SFGP – GT Thermodynamique , April 1, 2016

Production de données thermodynamiques par simulation moléculaire : quelles incertitudes ?

Marianna Yiannourakou, Benoit Leblanc, Xavier Rozanska, Alexander Mavromaras, Philippe Ungerer

Materials Design SARL, Montrouge, France

Acknowledgements : partners of ANR-funded projects Memobiol (2009-2013) and Predimol (2010-2014)

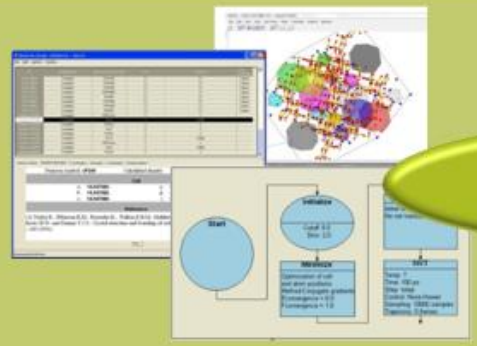
materials design®

CONTEXT

- General coverage of chemicals by thermodynamic data is low
  - More than 60 *million* pure substances filed in Chemical Abstracts
  - 26 *million* possible organics built with 11 C, N, O, F (Fink & Reymond, *JChInfModel*, 2007)
  - 140,000 substances declared by the chemical industry to comply with EU regulation REACH
  - 43,911 compounds in Dortmund Databank (VLE, LLE, SLE, enthalpic, Cp, volumetric,…)
  - ~ 3,000 compounds well characterized in DIPPR database (34 correlated properties)
  - Low coverage of properties at extreme T and P or toxic, hazardous compounds
- Expectations of chemical engineers vs prediction methods
  - Consistent prediction of multiple properties (ex. Cp, enthalpy, Tb, Psat, ...)
  - Sound principles, well-defined reference states
  - Sound parametrization , efficient validation
  - Small influence of numerical parameters, averaging procedures, ….
  - Ability to consider multifunctional molecules and mixtures with a limited number of parameters
- Which thermodynamic properties can be obtained by simulation from their molecular structure :
    - Quantum chemistry (DFT, Semi-Empirical methods, COSMO, QSPR)
    - Forcefield-based methods (Monte Carlo, Molecular dynamics)
  - What uncertainties ?

1 MEDEA SOFTWARE ENVIRONMENT

MedeA's Three Tier Architecture



User Interface

- Databases (exp/comp)
- Structure building
- Workflow creation
- Analysis

(Windows, Linux)

Local or remote

Job Server

- Job control

Task Servers

- Compute intensive tasks

VASP LAMMPS
GIBBS MOPAC

Molecular modeling applied to molecules

Quantum mechanics (QM)

- DFT (VASP)
- semi-empirical (MOPAC)

- Ideal gas properties
- Molecular and electronic structure

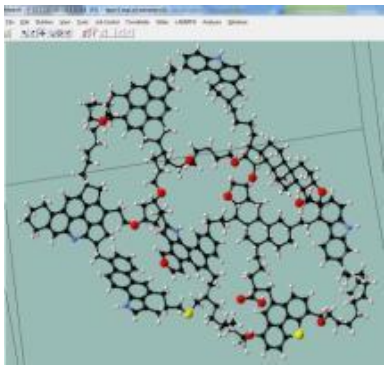
Classical statistical mechanics

- Molecular Dynamics (MD) LAMMPS
- Monte Carlo (MC) GIBBS
- Forcefields pcff+, TraPPE, AUA

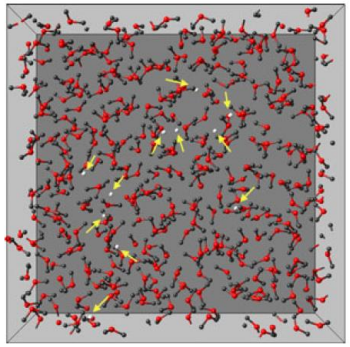
- Energy, enthalpy, density
- Free energy
- Structure & dynamics of matter

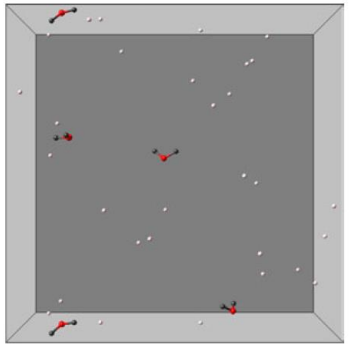
400-600 atoms
IR vibrational analysis

Up to 20000 atoms
nanoseconds



Kerogen fragment





Hydrogen-dimethylether

(periodic boundary conditions)

Property Calculation in MedeA®
with Forcefield Methods

Molecular Dynamics (MedeA®-LAMMPS*) Forcefields: AA (pcff+, OPLS,..)

- **Static Properties**
  - Density
  - Pressure
  - Cohesive Energy Density
- **Transport Properties**
  - Viscosity
  - Thermal Conductivity
  - Self-Diffusion Coefficient
- **Mechanical Properties**
  - Shear Modulus
  - Young's Modulus

Monte Carlo (MedeA®-GIBBS) Forcefields: TraPPE, AUA, pcff+**

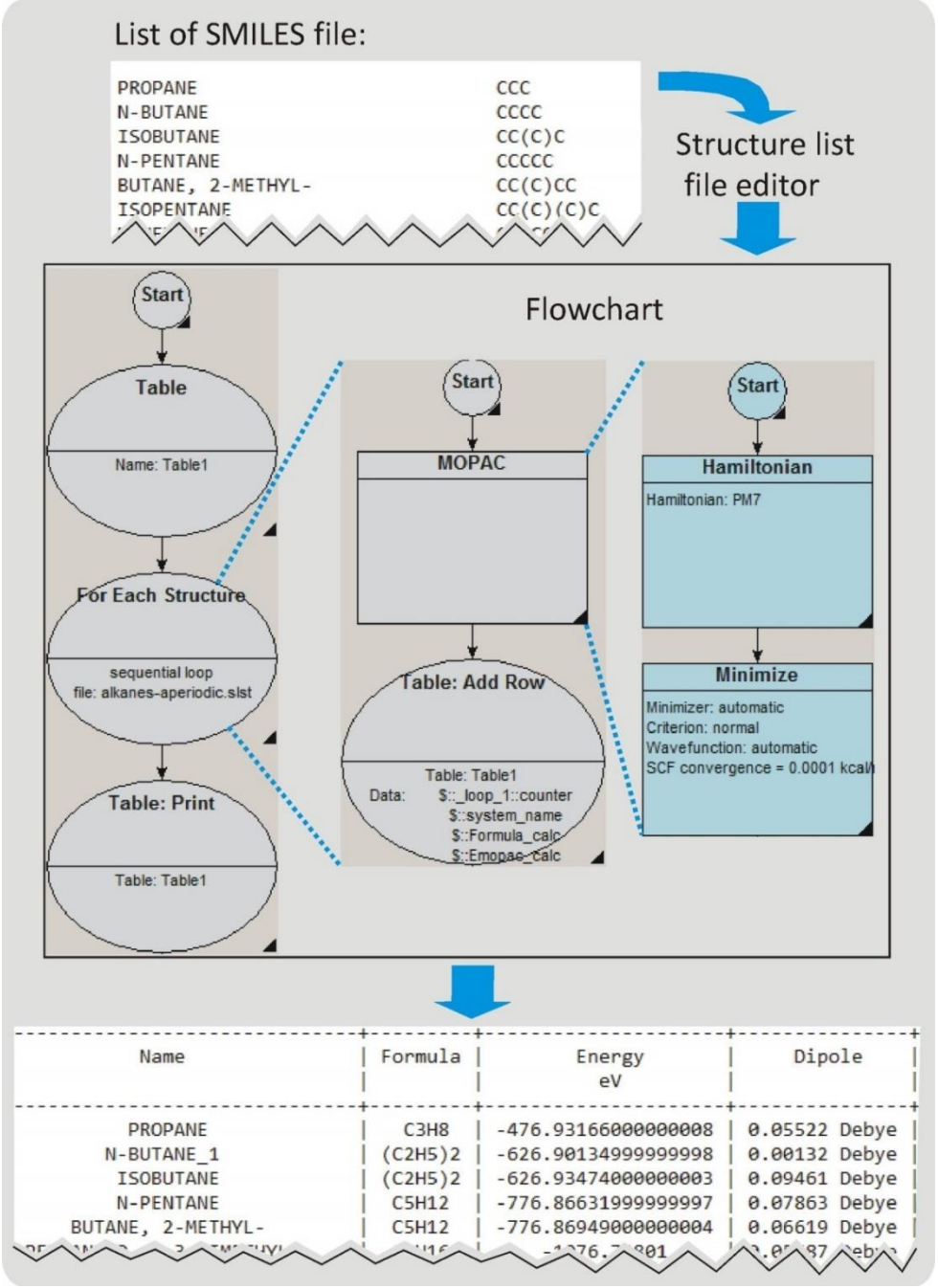
- **Static Properties**
  - Density
  - Pressure
  - Chemical Potential & Fugacity
  - Henry Solubility Constants
  - Boiling Point Temperature
  - Phase Equilibrium
  - Derivative Properties (residual heat capacity, isobaric thermal expansivity, isothermal compressivity, Joule-Thomson coefficient)
  - Adsorption Isotherms

* Large Scale Atomic/Molecular Massively Parallel Simulator ®, Sandia Corporation (2003)

** Gibbs v. 9.3, IFP-Energies Nouvelles, Rueil-Malmaison & Laboratory of Physical Chemistry, University Paris Sud -CNRS, Orsay

Automation of the preparation, processing and analysis of simulation

- Editor of structures list
  - Import/export structures from
    - crystallographic databases
    - SMILES formula (Openbabel)
    - conformer search (Openbabel)
    - a flowchart itself
  - Periodic and aperiodic structures
- Flowchart module: **Loop over all structures in the structure list**
- Integrated in the flowchart environment
  - Edition of structures
    - translation of atoms
    - supercell building
    - amorphous phase building
    - random atomic substitution
    - atomistic simulation
  - Loops over set of simulation variables and parameters
- Flowchart module: edition of personalized Table printing



2 THERMOCHEMICAL PROPERTIES

Set of organic molecules

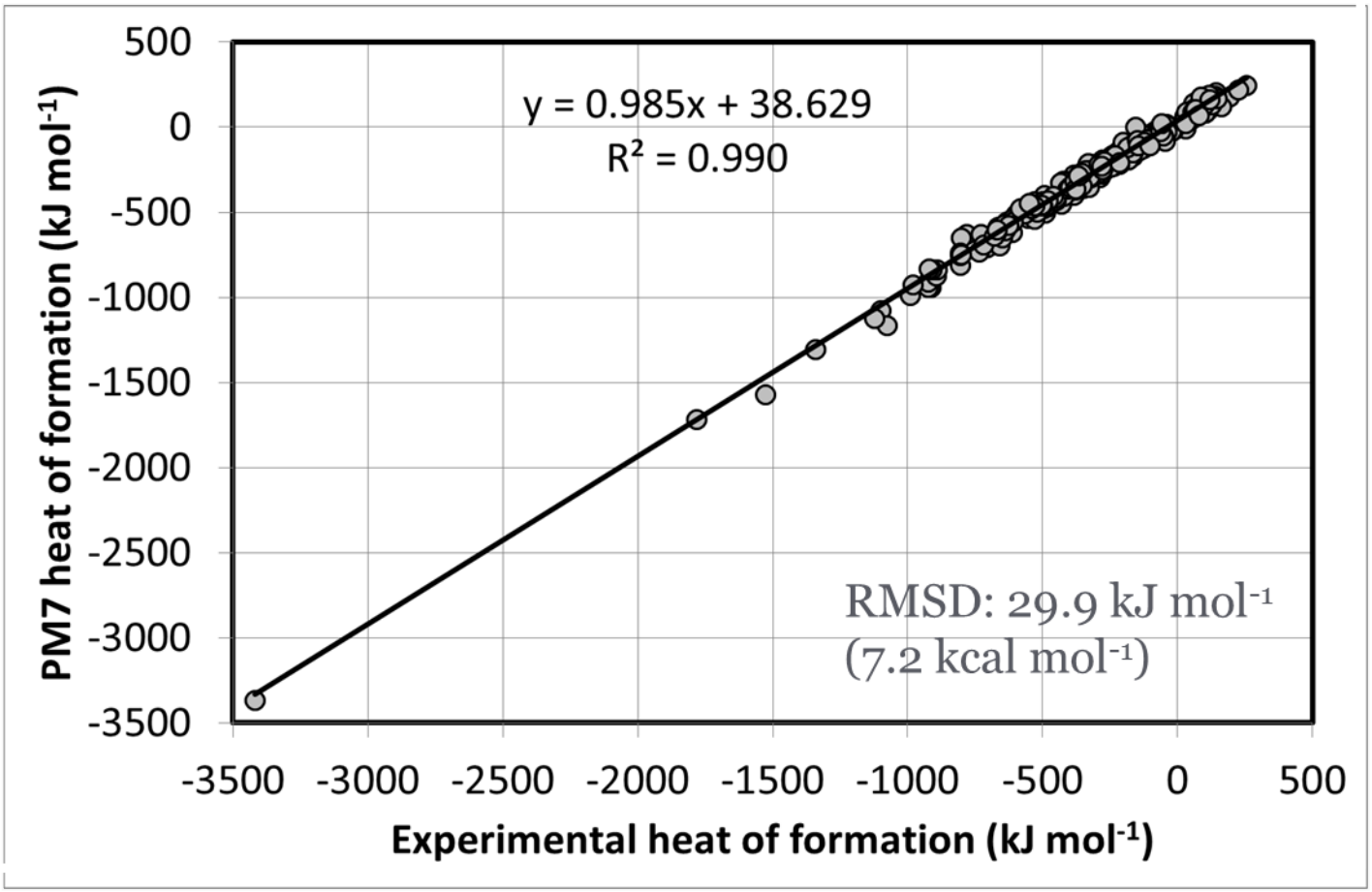
- 880 organic molecules SMILES formula are collected from DIPPR database
  - size C_1 to C_9 and covering 15 classes of organic compounds
  - subsets depending on availability of experimental data

MOLECULES CLASS	NUMBER OF MOLECULES	MOLECULES CLASS	NUMBER OF MOLECULES
Carboxylic Acids	39	Amines/Amides	127
Aldehydes	44	Halogenated	171
Alcohols	80	Esters	63
Polyols	38	Ethers	53
Alkanes	91	Ketones	41
Olefins	84	Peroxides	10
Alkylaromatics	14	Epoxides	16
Isocyanates	9	**TOTAL**	**880**

Rozanska et al., J. Chem Eng Data, 2014

Heats of formation of organic molecules

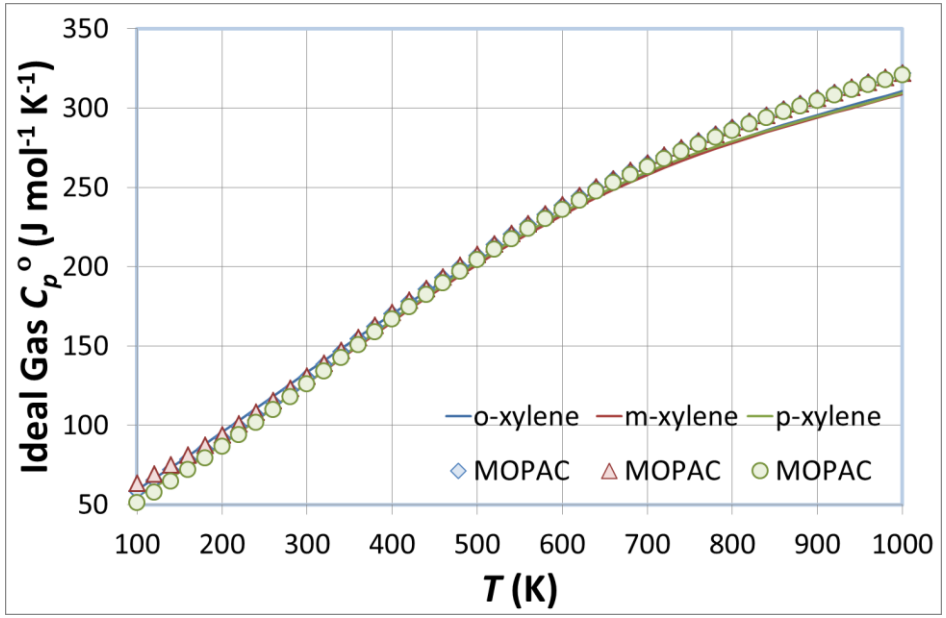
- Selected reference = DIPPR exp. data with error estimated to be lower than 5% - set of 428 values

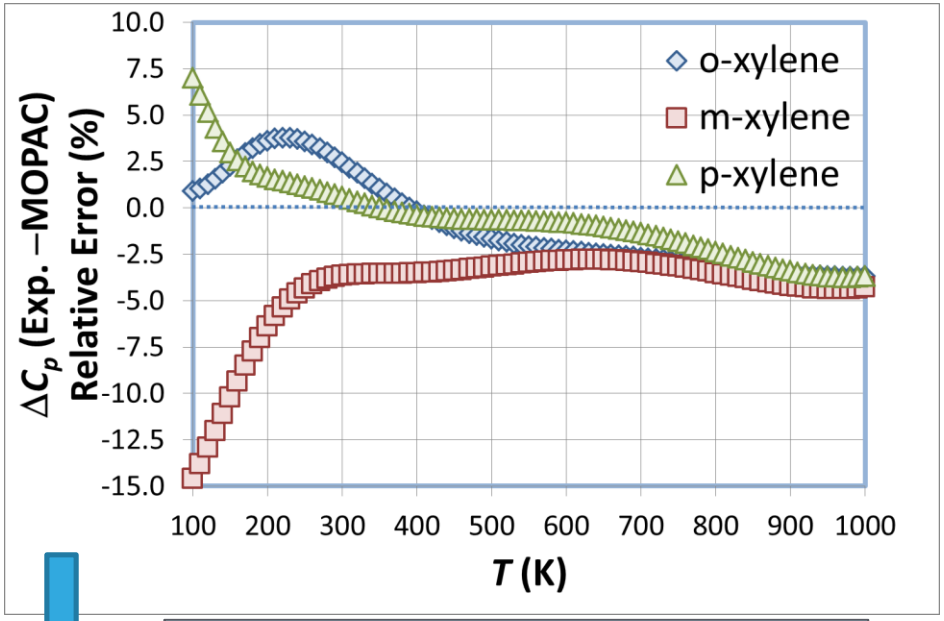


Source experimental data:DIADEM: The DIPPR Information and Data Evaluation Manager for the Design Institute for Physical Properties, Version 6.0.0, Database 2011

Ortho-, Meta-, and Para-Xylenes

- Molecules geometries are optimized with MOPAC(PM7) [1]
- Frequency and thermochemistry calculations follow
- Experimental data of C_p° from Poling *et al.*[2]



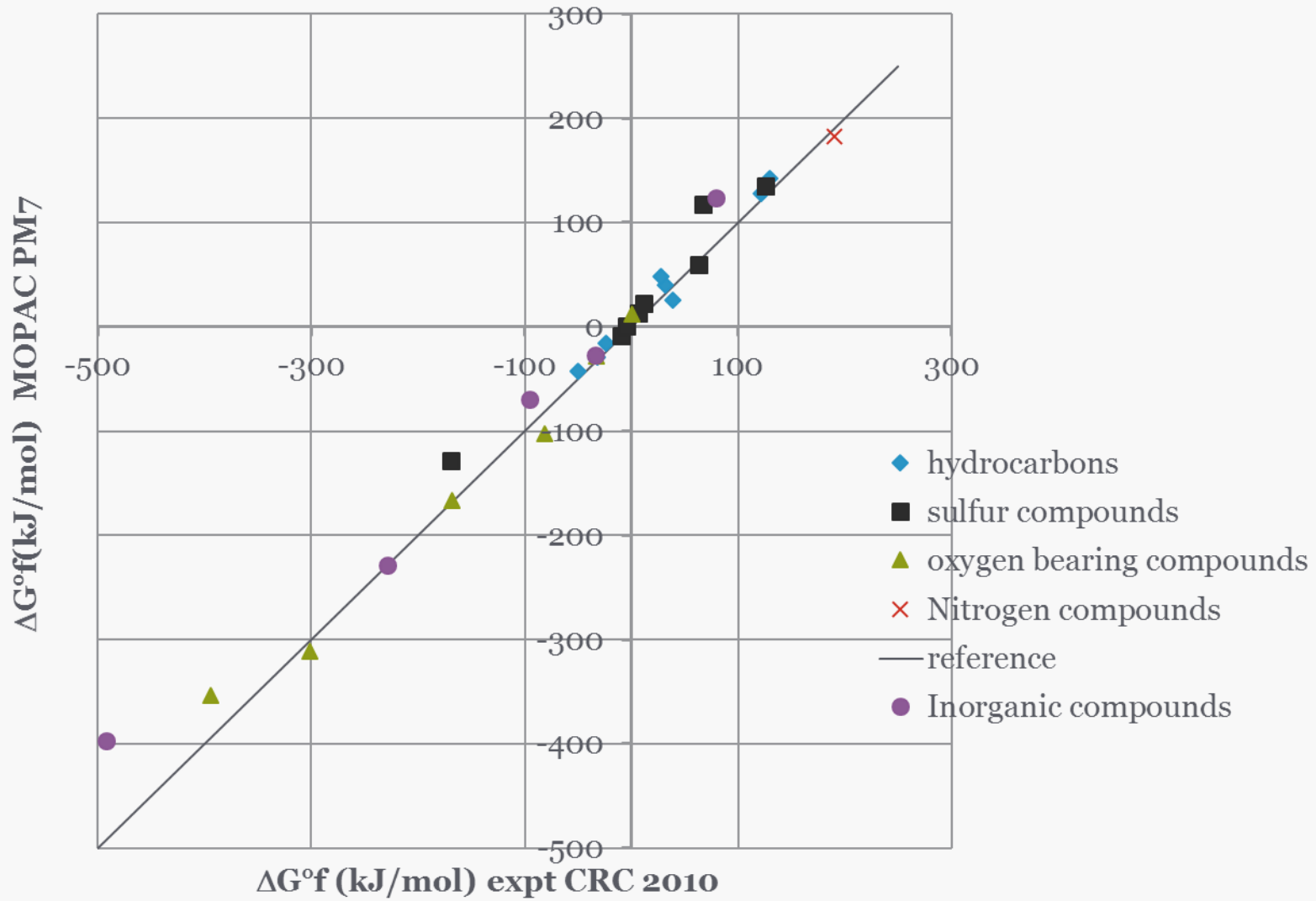


RMSD=2.7, 5.0, and 2.3% for o-, m-, and p-xylenes, resp.

[1]Stewart, J. P. (2013). J. Mol. Model. **19**: 1-32.

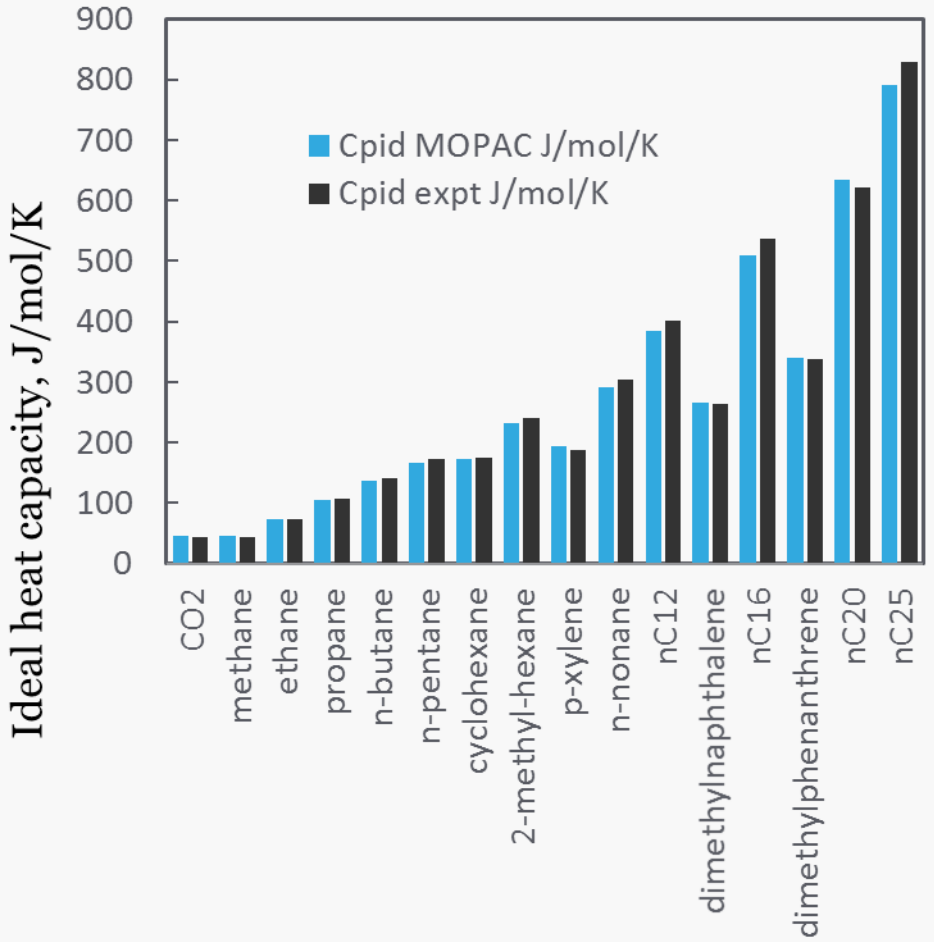
[2]Poling, B.E., Prausnitz, J.M., O'Connell, J.P. in "*The properties of gases and liquids – 5th edition*", McGraw-Hill, 2007.

Standard Gibbs free energy of formation of

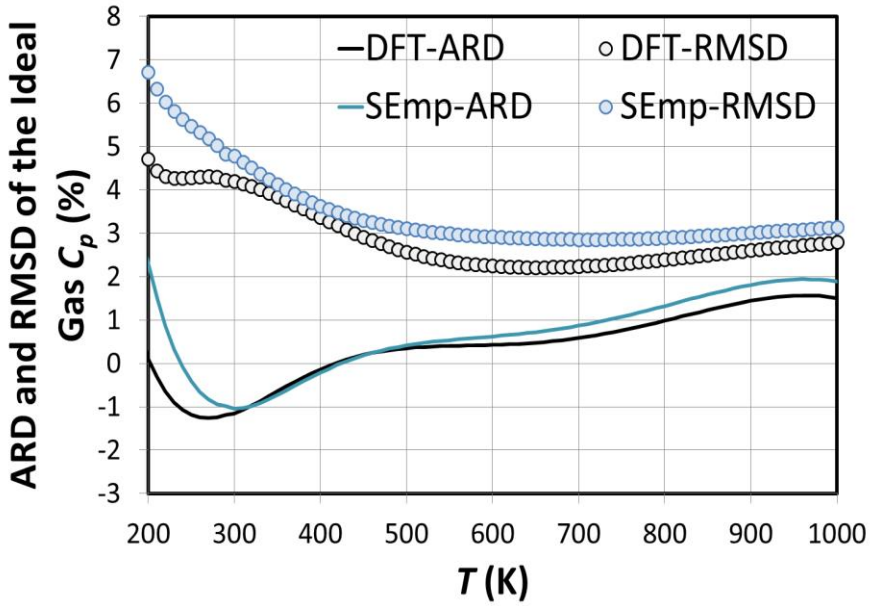


Ideal heat capacity of organic molecules

Prediction of ideal heat capacity Cp at 463K by quantum mechanics and vibrational analysis (MOPAC-PM7) AAD is 2.8%



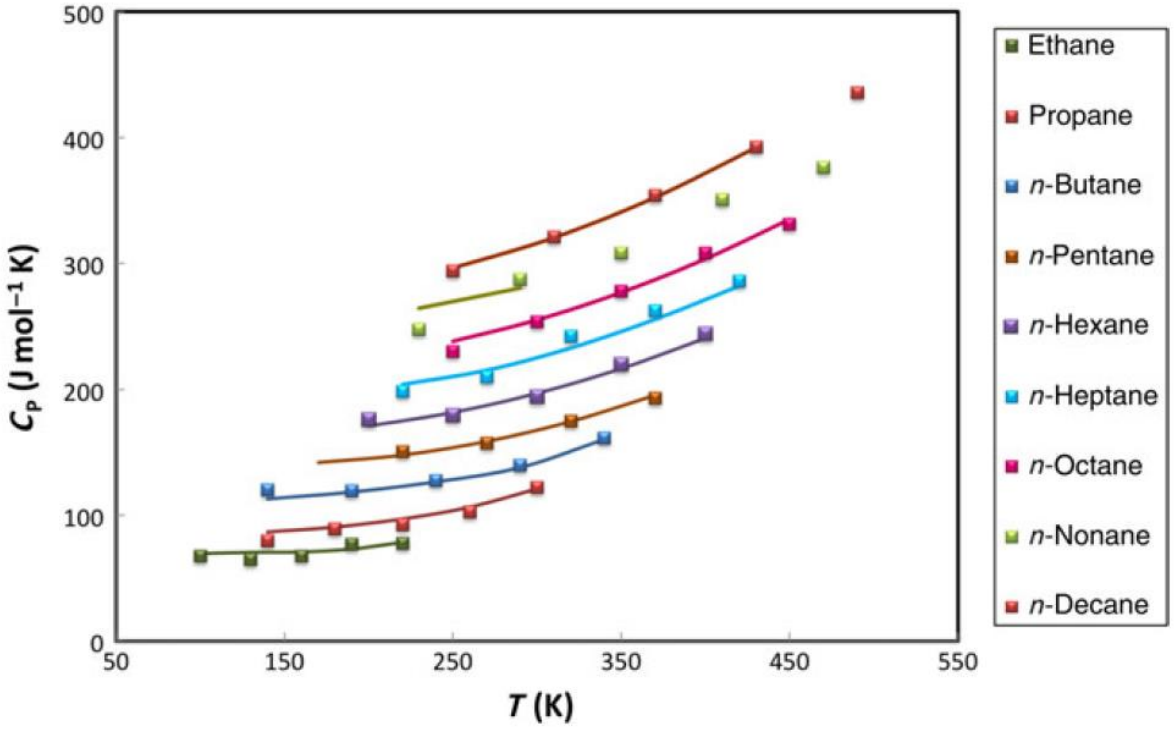
- Average relative deviation (ARD) and RMSD between 160 experimental ideal gas heat capacity (C_P) vs.
  - DFT - BP86/TZVP
  - Semi-empirical scaled PM7 (SEmp)



Source experimental data:The properties of gases and liquids, fifth international ed.; Poling et al. ; McGraw-Hill, Boston, 2007, pp. A.35-A.46. Thermodynamics Research Center (TRC) data bank, College Station, TX, USA ; NIST ; IUPAC

Heat capacity of liquid alkanes

- Total heat capacity is obtained with an uncertainty of 3 to 5% as the sum of :
  - Ideal heat capacity from vibrational analysis , using MedeA-MOPAC
  - Residual heat capacity from MedeA-GIBBS using forcefields (derivative property obtained from fluctuations, see *Lagache et al., PCCP, 2001*)



Inorganic molecules?

- Ideal gas heat capacity at T=298 K: RMSD of the average relative difference between PM7 and BP86/TZVP

Al_2O	$AlBr_3$	$AlCl_3$	AlF_3	AlO	AlS
Al_2Se	AlCl	AlF	AlI	AlOCl	AlSe
AlBr	$AlCl_2$	AlF_2	AlI_3	AlOF	$(Al_2O)_2$

Li (9) **22**	Be (8) 2					Li (9) 5					**B** (19) **17**	C (28) 10	N (13) 7	O
Na (10) 14	Mg (4) 13										**Al** (18) **17**	Si (22) 13	P (18) 12	S (23) 10
K (9) 11	Ca (5) 14	Sc (4) 9	**Ti** (15) **19**	V (5) 10	Cr (7) 12	**Mn** (4) **32**	Fe (10) 5	**Co** (4) **22**	Ni (11) 12	**Cu** (10) **19**	Zn (6) 6	Ga (14) 13	Ge (18) 10	As (10) 10	Se (14) 7
Rb (4) 11	Sr (5) 10	Y (3) 12	Zr (17) 2	Nb (5) 12	Mo (22) 11	Tc	Ru (3) 6	Rh	Pd	Ag (2) 7	Cd (7) 10	In (16) 8	Sn (13) 10	Sb (9) 11	Te (9) 15
Cs (9) 15	**Ba** (4) **19**	La	Hf (3) 13	Ta (11) 10	W (17) 15	Re	Os	Ir	Pt	Au (2) 1	Hg (8) 7	Tl (6) 5	Pb (13) 13	Bi (9) 11	Po

Element

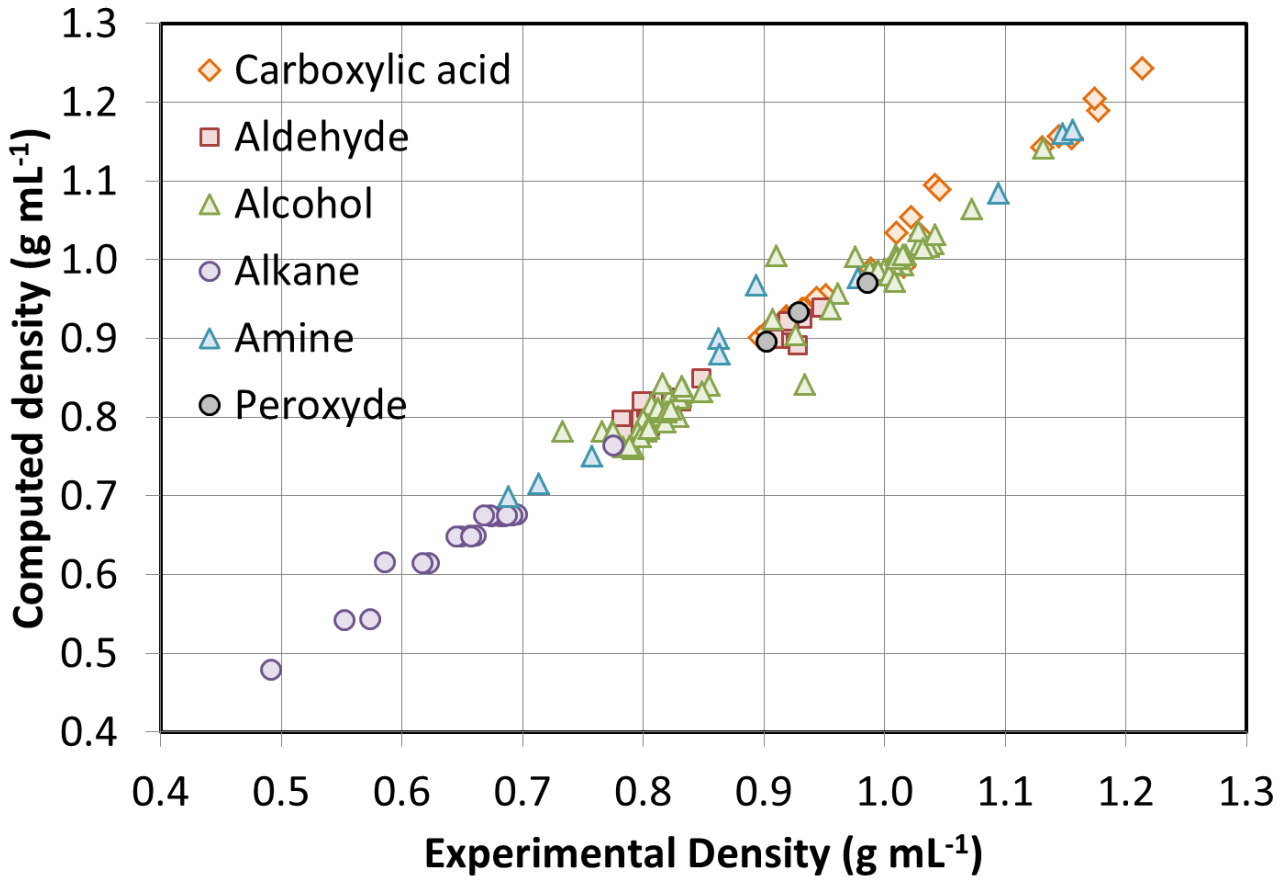
Number of molc.

RMSD (%)

Source for the 515 inorganic molecules: Knacke et al. *Thermochemical properties of inorganic substances*, Springer-Verlag, Berlin, 1991

Liquid density of organic compounds from molecular dynamics

- liquid density at 1 bar and 298 K for 174 compounds
- LAMMPS , pcff+ forcefield



Class	RMSD (%)
Carboxylic acid	1.4
Aldehyde	1.1
Alcohol	2.4
Alkane	2.0
Amine	2.6
Peroxide	0.8
All	2.1

Rozanska et al., JCED, 2014 (Experimental data: DIPPR Database 2011)

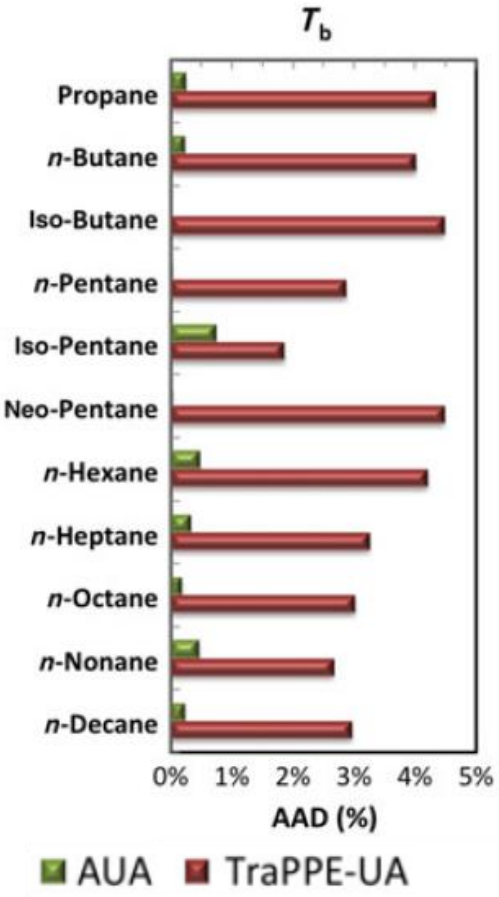
3 VAPOR-LIQUID EQUILIBRIA

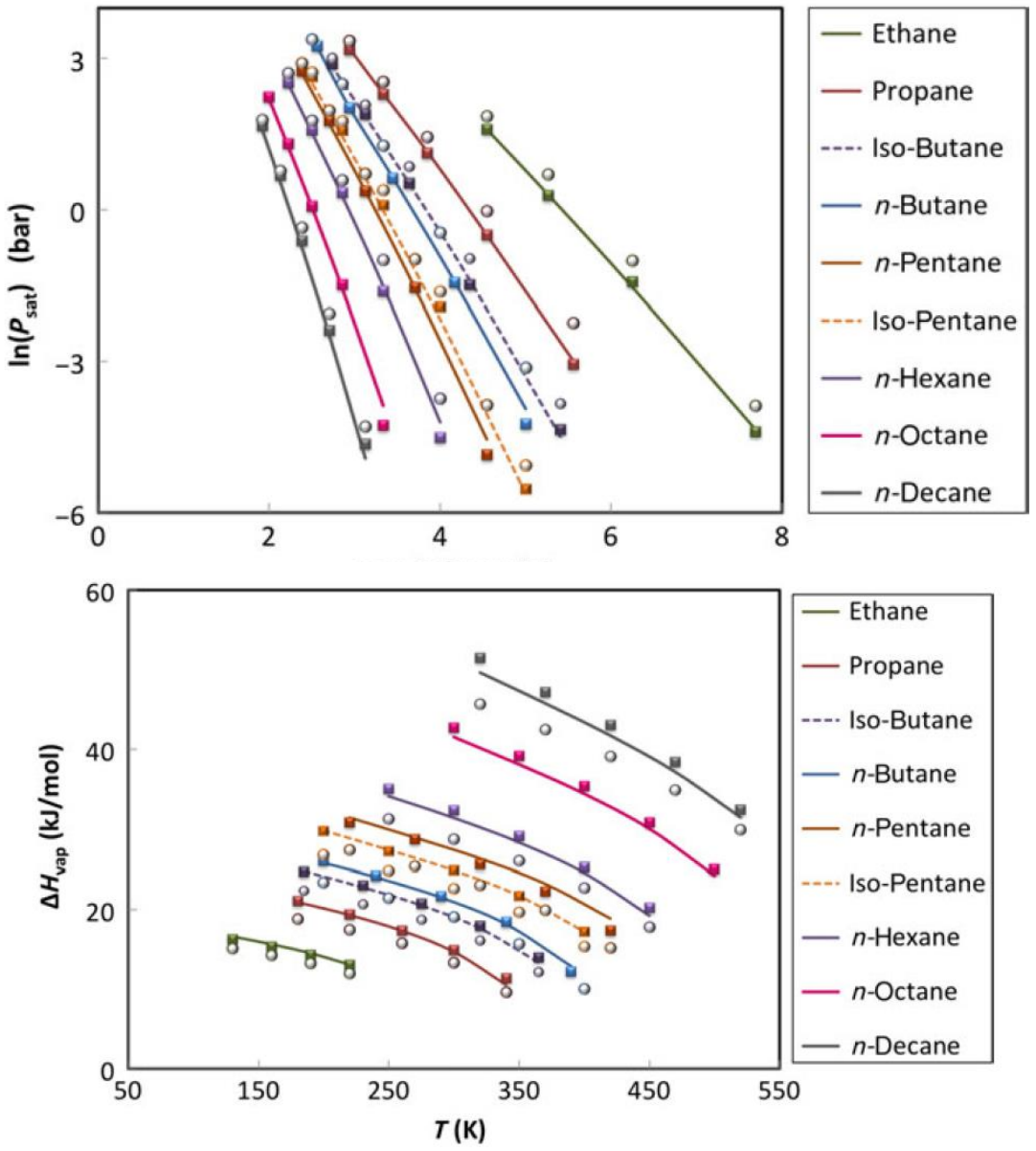
Yiannourakou et al., Molecular Simulation, 2013
Rozanska et al., J.Chem. Eng. Data 2014

Alkanes

Yiannourakou, M.,et al., 2013, ." Molecular Simulation **39**(14-15): 1165-1211.

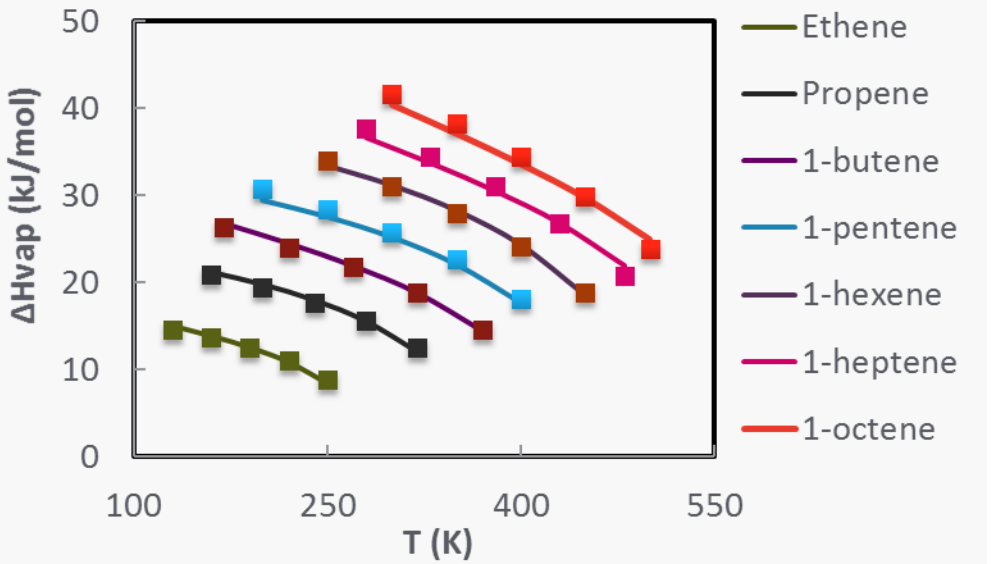
Open symbols : TraPPE
Filled symbols : AUA
LINES : DIPPR

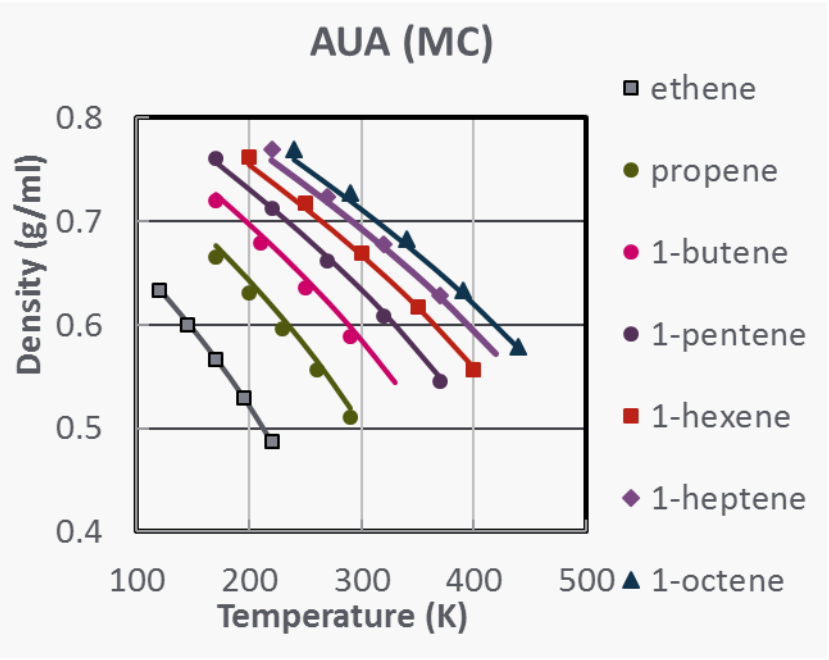


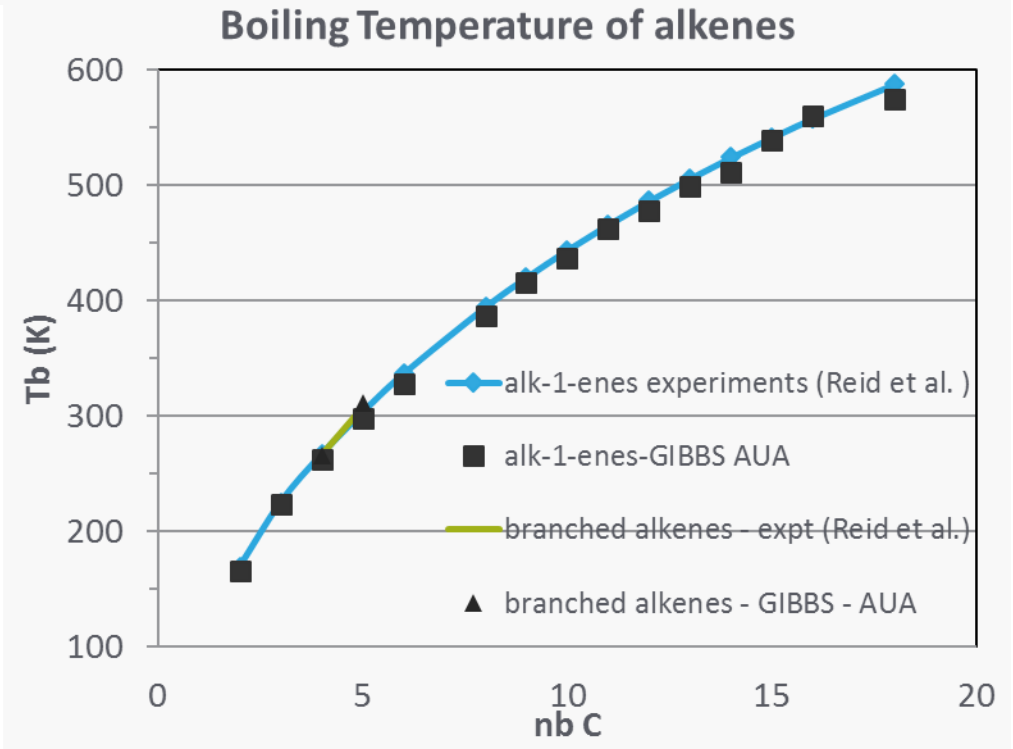


Olefins

	AAD on Tb, %	AAD on liq density all T <Tc
TraPPE C2-C8	**3.5 %**	**0.56 %**
AUA C2-C18	**0.81%**	**0.91 %**

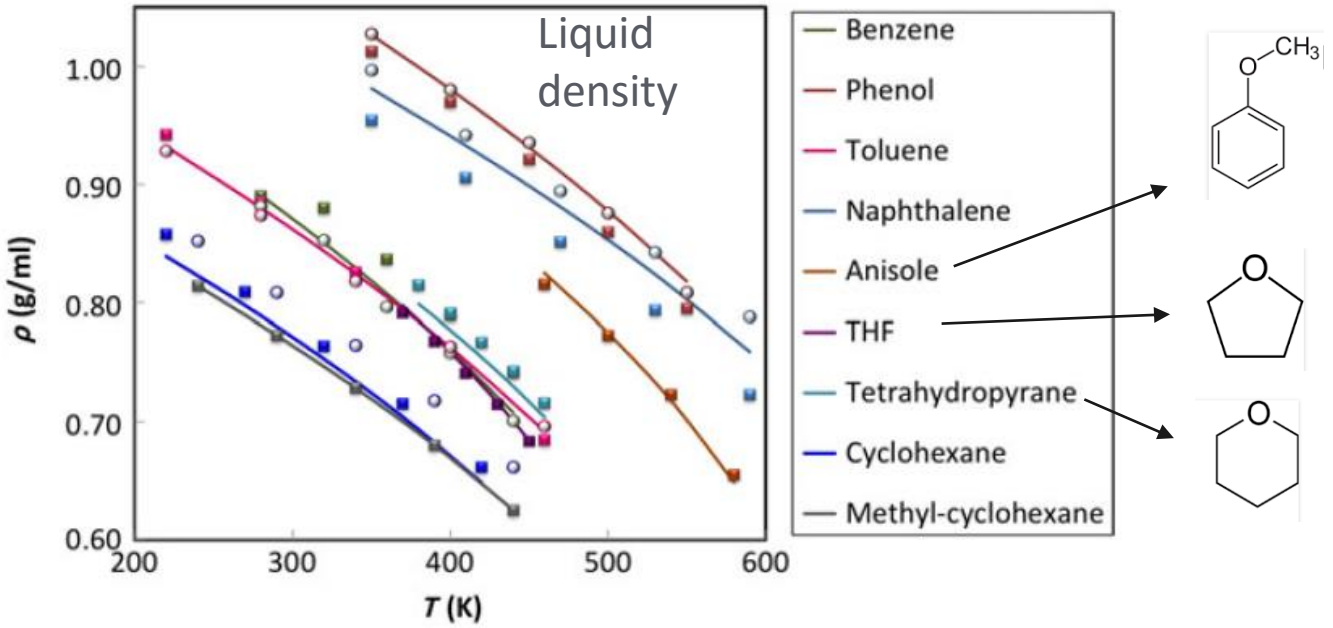


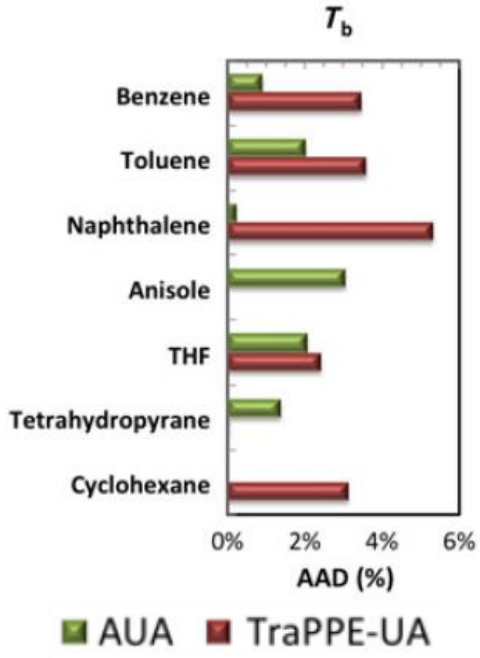


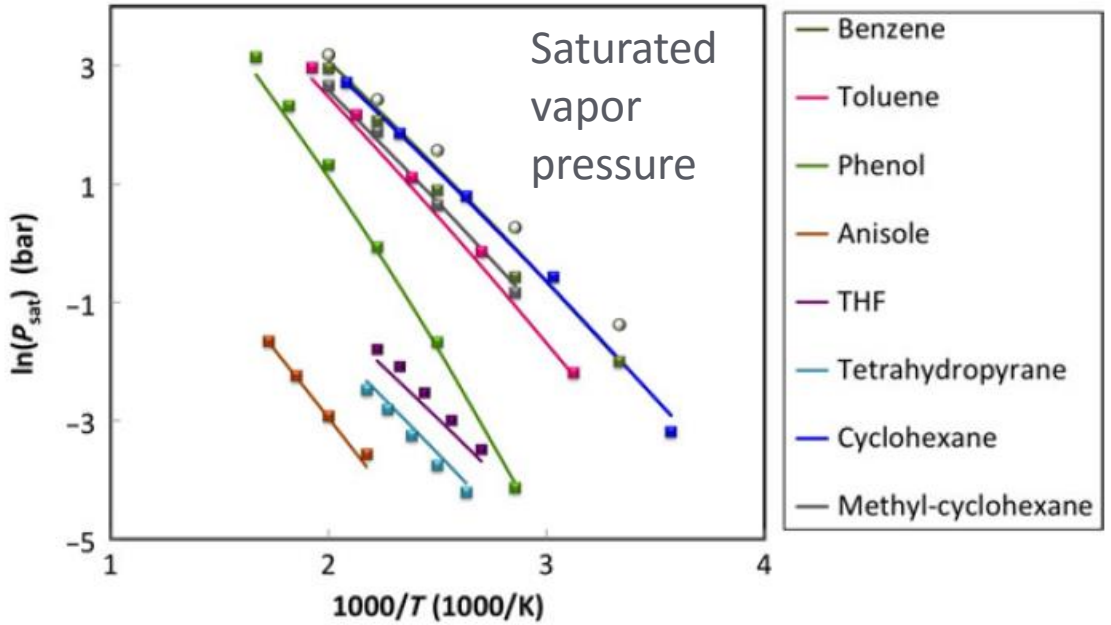


Cyclic and aromatic compounds

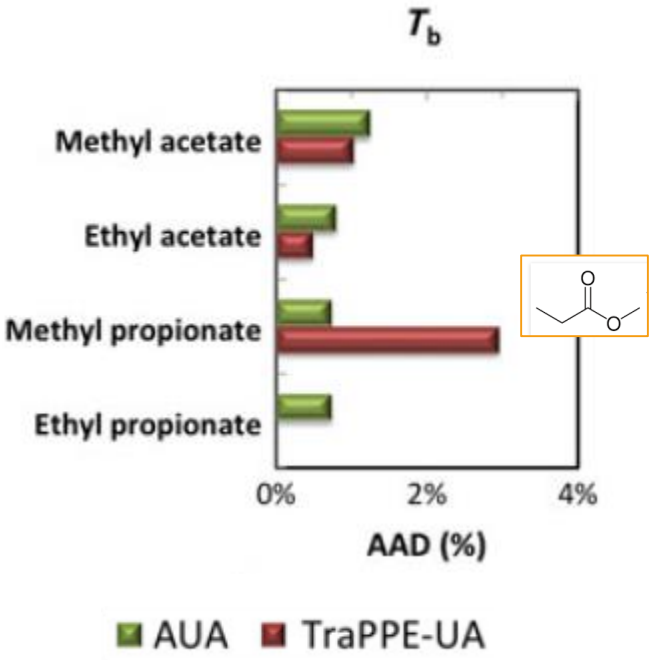
Open symbols : TraPPE
Filled symbols : AUA
LINES : DIPPR

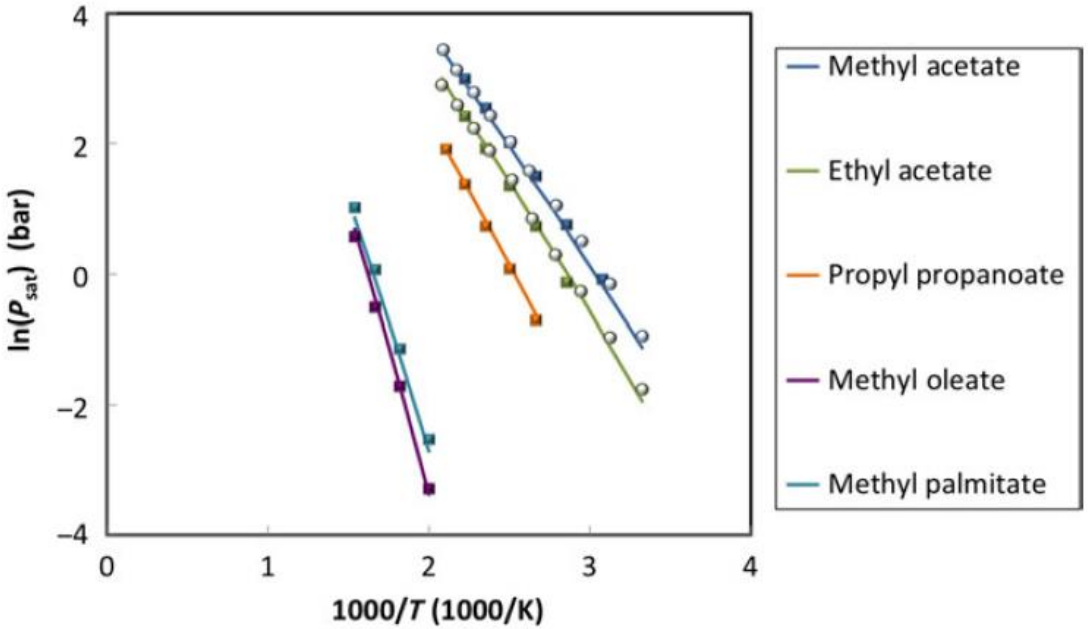




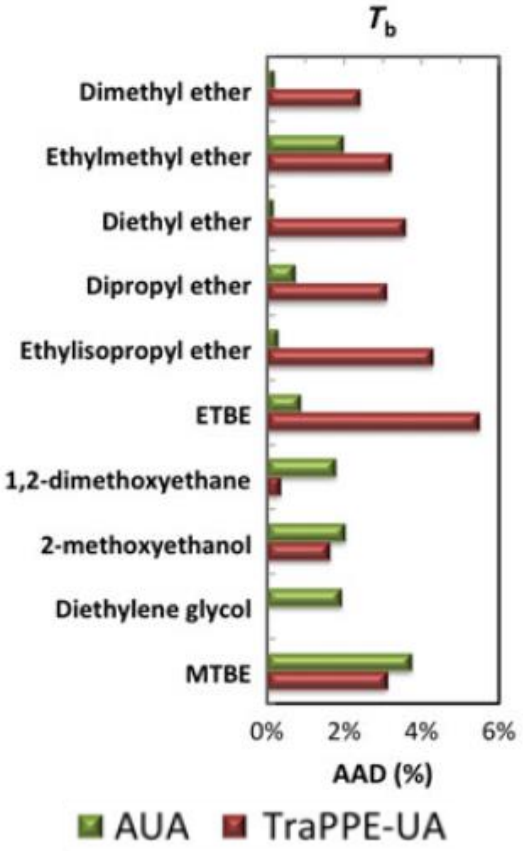


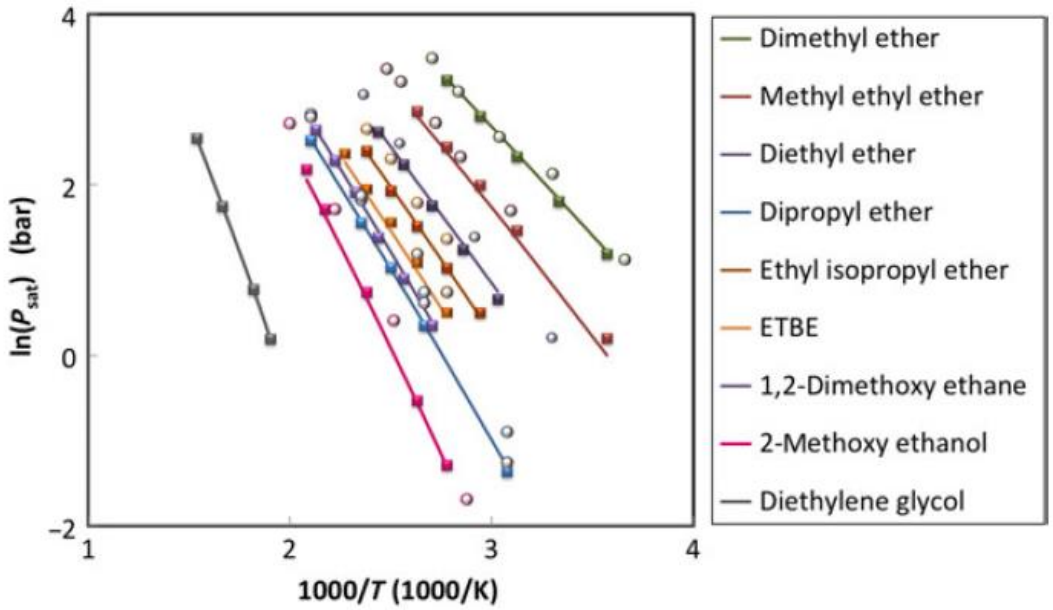
Esters


T_b
Methyl acetate
Ethyl acetate
Methyl propionate
Ethyl propionate
0%
2%
4%
AAD (%)
AUA
TraPPE-UA


4
2
0
−2
−4
0
1
2
3
4
ln(P_{sat}) (bar)
1000/T (1000/K)
Methyl acetate
Ethyl acetate
Propyl propanoate
Methyl oleate
Methyl palmitate

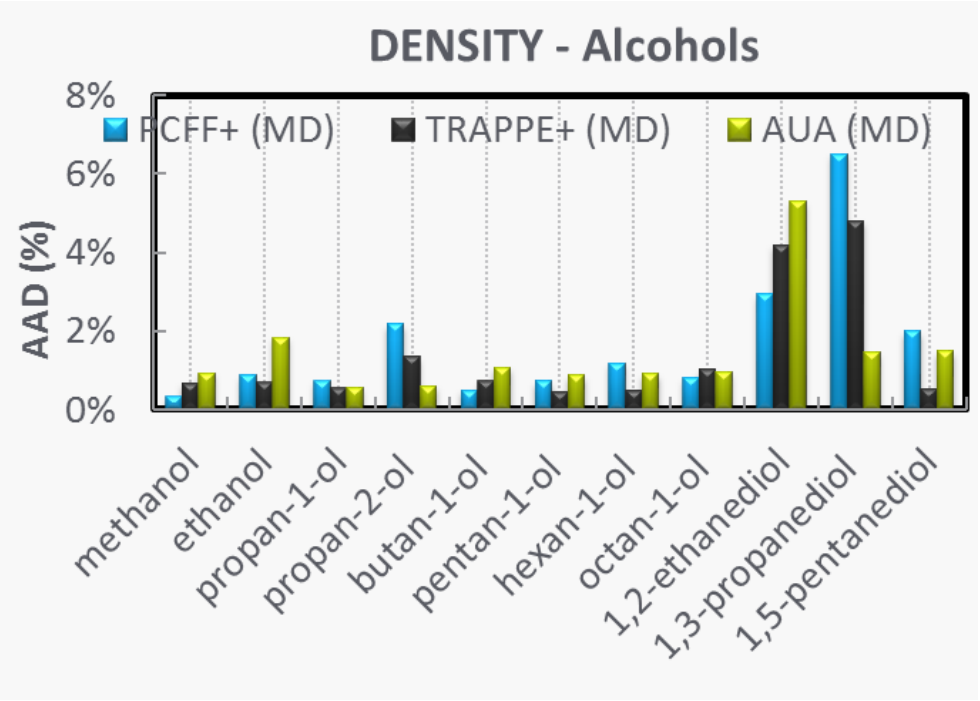
Ethers


T_b
Dimethyl ether
Ethylmethyl ether
Diethyl ether
Dipropyl ether
Ethylisopropyl ether
ETBE
1,2-dimethoxyethane
2-methoxyethanol
Diethylene glycol
MTBE
0%
2%
4%
6%
AAD (%)
AUA
TraPPE-UA

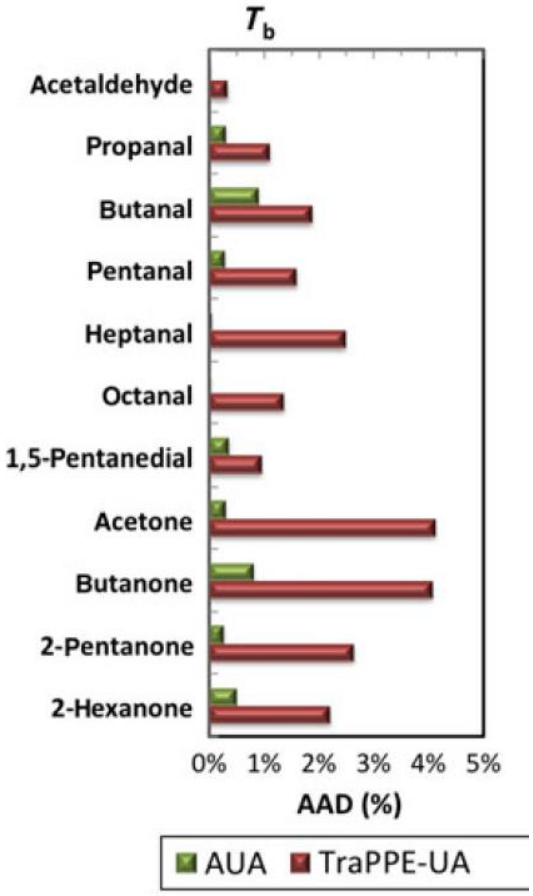

4
2
0
−2
1
2
3
4
ln(P_{sat}) (bar)
1000/T (1000/K)
Dimethyl ether
Methyl ethyl ether
Diethyl ether
Dipropyl ether
Ethyl isopropyl ether
ETBE
1,2-Dimethoxy ethane
2-Methoxy ethanol
Diethylene glycol

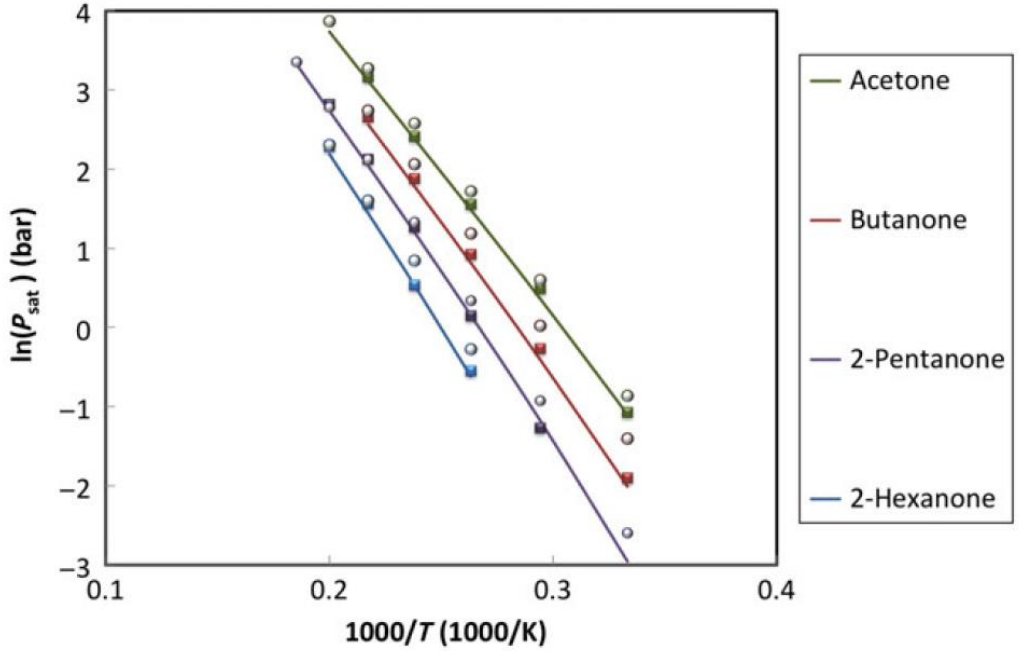
Alkanols, diols, triols

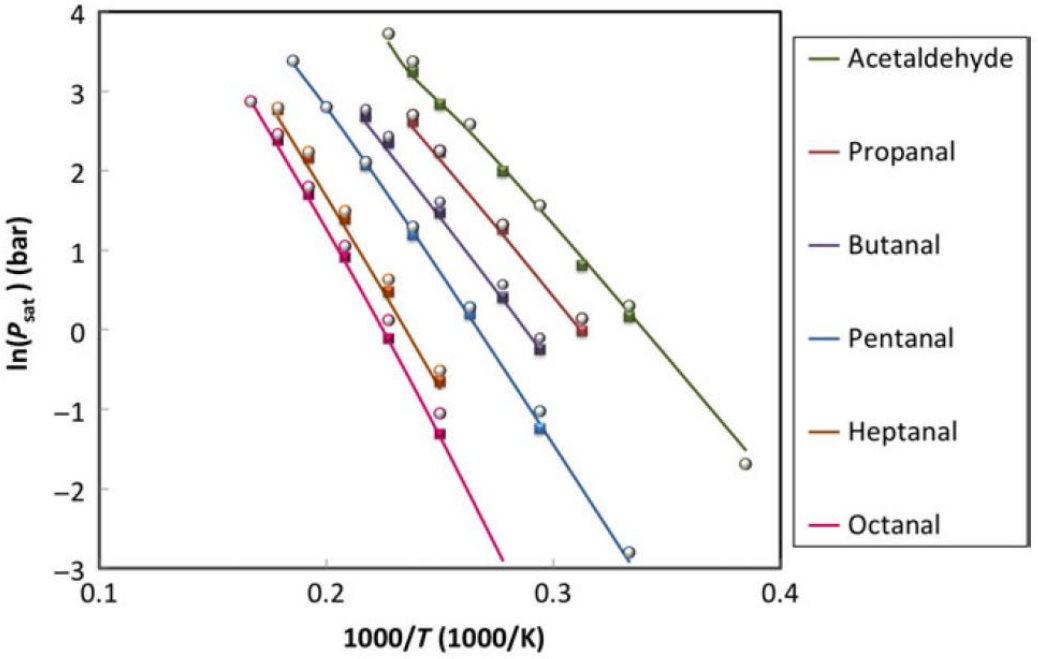
	AAD on Tb, %	AAD on liq density T <Tc
GIBBS - TraPPE	1.4 %	1.4 %
GIBBS - AUA	1.4% (2.0% incl glycols)	1.9 %
PCFF+		2.0% (3.0% with sorbitol,glycerol)
COSMO	2.4 %	



Ketones Aldehydes







QSAR

Why QSAR or QSPR ?

- Many properties cannot be computed from atomistic simulations (ex. Octane numbers of fuels, ecotoxicity, auto-ignition, kinetic rates in free radical mechanisms,….)
- Saves computing time when sampling is difficult (ex. Melting properties)
- Capitalize experimental data
- Simulation may generate useful descriptors (ex. Dipole moment, molecular size, saturation pressure, topology,…)

Current QSPR – QSAR in MedeA

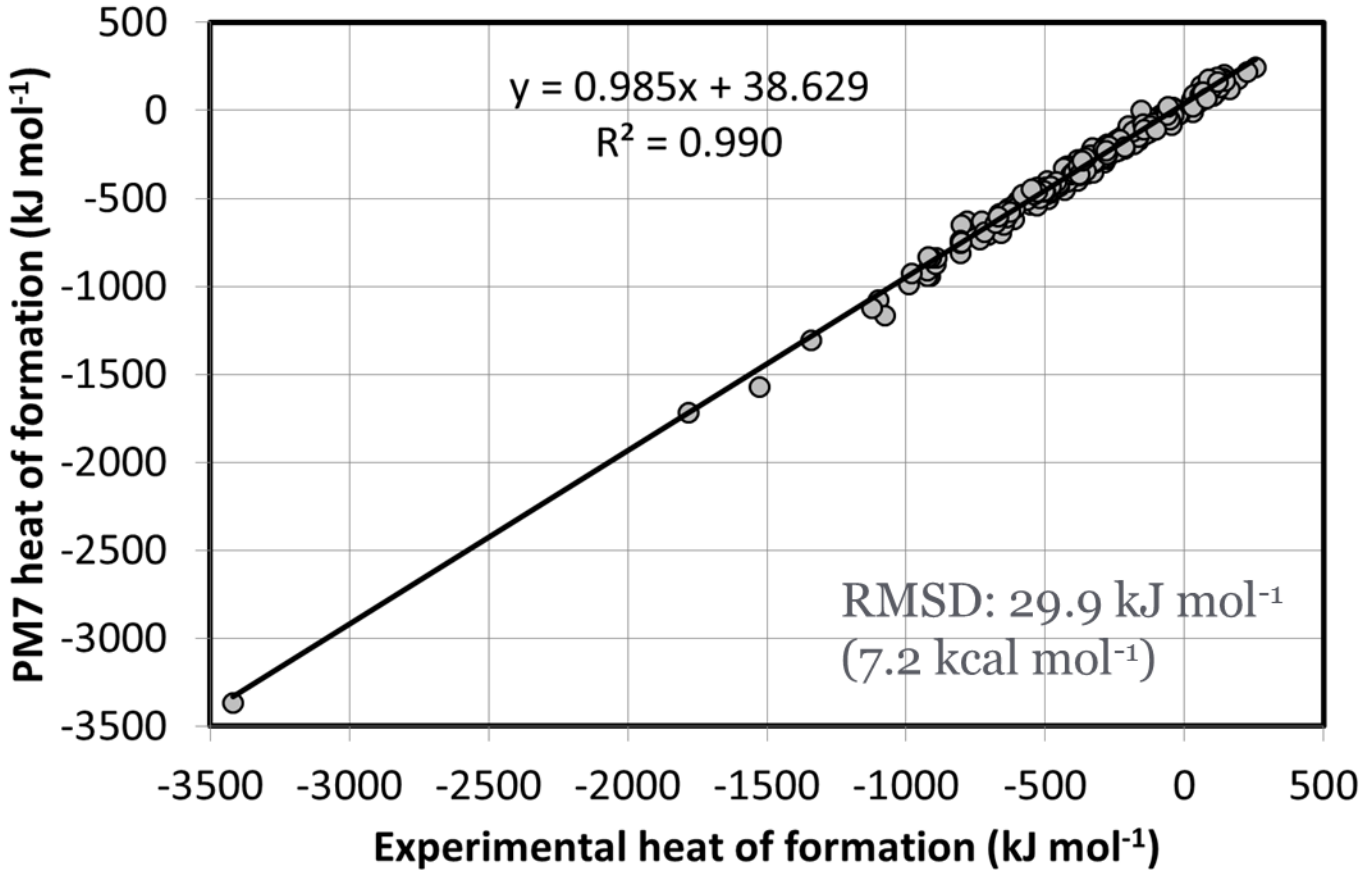
- P3C module -> properties of polymer materials using topological indicators *(Bicerano et al. Predition of polymer properties, 2002)*
- Designer correlations :
  - Define training set and validation set of data
  - Determination of descriptors from QM, MD , MC,…
  - Regress correlation parameters using standard spreadsheeting tools

Possible improvements in QSAR-QSPR developments :

- 1° include several conformers and multifunctional compounds in training set,
- 2° select functional forms with theoretical basis for large size molecules
- 3° include statistical uncertainties on either experiments or simulation results when regressing parameters (maximum likelihood criterion) and when evaluating correlation

Organic molecules

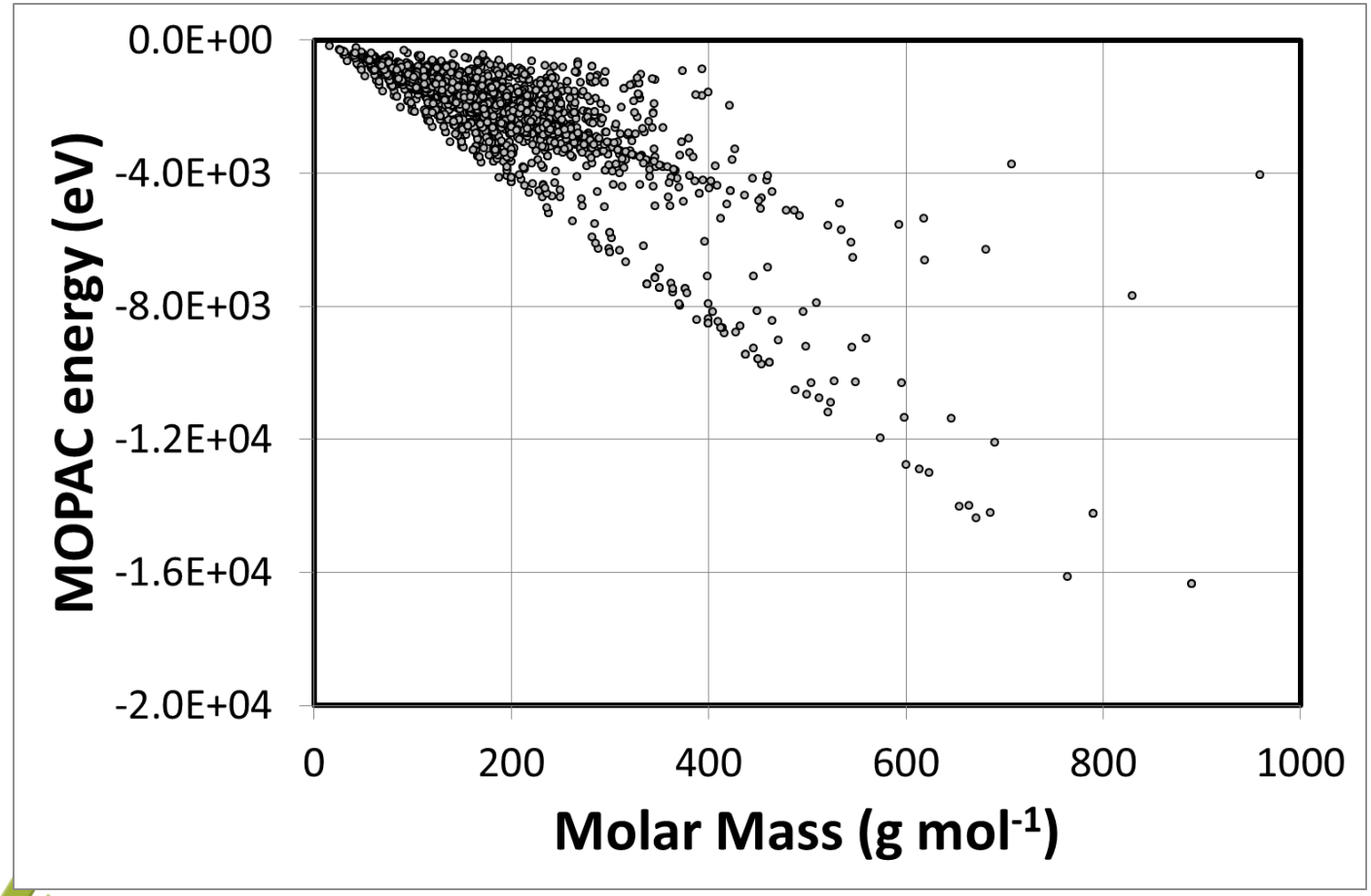
- **Heat of formation** - DIPPR exp. data with error estimated to be lower than 5% - set of 428 values



Source experimental data:DIADEM: The DIPPR Information and Data Evaluation Manager for the Design Institute for Physical Properties, Version 6.0.0, Database 2011

Stability and efficiency

- MOPAC geometry optimiz. and vibrational analysis flowchart
- ~24 hours to compute the set of 5869 molecules from EPI suite (2 proc.)
  - unsuccessful for 7 molc. but solved 'manually' (optimizer)



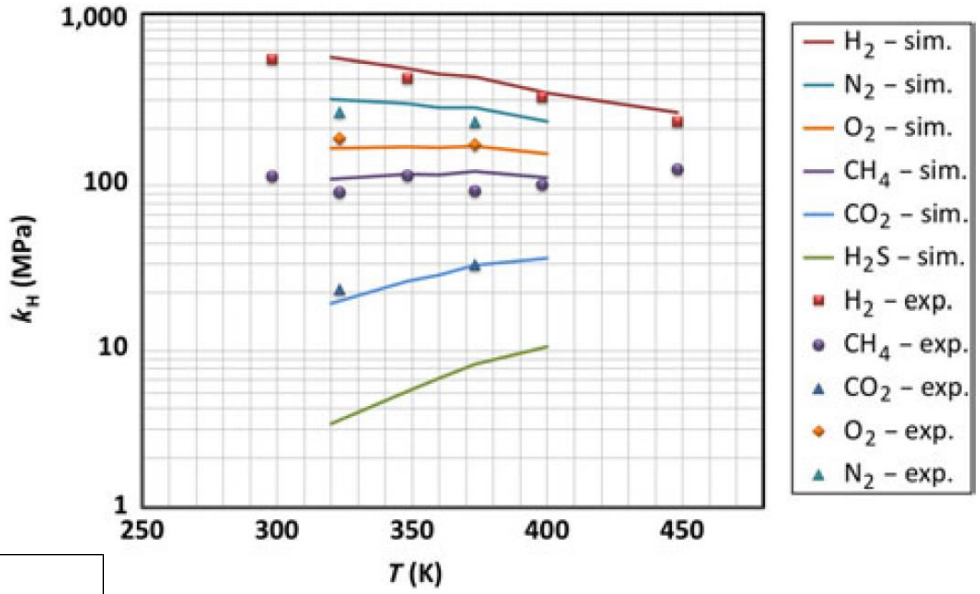
- Total energy
- ΔH°_{f}
- LUMO/HOMO energies
- IR/Raman frequencies
- $C_{pid}(T)$, S°, H_{id}, ΔG°_{f}
- COSMO volume and surface
- Dipole
- Charges

4 VLE OF MIXTURES

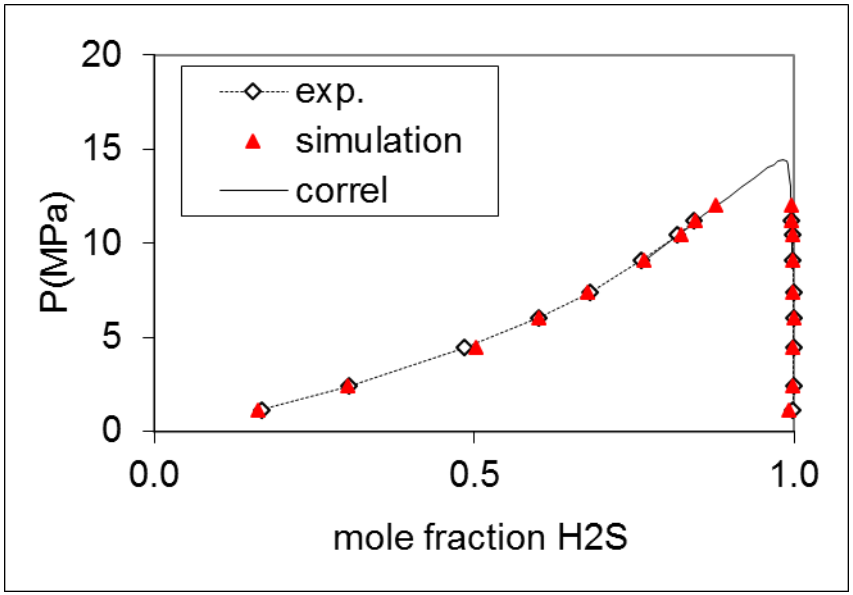
MedeA-GIBBS prediction of VLE/VLLE with gases

Henry solubility constants of gases in ethanol, AUA + literature forcefields for gases

(*Boutard et al., AIChE J., 2004*)



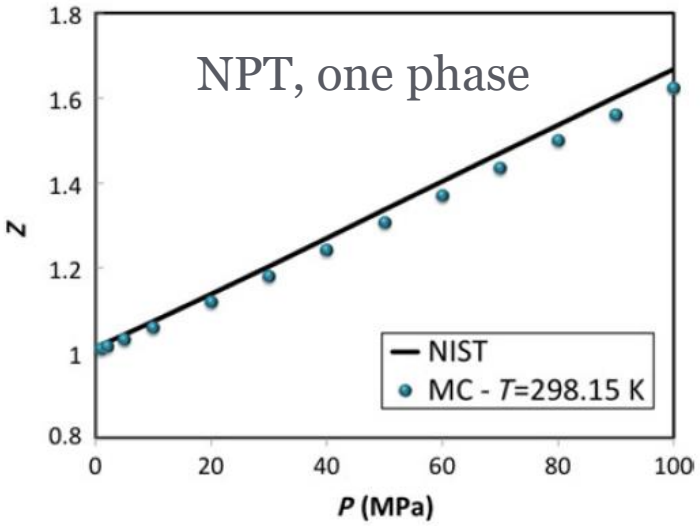
H2S-nC15, 422.3 K
(data from Laugier et al., 1995)

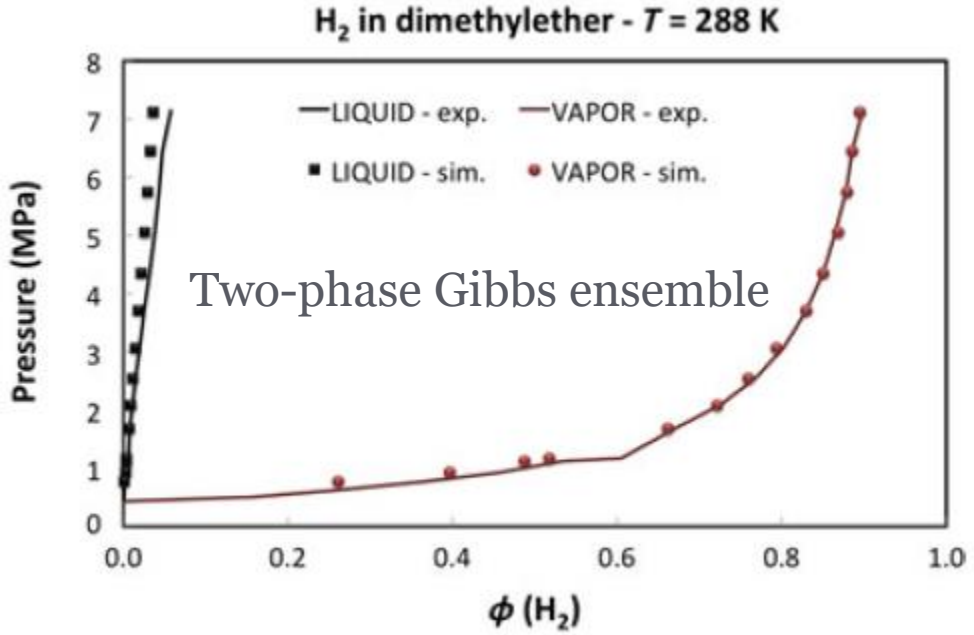


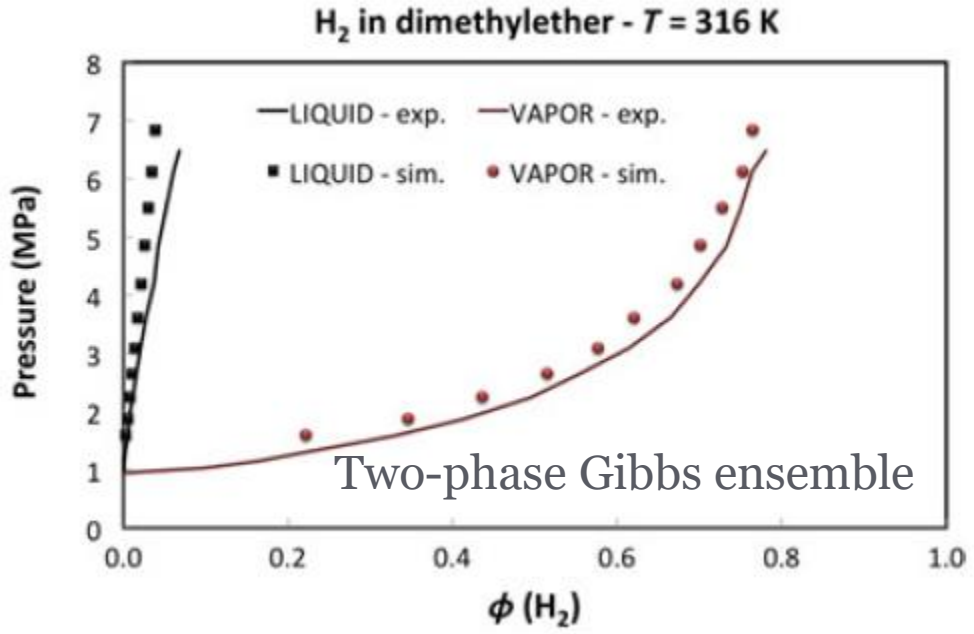
Supercritical H2S + long chain alkane
GEMC + near-critical extrapolation with scaling laws

High pressure phase diagram of H2-dimethylether

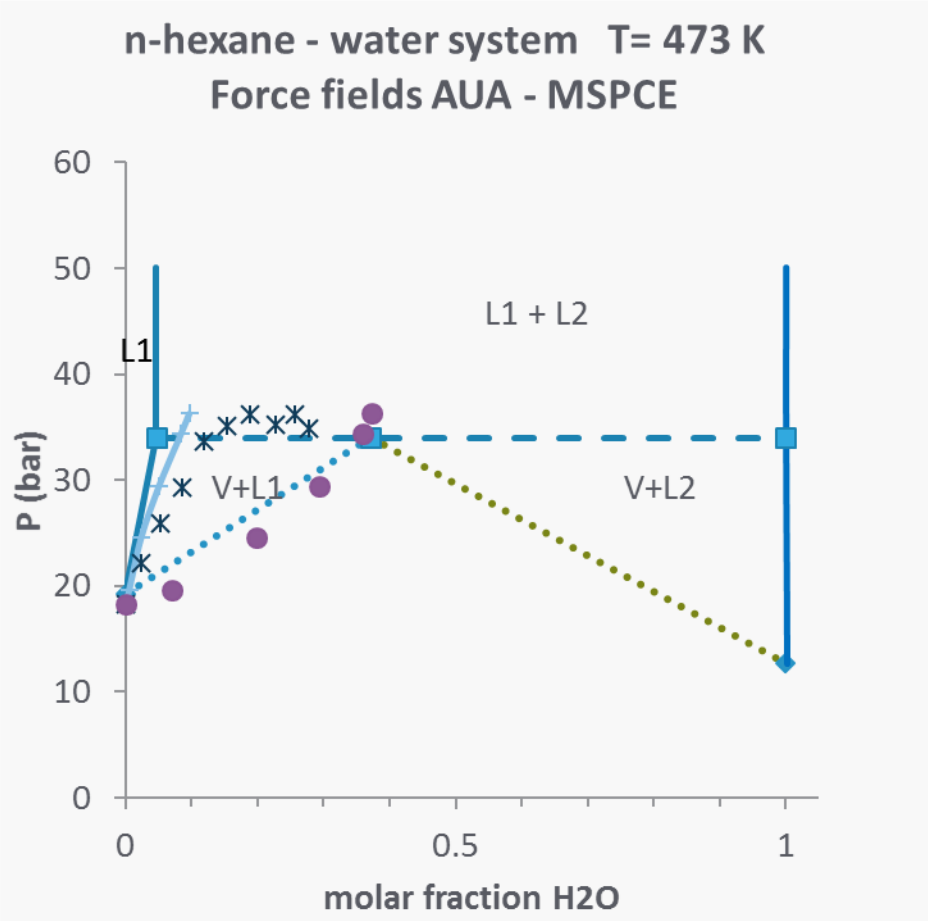
MedeA-GIBBS
AUA forcefield (ether)
Darkrim forcefield
(H2)

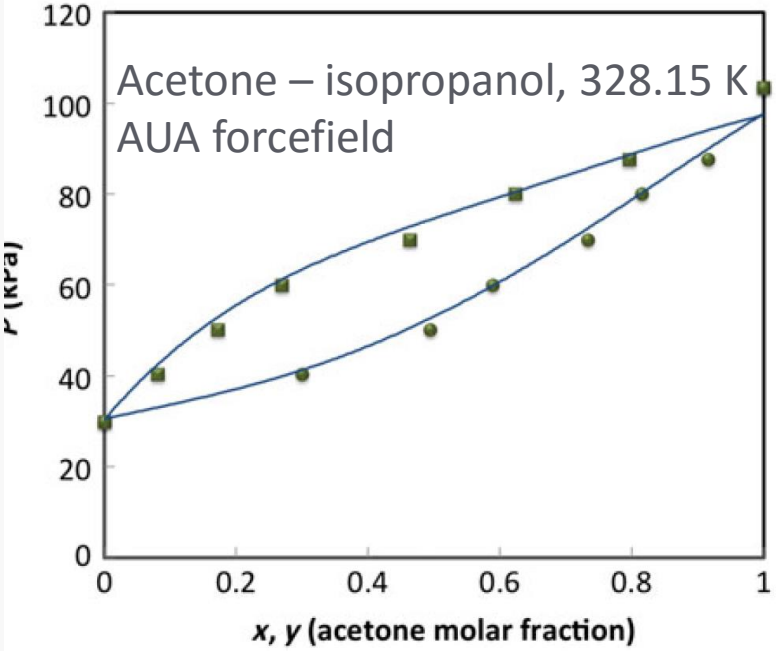






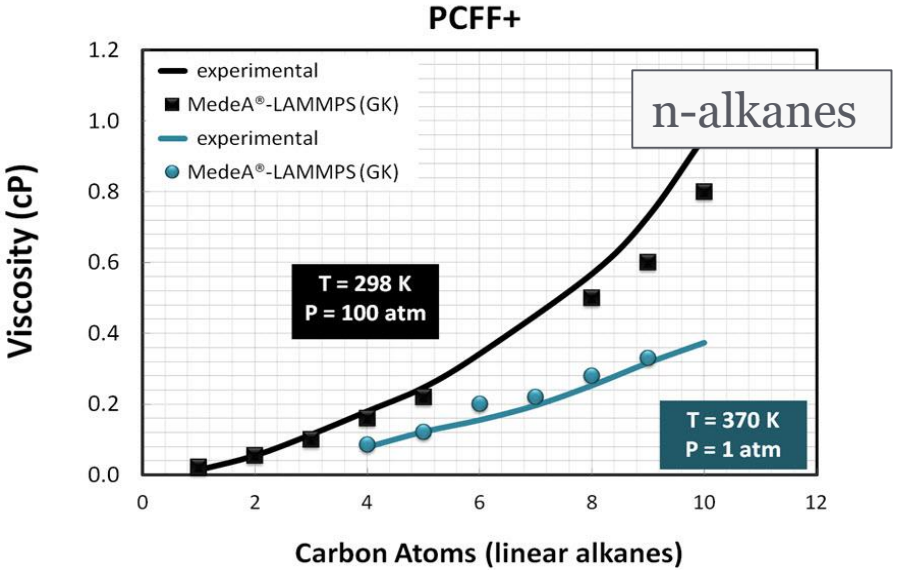
VLE and VLLE with MEdeA-GIBBS

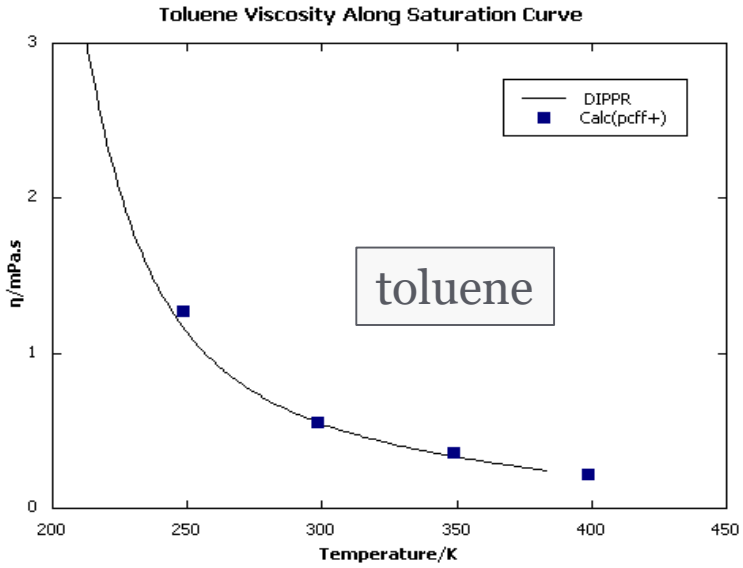




Transport properties (viscosity, diffusivity, heat conductivity) from forcefield-based methods

- Good viscosity predictions (5-20%) with All Atoms force fields using LAMMPS
- Good extrapolation capability in T and P
- Convergence with high viscosity liquids (>20 mPa.s) requires long computing times (days) and/or supercomputers, either with equilibrium or non-equilibrium molecular dynamics
- High throughput predictions appear feasible with similar approach as liquid density





Ungerer et al., Molec. Simulation, 2014

6 CONCLUSIONS

Conclusions

1. **Achievements**
   - User-friendly structure building and initialization, automated forcefield assignment, flowcharts
   - Automated post-processing, convergence control
   - High throughput simulations Structure list editor (SMILES code, conformer search, import/export)
   - Graphical flowcharts → user-friendly access to Quantum mechanics, Forcefield assignment, Molecular dynamics, Monte Carlo methods
   - Web-based job server → access to distant parallel computing resources
   - Automation of the simulation preparation, submission, processing, and collection of data tested on set of up to ~6000 molecules up to 11 carbon atoms
2. **Applications to molecular properties**
   - Liquid density : good performance of MD with pcff+ forcefield (AAD ~ 2%) and TraPPE as well
   - VLE properties : good performance on Tb, Psat, ΔH_{vap} with MedeA-GIBBS
   - Thermochemical properties : good performance of Semi-empirical QM for organics (avg. abs deviation ~3% on C_{Pid} at 300-1000 K, ; 30 kJ mol^{-1} on $\Delta H°_f$)
   - Transport properties : mature applications if relaxation times compatible with MD (
   - QSAR, QSPR : high productivity computation of descriptors for correlation development
3. **Remaining challenges :**
   - Large molecular weights , large number of conformers
   - Transferable forcefields for phase equilibria
   - Volumetric properties, transport properties : High efficiency with MD, wider scope expected from further extensions of Forcefields
   - VLE properties, : High throughput possible with increasing automation of Monte Carlo methods and extension of forcefield parametrization.