AI
&
material search

Fraunhofer SCAI

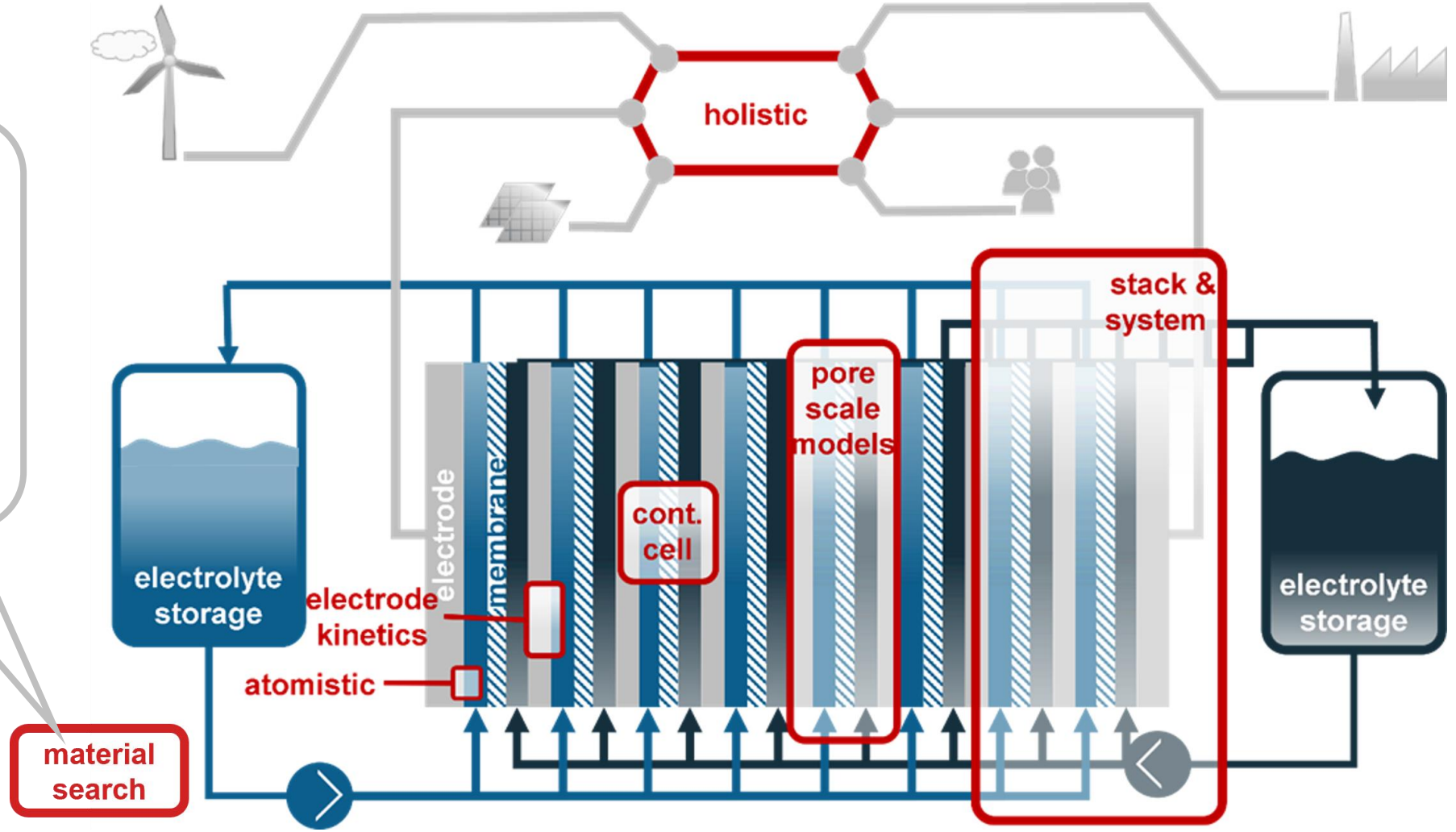
Workshop 2023

St.Augustin, 21.09.2023

SONAR at a glance

task WP1:

tools to identify
organic
affordable
redox-active
materials



Fraunhofer
DTU Danmarks Tekniske Universitet
cnrs CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
zhaw
KIT Karlsruher Institut für Technologie
UNSW SYDNEY

Data driven models

search

organic chemical space ~10^{23-60}

for high energy density at low cost

speed & focus

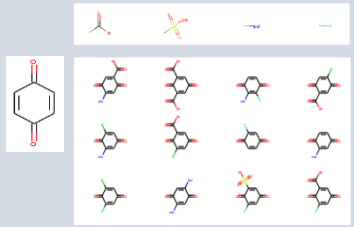
efficient down-selection

cheap filters first

then expensives

all fast

virtual library



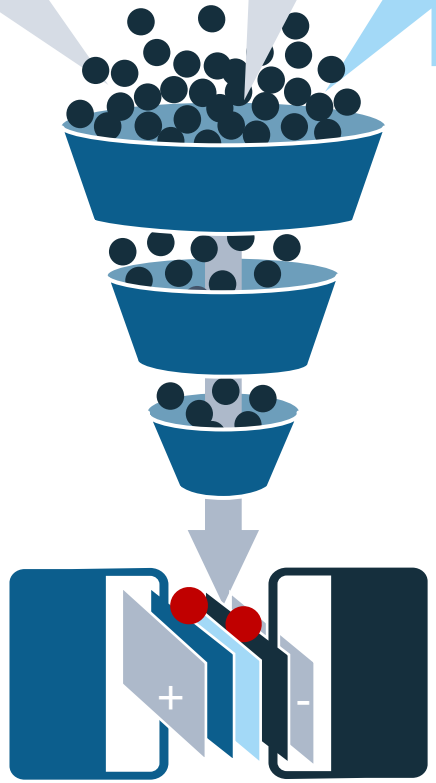
vendor compounds

generative models for new candiates

- VAE
- diffusion models
- transformers

predictive models for screening

1. compounds
   - synthetic accessibility
   - solubility
   - redox potential
2. negolyte/posolyte pairs
   - basic cell performance
   - hand over to ZHAW



Fraunhofer

DTU Danmarks Tekniske Universitet

cnrs CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

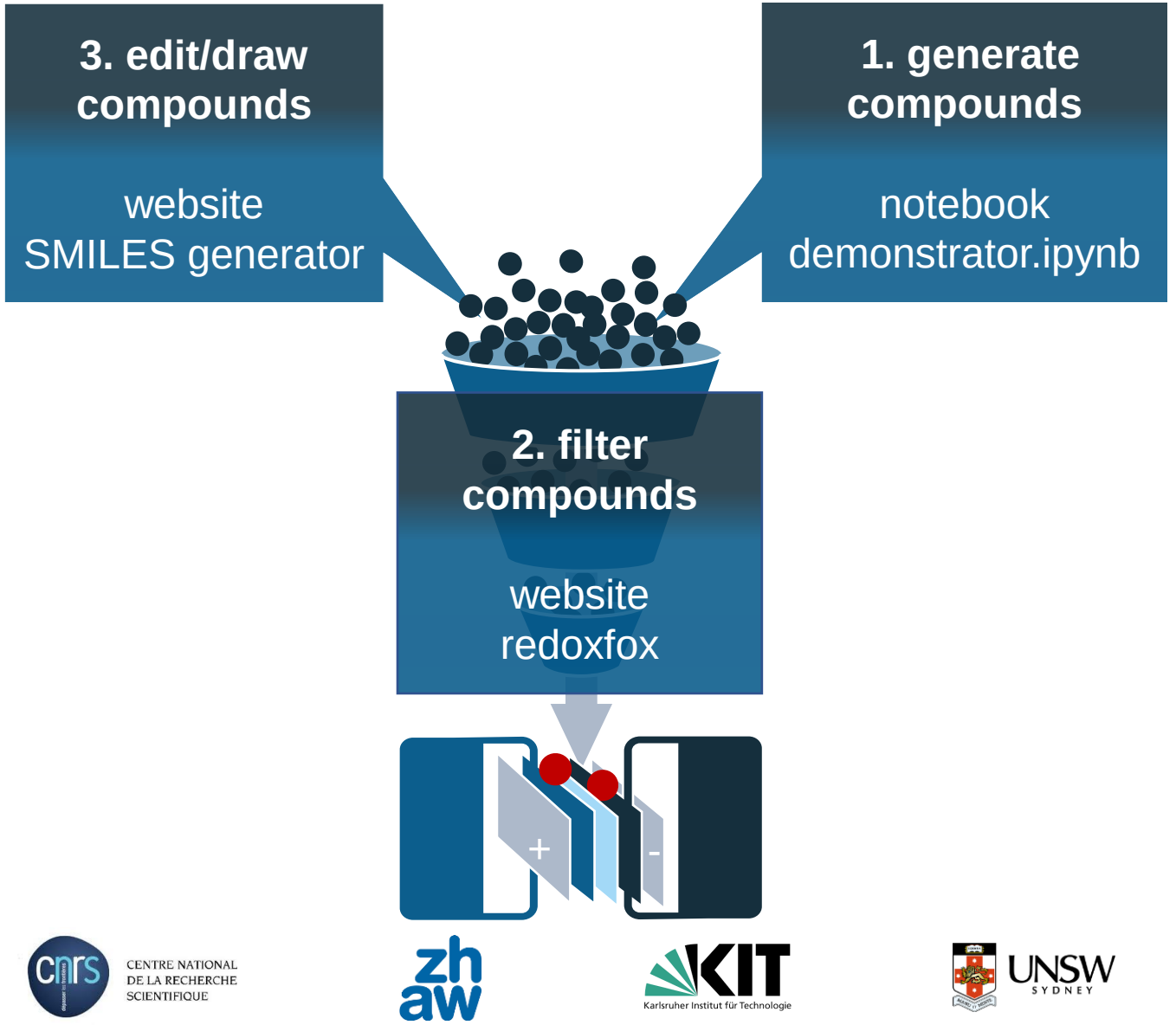
zhaw

KIT Karlsruher Institut für Technologie

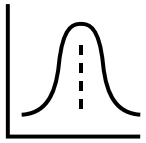
UNSW SYDNEY

Hands on session

+ page 20 in handout
+ jupyter notebook
+ webinterface
  – chromium-browser &



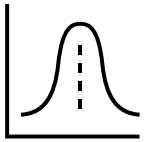
Background



Underlying Distribution

Dataset: (x,y) pairs
x - Molecule
y - Property

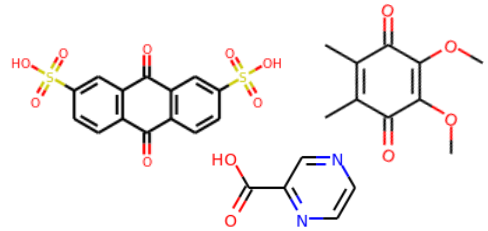
Train



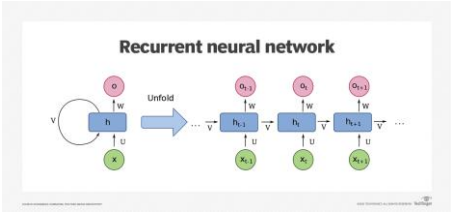
Approximate
Unconditional: x ~ G
Conditional: x ~ G(y)

Generator **G**

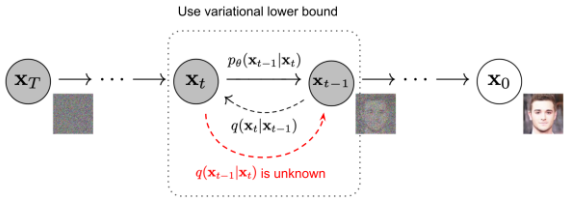
Sample



RNNs



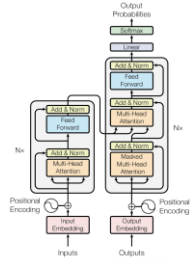
Diffusion



Transformer

BERT

Encoder



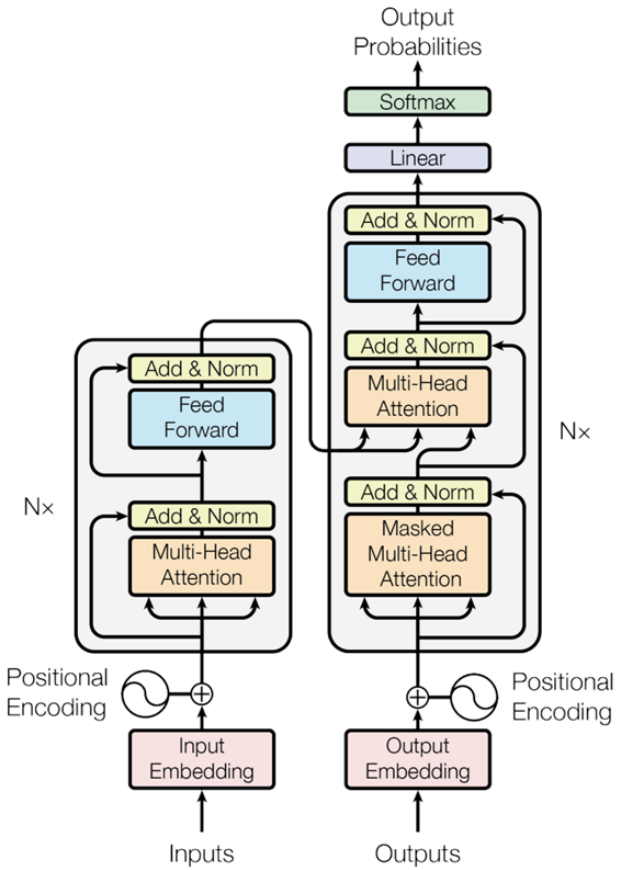
GPT

Decoder

Transformer

BERT

Encoder



GPT

Decoder

Next Token

Softmax

Add & Layer Normalization

Feed Forward

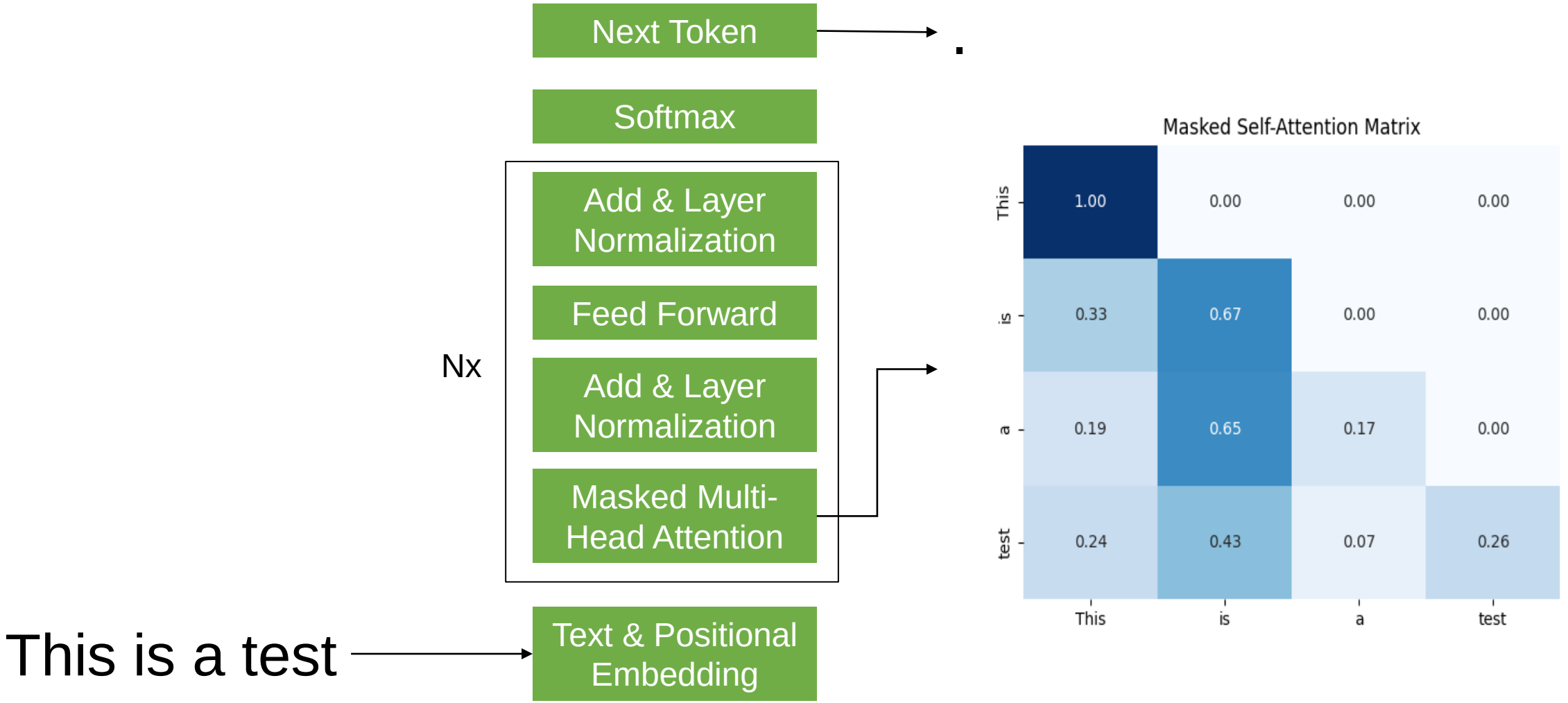
Add & Layer Normalization

Masked Multi-Head Attention

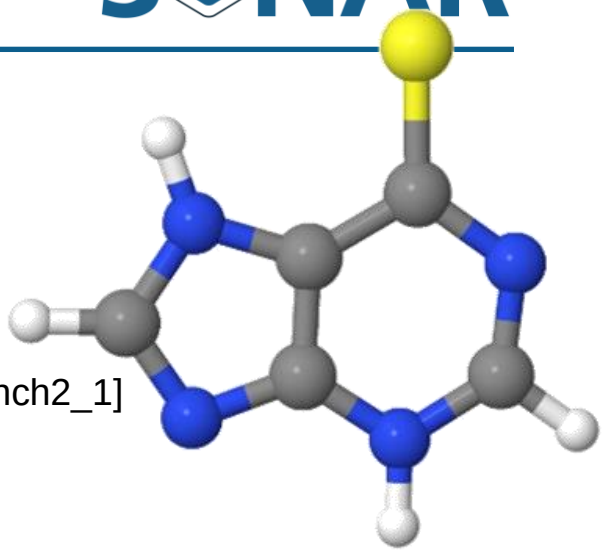
xN

Text & Positional Embedding

Transformer – Basic Generation



Molecule representation



2D

InChi InChI=1S/C5H4N4S/c10-5-3-4(7-1-6-3)8-2-9-5/h1-2H,(H2,6,7,8,9,10)

SMILES C1=NC2=C(N1)C(=S)N=CN2

SELFIES [C][=N][C][=C][Branch1_1][Ring2][N][Ring1][Branch1_1][C][Branch1_2][C][=S][N][=C][N][Ring1][Branch2_1]

3D

Coordinates

C	-2.28752	-0.09279	0.00131
N	-1.52093	-1.16587	0.00160
C	-0.26385	-0.66545	0.00005
C	-0.23058	0.70099	-0.00110
N	-1.55044	1.06537	0.00000
⋮	⋮	⋮	⋮

Descriptors

- matter of choice

→ meaningful

→ Cheap

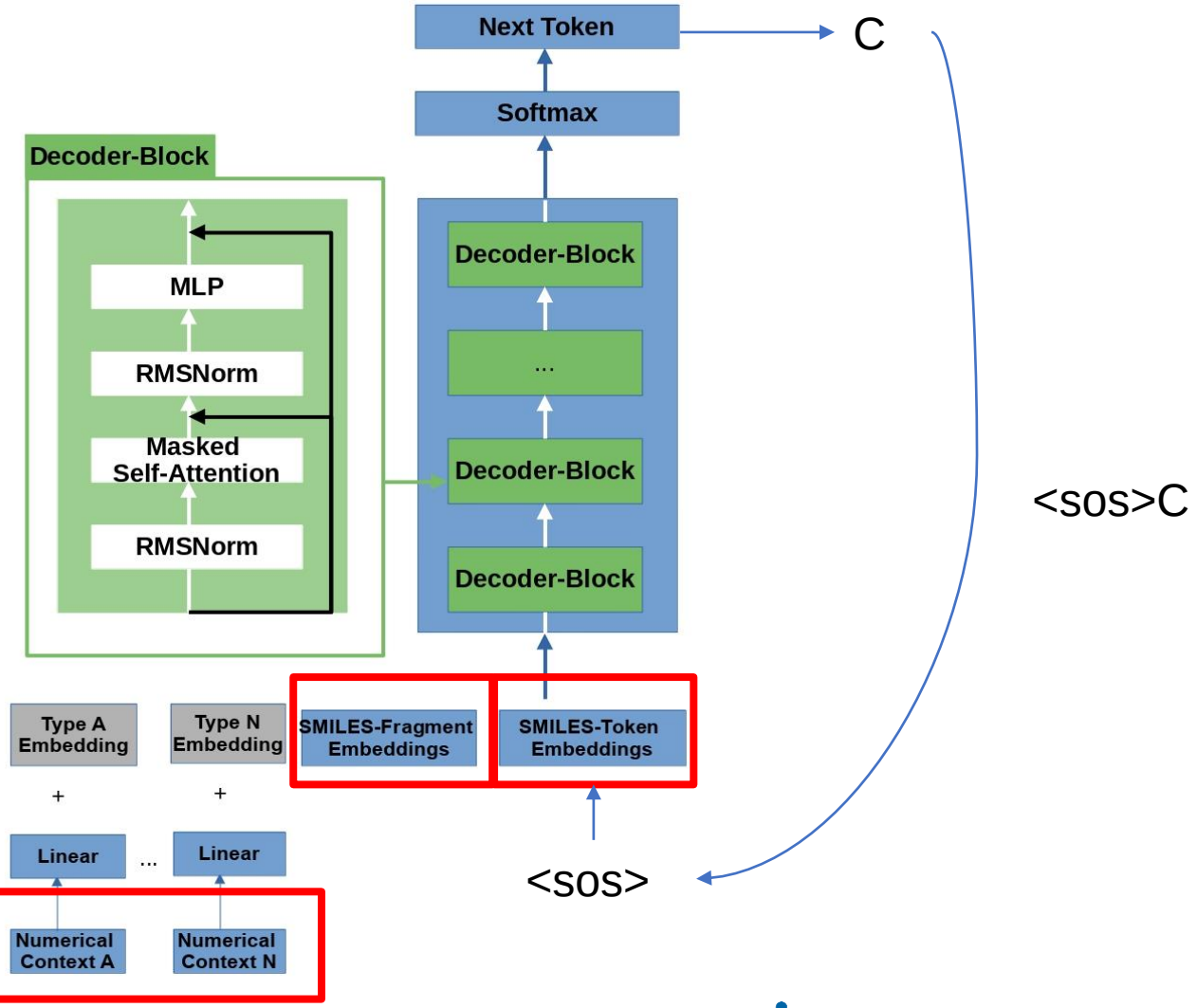
→ LogP, SAScore and Molecular Weight

Fraunhofer DTU Danmarks Tekniske Universitet cnrs CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE zhaw KIT Karlsruher Institut für Technologie UNSW SYDNEY

Molecule representation – Conditions

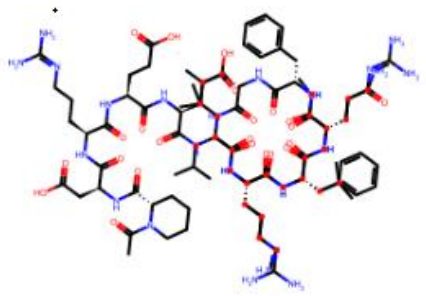
- **LogP** [-4, 8]: How hydrophobic is a molecule?
  - Negative: soluble in water
  - Positive: attracted to non-polar environments
- **SAScore** [0,10]: ease(0) or difficulty(10) to synthesize a molecule
- **Molecular Weight** [0, 10]: proxy for the size of a molecule
  - We divide it by the constant 100 to „normalize“
- **SMILES Fragment**: Can be a full molecule or a part of a SMILES e.g. „c1cccc1“ or „csc“

Transformer – Molecule Generation

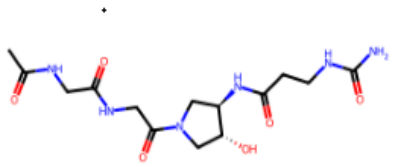


LogP
SAScore
Molecular Weight

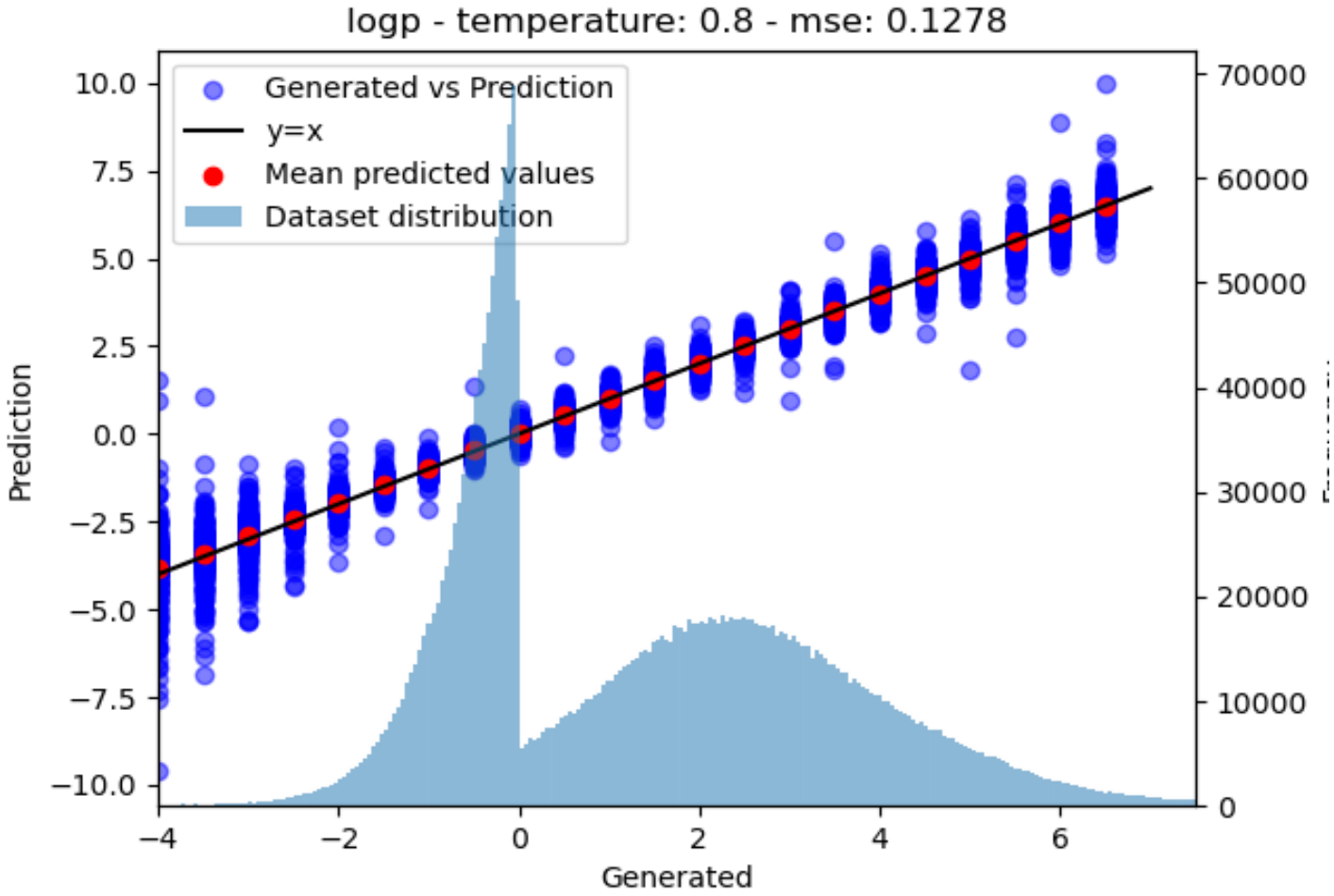
Transformer – Performance LogP

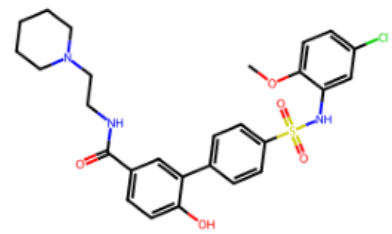


logp: -4.0 vs -4.65

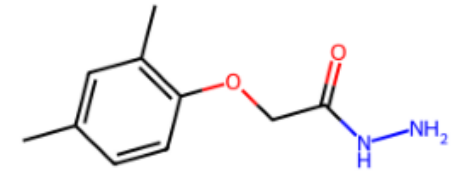


logp: -4.0 vs -4.01



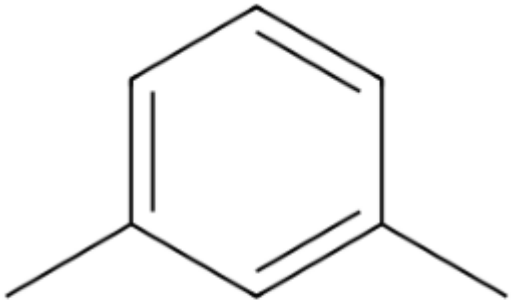


logp: 4.5 vs 4.74

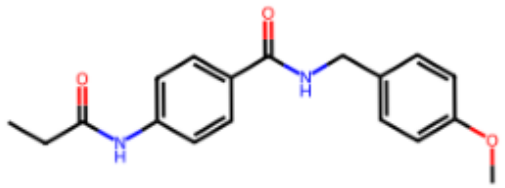


logp: 0.5 vs 0.67

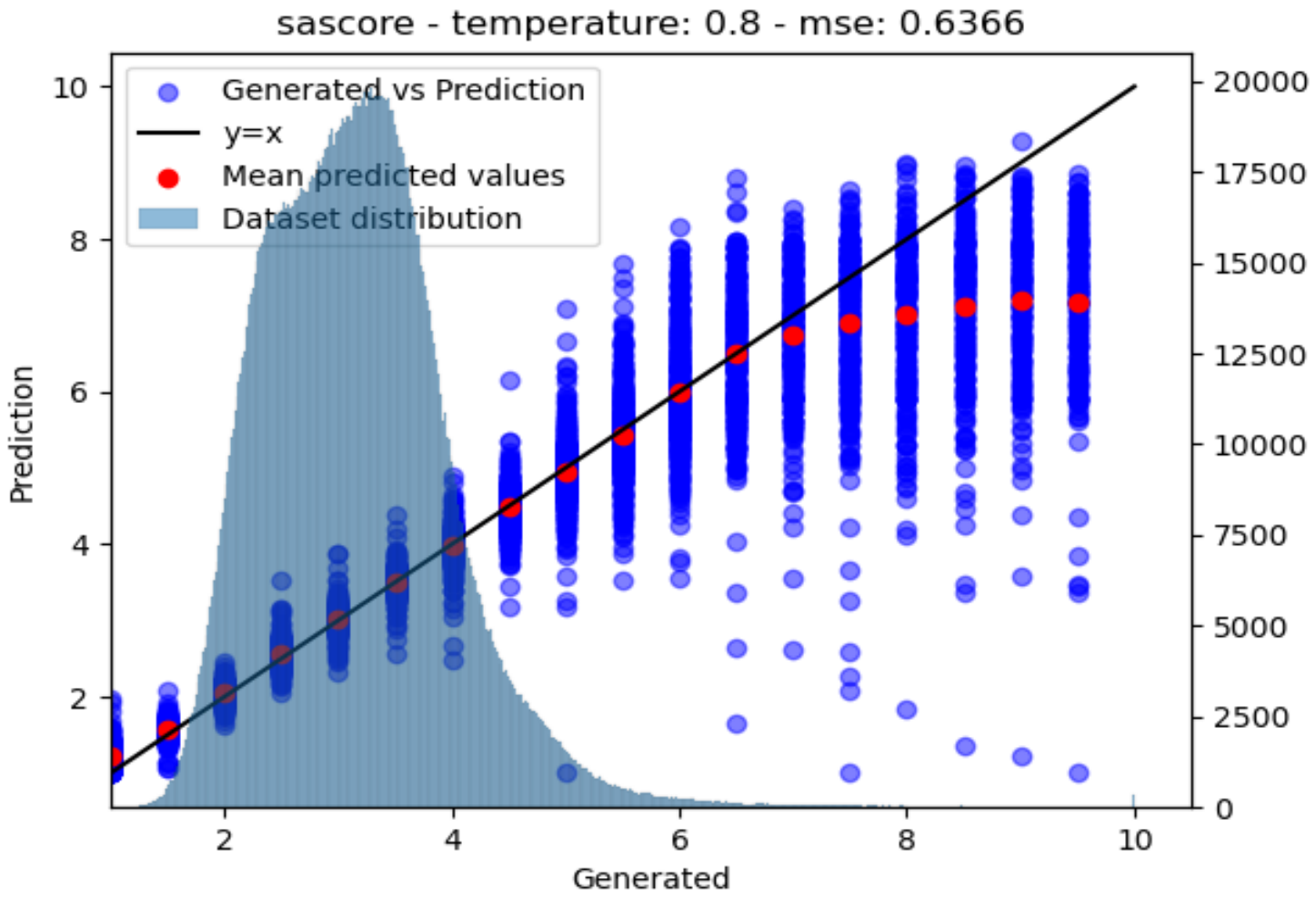
Transformer – Performance SAScore

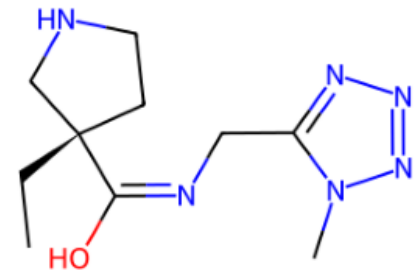


sascore: 0.5 vs 1.14

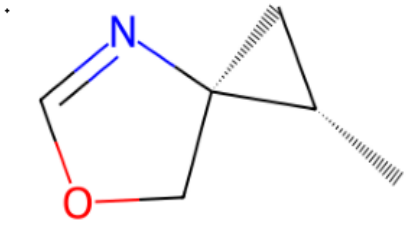


sascore: 1.5 vs 1.54



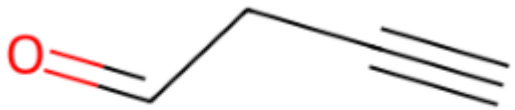


sascore: 4.0 vs 3.93

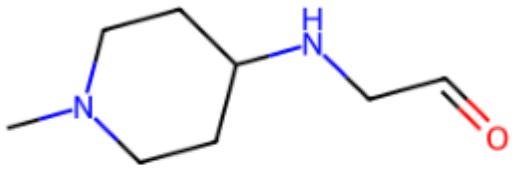


sascore: 5.5 vs 5.53

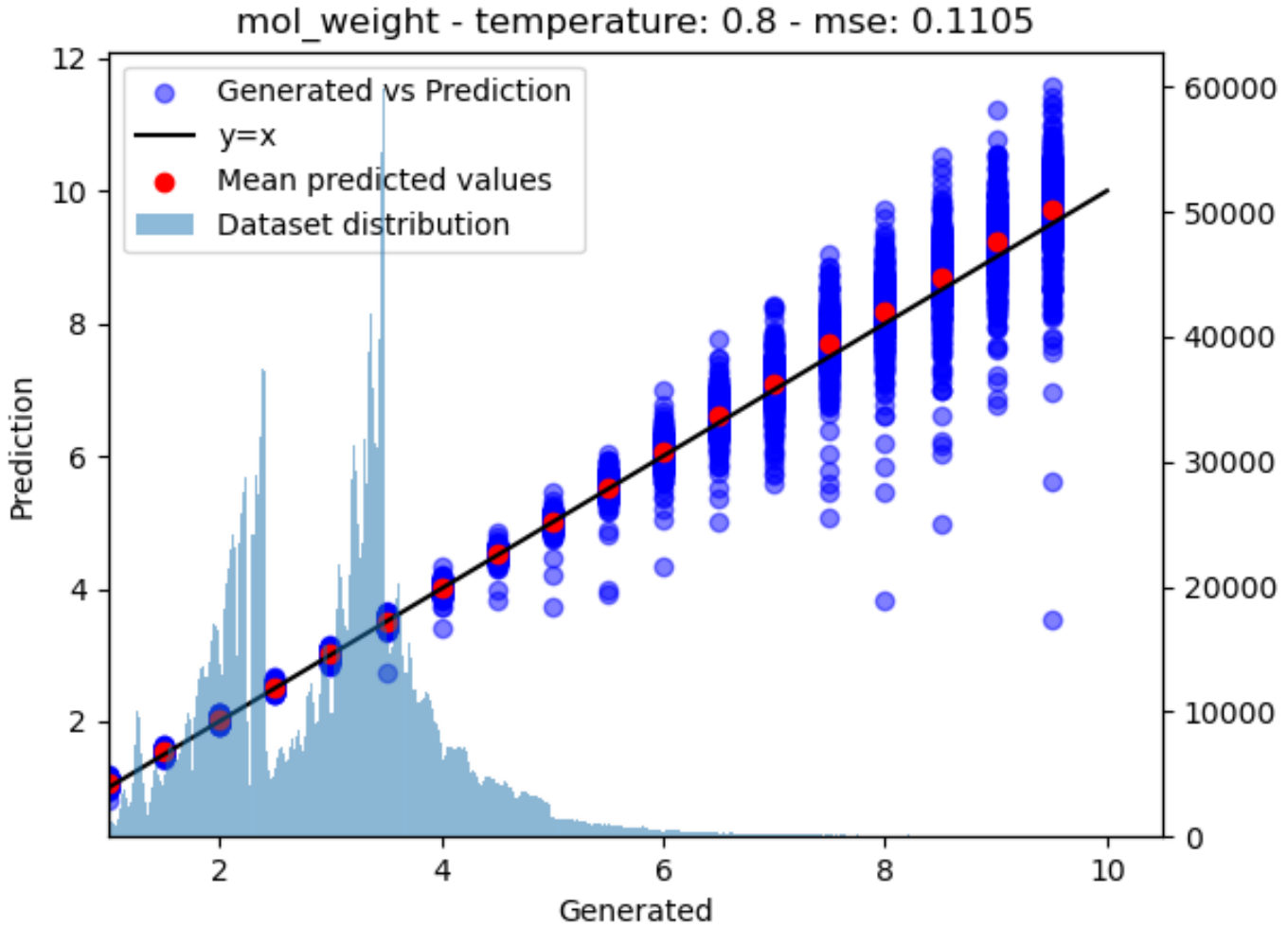
Transformer – Performance Molecular Weight

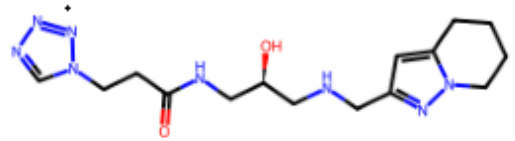


mol_weight: 0.5 vs 0.68

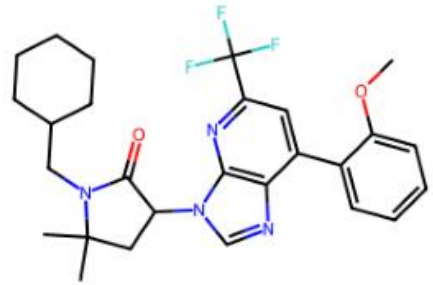


mol_weight: 1.5 vs 1.56



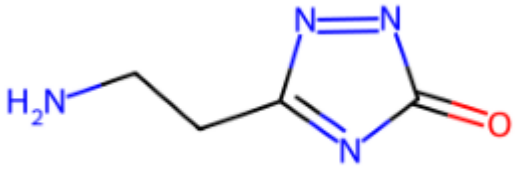


mol_weight: 3.5 vs 3.48



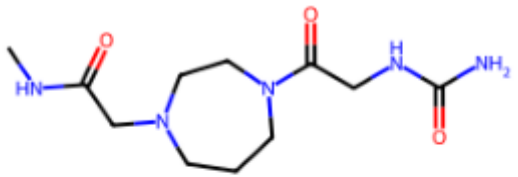
mol_weight: 5.0 vs 5.0

Transformer – Performance Multi Condition



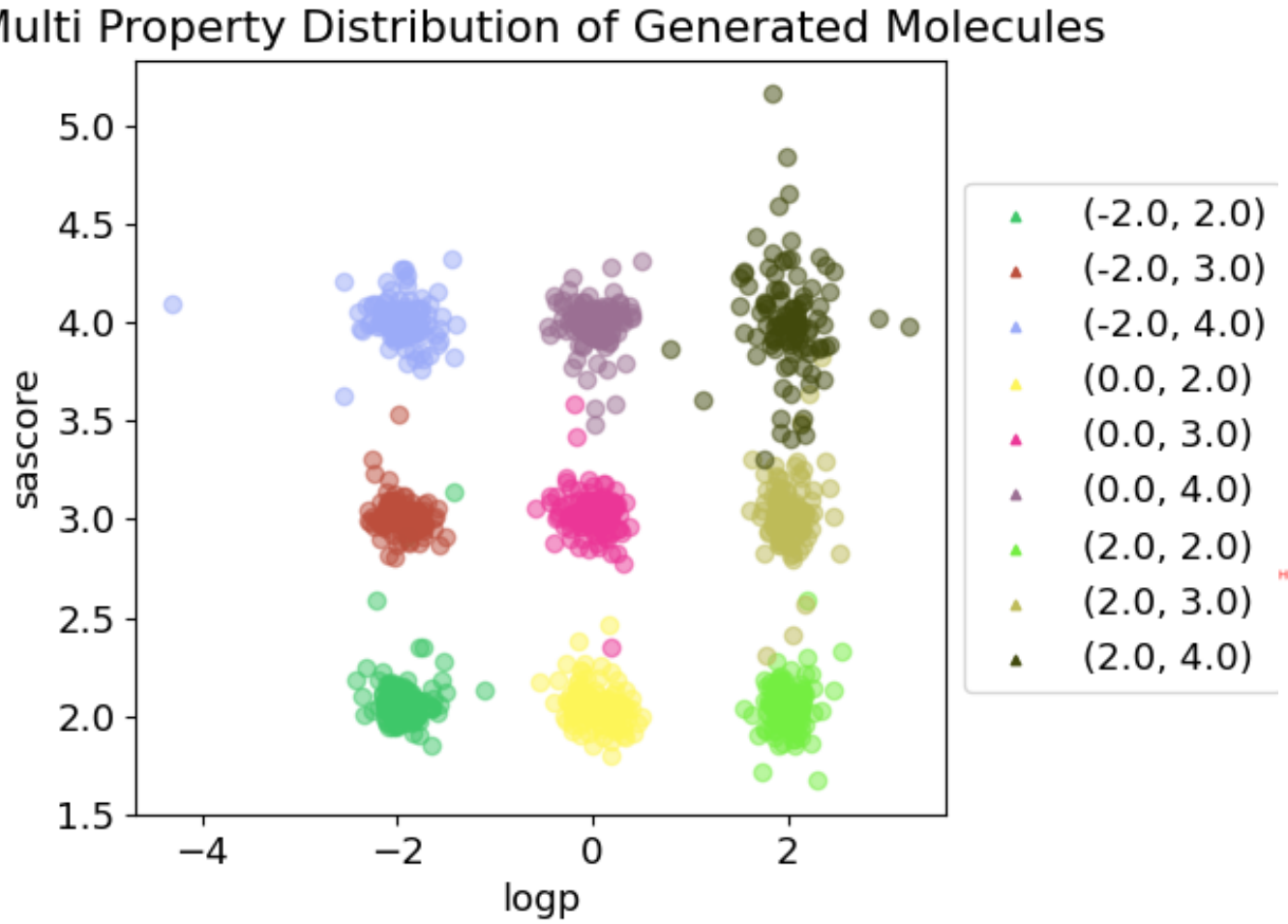
logp: 0.5 vs 0.32

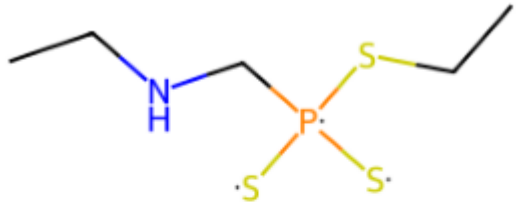
sascore: 4.0 vs 4.12



logp: -2.0 vs -2.06

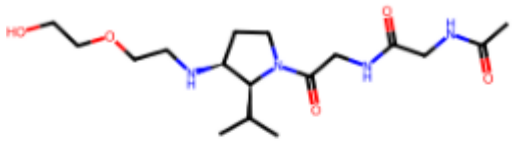
sascore: 2.0 vs 2.13





logp: 3.0 vs 3.46

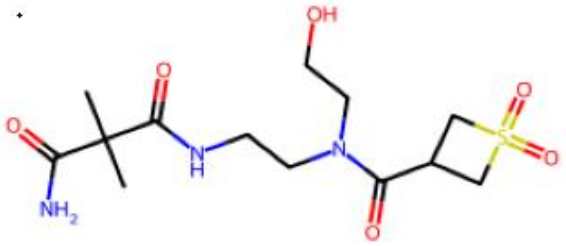
sascore: 5.5 vs 4.45



logp: -1.5 vs -1.54

sascore: 3.5 vs 3.53

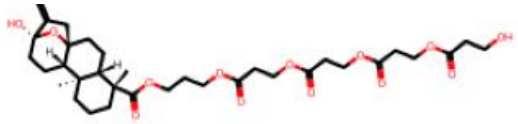
Transformer – Performance Multi Condition



logp: -2.5 vs -2.52

sascore: 3.0 vs 2.95

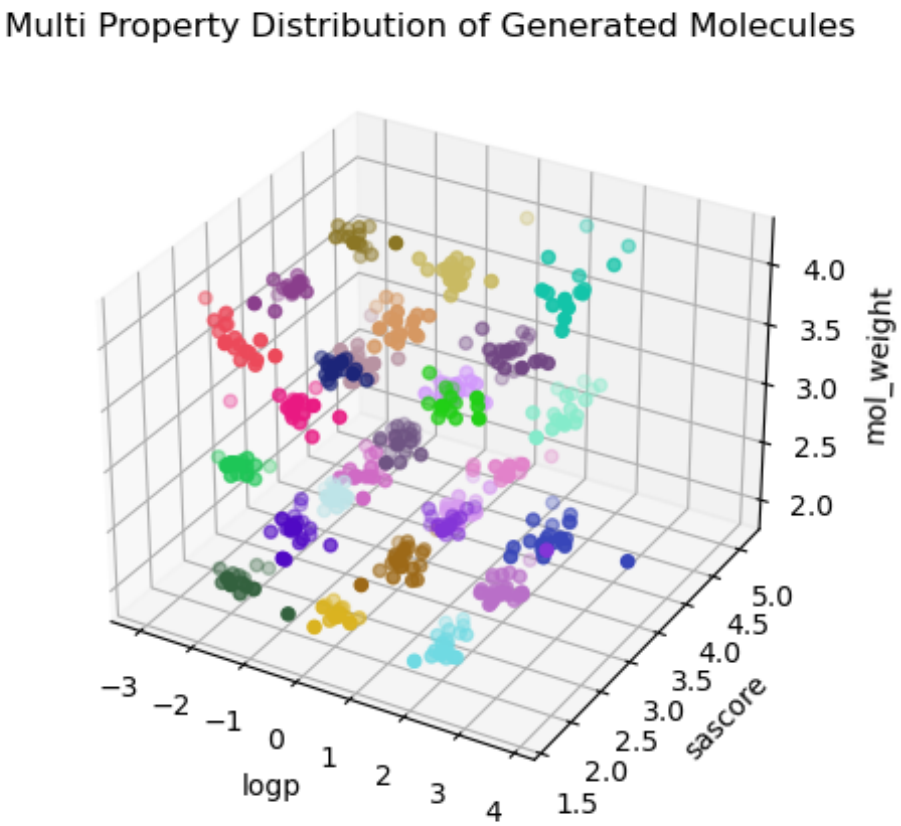
mol_weight: 3.5 vs 3.49

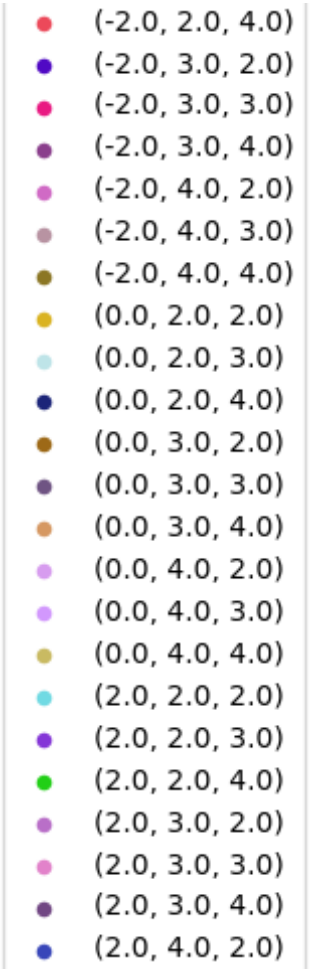


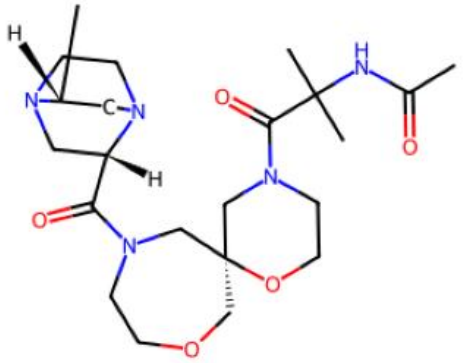
logp: 3.0 vs 3.07

sascore: 6.0 vs 6.07

mol_weight: 6.5 vs 6.66



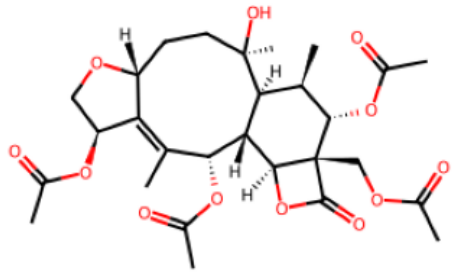




logp: -1.0 vs -1.25

sascore: 5.5 vs 5.61

mol_weight: 4.5 vs 4.51



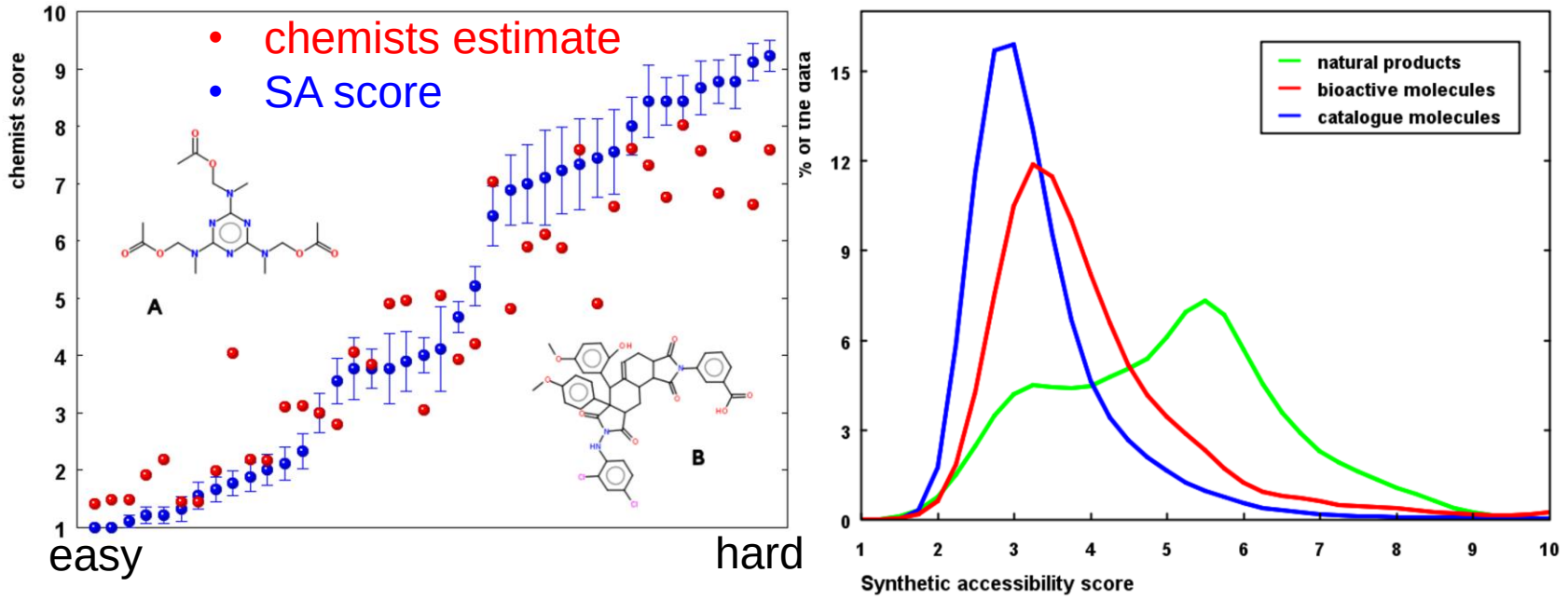
logp: 2.0 vs 1.4

sascore: 6.0 vs 5.92

mol_weight: 5.5 vs 5.66

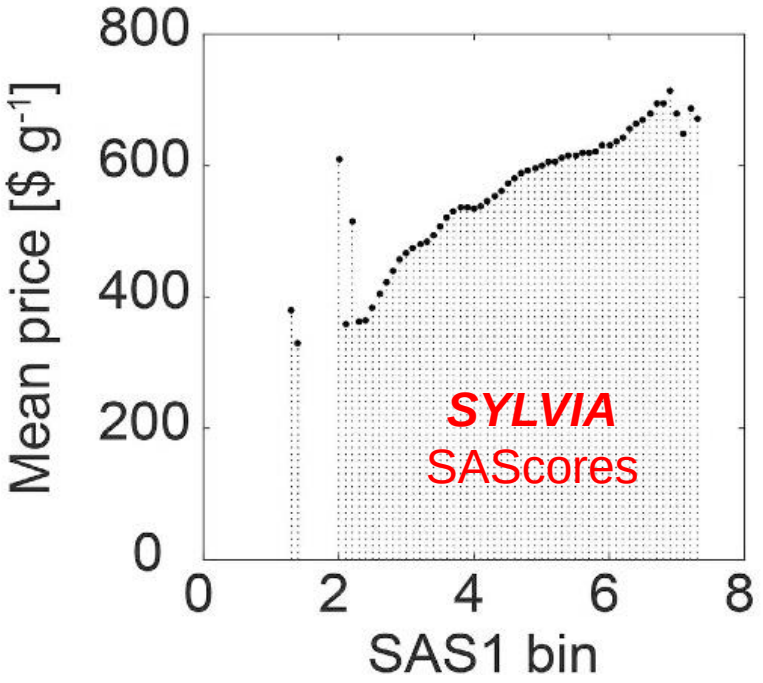
Synthetic accessibility score

Estimate SA score of drug-like molecules based on molecular complexity and fragment contributions



Molecular descriptor data explain market prices of a large commercial chemical compound library

Polanski et al, 2016

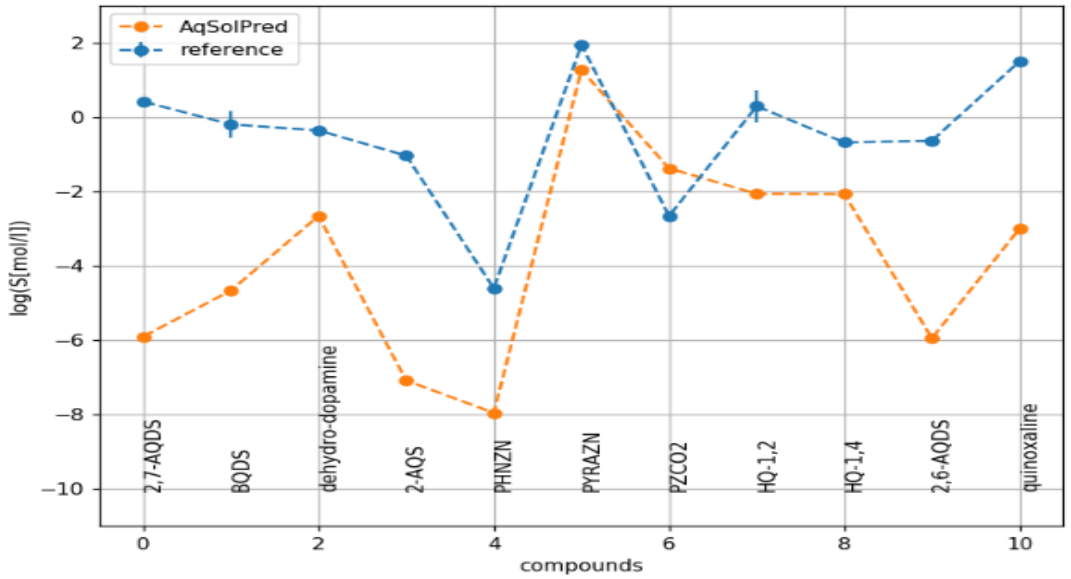


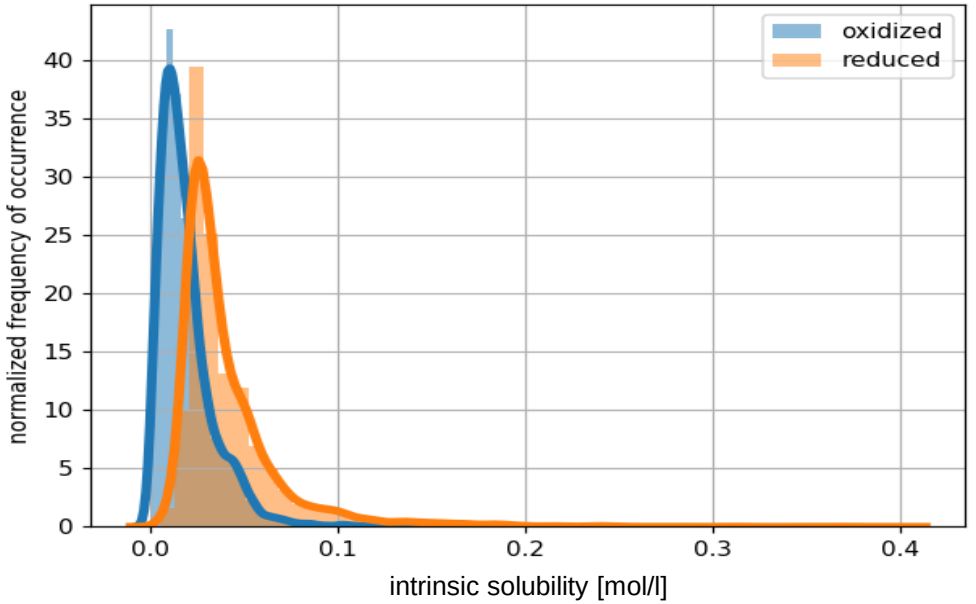
Ertl & Schuffenhauer **2009**. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, *1*, 1-11.

Solubility by AqSolPred v1.0

consensus model (Sorkun, Vianney, Koelman & Er **2021,** Pushing the limits of solubility prediction via quality-oriented data selection' *science* 24.1 : 101961)

trained on ~10k experimental data (Sorkun, Khetan & Er. **2019,** 'qSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds, Scientific data 6.1, 1-8)





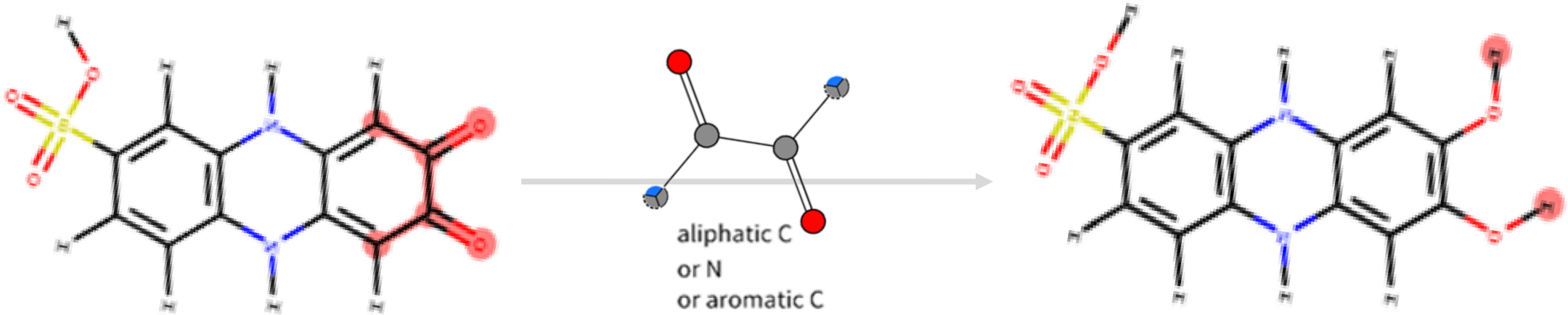
→ *intrinsic* solubility prediction neglects (de)protonation processes

→ oxidized states considered less soluble than reduced ones

Redox potential via ΔH

$$\mathbf{A_{ox}} + \mathbf{H_2} \rightarrow \mathbf{A_{red}} \qquad \Delta G \cong \Delta H_{G4(MP2)} \qquad E^o = -\frac{\Delta G}{zF}$$

neutral, organic, closed shell compounds in gas phase



SMARTS reaction templates

[OX1:1]=[CX3:2](-,:[C,N,n,c:3])-,:[CX3:4](-,:[C,N,n,c:5])=[O:6].[H:7]-[H:8]

>> **[H:7]-[O:1]-[C:2]([C,N,n,c:3])=[C:4]([C,N,n,c:5])-[O:6]-[H:8]**

SMARTSviewer smartsview.zbh.uni-hamburg.de, ZBH Center for Bioinformatics, University of Hamburg

Fraunhofer DTU Danmarks Tekniske Universitet cnrs CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE zhaw KIT Karlsruher Institut für Technologie UNSW SYDNEY

Predicting ΔH

SONAR

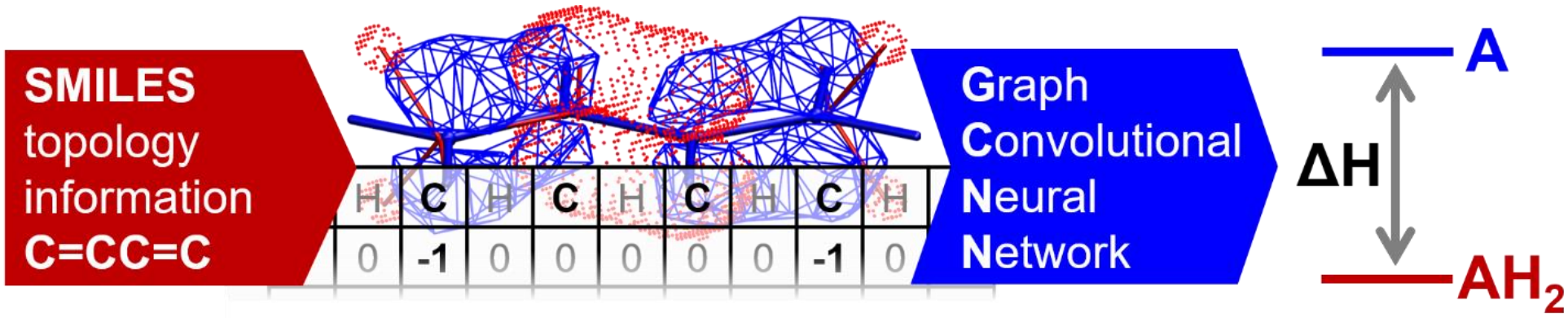
QM9

134 k
molecules +
heat of atomization
(Ramakrishnan (B3LYP), 2014,
Kim: G4(MP2), 2019)

define
reactants & products

qm9

45 k
A/AH_2 pairs +
heat of reaction



working with 2D structures: atomic features + bond graph

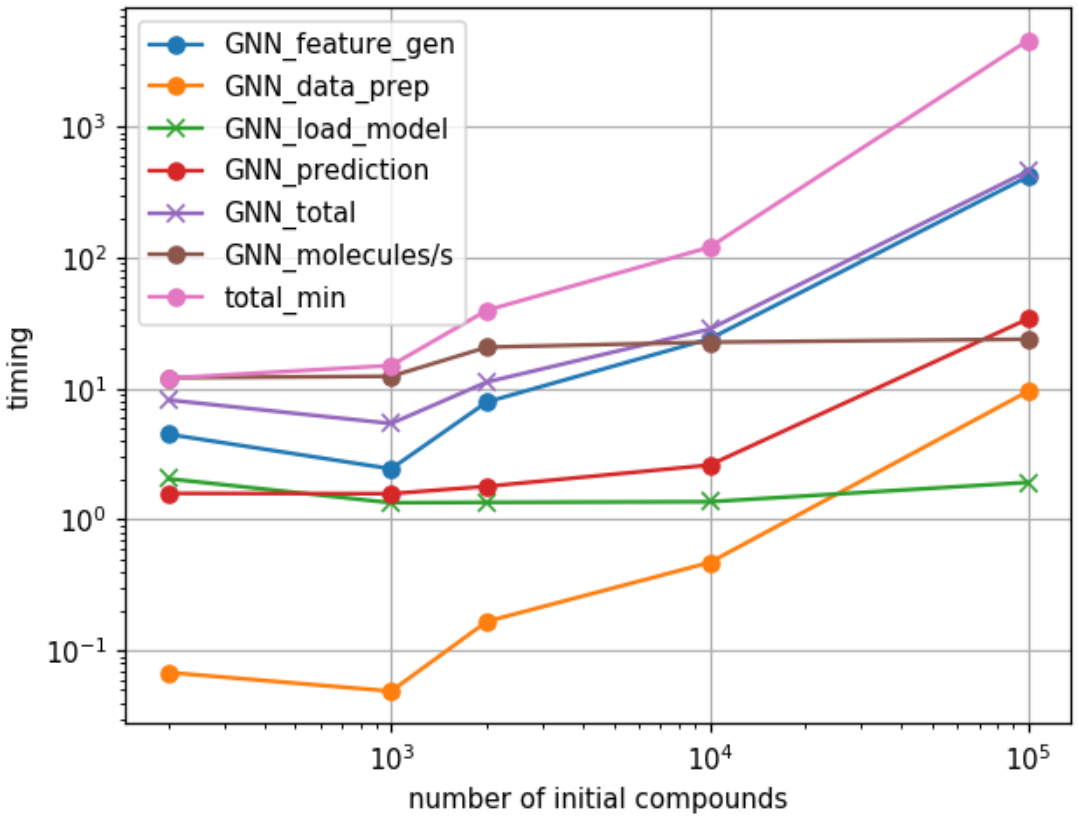
- generate differential features from 2D
- train model to heat of reaction

→ predict heat of reaction for new mols

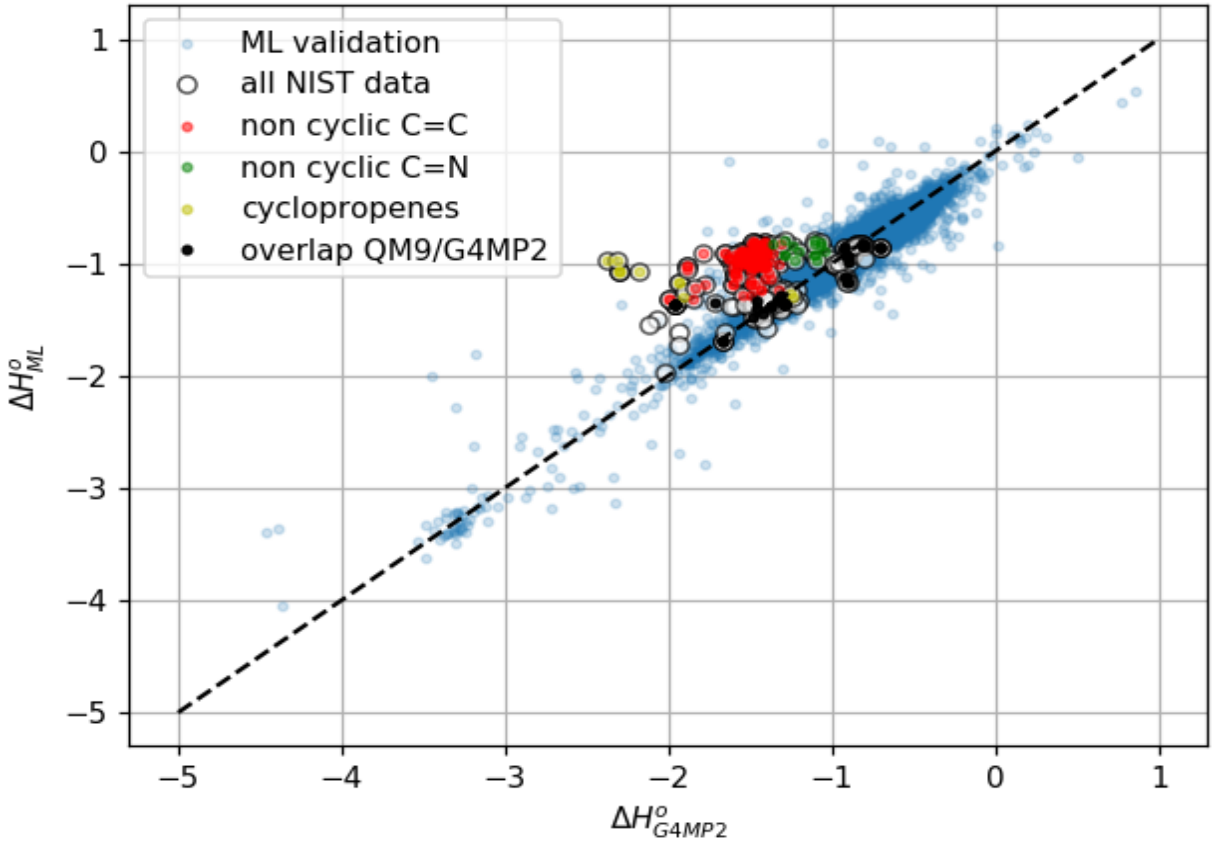
Barker, Berg, Hamaekers & Maass, **2021:** Rapid prescreening of organic compounds for redox flow batteries: a graph convolutional network for predicting reaction enthalpies from SMILES, *Batteries & Supercaps*, 4(9), 1482-1490.

Fraunhofer · DTU Danmarks Tekniske Universitet · cnrs Centre National de la Recherche Scientifique · zhaw · KIT Karlsruher Institut für Technologie · UNSW Sydney

Model performance



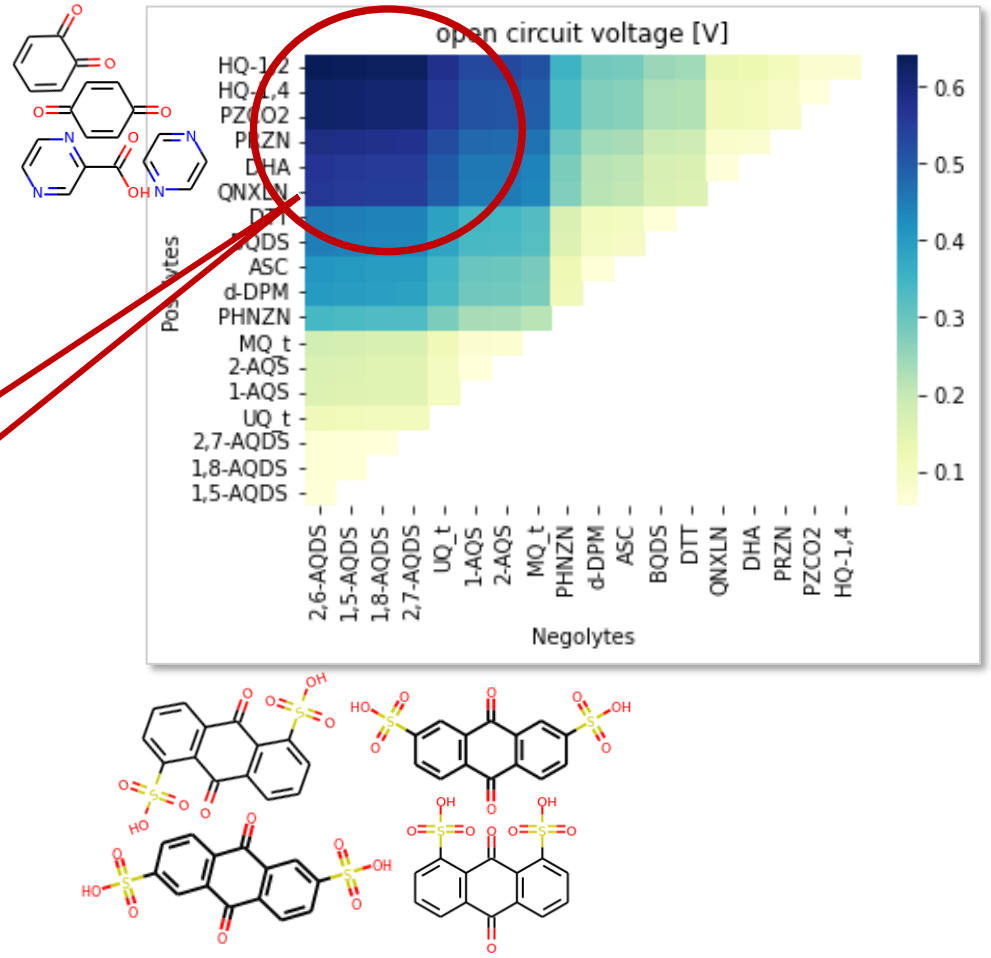
- 50 mols/sec – GNN feature generation most time consuming step
- 10e+5 compounds ≤ 2 hours on std desktop PC (4 cores, 8 GB memory)



- limitation of the chemical space represented within the training set

→ provide larger training set & train again

Full cell performance: OCV



reasonable matches

Fraunhofer

DTU Danmarks Tekniske Universitet

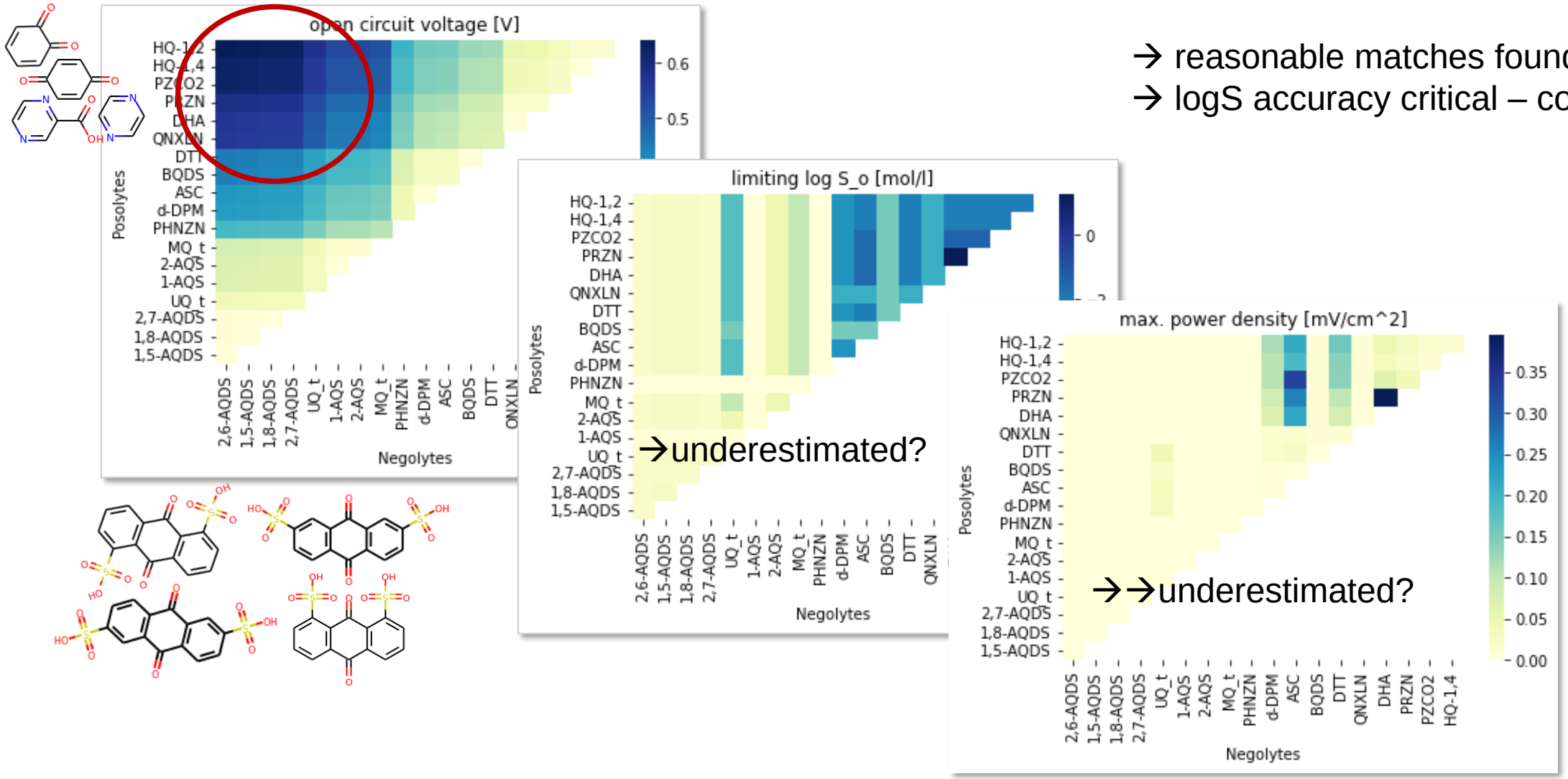
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

zhaw

KIT Karlsruher Institut für Technologie

UNSW SYDNEY

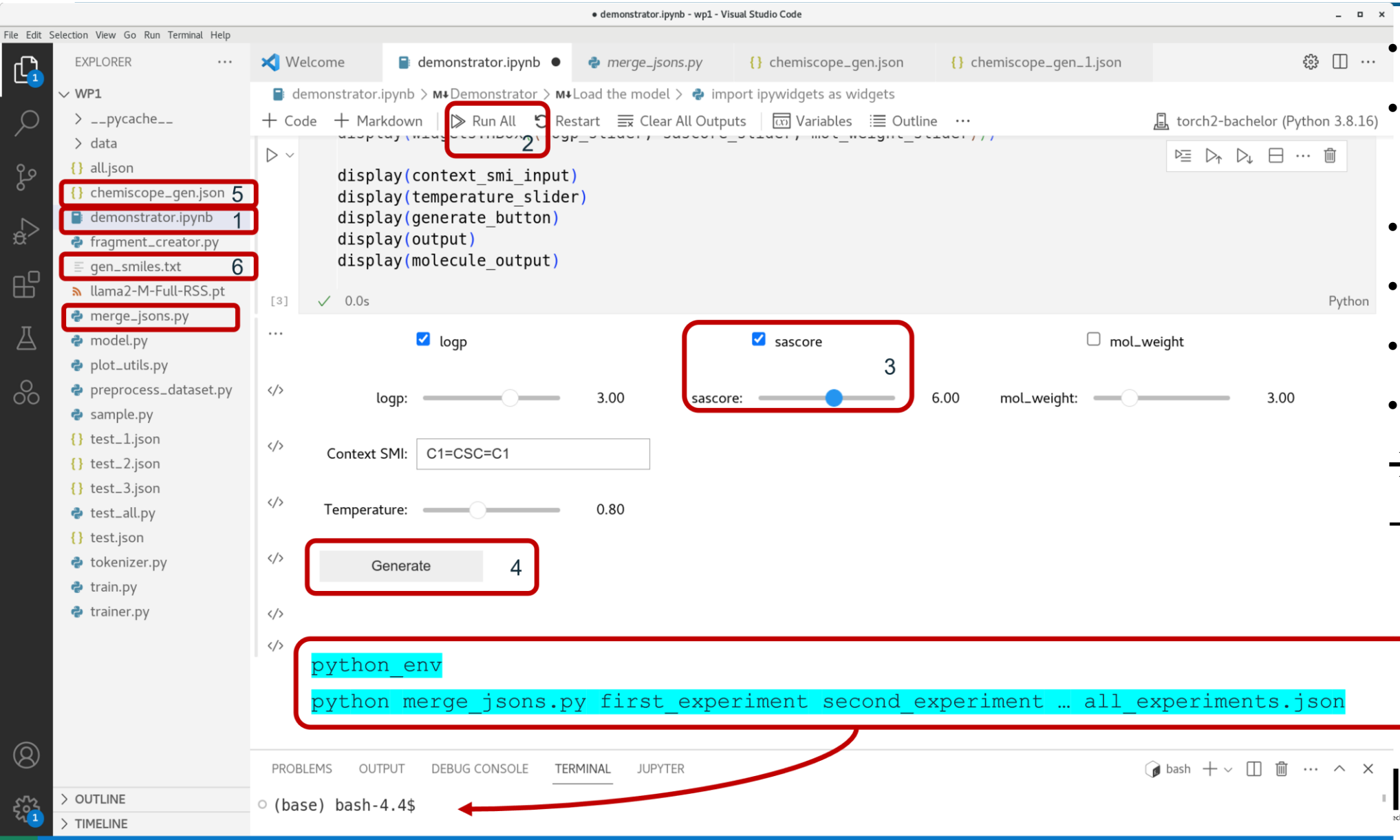
Full cell performance



→ reasonable matches found
→ logS accuracy critical – consider pH

Fraunhofer
DTU Danmarks Tekniske Universitet
CNRS Centre National de la Recherche Scientifique
zhaw
KIT Karlsruher Institut für Technologie
UNSW Sydney

Generate compounds

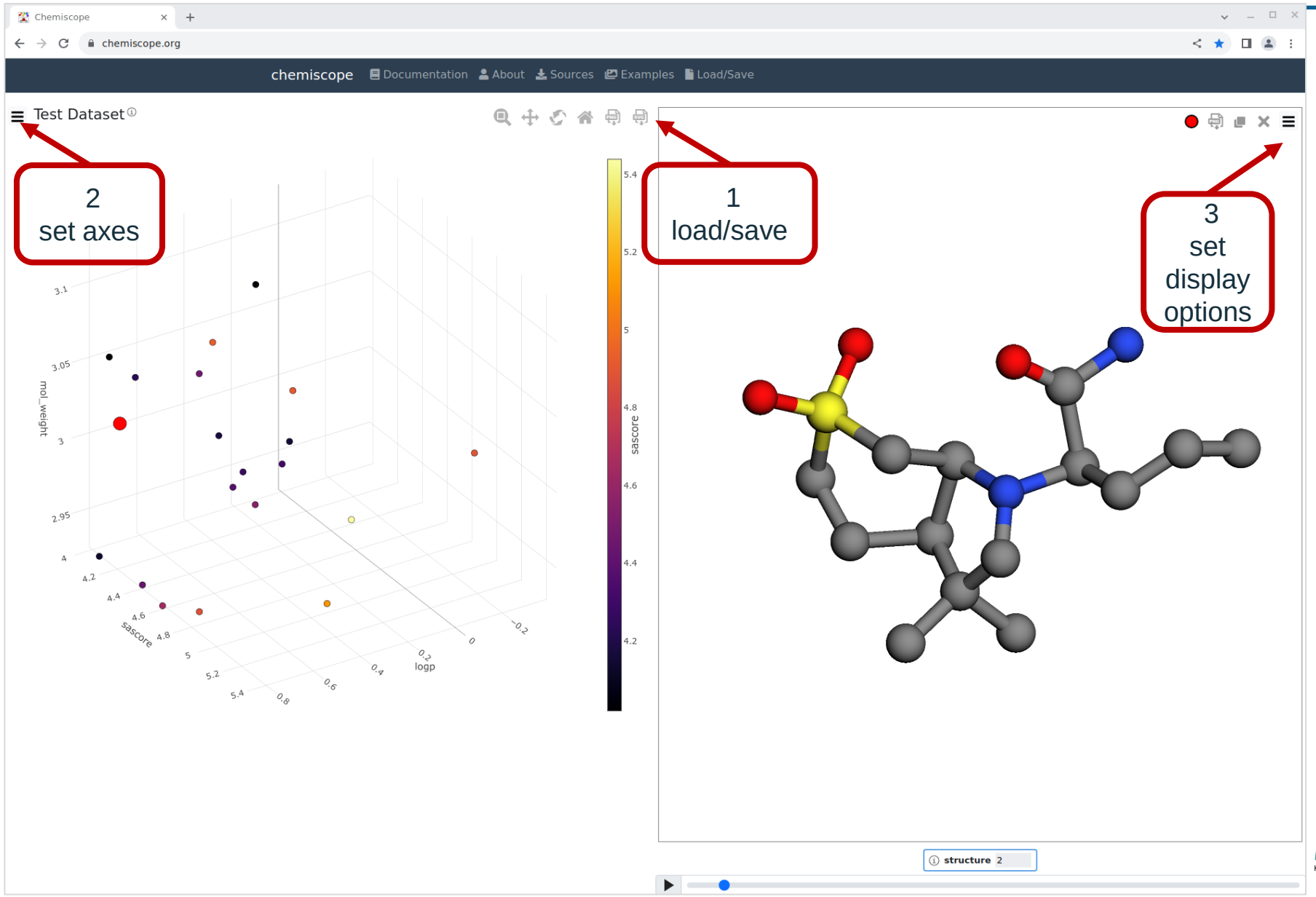


- code .
- open *demonstrator.ipynb*
- **>> Run All**
- scroll down
- [specifiy targets]
- **Generate**

→ scroll down

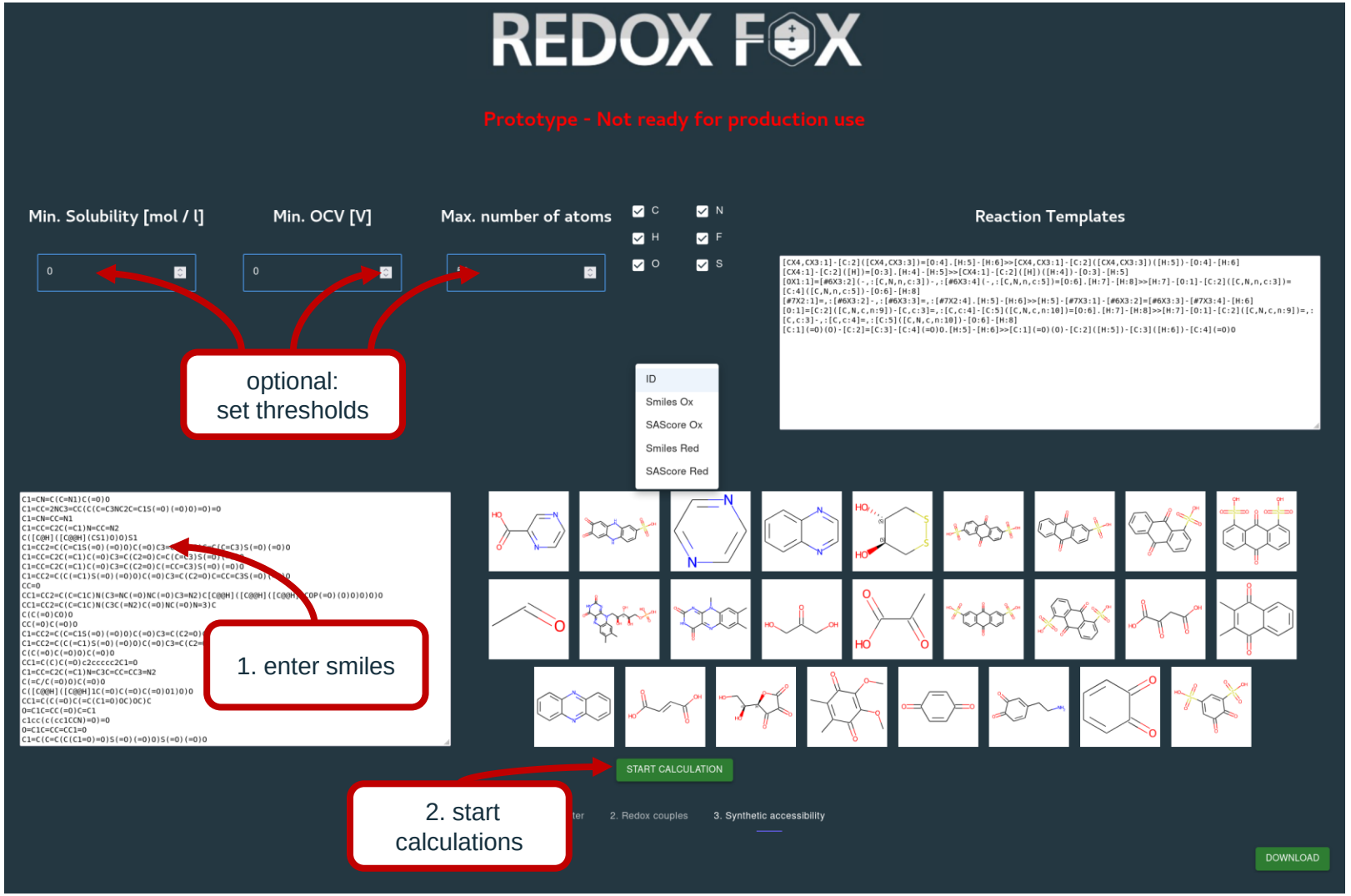
→ continue further analyses with output *chemiscope_gen.json* and/or *gen_smiles.txt*

View results



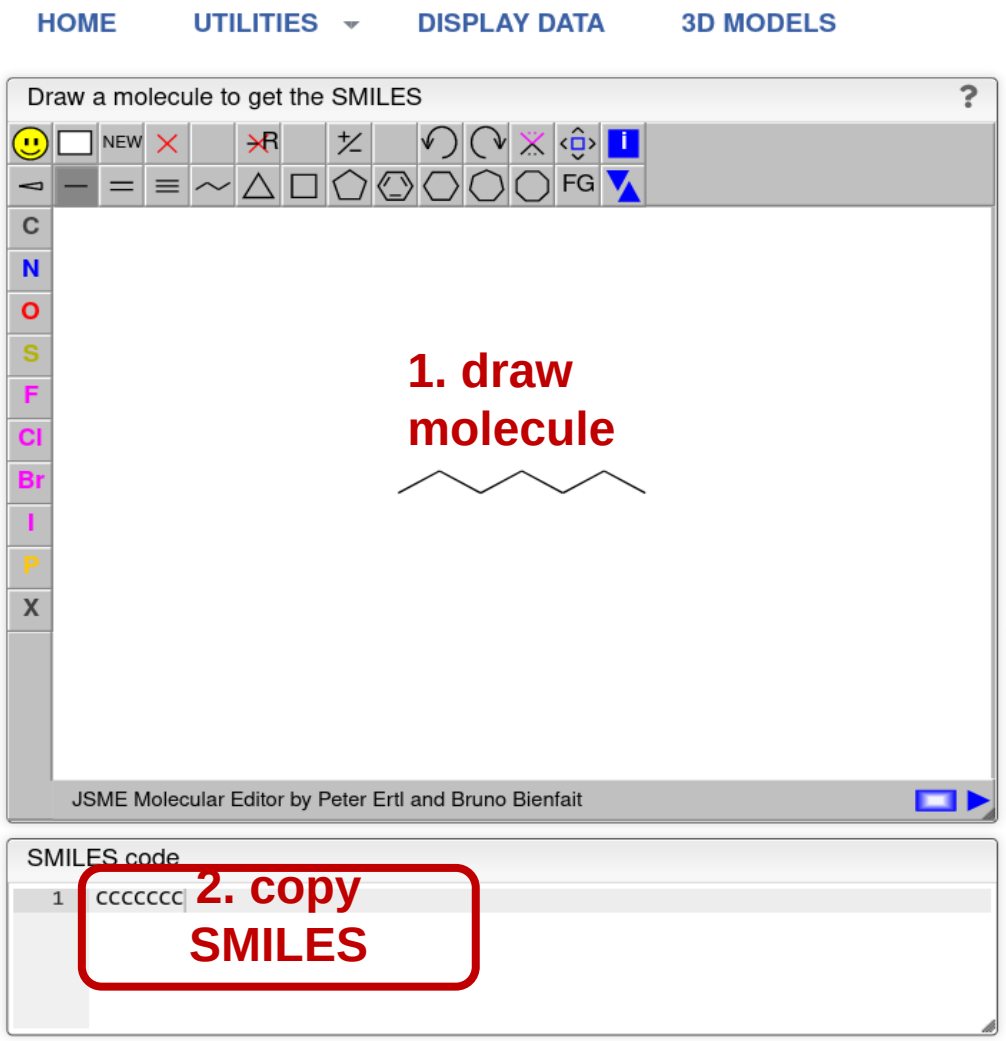
- open Tab Chemiscope in chromium-browser &
- upload *chemiscope_gen.json*

Screening



- open Tab **Redoxfox** in chromium-browser &
- cat gen_smiles.txt
- copy & paste
- [apply filters?]
- start calculation
- scroll down
- choose **OCV** as Y-label
- click on point of interest in diagram → data row

Define target molecule/fragment

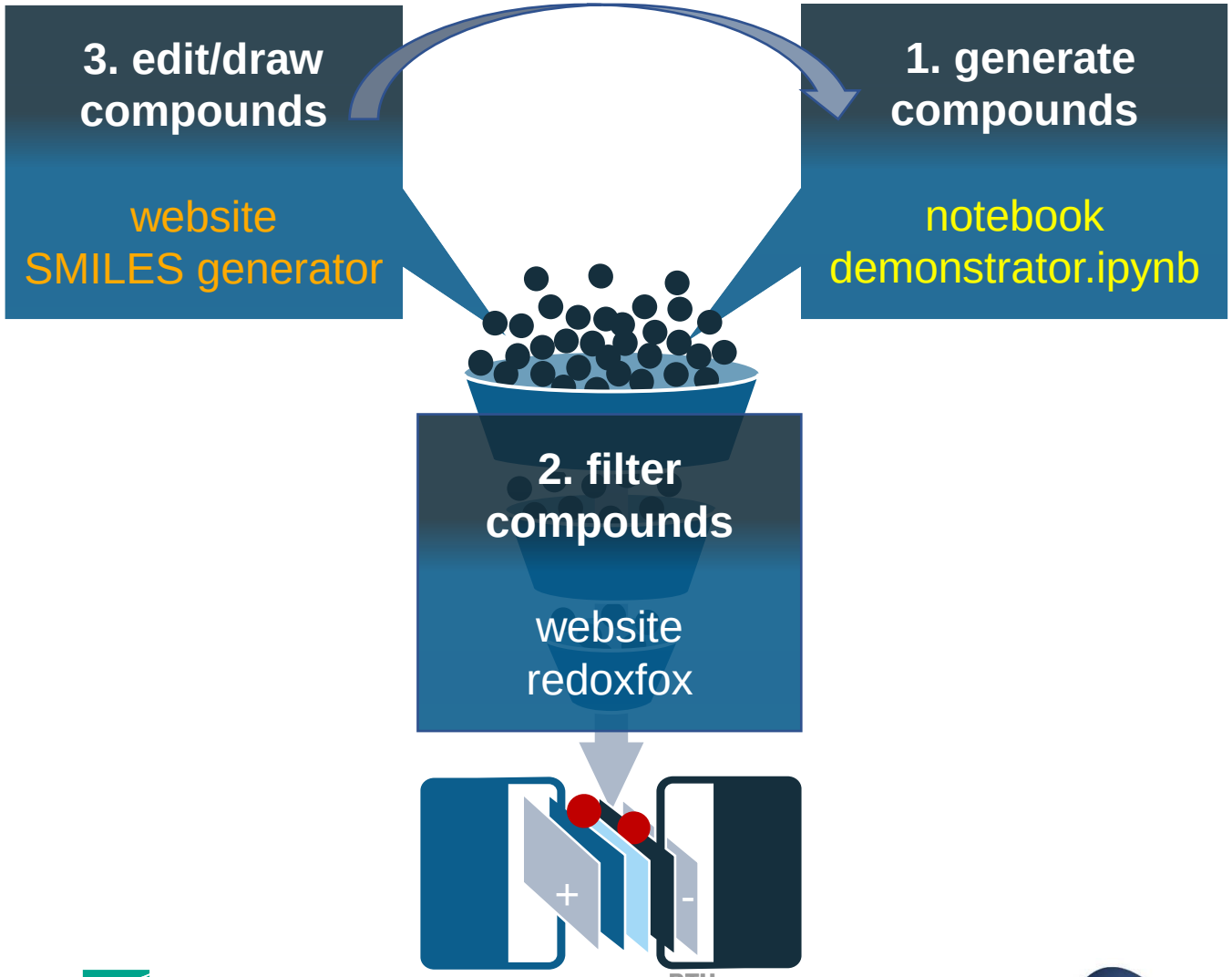


paste into

a) demonstrator

b) redoxfox

Play around…



Fraunhofer

DTU Danmarks Tekniske Universitet

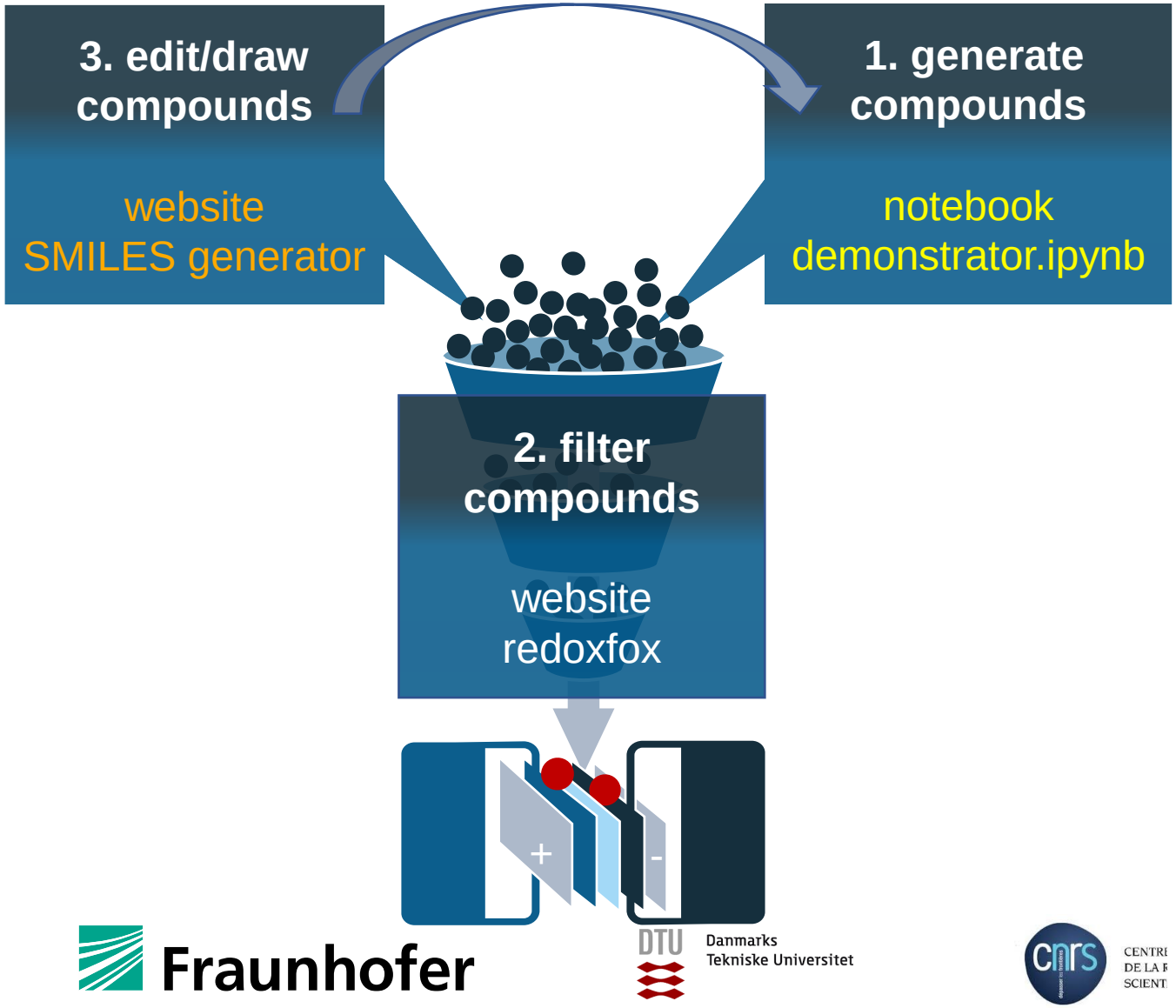
cnrs CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

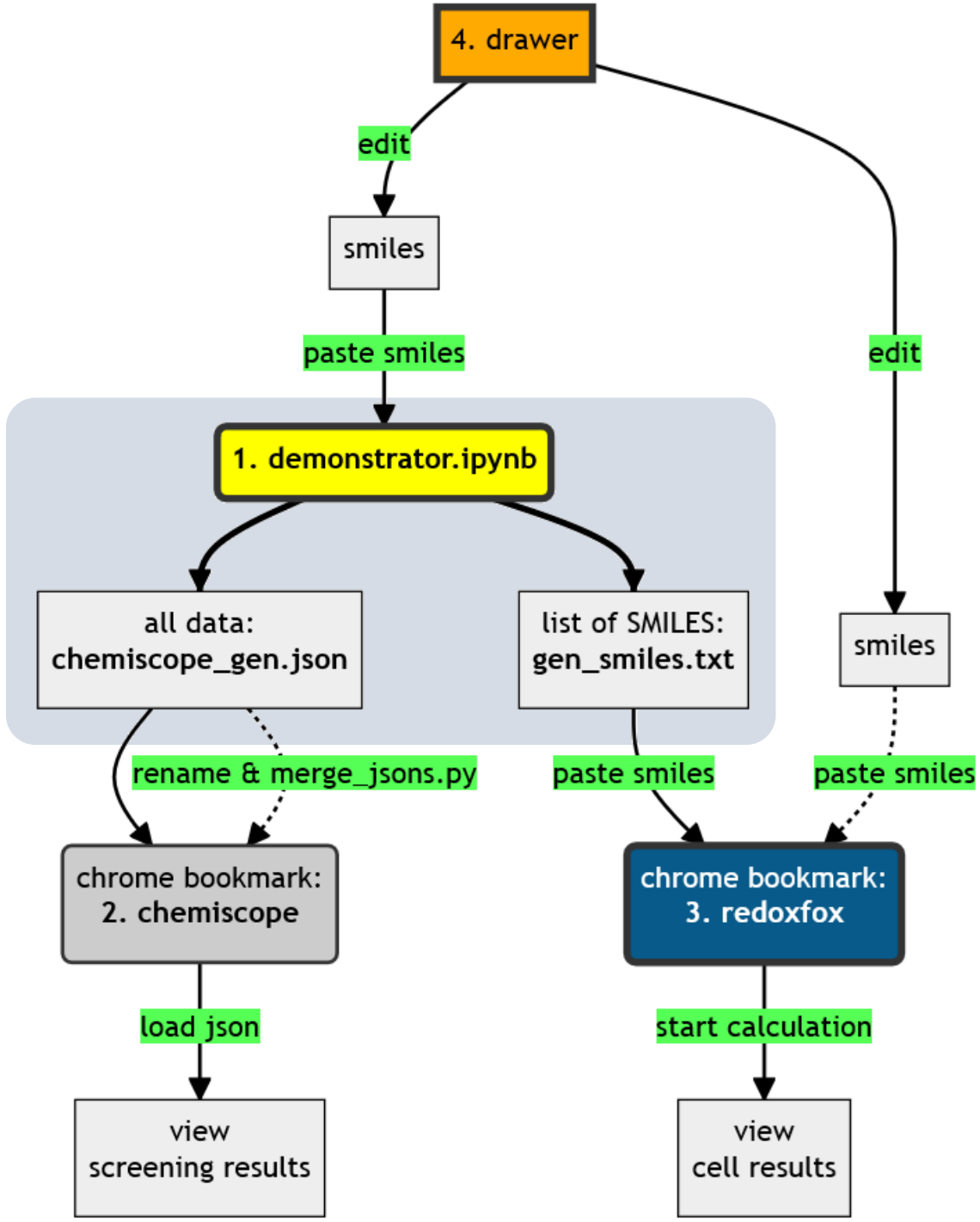
zhaw

KIT Karlsruher Institut für Technologie

UNSW SYDNEY

Play around…





Fraunhofer

DTU Danmarks Tekniske Universitet

cnrs CENTRE DE LA R SCIENT