Curation of Big Data for Atmospheric Science

Vitus Besel[1*], Milica Todorović[2], Theo Kurtén[1], Patrick Rinke[3], Hanna Vehkamäk

[1]Institute for Atmospheric and Earth System Research, University of Helsinki
[2]University of Turku
[3]Aalto University

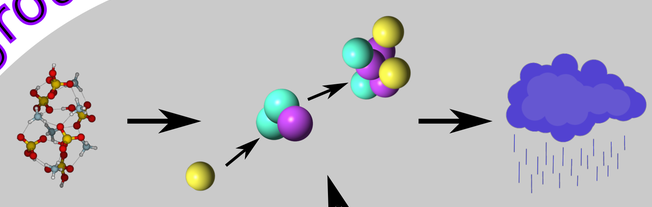
HELSINGIN YLIOPISTO

This presentation participates in OSPP

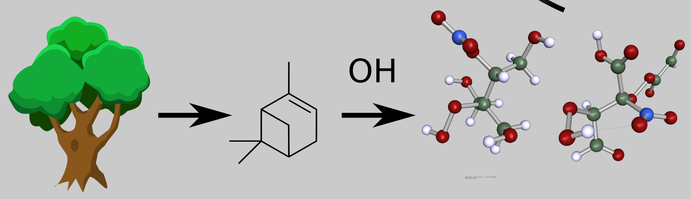
EGU

Outstanding Student & PhD candidate Presentation contest

Background



II Particle growth in the atmosphere is heavily driven by LVOC condensation.[1]



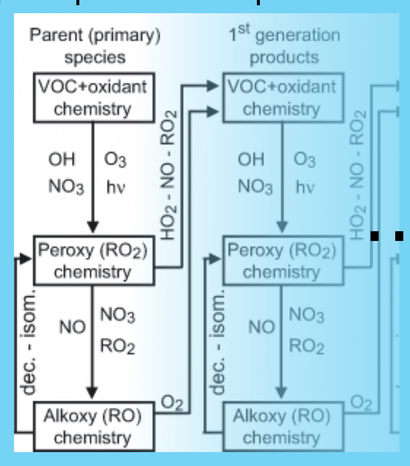
I Vegetation emits organic molecules which are rapidly oxidized on air, resulting in low-volatile organic compounds (LVOC).

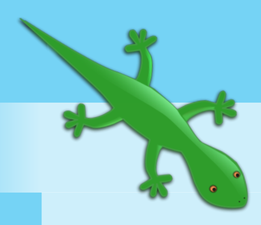
III There are tens of thousands LVOC species in the atmosphere → Key property is their **Saturation Vapor Pressure** (**pSat**) as measure for their affinity to condense.

Objectives:
- Collect possible LVOC in dataframe
- Compute pSat efficiently
- Train Machine Learning model on pSat

Molecule Generation: Gecko-A

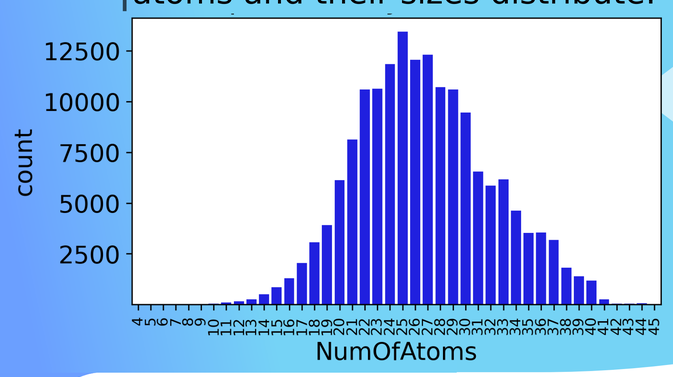
I GECKO-A is a chemical mechanism simulating the oxidation of organic atmospheric compounds.[2]





II Parentspecies Decane, Toluene and α-pinene result in SMILES strings of **166k** HOM molecules.
E.g. "C(=O)(O)C(OO)"

III The created species have in median 7 C-, 10 O- and 1 N-atoms and their sizes distribute:



Next: **pSat** is calculated in two steps.

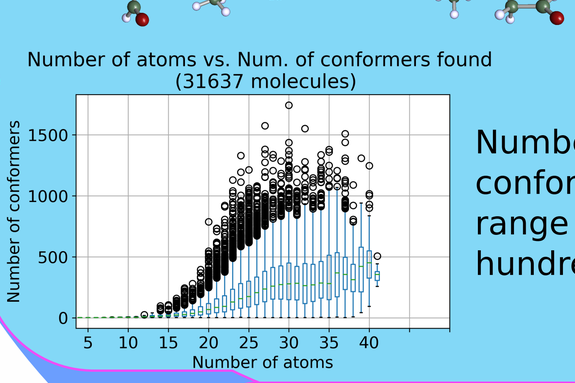
COSMO

II COSMOconf[3] finds all conformers

INPUT: SMILES
C(=O)(O)C(C(ON(=O)(=O)))OOC(OO)(C(ON(=O)

- Conformer generation
- Repeated optimization (Density Functional Theory)
- Sorting

Output: Conformers structures + Energies



Number of conformers range in the hundreds.

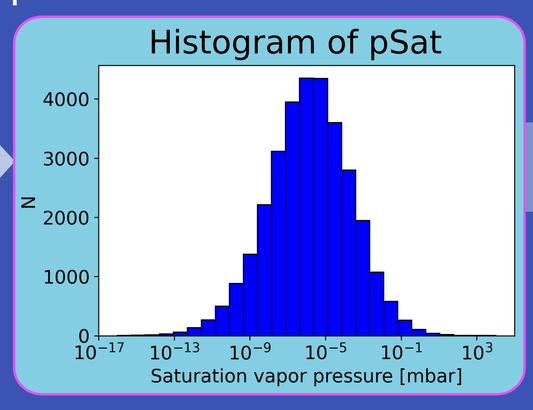
II COSMO*therm* calculates pSat

COSMO*therm*[3,4] simulates molecule in solvent of itself and calculates thermodynamical properties (such as **pSat**);
Conformers are weighted by energy.

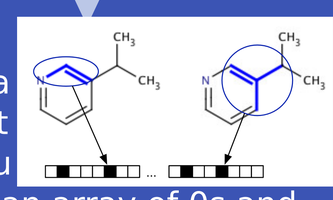
CosmoConf is computationally very expensive. Calculation of 1 HOM takes 3h - 30h on 128 CPU cores.

Machine Learning

I At this point 32k HOM have been labelled:



II Topological Fingerprint is a descriptor that hashes molecule groups into an array of 0s and 1s depending on the presence of the group. It is the same for all conformers and maps a molecule to a pSat.
E.g. [0,1,1,1,0,0,1,0,...]



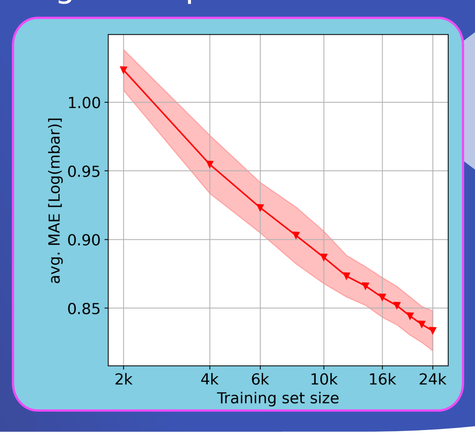
Summary

I GeckoQ[5] database contains 31,637 molecules and 7.26 Mio. conformers and is publicly available. The data contains thermodynamic properties, structural attributes and functional group counts.

II We can predict pSat with 0.83 log units(mbar) accuracy - compared to COSMO accuracy of 0.5 log units this is quite good!

PS: We also built a Merlin workflow framework to execute hundreds of multilayered DFT calculations on a supercomputer. If you are interested, ask!

III We can train a Gaussian Process Regression (GPR) Machine Learning model using the pSat and the Topological Fingerprint:
→ Relationship between Topological fingerprint and vapor pressure can be modelled
→ The more data, the better



Future Work:
- Most interesting for atmospheric science are LVOC that are sure to condense: extremely low-volatile compounds (ELVOC)
- We want to identify ELVOC and improve ELVOC pSat predictions, by employing **Active Learning**

Acknowledgments:
This work is part of the Virtual Laboratory for Molecular Level Atmospheric Transformations (VILMA). We also thank CSC - Finnish IT Centre for access to computer clusters and awarding the Mathi Grand Challenge.

Sources:
[1] Bianchi, F. et al., *Chem. Rev.*, 119, 3472–3509, 2019.
[2] Aumont et al., *Atmos. Chem. and Phys.*, 5, 2497-2517, 2005
[3] BIOVIA COSMOtherm/COSMOconf, Release 2021; Dassault Systèmes. http://www.3ds.com
[4] Klamt, A. *J. Phys. Chem.* 99, 2224 (1995)
[5] Besel, V. Etsin repository: GeckoQ (Version 1), 2023