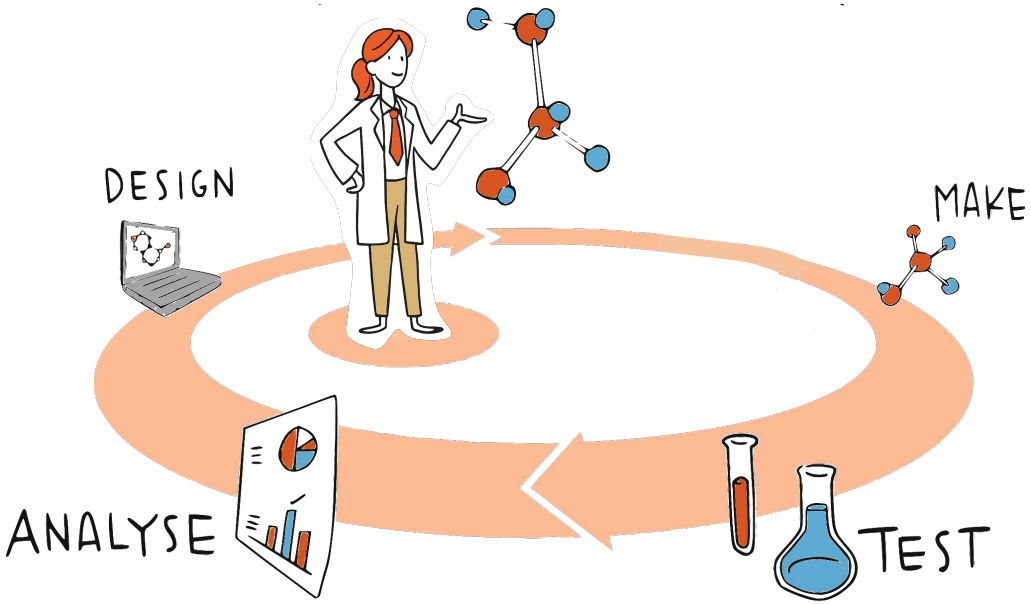
Beyond Performance: How design choices shape chemical language models

AiChemist Summer School, Copenhagen
15th June 2026,
Inken Fender, University of Bern, Switzerland

Why should we care about chemical language models?
The Design-Make-Test-Analyse-Cycle (DMTA)

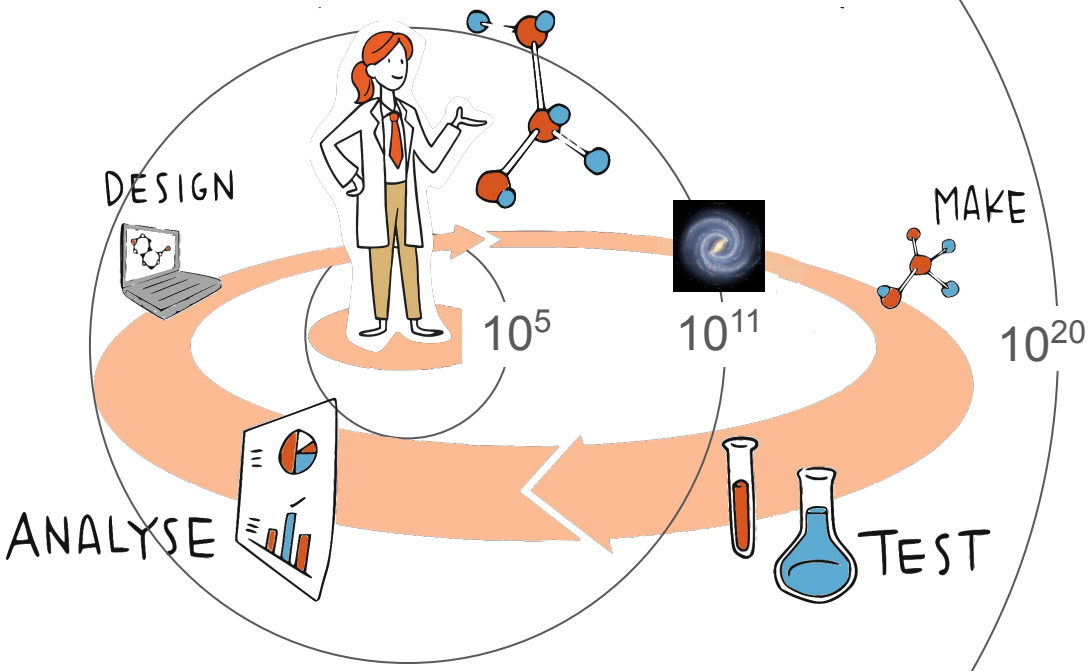


Chemical space: >10^{60} molecules!

Adapted from Schneider et al., 2019: https://www.nature.com/articles/s41573-019-0050-3/figures/1, Image courtesy of Jack Burgess, Jack Burgess Studios.

Why should we care about chemical language models?

The Design-Make-Test-Analyse-Cycle (DMTA)

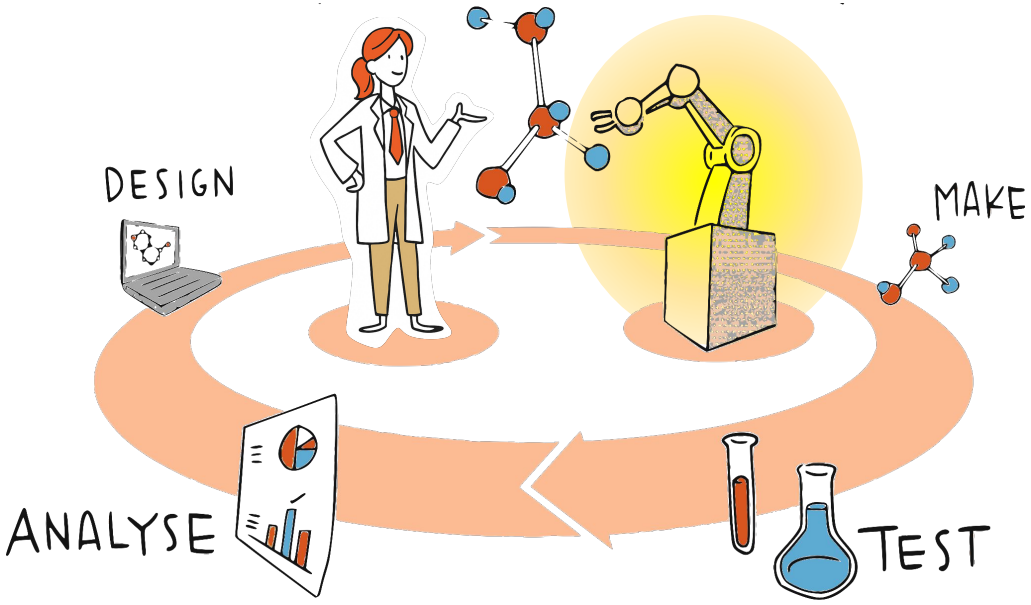


Chemical space: >10^{60} molecules!

10^{40}

https://www.biosolveit.de/wp-content/uploads/2024/02/BioSolveIT_ChemicalSpaces-scaled.jpg
Schneider et al., 2019: https://www.nature.com/articles/s41573-019-0050-3/figures/1
NASA/JPL-Caltech/R. Hurt (SSC/Caltech)

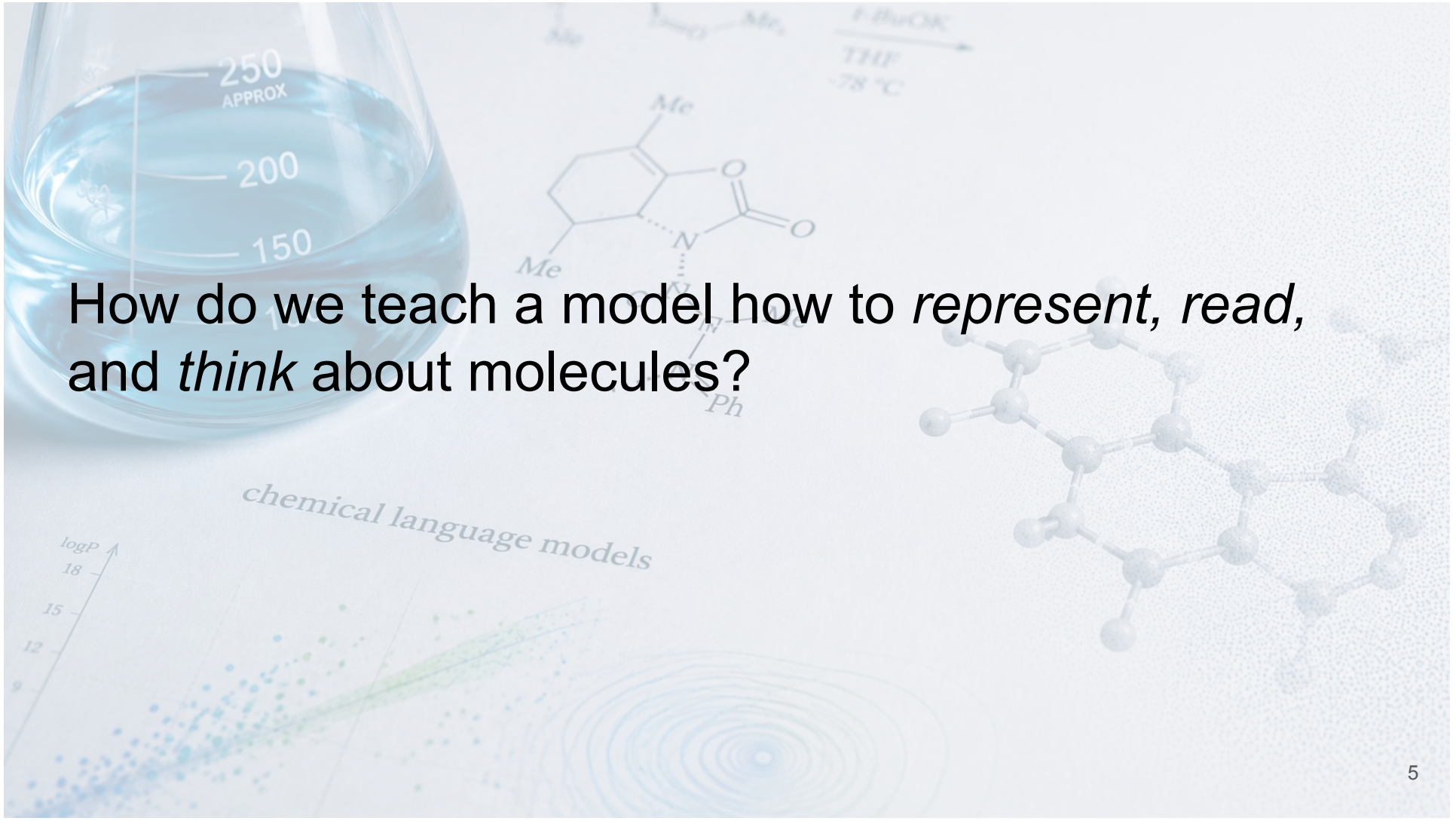
Why should we care about chemical language models?
The Design-Make-Test-Analyse-Cycle (DMTA)



Chemical Language Models:

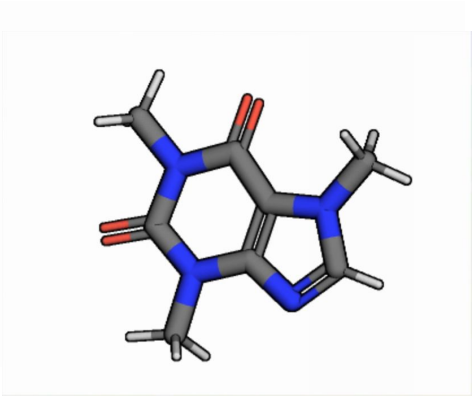
- property prediction
- reaction prediction
- retrosynthesis
- de novo molecule generation

https://www.nature.com/articles/s41573-019-0050-3/figures/1

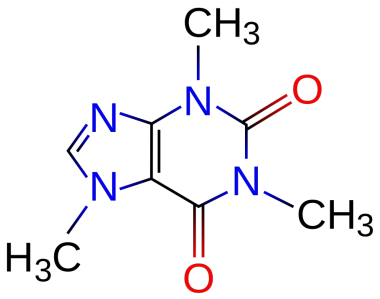


How do we teach a model how to *represent, read,* and *think* about molecules?

What do we need to train a **chemical language** model?



Caffeine



PubChem Caffeine (Compound)

2 Names and Identifiers

2.1 Computed Descriptors

2.1.1 IUPAC Name

1,3,7-trimethylpurine-2,6-dione

Computed by Lexichem TK 2.7.0 (PubChem release 2025.04.14)

PubChem

2.1.2 InChI

InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Computed by InChI 1.07.2 (PubChem release 2025.04.14)

PubChem

2.1.3 InChIKey

RYYVLZVUVIJVGH-UHFFFAOYSA-N

Computed by InChI 1.07.2 (PubChem release 2025.04.14)

PubChem

2.1.4 SMILES

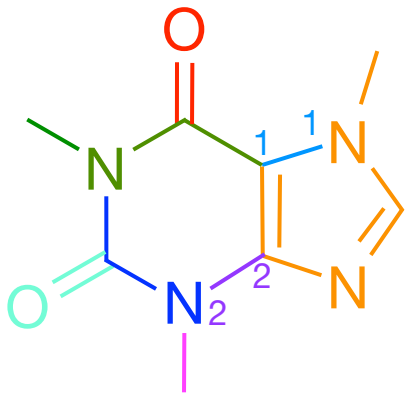
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Computed by OEChem 2.3.0 (PubChem release 2025.04.14)

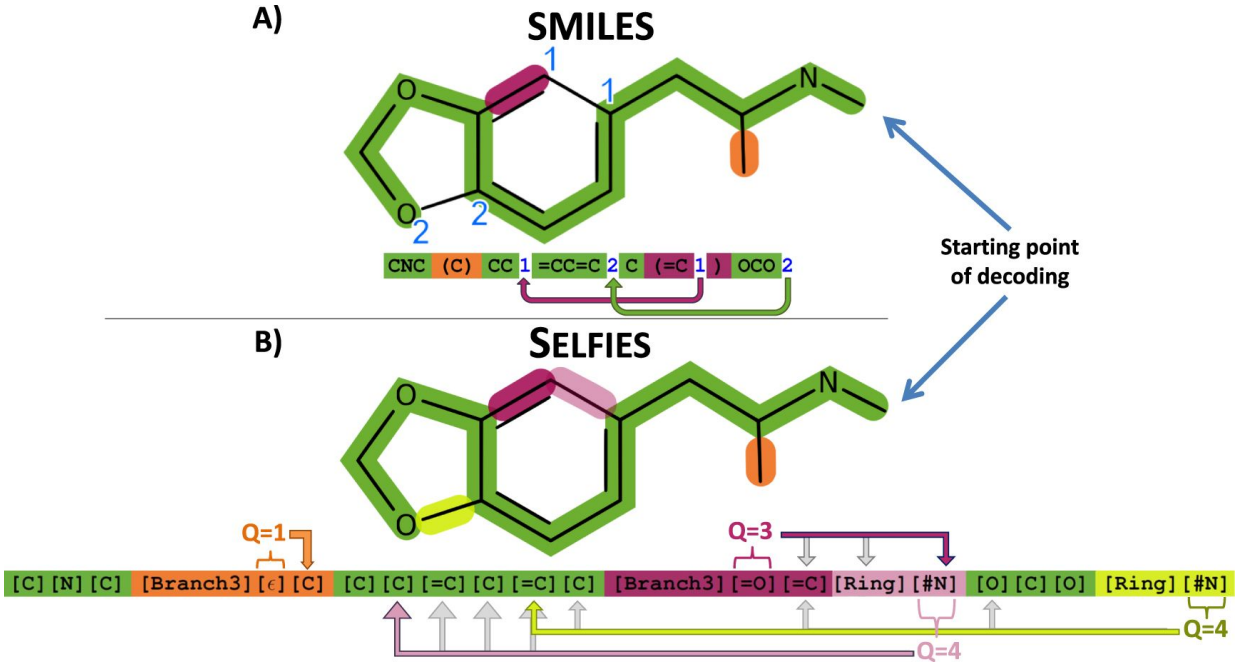
PubChem

SMILES = **Simplified Molecular Input Line Entry System**

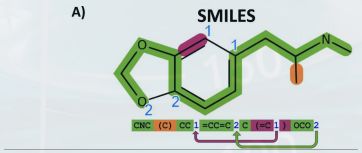
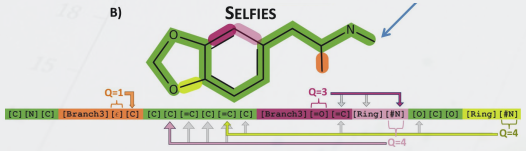
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

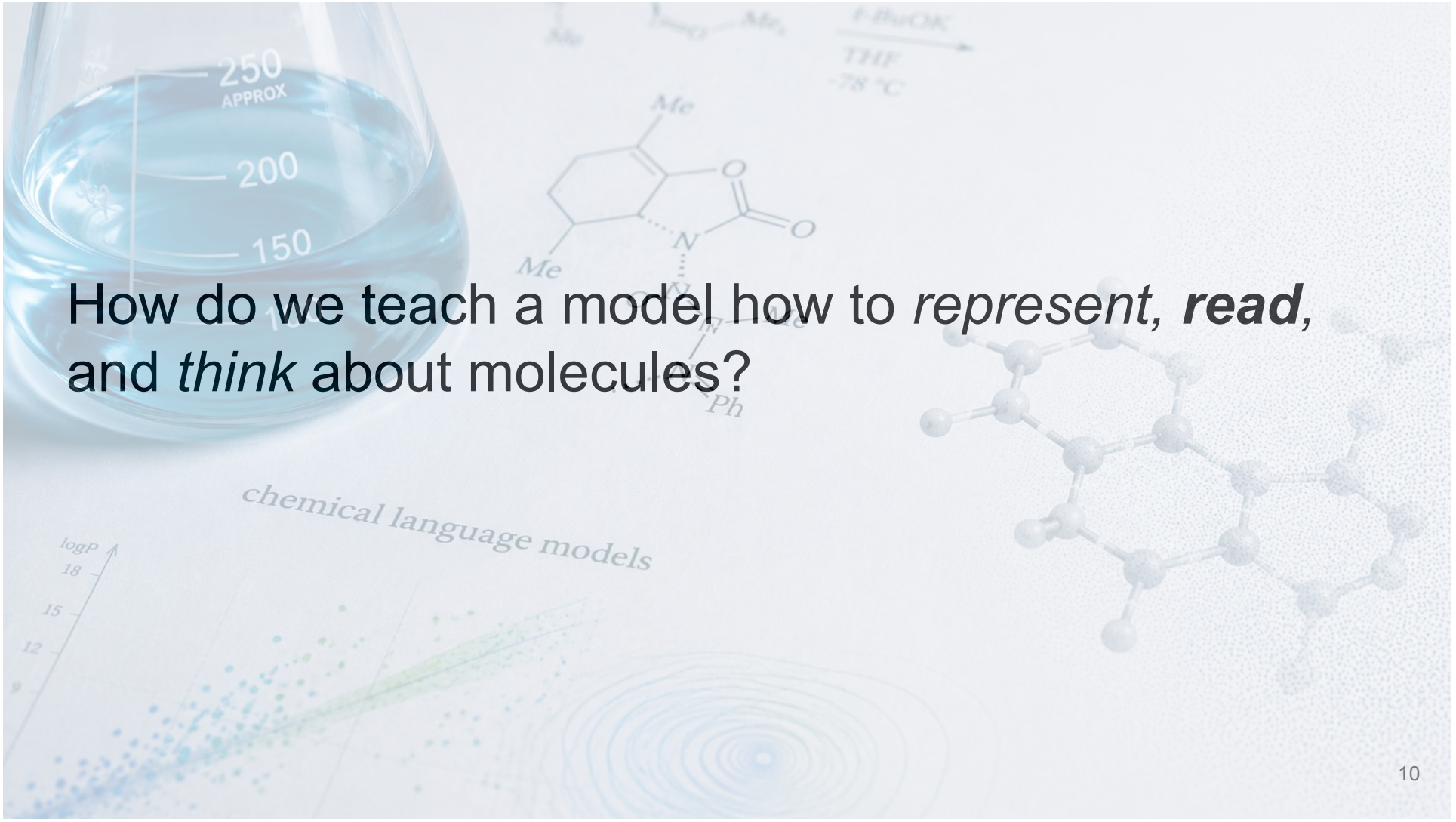


Not all SMILES are valid, but SELFIES are



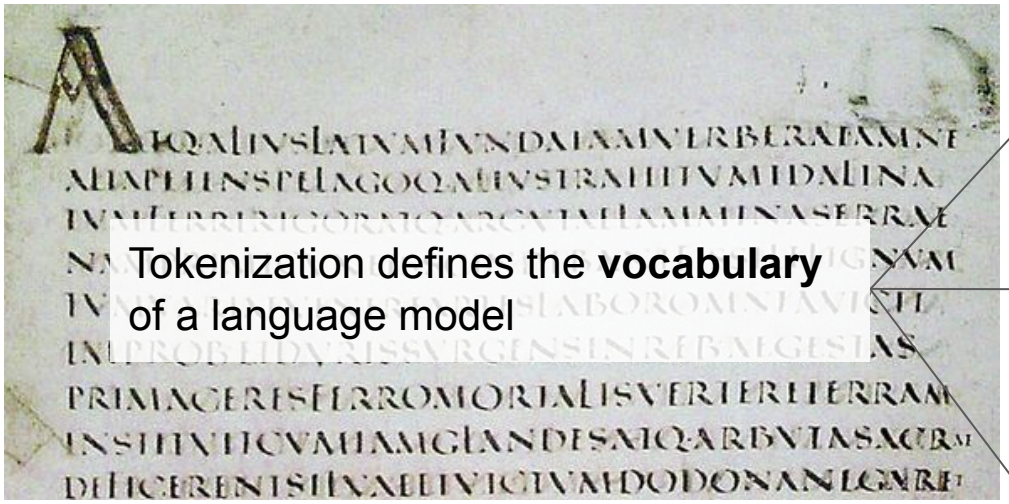
Krenn et al., 2020

REPRESENT	READ	THINK
SMILES  SELFIES  		



How do we teach a model how to *represent*, ***read***, and *think* about molecules?

Tokenization controls how the model balances detail vs. generalization



https://en.wikipedia.org/wiki/Scriptio_continua#/media/File:Vergilius_Augusteus,_Georgica_141.jpg

Character-level tokenization

‘T’ ‘o’ ‘k’ ‘e’ ‘n’ ‘i’ ‘z’ ‘a’ ‘t’ ‘i’ ‘o’ ‘n’ ‘ ‘d’ ‘e’ ‘f’ ‘i’ ‘n’ ‘e’ ‘s’ ‘t’ ‘h’ ‘e’ ‘**v’ ‘o’ ‘c’ ‘a’ ’b’ ‘u’ ‘l’ ’a’ ‘r’ ‘y’ ‘**a’ …

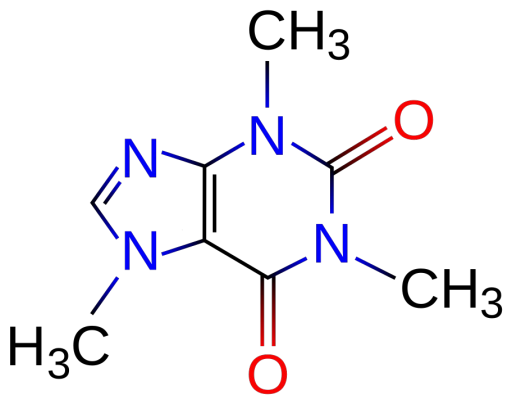
Subword-level Tokenization

‘Token’ ‘ization’ ‘ defines’ ‘the’ ‘**vocabulary’ ‘**of’ ‘a’ ‘language’ model’

Word-level Tokenization

‘Tokenization’ ‘ defines’ ‘the’ ‘**vocabulary’ ‘**of’ ‘a’ ‘language’ model’

Tokenization controls how the model balances detail vs. generalization



Cn1c(=O)c2c(ncn2C)n(C)c1=O

Atomwise tokenization

‘C’ ’n’ ’1’ ’c’ ’(‘ ’=’ ’O’ ’)’ ’c’ ’2’ ’c’ ’(‘ ’n’ ’c’ ’n’ ’2’ ’C’ ’)’ ’n’ ’(‘ ’C’ ’)’ ’c’ ’1’ ’=’ ’O’

SentencePiece tokenization

‘Cn1c’ ‘(=O)’ ‘c2c(ncn2’ ‘C)n(C)’ ‘c1=O’

https://de.wikipedia.org/wiki/Datei:Caffeine.svg

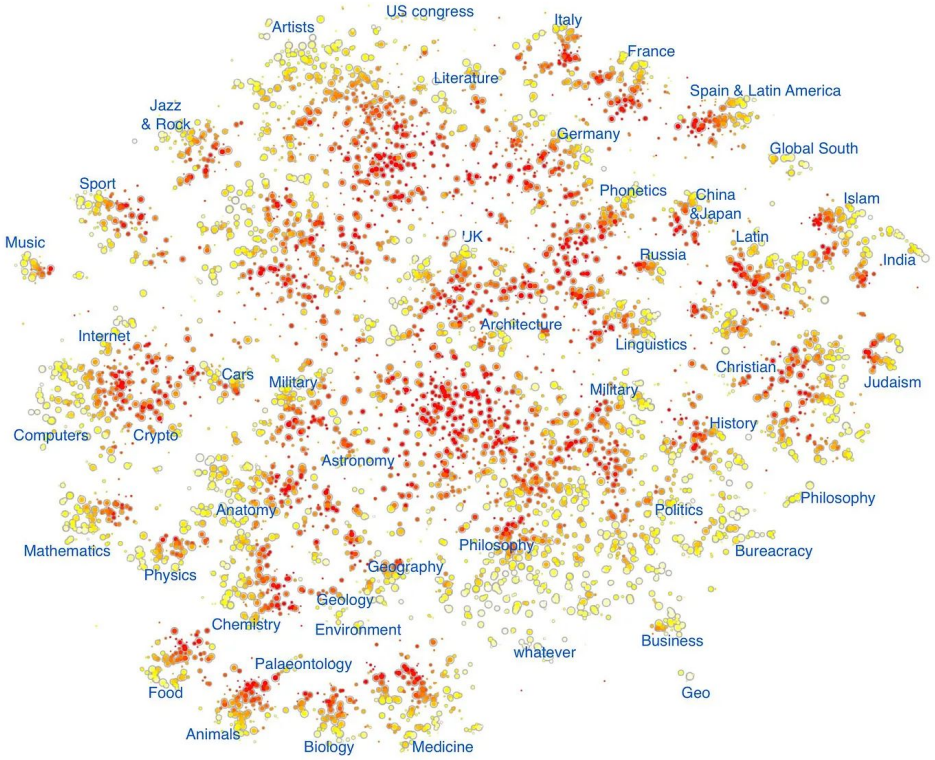
Token IDs map to embeddings

Assign Token IDs

‘Token’ = 0
‘ization’ = 1
defines’ = 2
‘the’ = 3
‘vocabulary’ = 4
‘of‘ = 5 …

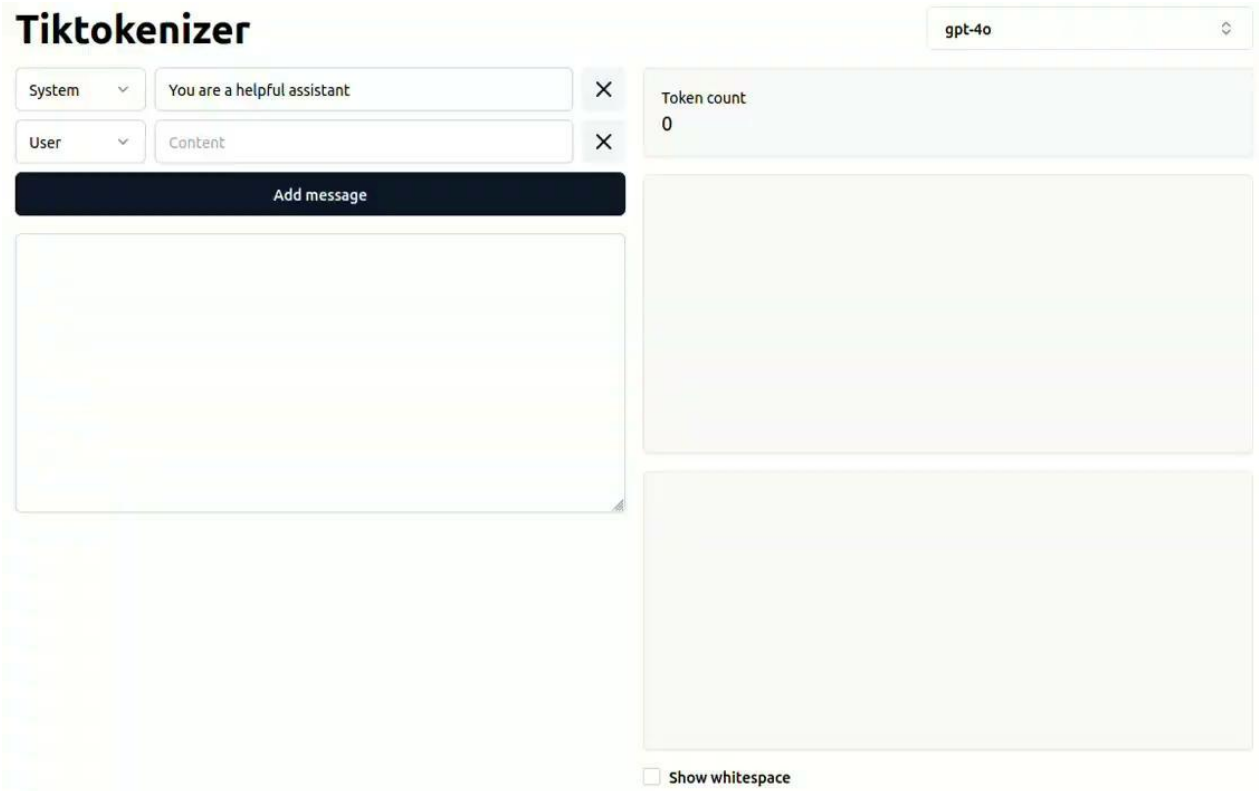
IDs link to embeddings

0
1
2
3
…

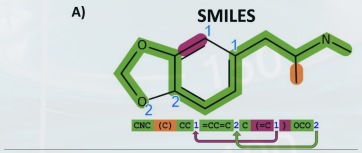
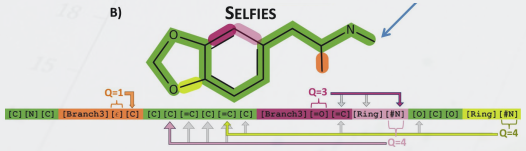


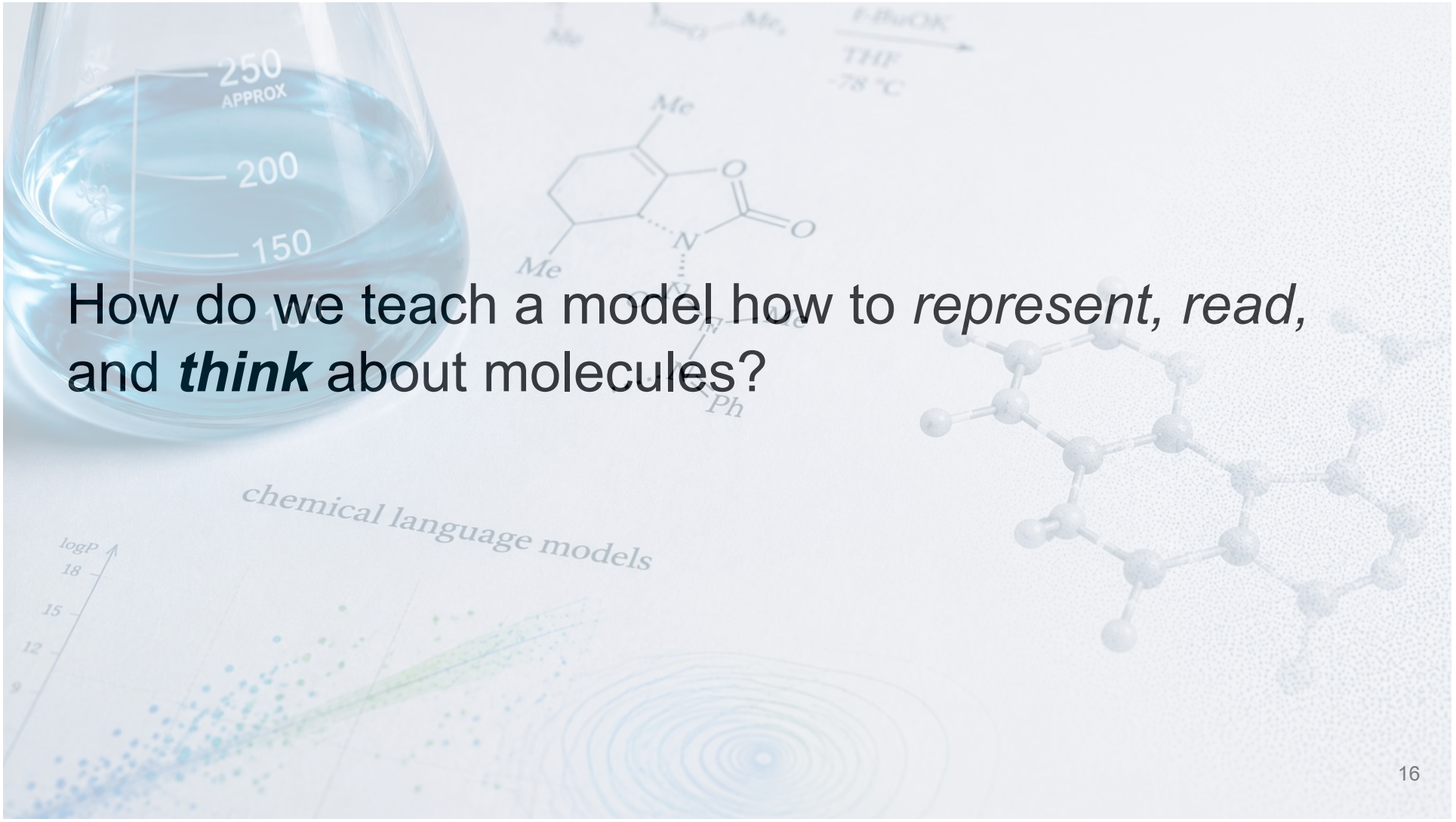
https://medium.com/@oxburybill/using-topology-to-help-visualise-word-vectors-5e3db4856b1a

Token IDs map to embeddings



https://tiktokenizer.vercel.app/?model=gpt-4o

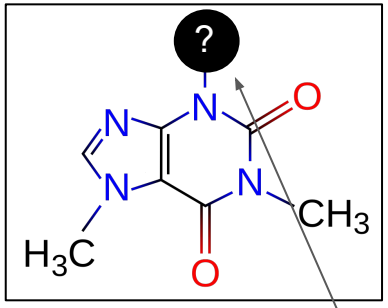
REPRESENT	READ	THINK
SMILES 	Atomwise tokenization **‘C’ ’n’ ’1’ ’c’ ’(‘ ’=’ ’O’ ’)’ ’c’ ’2’ ’c’ ’(‘ ’n’ ’c’ ’n’ ’2’ ’C’ ’)’ ’n’ ’(‘ ’C’ ’)’ ’c’ ’1’ ’=’ ’O’**	
SELFIES 	SentencePiece tokenization **‘Cn1c’ ‘(=O)’ ‘c2c(ncn2’ ‘C)n(C)’ ‘c1=O’**	



How do we teach a model how to *represent, read,* and ***think*** about molecules?

Models learn to predict masked token probabilities

This summer school talk is _____ .



Cn1c(=O)c2c(ncn2C)n(___)c1=O

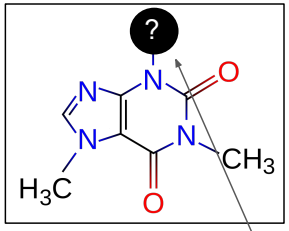
mildly entertaining	0.3 ✔	OH	0.01
peer-reviewed	0.3	CH_3	0.99 ✔
making me hungry	0.2		

RoBERTa

Models learn to predict masked token probabilities

MASKING

This summer school talk is _____ .



Cn1c(=O)c2c(ncn2C)n(___)c1=O

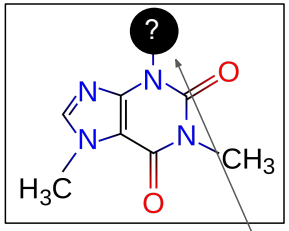
mildly entertaining	0.7 ✔	OH	0.01
peer-reviewed	0.1	CH_3	0.99 ✔
making me hungry	0.2		

RoBERTa

…or to correct *corrupted* sequences

MASKING

This summer school talk is _____ .



Cn1c(=O)c2c(ncn2C)n(___)c1=O

mildly entertaining	0.7 ✔	OH	0.01
peer-reviewed	0.1	CH_3	0.99 ✔
making me hungry	0.2		

RoBERTa

CORRUPTION

summer school ___… This .

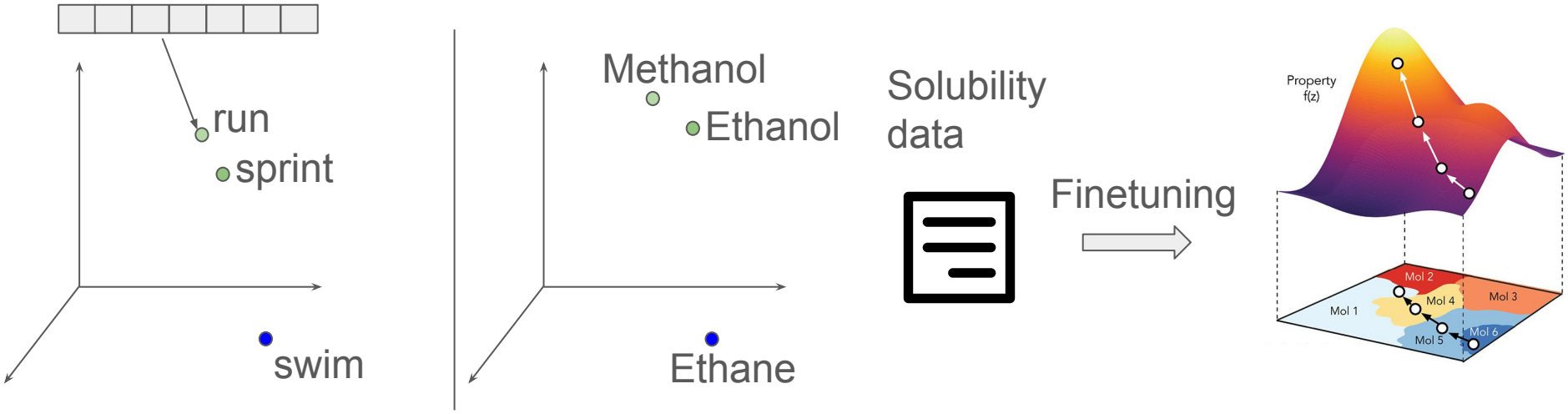
This summer school talk is mildly entertaining.

(=O)c2c(ncn2C)n(___)....*Cn1c*

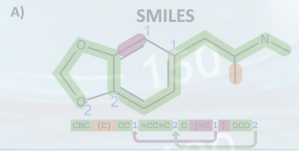
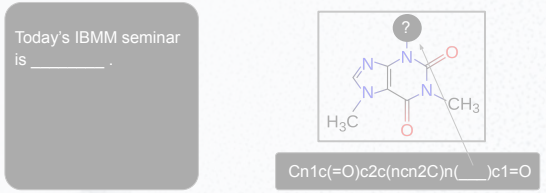
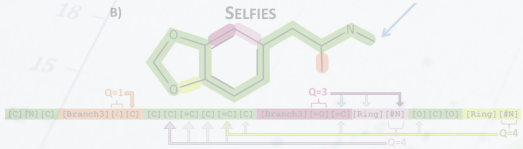
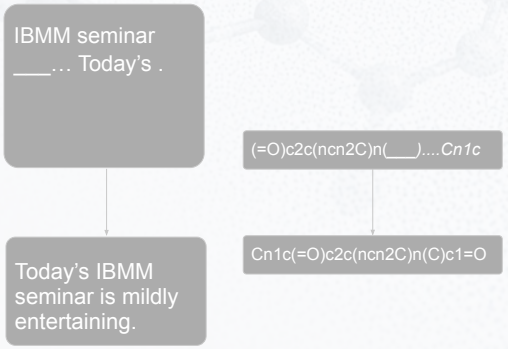
Cn1c(=O)c2c(ncn2C)n(C)c1=O

BART

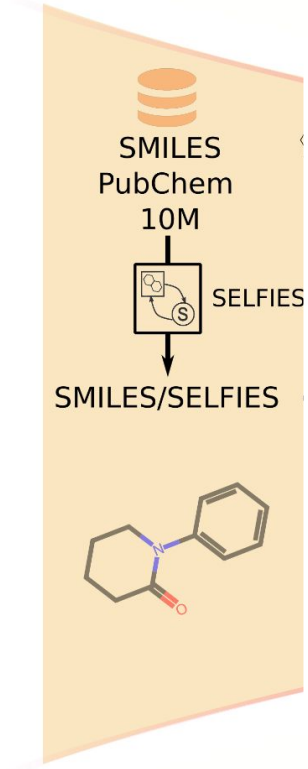
Training and *latent* space



https://pubs.acs.org/cms/10.1021/acscentsci.7b00572/asset/images/large/oc-2017-00572f_0001.jpeg

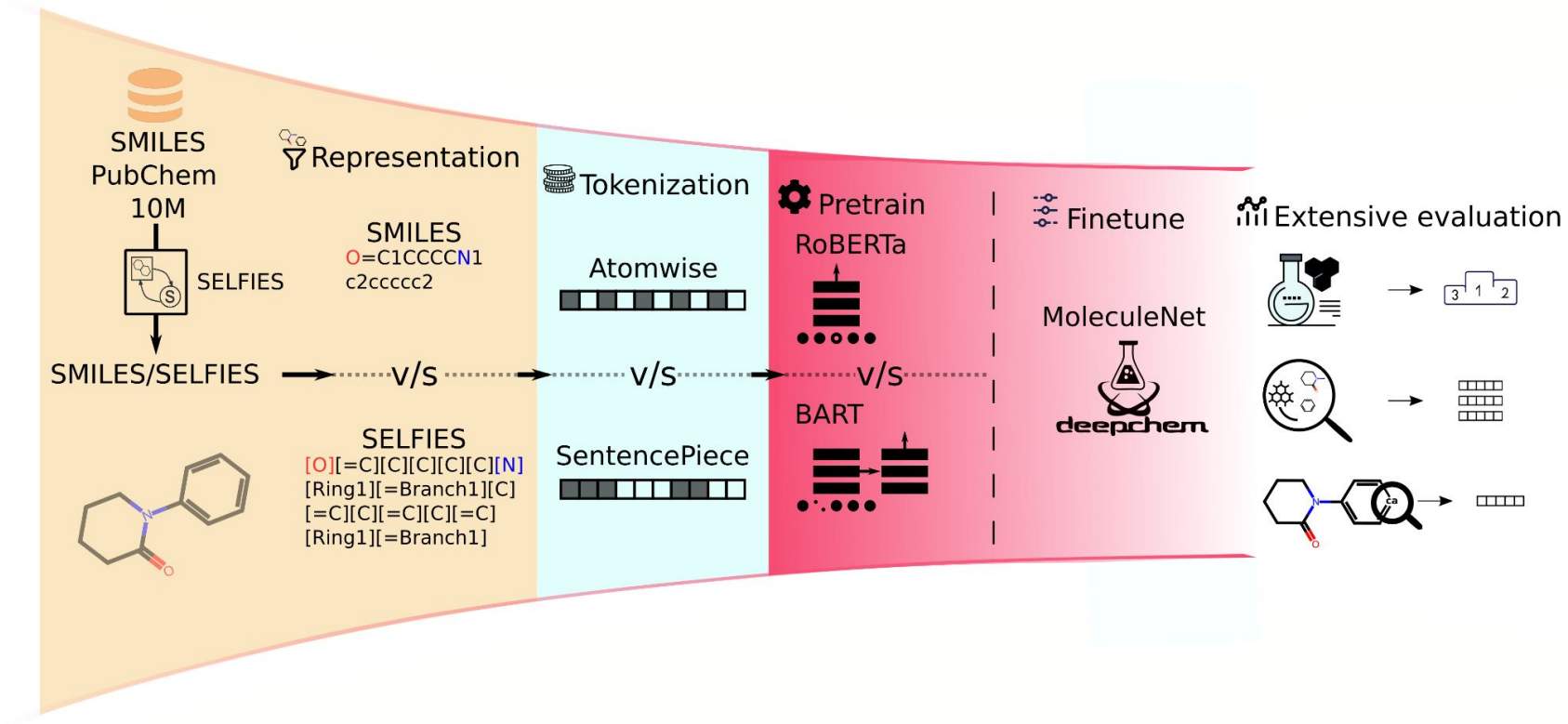
REPRESENT	READ	THINK
SMILES	Atomwise tokenization	RoBERTa
	‘C’ ’n’ ’1’ ’c’ ’(‘ ’=’ ’O’ ’)’ ’c’ ’2’ ’c’ ’(‘ ’n’ ’c’ ’n’ ’2’ ’C’ ’)’ ’n’ ’(‘ ’C’ ’)’ ’c’ ’1’ ’=’ ’O’	
SELFIES	SentencePiece tokenization	BART
	‘Cn1c’ ‘(=O)’ ‘c2c(ncn2’ ‘C)n(C)’ ‘c1=O’	

How do design choices shape chemical language models?

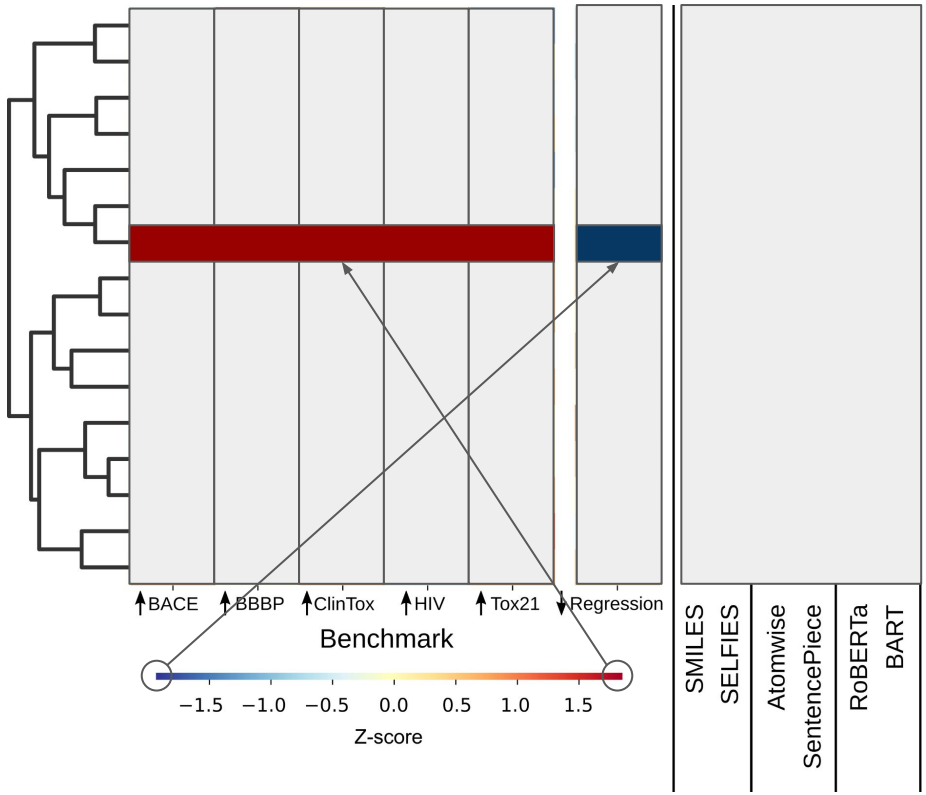


Jannik Gut

Jannik Gut

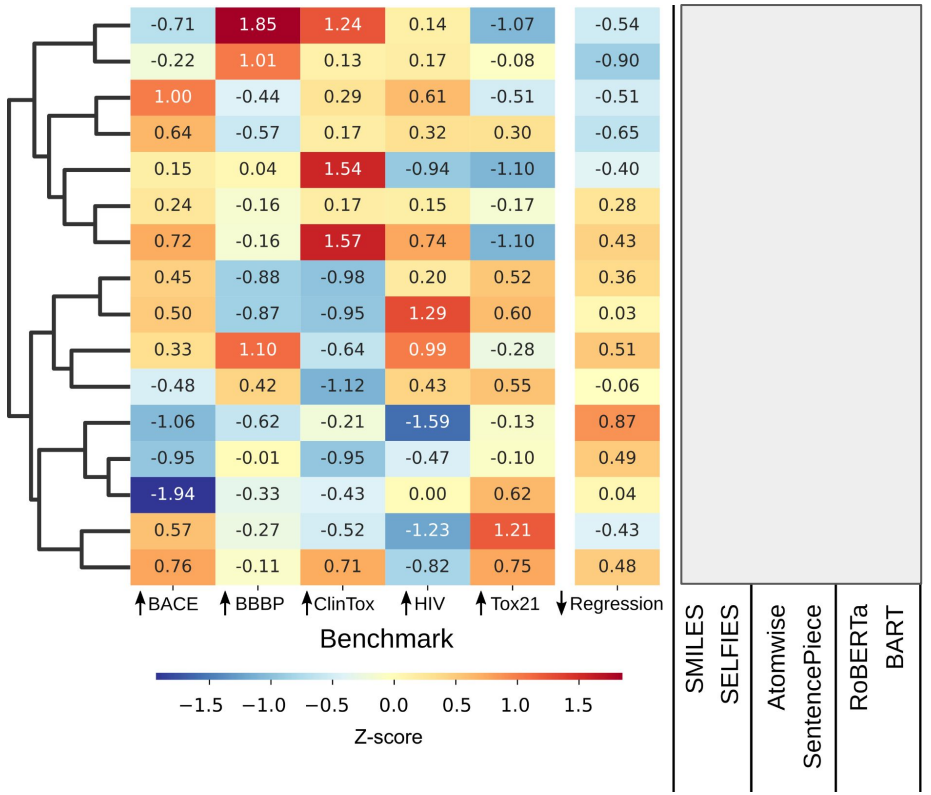

SMILES
PubChem
10M
SELFIES
SMILES/SELFIES
Representation
SMILES
O=C1CCCCN1
c2ccccc2
v/s
SELFIES
[O][=C][C][C][C][C][N]
[Ring1][=Branch1][C]
[=C][C][=C][C][=C]
[Ring1][=Branch1]
Tokenization
Atomwise
v/s
SentencePiece
Pretrain
RoBERTa
v/s
BART
Finetune
MoleculeNet
deepchem
Extensive evaluation
3 1 2

Comparing performance on benchmarks



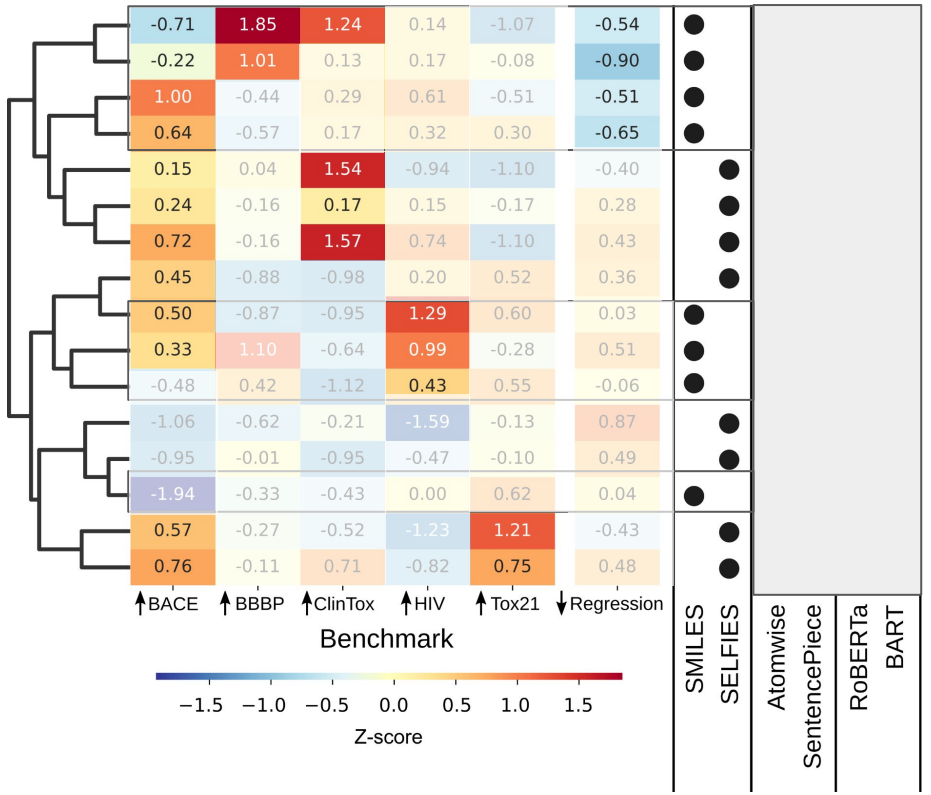
Comparing performance on benchmarks

- performance between all models is comparable → all models learn & predict well



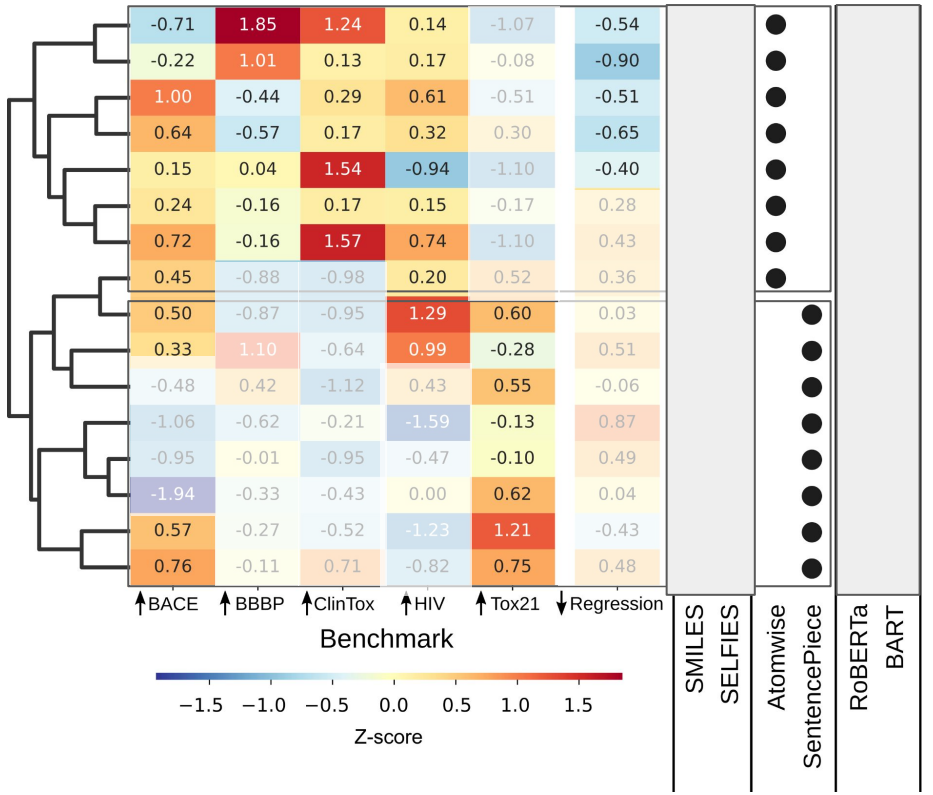
Comparing performance on benchmarks

- performance between all models is comparable → all models learn & predict well
- SMILES vs. SELFIES



Comparing performance on benchmarks

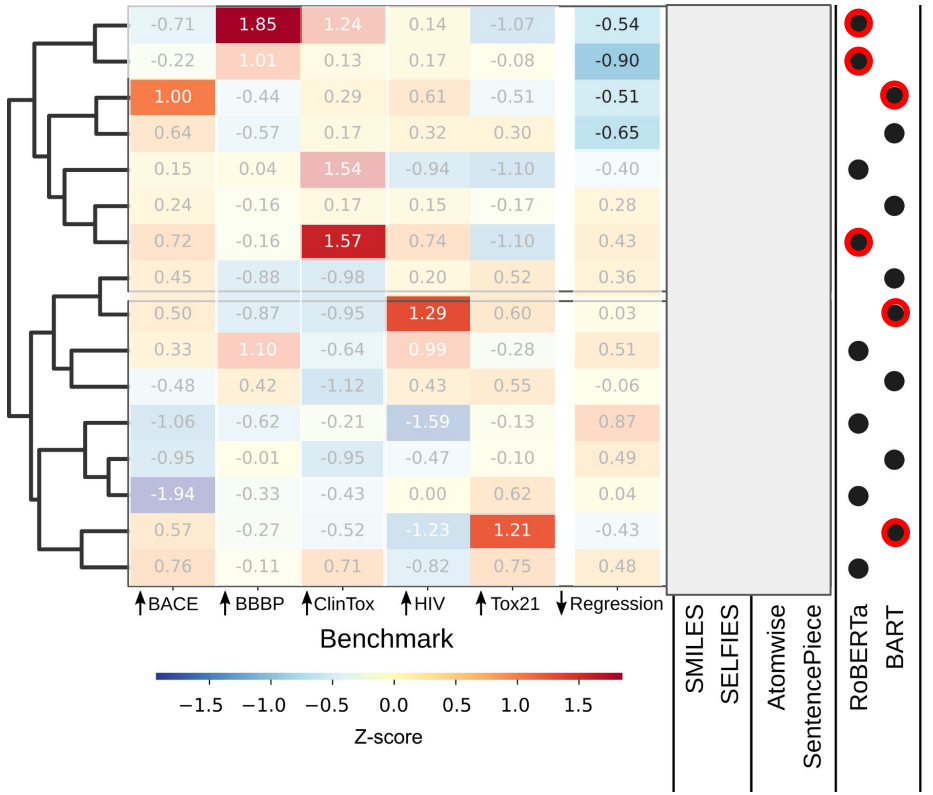
- performance between all models is comparable → all models learn & predict well
- SMILES vs. SELFIES
- Atomwise > Sentencepiece



Comparing performance on benchmarks

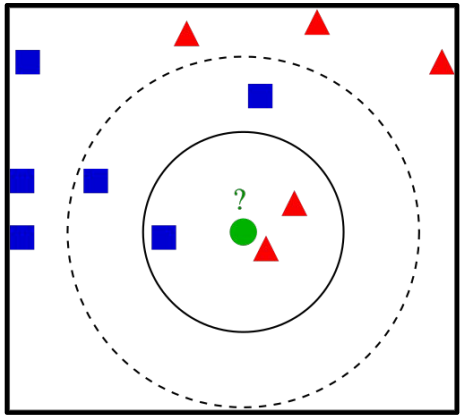
- performance between all models is comparable → all models learn & predict well
- SMILES vs. SELFIES
- Atomwise > Sentencepiece
- BART vs. RoBERTa

→ subsequent focus on atomwise models only



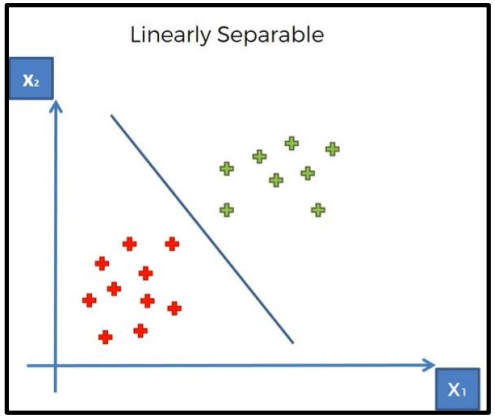
What is the shape of the latent space?

local?



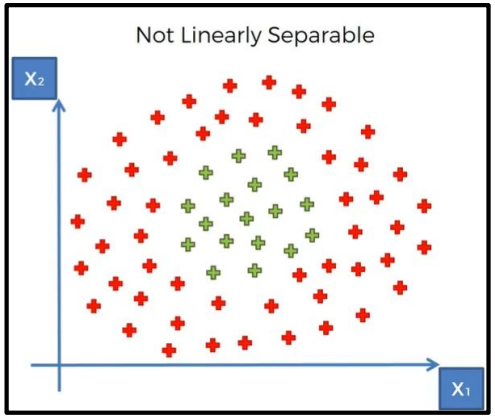
KNN

simple?



Linear SVM

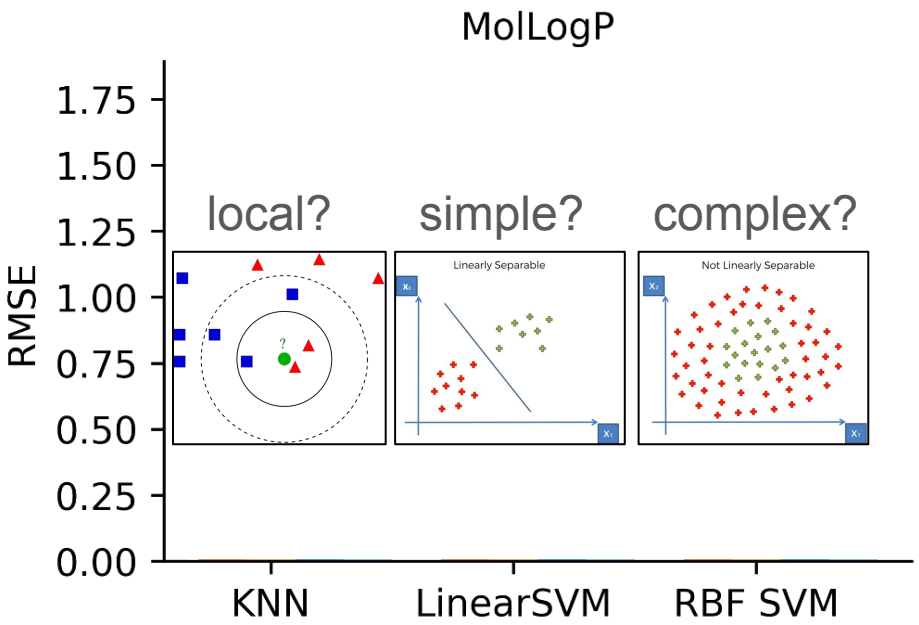
complex?



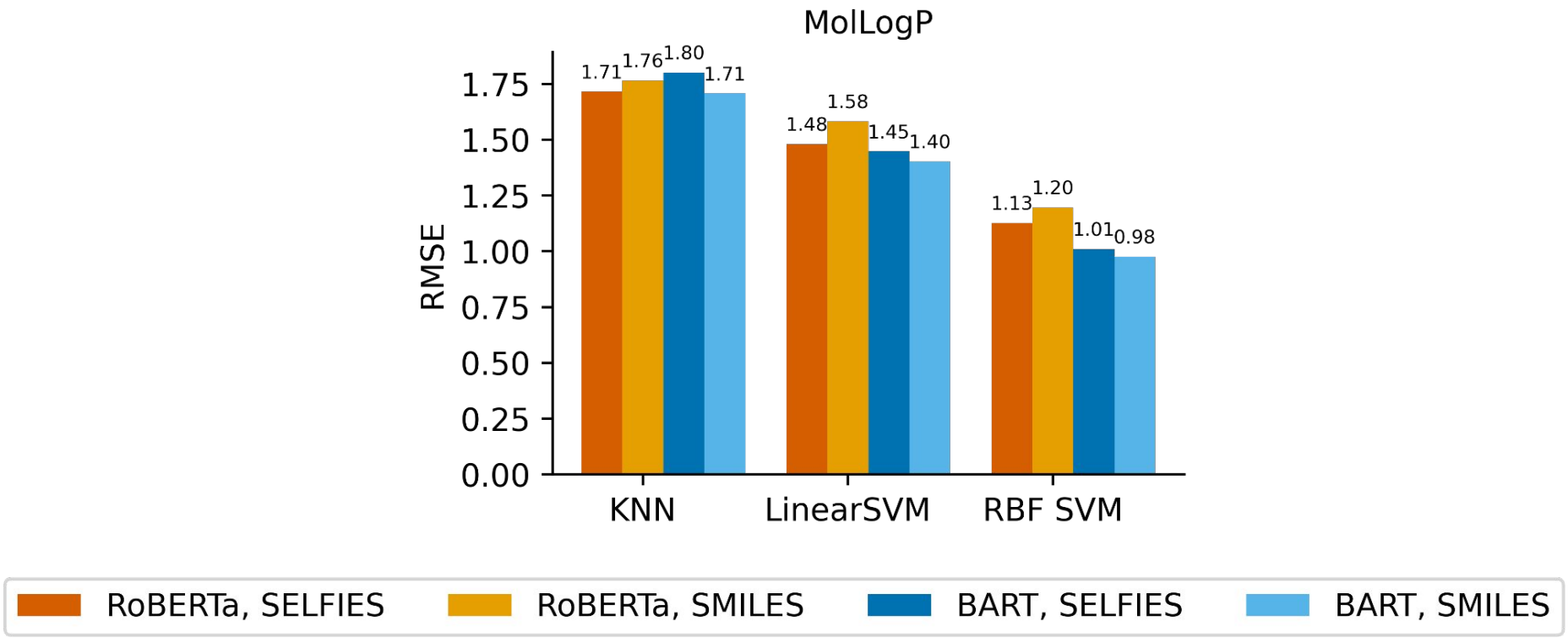
RBF SVM

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg
https://medium.com/@suvigya2001/the-gaussian-rbf-kernel-in-non-linear-svm-2fb1c822aae0

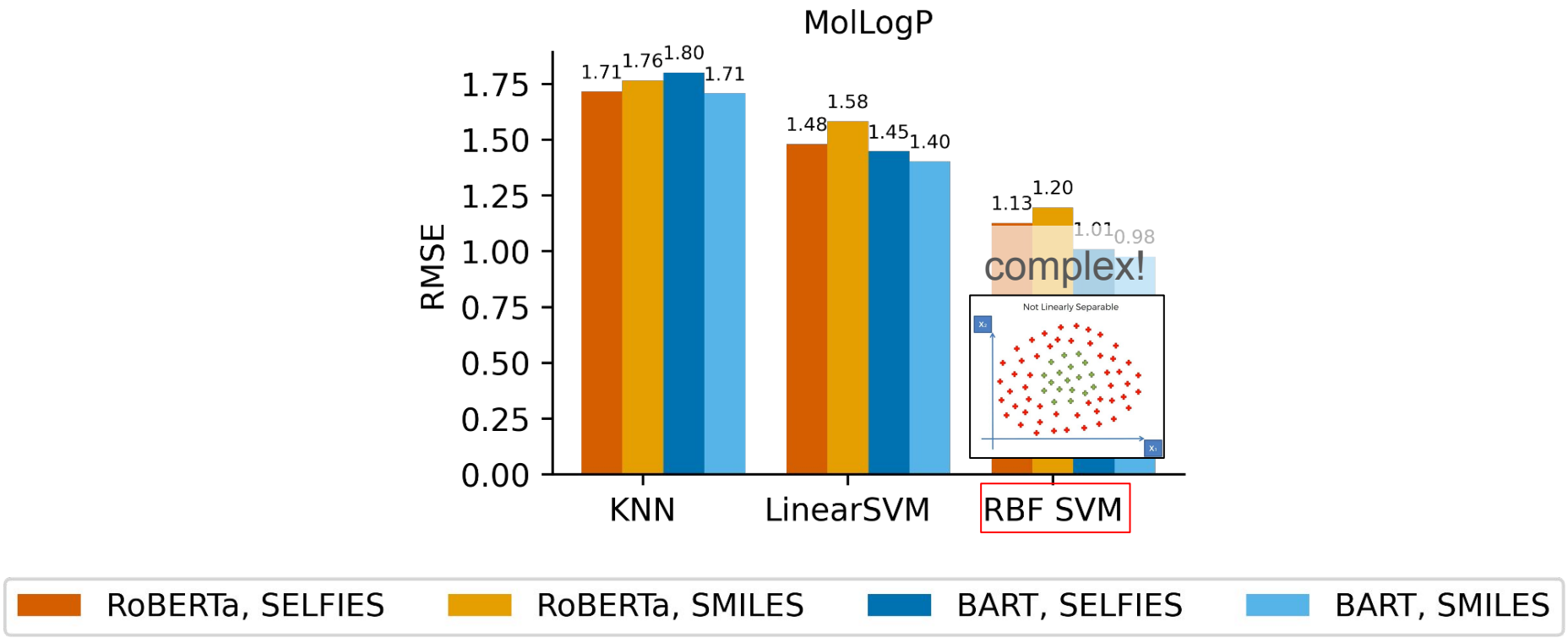
How do latent spaces between models differ?



How do latent spaces between models differ?

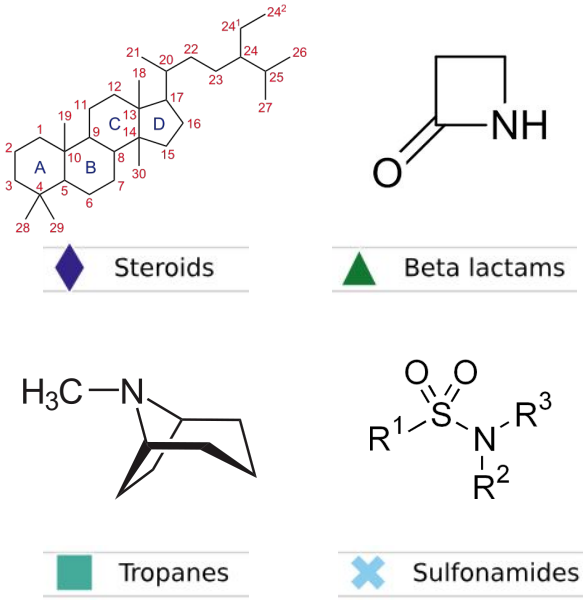


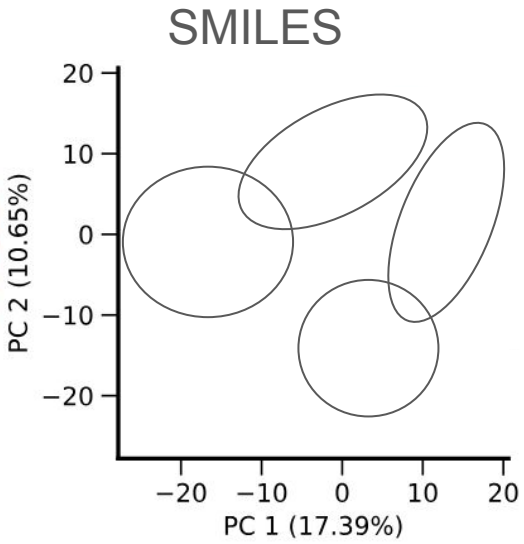
How do latent spaces between models differ?

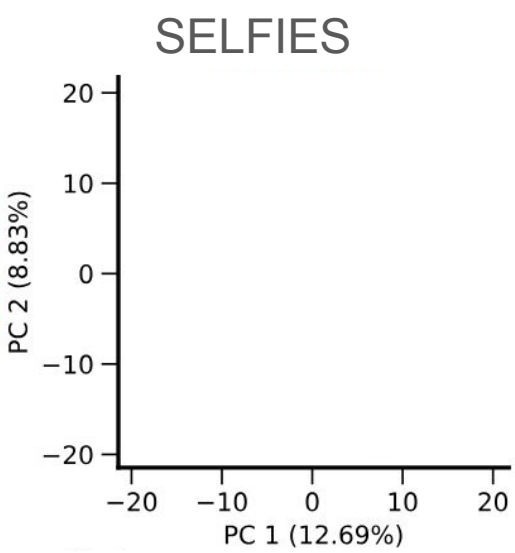


How does latent space differ?

64 molecules from:



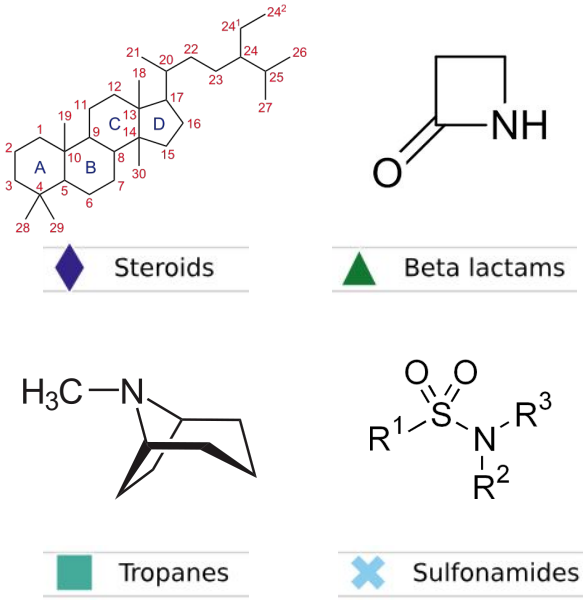


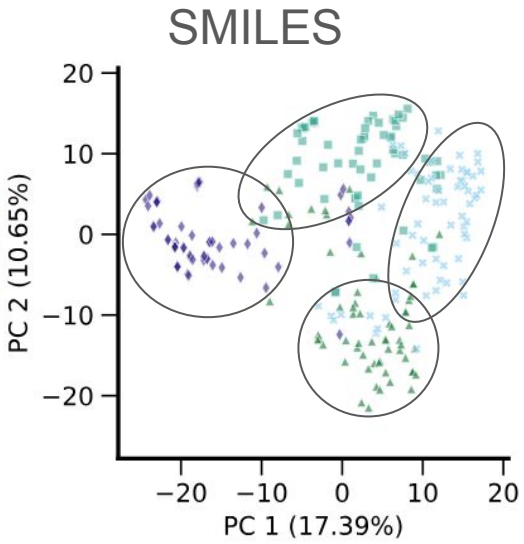


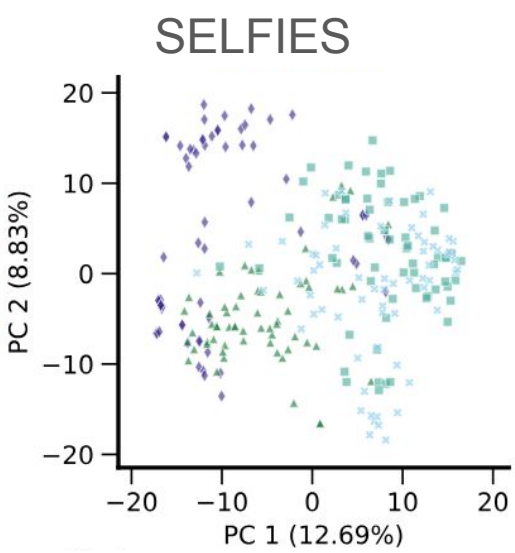
https://de.wikipedia.org/wiki/Datei:Trimethyl_steroid-nomenclature.svg, https://en.wikipedia.org/wiki/%CE%92-Lactam, https://upload.wikimedia.org/wikipedia/commons/thumb/a/a9/Tropan_-_Tropane.svg/490px-Tropan_-_Tropane.svg.png,

How does latent space differ?

64 molecules from:

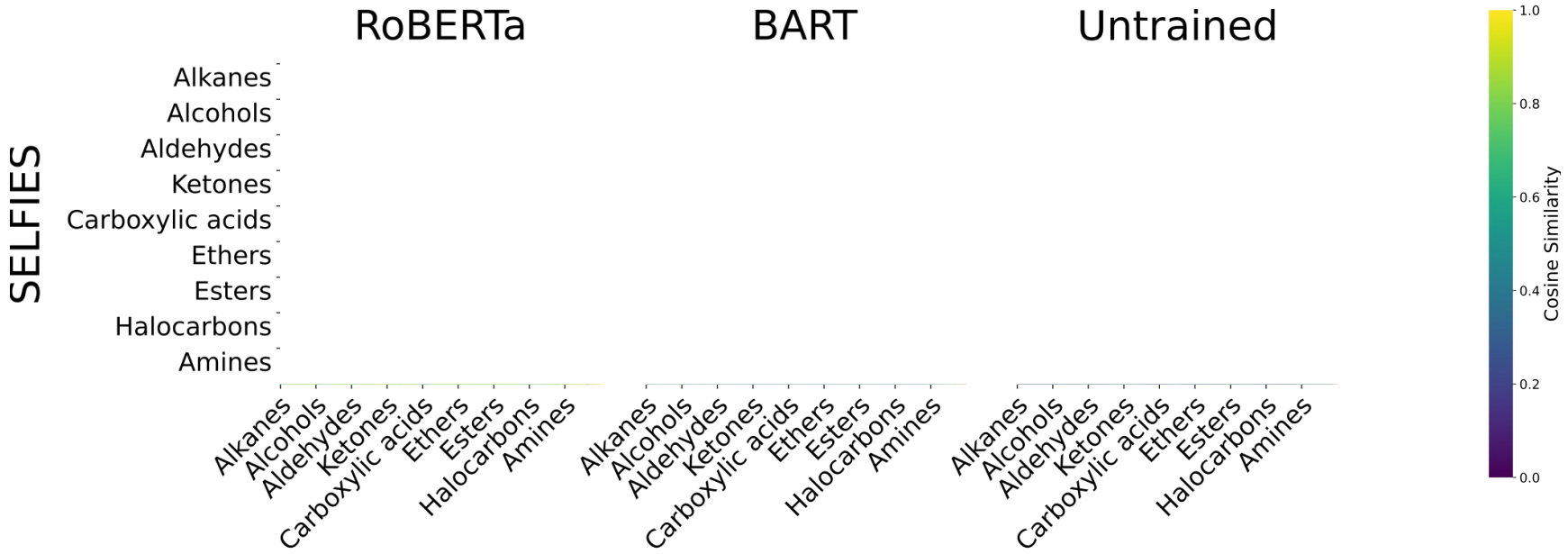




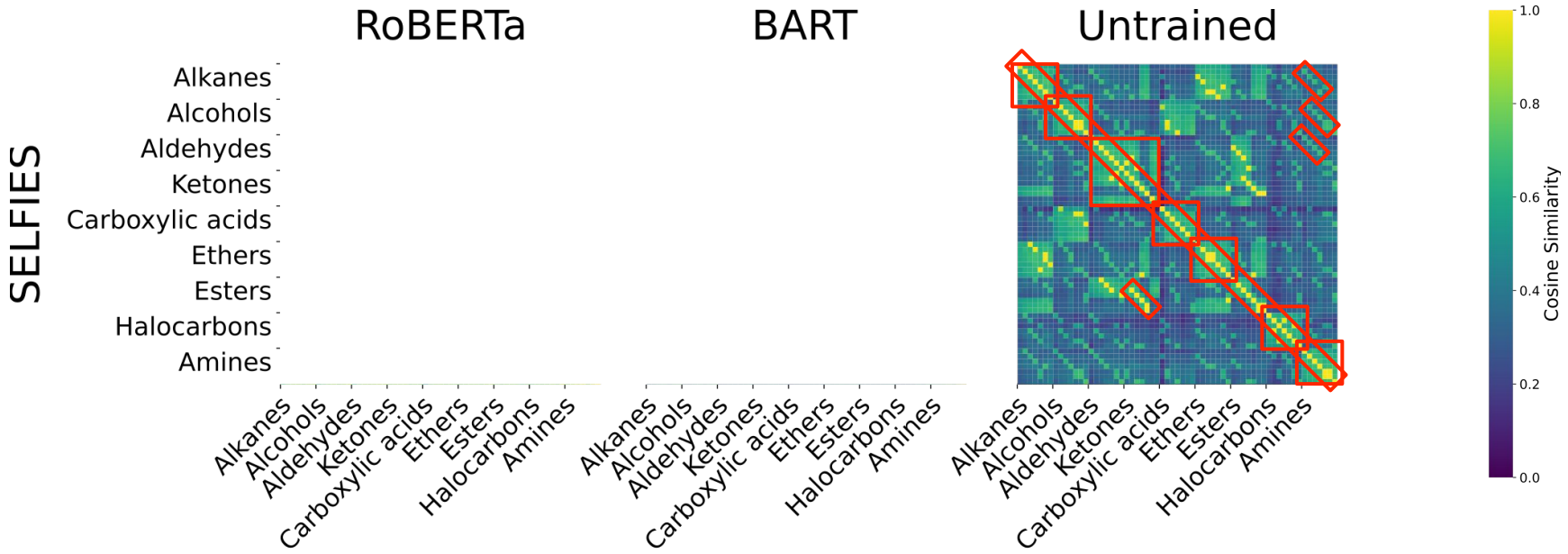


https://de.wikipedia.org/wiki/Datei:Trimethyl_steroid-nomenclature.svg, https://en.wikipedia.org/wiki/%CE%92-Lactam, https://upload.wikimedia.org/wikipedia/commons/thumb/a/a9/Tropan_-_Tropane.svg/490px-Tropan_-_Tropane.svg.png,

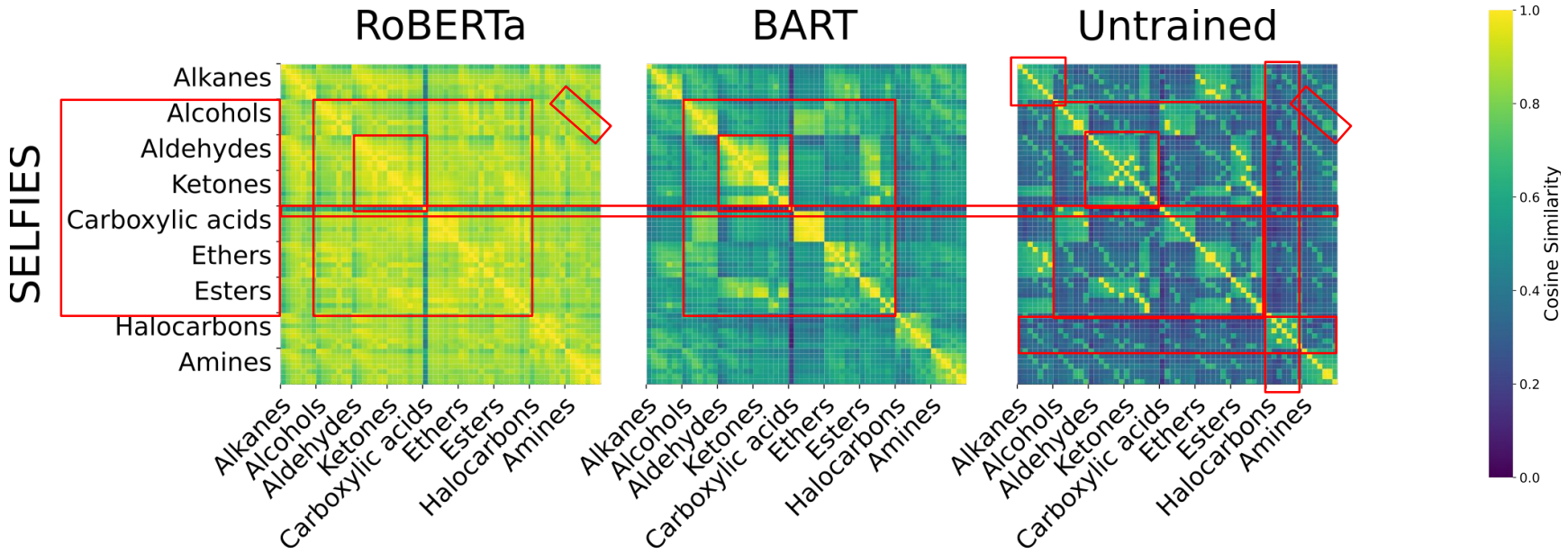
Does understanding of molecule similarities differ?



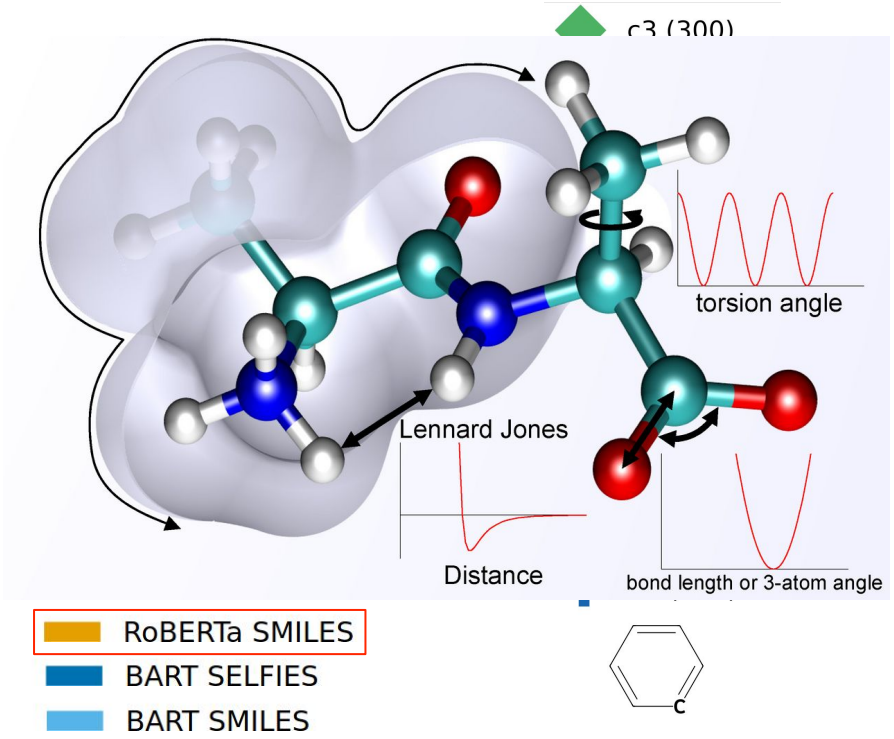
Does understanding of molecule similarities differ?

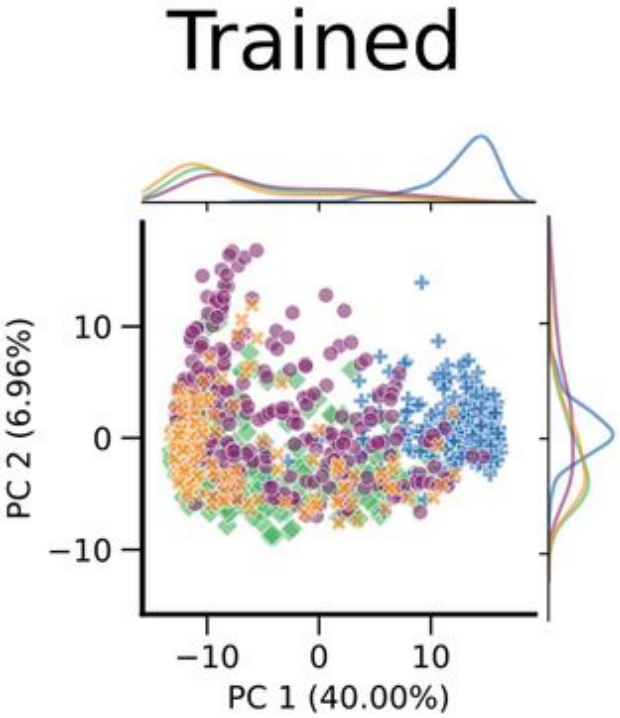


Does understanding of molecule similarities differ?

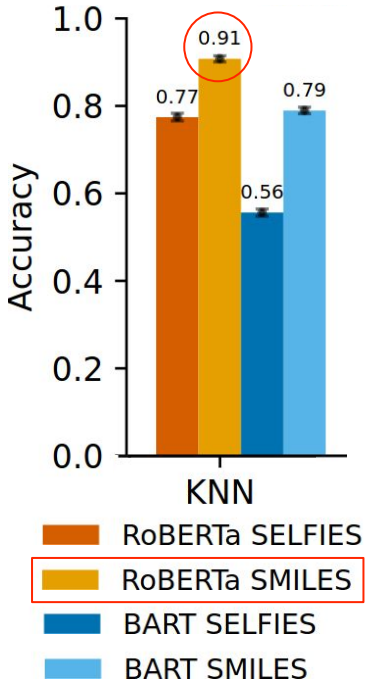


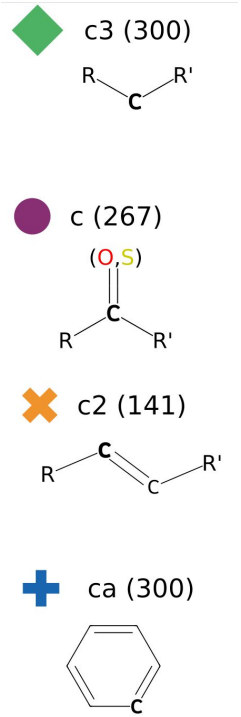
What about atoms?

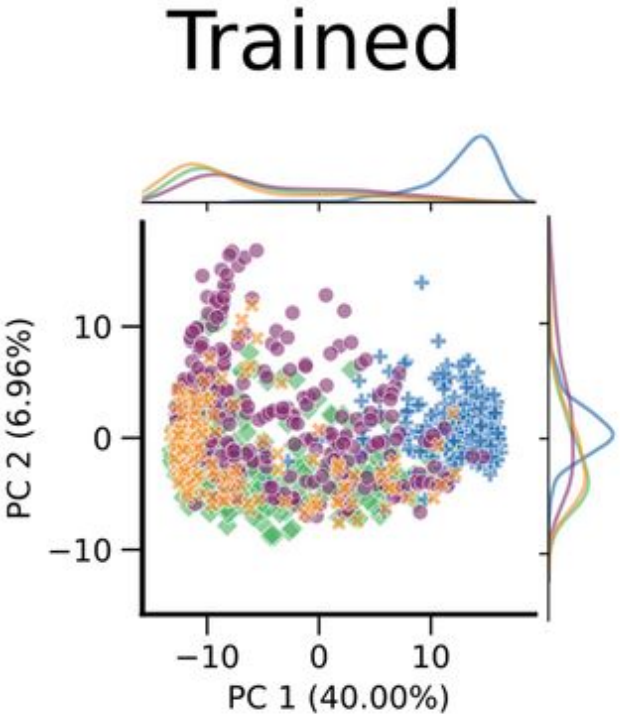




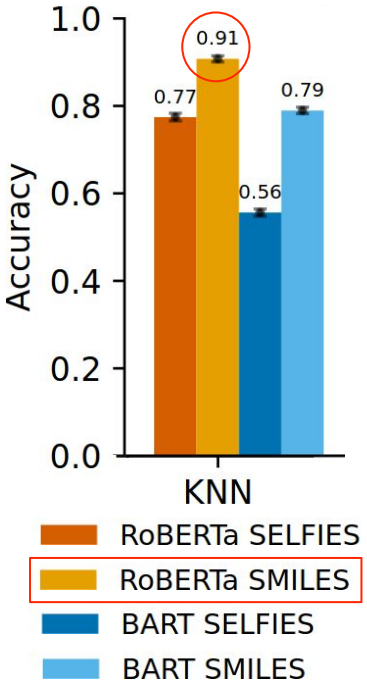
What about atoms?

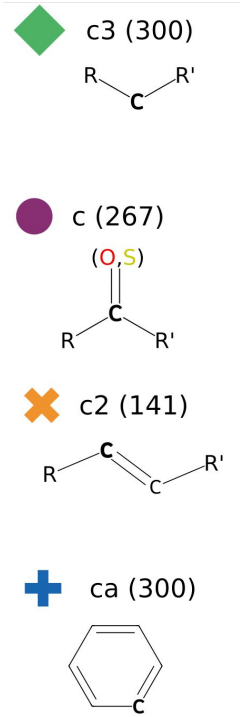


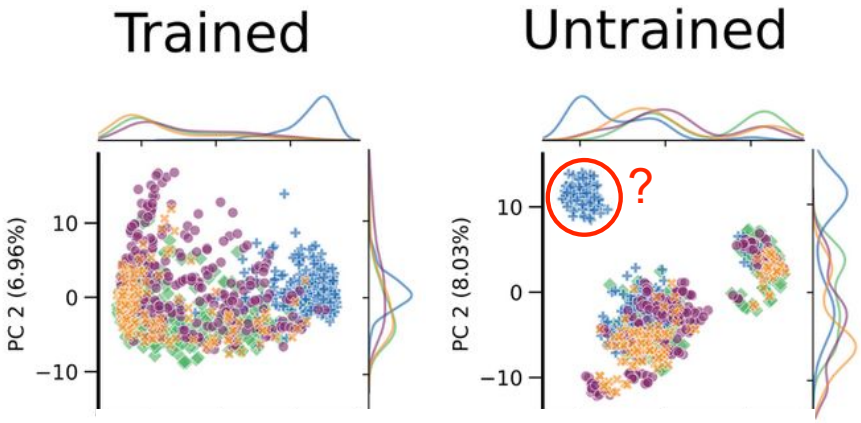




What about atoms?

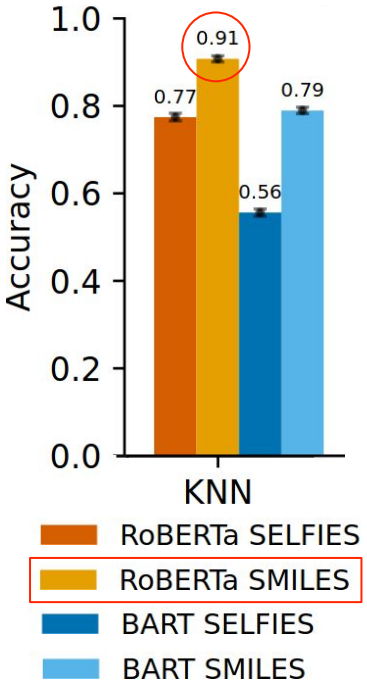


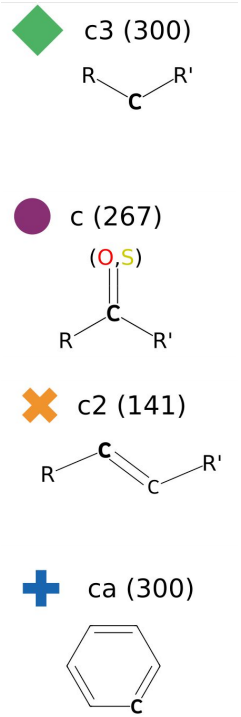


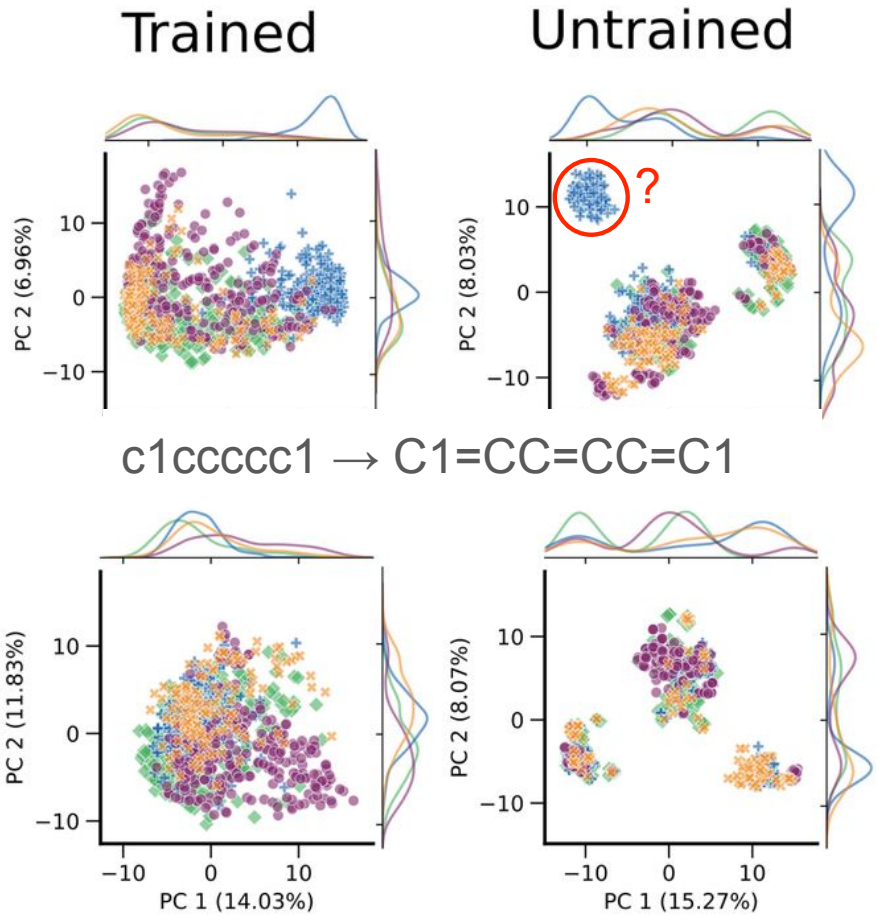


c1ccccc1 → C1=CC=CC=C1

What about atoms?







There may not be a single best configuration, but..

- The way the models leverage representations and how and what details they learn differs
- There are **three main design choices** to create a CLM:
  - Representation
  - Tokenization
  - Model
- Latent space tells us how the model **internally understands chemistry**

Acknowledgements

This summer school talk is peer-reviewed .

The paper:
https://doi.org/10.1186/s13321-025-01099-w

CHECK OUT OUR BLOG POST!

Jannik Gut

Prof. Dr. Thomas Lemmin
Noah Kleinschmidt
Calvin Klein
Lea Brönnimann
Giulia Peteani
Symela Lazaridi
Axel Giottonini

UNIVERSITÄT BERN