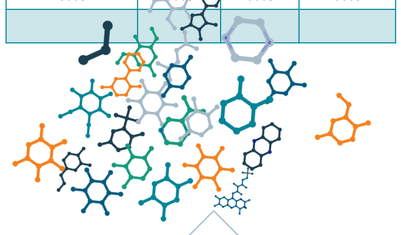
Artificial intelligence to discover new chemistries

Maass, A., Dobberstein, N., Hamaekers, J.

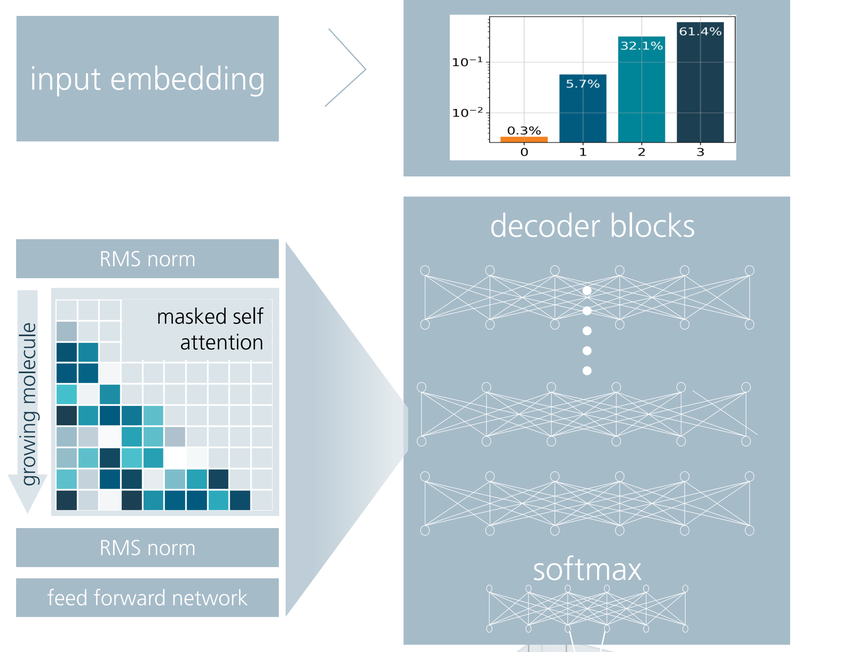
Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven 1, D-53757 Sankt Augustin, Germany

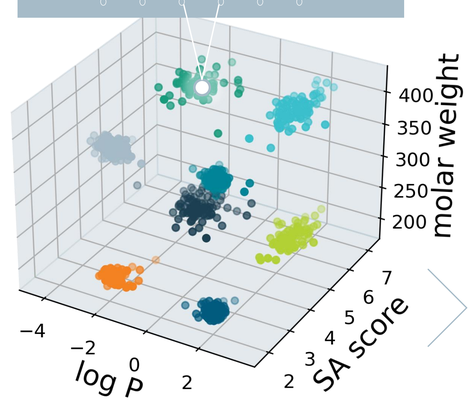
Generative AI for de novo molecular design: LLamol[1]

	Numerical context		
SMILES	logP	SA score	molar weight
C/C=C/CC	1.972	2.88	70
CCC=C(C)C	2.363	2.63	84
….	….	….	….



data set	size
ZINC 15 [3]	5M
QM9 [4,5]	134k
RedDB [6]	31
OPV [7]	91k
PubchemQC [8,9]	5.3M
CEP [10] subset [11]	20k
ChEMBL [12-15]	2.3M
OrganiX13	**12.5 M**





- **generative transformer model** based on LLama2 architecture[2]
  - 8 decoder blocks
  - ~15 M parameters
  - 2 days of training on Nvidia A100 GPU
  - trained on 12.5 M superset of organic compounds
  - new training method **Stochastic Context Learning** for maximum flexibility and robustness
- model handles **single- and multi-conditional** organic molecule generation with up to four conditions
  - valid molecular structures in SMILES notation
  - control creativity via 'temperature' parameter
  - incorporates numerical and/or SMILES sequence
  - easily expandable with new properties

https://github.com/Fraunhofer-SCAI/llamol

condition	novelty %	Uniqueness %	Validity %	MAE
None	89.7	99.9	99.5	-
logP	85.5	99.7	99.4	0.2
SAscore	85.4	98.8	82.1	0.4
molar weight	84.3	99.6	99.4	4

1. Dobberstein N, et al (2023) LLamol: A Dynamic Multi-Conditional Generative Transformer for De Novo Molecular Design. https://arxiv.org/abs/2311.14407
2. Touvron H, et al (2023) Llama 2: open foundation and fine-tuned chat models. http://arxiv.org/abs/2307.09288arXiv:2307.09288
3. Sterling T, Irwin JJ (2015) ZINC 15 - ligand discovery for everyone. https://doi.org/10.1021/acs.jcim.5b00559
4. Ramakrishnan R, et al. (2014) Quantum chemistry structures and properties of 134 kilo molecules. Scientific Data 1. https://doi.org/10.1038%2Fsdata.2014.22
5. Ruddigkeit L, et al. (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. https://doi.org/10.1021%2Fci300415d
6. Sorkun E, et al (2022) RedDB, a computational database of electroactive molecules for aqueous redox flow batteries. https://doi.org/10.1038%2Fs41597-022-01832-2
7. John PCS et al (2019) Message-passing neural networks for high- throughput polymer screening. https://doi.org/10.1063%2F1.5099132
8. Nakata M, et al (2020) PubChemQC PM6: data sets of 221 million molecules with optimized molecular geometries and electronic properties. https://doi.org/10.1021/acs.jcim.0c00740
9. Nakata M, Shimazaki T (2017) PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. https://doi.org/10.1021/acs.jcim.7b00083
10. Hachmann J et al (2011) The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. https://doi.org/10.1021%2Fjz200866s
11. Duvenaud D, et al (2015) Convolutional networks on graphs for learning molecular fingerprints. http://arxiv.org/abs/1509.09292arXiv:1509.09292
12. Zdrazil B et al (2023) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. https://doi.org/10.1093/nar/gkad1004
13. Blackshaw J et al (2009) CHEMBL database release 31. https://doi.org/10.6019/chembl.database.31
14. Davies M et al (2015) Chembl web services: streamlining access to drug discovery data and utilities. https://doi.org/10.1093/nar/gkv352
15. Jupp S et al (2014) The ebi rdf platform: linked open data for the life sci-ences. https://doi.org/10.1093/bioinformatics/btt765

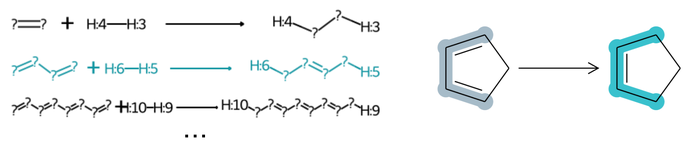
Predictive AI for estimating the redox potential[11]

- chemistry:
  proton coupled electron transfer only: net addition of H_2

$$A_{ox} + H_2 \longrightarrow A_{red} + \Delta E_{DFT}$$

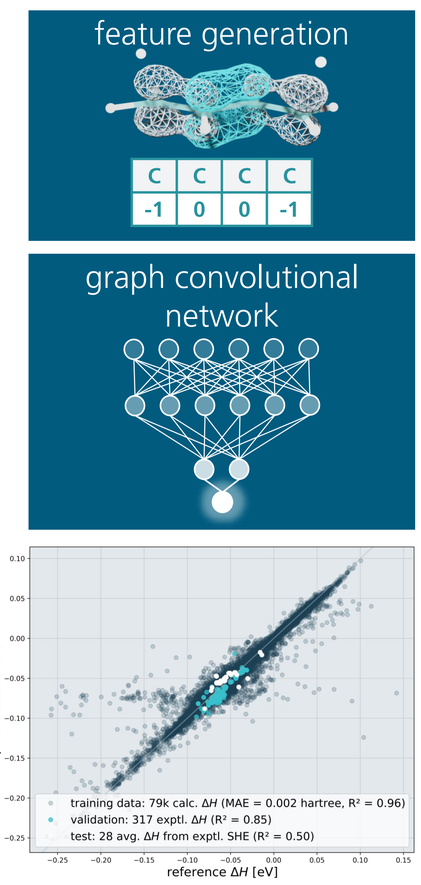
$$E^o = -\frac{\Delta G}{zF} \qquad \Delta G \cong \Delta H_{G4(MP2)} \cong \Delta E_{DFT}$$

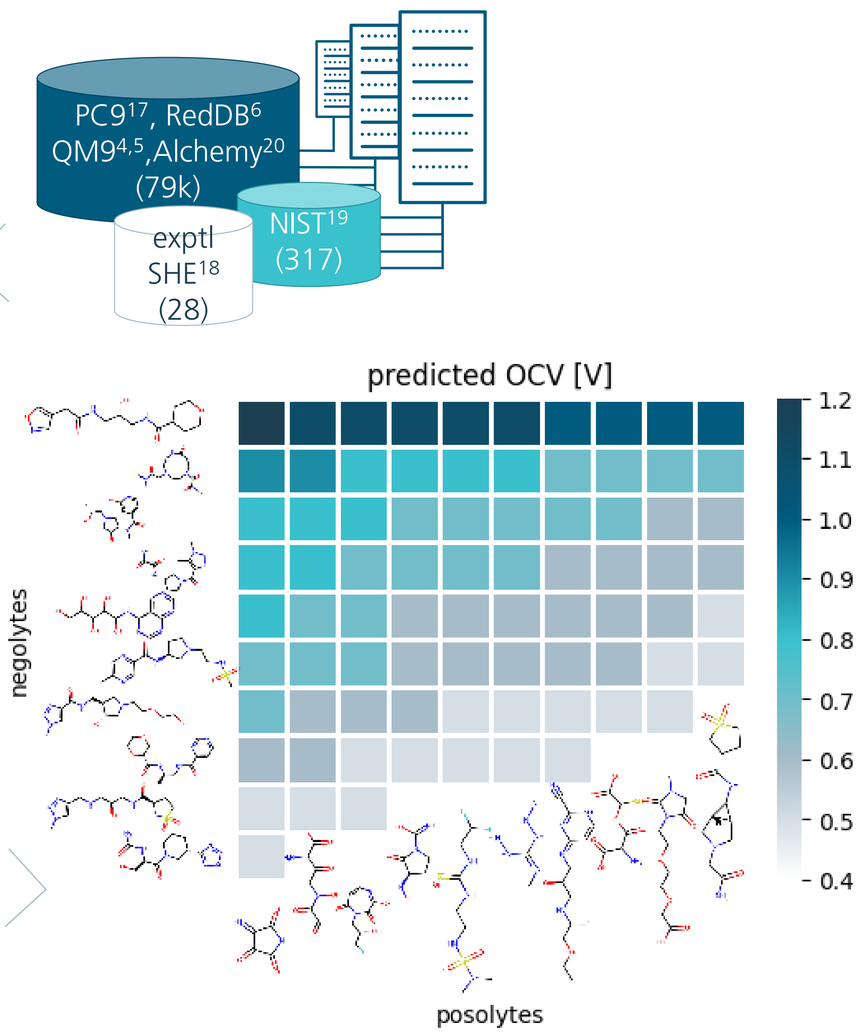
- construct from given molecule A_{ox} reduced form A_{red} via generic reaction templates



- 70 k redox pairs found in precomputed data sets
  - determine
    - **differential features** between A_{ox} and A_{red}
    - **energetic difference** between A_{ox} and A_{red}
- train **graph convolutional network** on molecular features and energy differences
- **predict energy difference** for novel compounds
  - ~50 predictions/sec
  - general trend reproduced[14,15]
  - identify most promising posolyte/negolyte candidates by max OCV

www.redoxfox.scai.fraunhofer.de





11. Barker J, et al. (2021). Rapid prescreening of organic compounds for redox flow batteries: A graph convolutional network for predicting reaction enthalpies from SMILES. https://doi.org/10.1002/batt.202100059
12. Glavatskikh M, et al. (2019). Dataset's chemical diversity limits the generalizability of machine learning predictions. https://doi.org/10.1186/s13321-019-0391-2
13. Maass, A. (2023) SONAR -- experimental redox potentials for organic compounds undergoing 2-electron/2-proton transfer reactions, https://zenodo.org/doi/10.5281/zenodo.10200227
14. Maass, A. (2024) heat of hydrogenation for diverse organic compounds -- experimental and calculated data for 166 unique reactions, https://zenodo.org/doi/10.5281/zenodo.10820775
15. Chen G, et al (2019). Alchemy: A quantum chemistry dataset for benchmarking ai models. https://arxiv.org/abs/1906.09427

Summary

- prototypic machinery for AI-based creation and evaluatipm of new chemistries in place
- improvements planned
  - generative model: respond to more conditions, train on more relevant properties (e.g. HOMO/LUMO gap)
  - predictive model: train on larger & standardized data; train new model on other properties (e.g. solubility)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 875489.