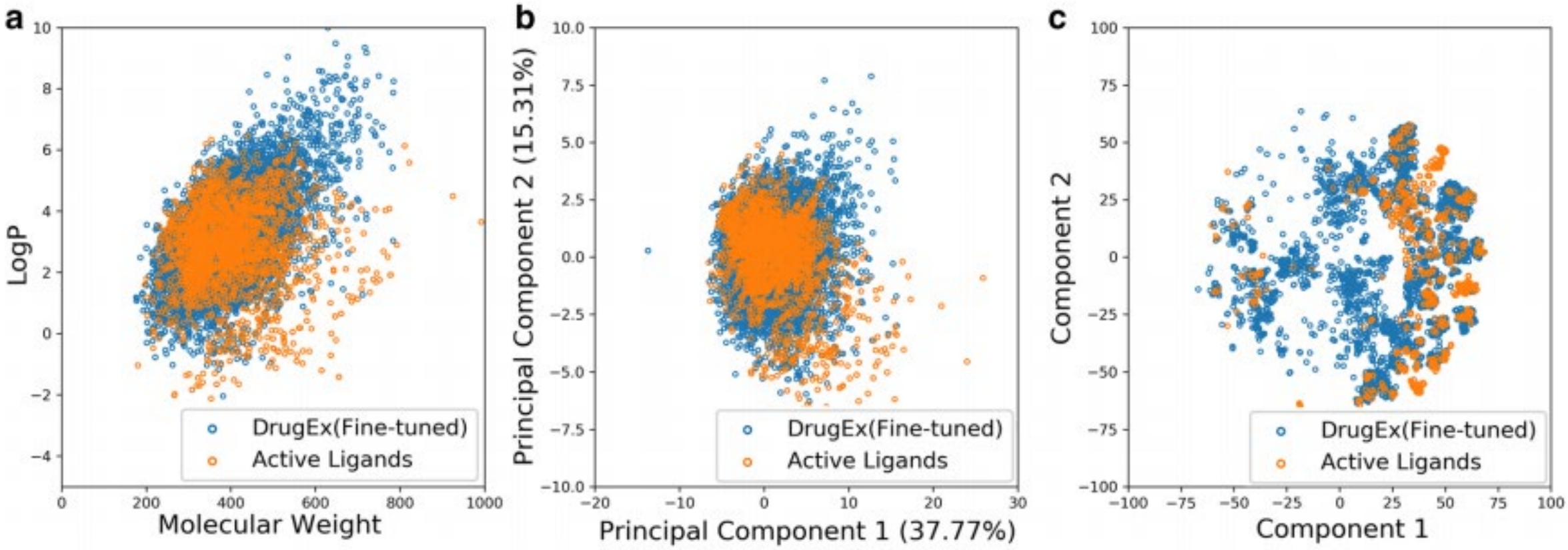
Liu, X. et al, (**2019**), J Cheminf, 10.1186/s13321-019-0355-6

Generated molecules chemically similar



Liu, X. et al, (**2019**), J Cheminf, 10.1186/s13321-019-0355-6

New A2A ligands

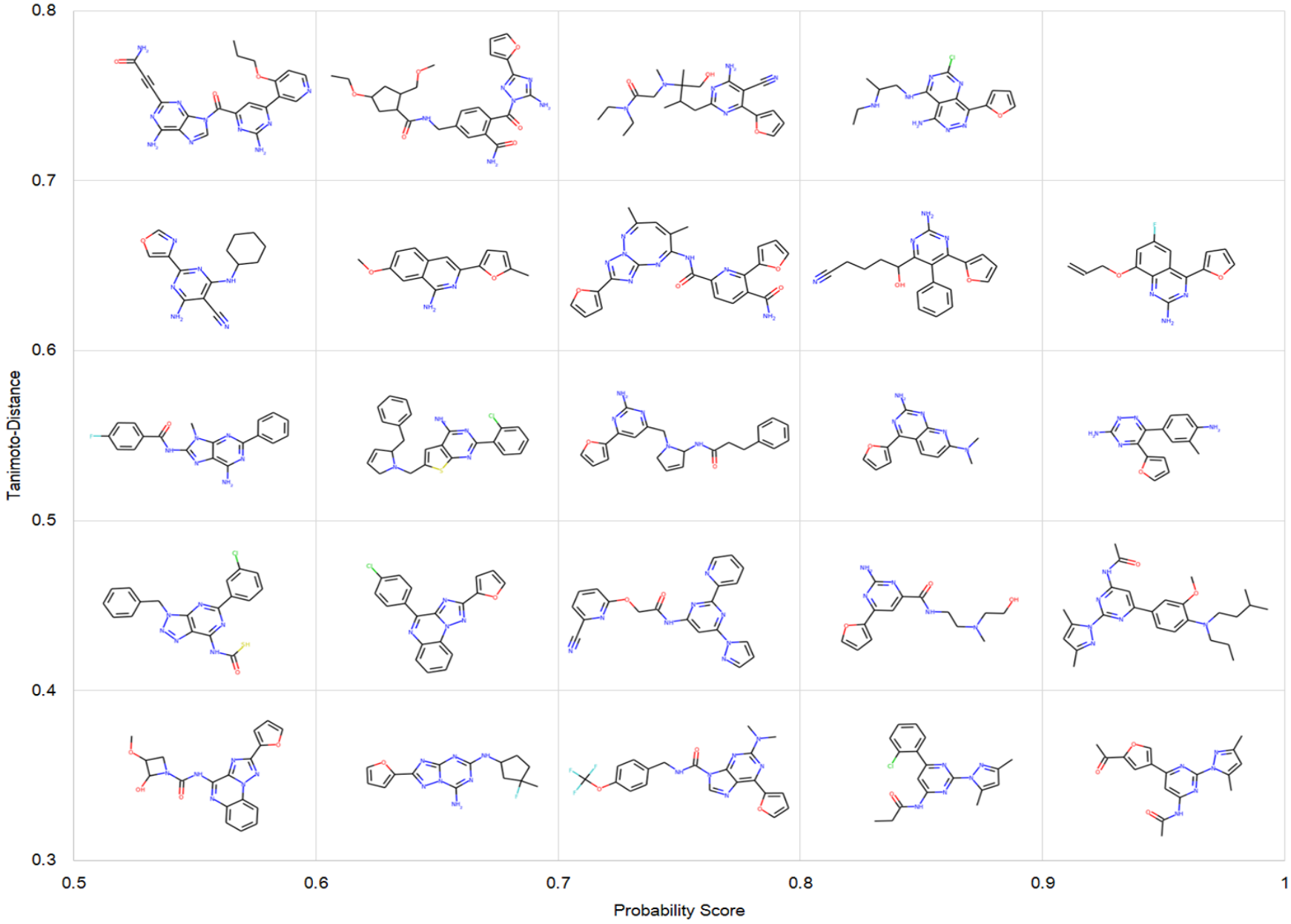


logP~MW **PCA (PhysChem)** **t-SNE (Fingerprints)**

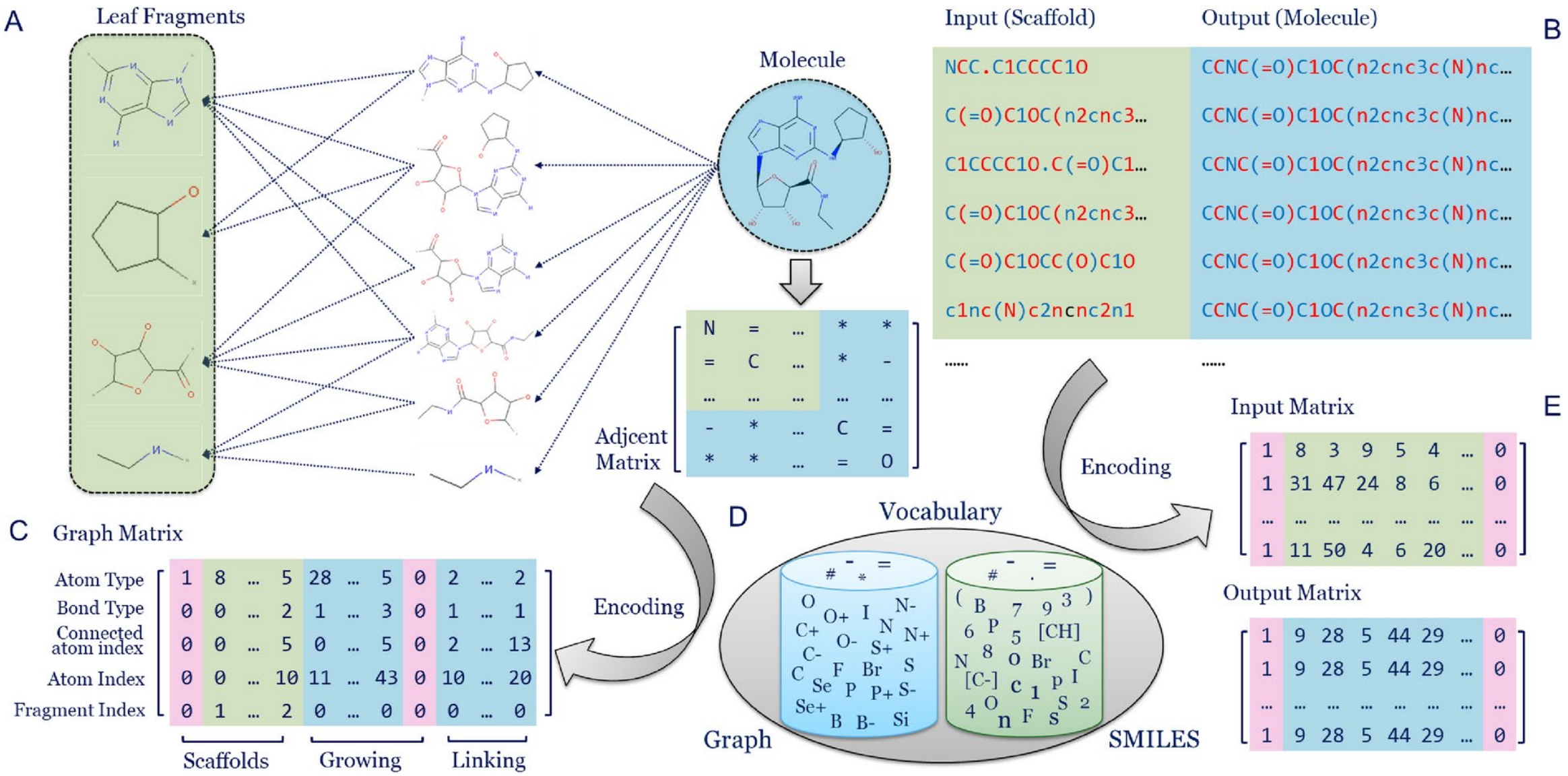
Liu, X. et al, (**2019**), J Cheminf, 10.1186/s13321-019-0355-6

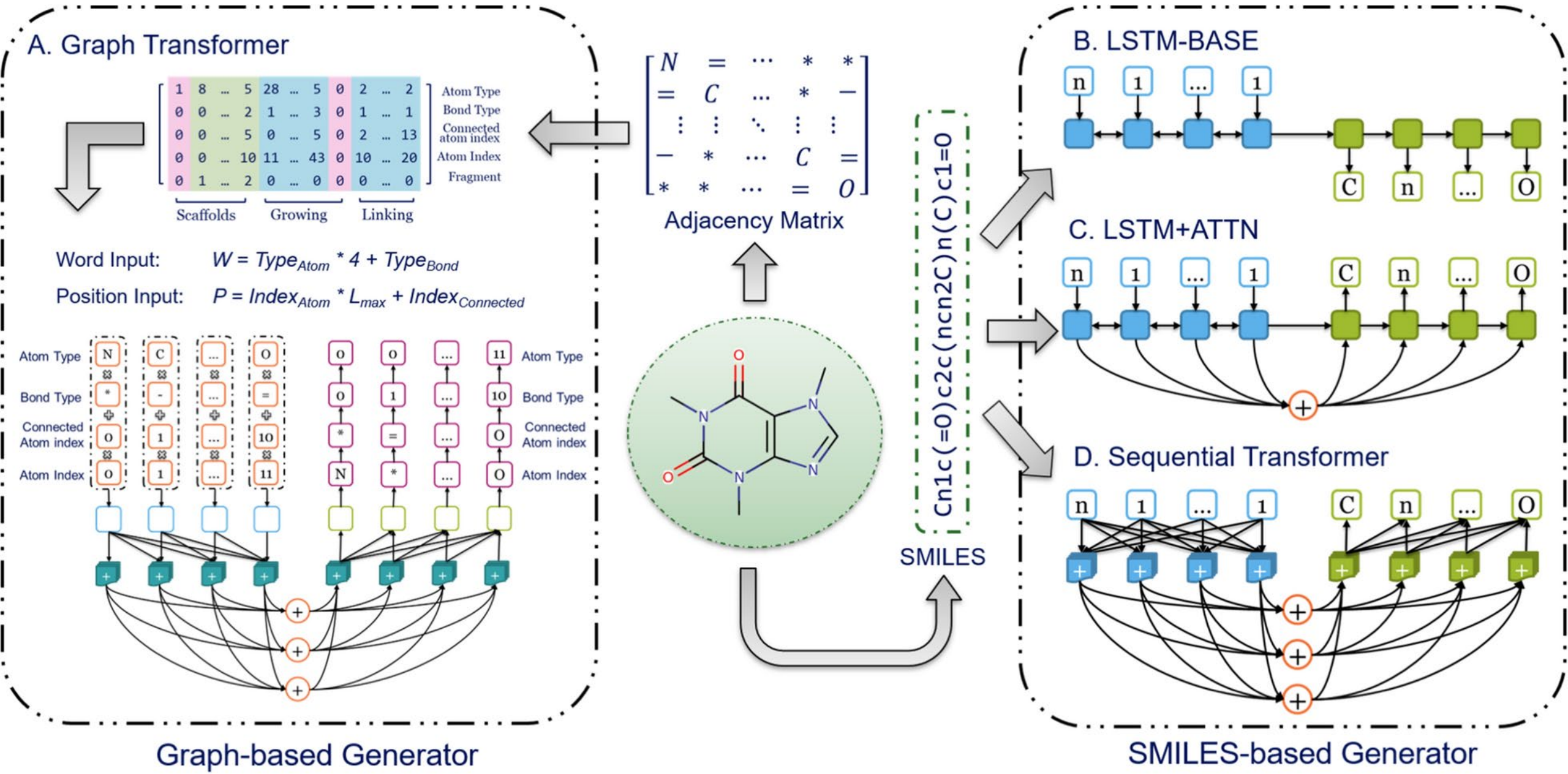
Also more complex chemical features are generated

		Fused Ring	Furan Ring	Benzene Ring
DrugEx (Pre-trained)		9.12%	82.32%	61.48%
DrugEx (Fine-tuned)		**60.69%**	**66.35%**	**65.62%**
REINVENT		0.20%	95.26%	61.98%
ORGANIC		0.02%	99.96%	39.45%
Pre-trained		24.22%	4.51%	63.31%
Fine-tuned		76.33%	23.82%	72.85%
ZINC		26.66%	3.86%	63.97%
A2AR	Active	79.09%	40.29%	75.33%
	Inactive	76.73%	9.33%	70.88%


Tanimoto-Distance
0.8
0.7
0.6
0.5
0.4
0.3
Probability Score
0.5
0.6
0.7
0.8
0.9
1

Graph based (this just in)




A. Graph Transformer
Atom Type
Bond Type
Connected atom index
Atom Index
Fragment
Scaffolds
Growing
Linking
Word Input: $W = Type_{Atom} * 4 + Type_{Bond}$
Position Input: $P = Index_{Atom} * L_{max} + Index_{Connected}$
Adjacency Matrix
SMILES
Cn1c(=O)c2c(ncn2C)n(C)c1=O
B. LSTM-BASE
C. LSTM+ATTN
D. Sequential Transformer
Graph-based Generator
SMILES-based Generator

Ongoing work

- Select molecules for follow up by synthesis

- Multi-objective ranking to identify best molecules from batch of ~ 10.000
  - Maximize on-target affinity (e.g. adenosine A2A)
  - Minimize off-target affinity (e.g. hERG, other adenosine receptors)
  - Maximize Quantitative Estimation of Drug-likeness (QED)
  - Estimate synthetic accessibility (currently SA Score)

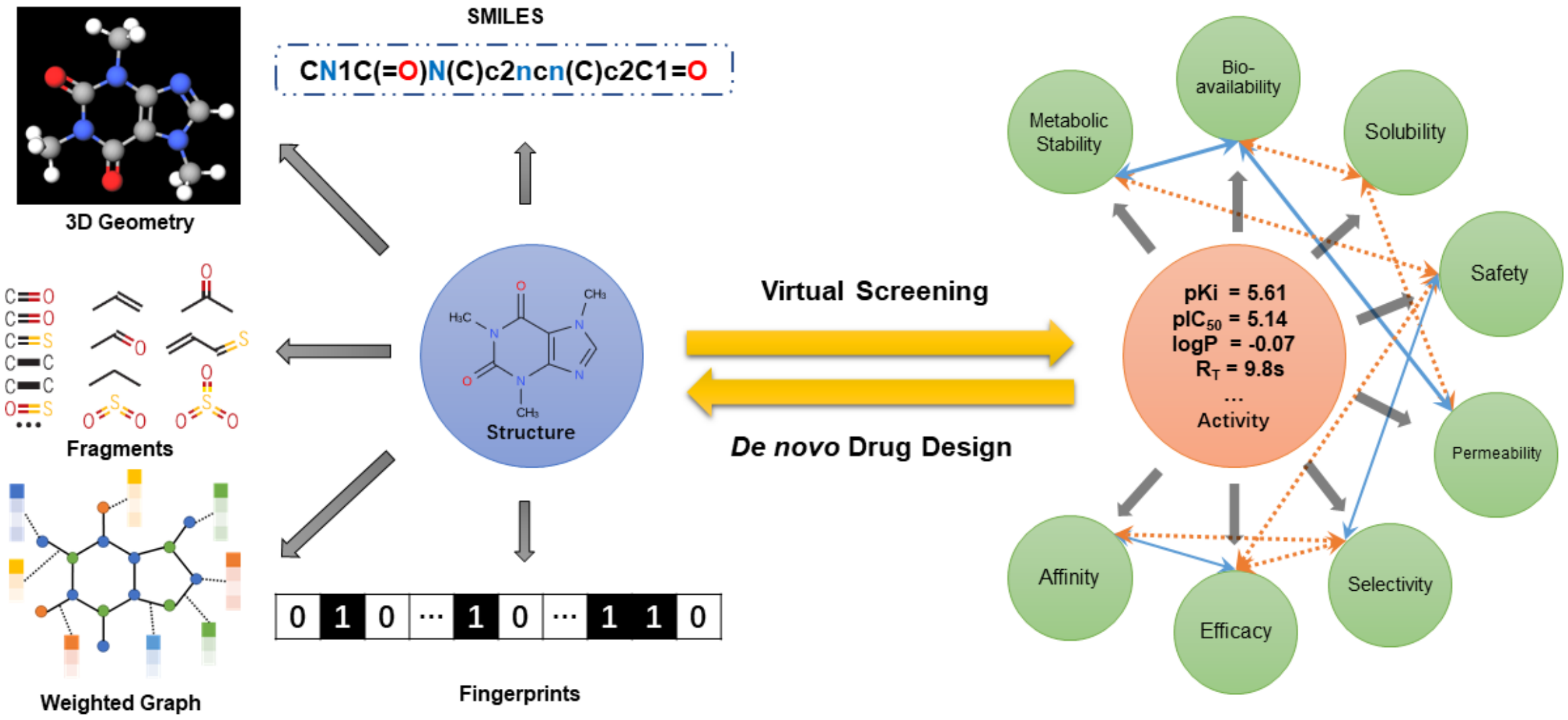
Take home messages de novo generation

- Machine learning can be used to learn the grammar of molecules in SMILES. It does this by learning the probabilities that in SMILES certain atoms follow other atoms in a sequence.

- After training the algorithm can be used to suggest new molecules that resemble the training set but are not the same.

- The power is in the numbers, 10, 100, 1000000 molecules can be generated

AI approaches in a ligand based world..



Liu, et al, Artificial Neural Networks, Methods in Molecular Biology, (2012), /10.1007/978-1-0716-0826-5_6

Proteochemometrics (PCM) de novo generation (DrugEx)

Willem Jespers