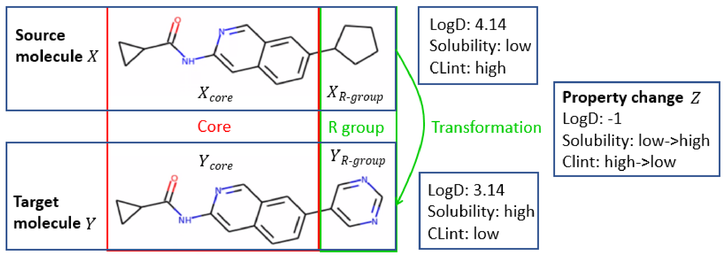
Transformer Neural Network for Structure Constrained Molecular Optimization

Jiazhen He[1] (jiazhen.he@astrazeneca.com), Felix Mattsson[1], Marcus Forsberg[1],
Esben J. Bjerrum[1], Ola Engkvist[1], Eva Nittinger[2], Christian Tyrchan[2] and Werngard Czechtizky[2]

[1]Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden [2]Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I) BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

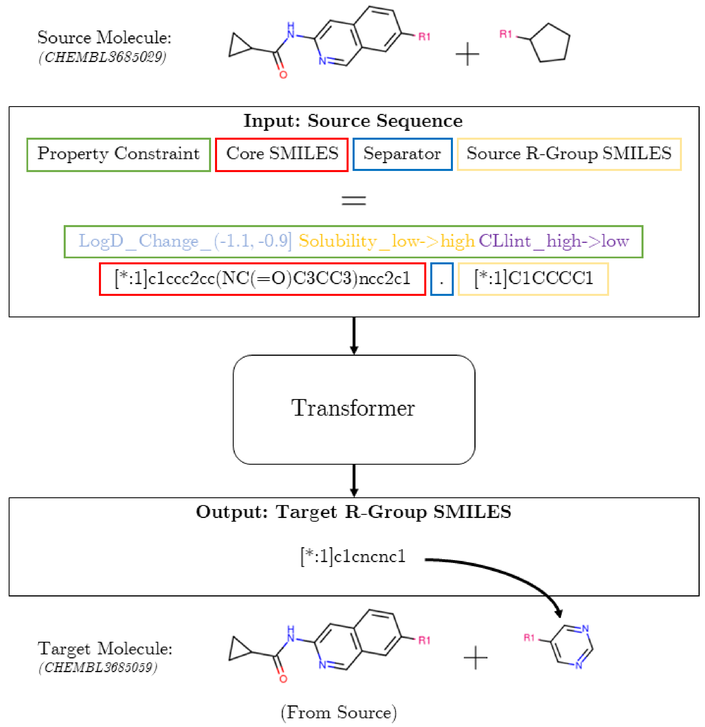
Introduction

- A drug requires **a balance of multiple properties**, *e.g.*, physicochemical properties, ADMET (absorption, distribution, metabolism, elimination and toxicity) properties, safety and potency against its target.
- **Challenge**: large chemical space (*i.e.*, 10^{23}-10^{60}) to search and explore.
- **Molecular optimization**: a promising molecule needs to be optimized towards desirable properties.
- **Chemist's approach**: matched molecular pairs (MMPs)
  - Keeping the core constant, substituting the R-group



- **Aim**: train a Transformer model to mimic the chemist's approach of using MMPs for molecular optimization.

Methods



- Molecules are represented by SMILES [1] strings.
- Three ADMET properties, *logD*, *solubility* and *clearance* are optimized simultaneously.
- Our model, **Transformer-R**
  - is trained on a set of MMPs extracted from ChEMBL together with the property changes between source and target molecules.
  - generates **R-groups** given the starting molecule (with core and R-group specified) and the specified desirable properties.
- The **goal** is to generate molecules which
  - have the desirable properties specified in the input (see **Desirable** metric in Results)
  - have small and single transformation applied to the starting molecule (see **MMP33** metric in Results)
  - keep the core specified in the input (see **Unchanged Core** metric in Results)

Results

Baselines for comparison of our model **Transformer-R** performance: **1) Transformer**: generates the whole molecule at once [2] and **2) Enumeration**: full enumeration of all R-groups extracted from ChEMBL MMP data set.

Test set	Method	Metric				
		Desirable	MMP33	Unchanged Core	Unseen Trans.	Novel R-groups
Test-Original	Transformer-R	**58.97%**	**97.67%**	**100.00%**	53.92%	**4.30%**
	Transformer	56.14%	90.45%	69.10%	51.31%	3.99%
	Enumeration	16.93%	77.85%	**100.00%**	**96.62%**	0.00%
Test-Core	Transformer-R	**56.76%**	**97.42%**	**100.00%**	32.37%	2.14%
	Transformer	55.61%	86.82%	44.60%	34.76%	**2.27%**
	Enumeration	18.64%	77.93%	**100.00%**	**98.36%**	0.00%
Test-Property	Transformer-R	**42.90%**	**97.57%**	**100.00%**	57.84%	**4.66%**
	Transformer	41.75%	90.69%	62.25%	57.98%	4.25%
	Enumeration	15.91%	81.19%	**100.00%**	**96.65%**	0.00%

Observations:

- Slight improvement in Desirable
- Much more improvement in MMP33
- Always guarantee the core constant

Significance in preventing and combating pandemics

Our model could accelerate the process of optimizing antiviral drug candidates in terms of various properties of interest, *e.g.*, pharmacokinetics.

- Existing drugs can be identified as lead and chemically modified to improve specific properties. For example, ivermectin has been reported to show *in vitro* antiviral activity against SARS-CoV-2, but have pharmacokinetic problems such as high cytotoxicity and low solubility.

References

[1] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[2] J. He, H. You, E. Sandström, E. Nittinger, E. J. Bjerrum, C. Tyrchan, W. Czechtizky, and O. Engkvist, "Molecular optimization by capturing chemist's intuition using deep neural networks," *Journal of cheminformatics*, vol. 13, no. 1, pp. 1–17, 2021.