materials design®

Forum TERATEC – Ecole Polytechnique
Mercredi 24 Juin 2015

Calculs à haut débit de propriétés moléculaires dans l'environnement MedeA

Xavier Rozanska, Philippe Ungerer, Benoit Leblanc, Erich Wimmer
Materials Design, S.A.R.L., Montrouge, France

Context

- REACH (Registration, Evaluation, Authorization and restriction of CHemicals) – Chemical industry
  - 17 physico-chemical properties of several 1,000s of chemical compounds need to be determined
  - it can be done ***experimentally*** but also ***computationally***
- ANR PREDIMOL Objectives
  - To **demonstrate** whether REACH properties can be obtained by simulation by providing **only** molecular structures (and formulation)
  - Different methods were **evaluated**
    - Quantum chemistry (DFT, Semi-Empirical methods->QSPR, COSMO-RS)
    - Monte Carlo and molecular dynamic methods
  - Scientific validation and regulatory acceptance
- **Contributions of Materials Design**
  - **Automation** of the software procedures,
  - **Optimization** of simulation protocols and parameters
  - **High-throughput** calculations: typical set of 1k molecules

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

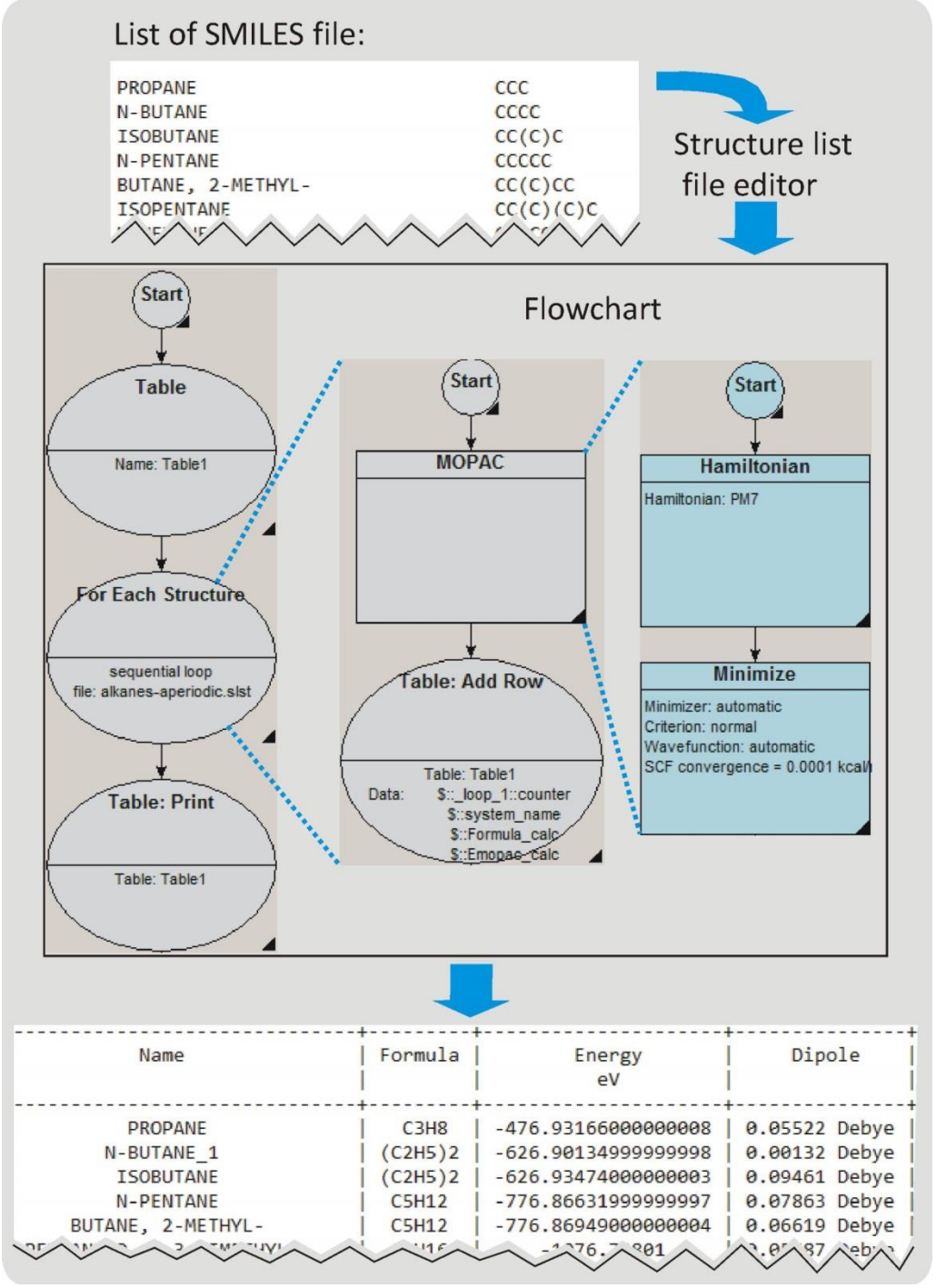
Task Servers

- Compute intensive tasks

VASP LAMMPS
GIBBS MOPAC

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 structures list**
- Integrated in the flowchart environment
  - Manipulation of structures
    - translation of atoms
    - supercell building
    - periodic/aperiodic
    - amorphous phase building
    - random atomic substitution
    - atomistic simulation
  - Loops over set of simulation variables and parameters
- Flowchart module: personalized Table printing



2. Models

Set of molecules

- 795 organic molecules SMILES formulas are collected from DIPPR database + experimental data when available
  - size C_1 to C_9 and covering 15 classes of organic compounds
    - 151 ketones and esters
    - 147 halogenated hydrocarbons
    - 111 amines and amides
    - 85 alkanes
    - 74 olefins
    - 7 alkylaromatics
    - 72 (mono) alcohols
    - 43 aldehydes
    - 37 polyols
    - 35 carboxylic acids
    - 15 oxanes
    - 10 peroxides
    - 8 isocyanates
- 515 inorganic gas molecules covering the entire periodic table (H to Bi) from Knacke *et al*. *Thermochemical properties of inorganic substances*, Springer-Verlag, Berlin, 1991

3. Methods

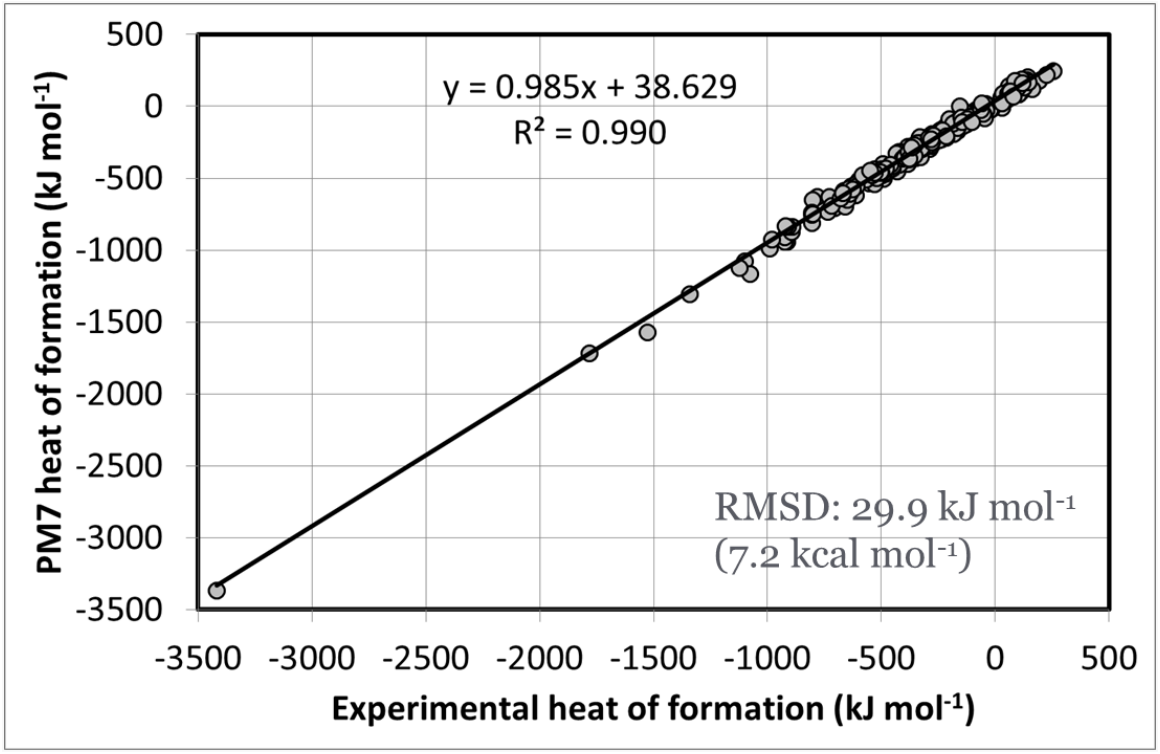
Methods

- Targets – Molecular frequencies of vibration and thermochemical properties within the rigid body harmonic approximation
  - DFT calculations - Gaussian/Turbomole
    - BP86/TZVP (DFT)
    - B3LYP/TZVP
  - Semi-empirical calculations –MedeA-MOPAC
    - PM7 (SEmp)
- Further details :
  - Rozanska X., Stewart J. J. P., Ungerer P., Leblanc B., Freeman F., Saxe P., Wimmer E. *J. Chem. Eng. Data* **2014**, *59*, 3136-3143.
  - Rozanska X., Ungerer P., Leblanc B., Saxe P., Wimmer E. *Oil Gas Sci. Technol.* **2014**

4. Molecular thermochemical properties

Organic molecules

- Heat of formation - selecting DIPPR exp. data with error lower than 5% - set of 428 values



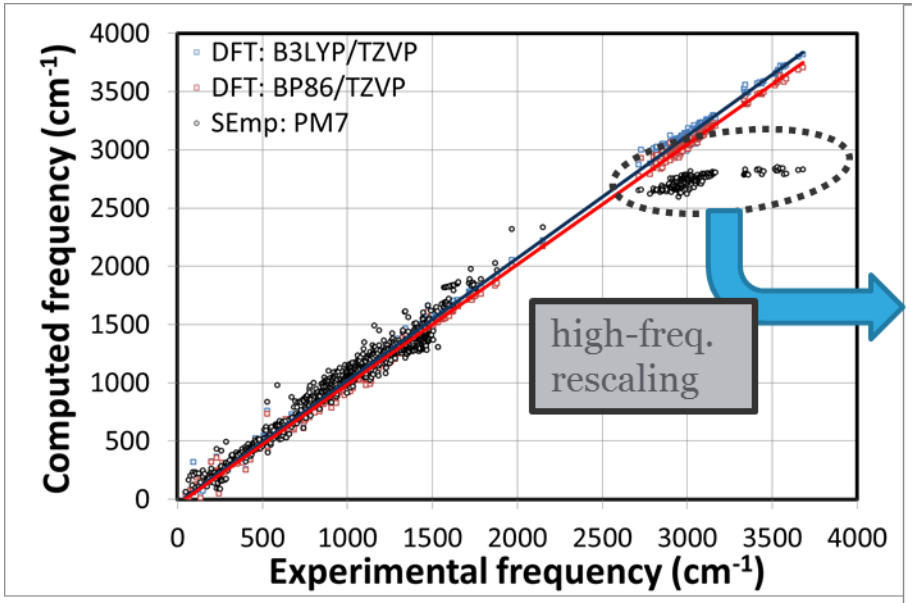
PM7 Average Unsigned Error for 1366 organic molc.: 17 kJ mol^{-1}

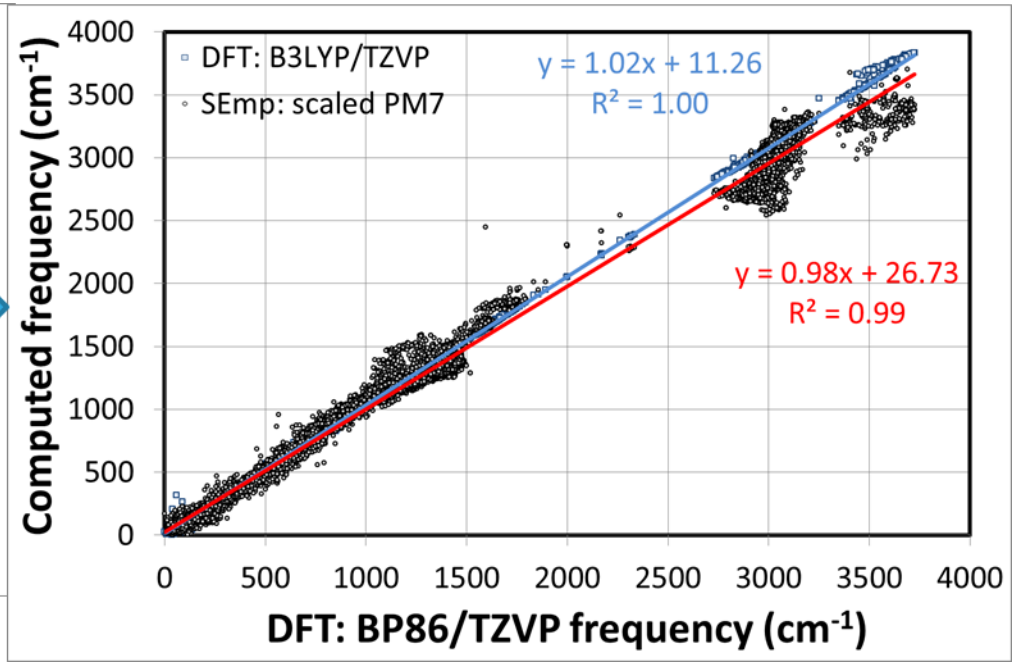
Stewart *J. Mol. Model.* **2013**, *19*, 1-32.

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

Organic molecules

- Comparison of the computed frequencies of vibrations : Experimental vs. computed for 52 organic molecules (1,135 values)

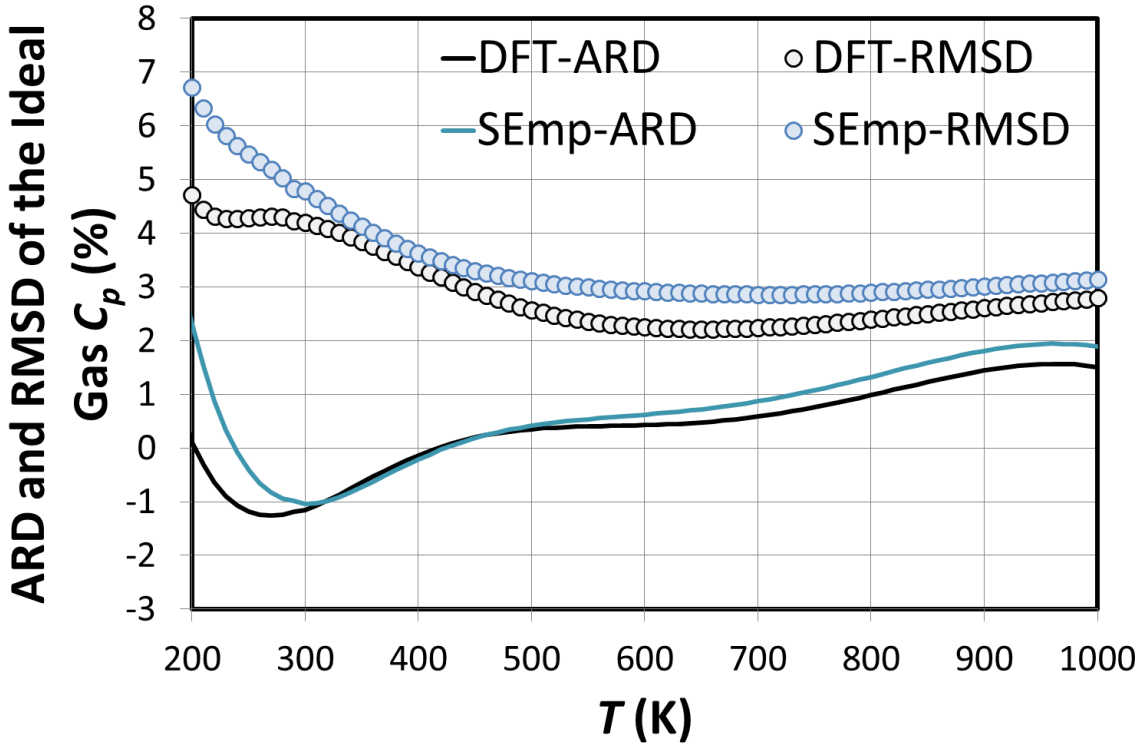




Source experimental data : NIST Chemistry WebBook; Mallard, W.G.; Linstrom, P.J., Eds. NIST Standard Reference Database Number 69; National Institute of Standards and Technology: Gaithersburg, MD, 2011, (http://webbook.nist.gov)

Organic molecules

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



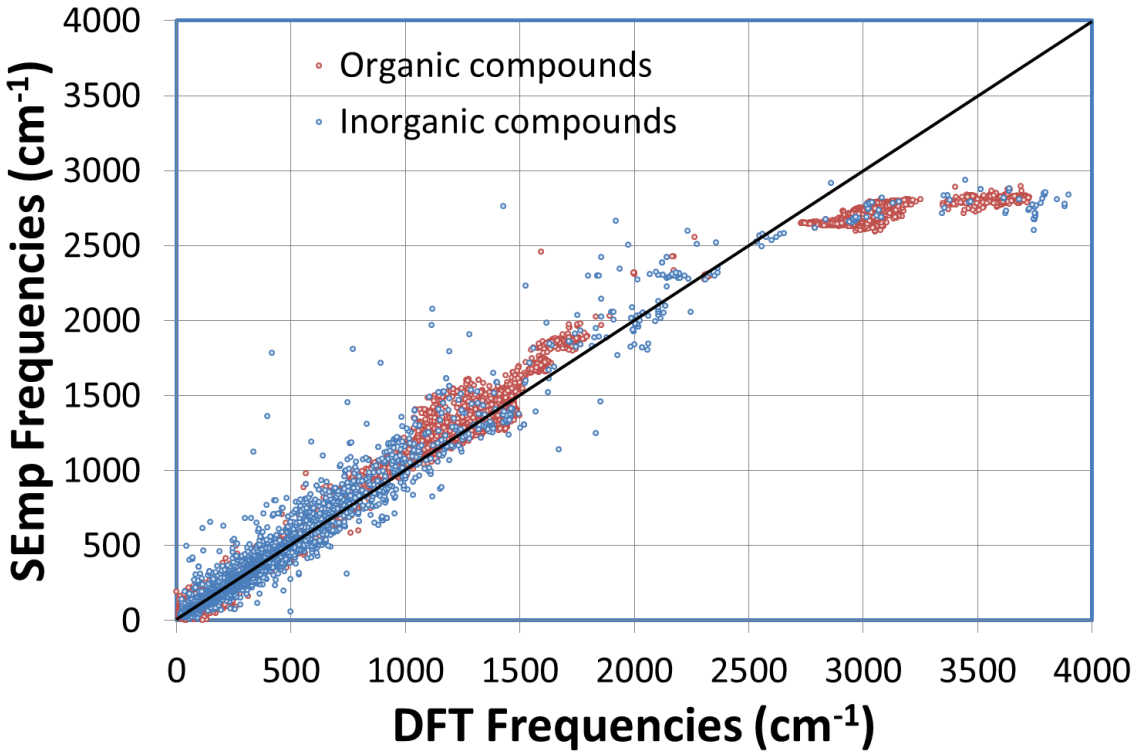
SEmp vs. DFT

ΔC_p :
T=300 K, RMSD=4.0%
T=1000 K, RMSD=1.0%
$\Delta S°$:
T=300 K, RMSD=6.5%
T=1000 K, RMSD=5.0%
$\Delta G°$:
T=300 K, RMSD=15 kJ mol^{-1}
T=1000 K, RMSD=30 kJ mol^{-1}

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.

Inorganic molecules

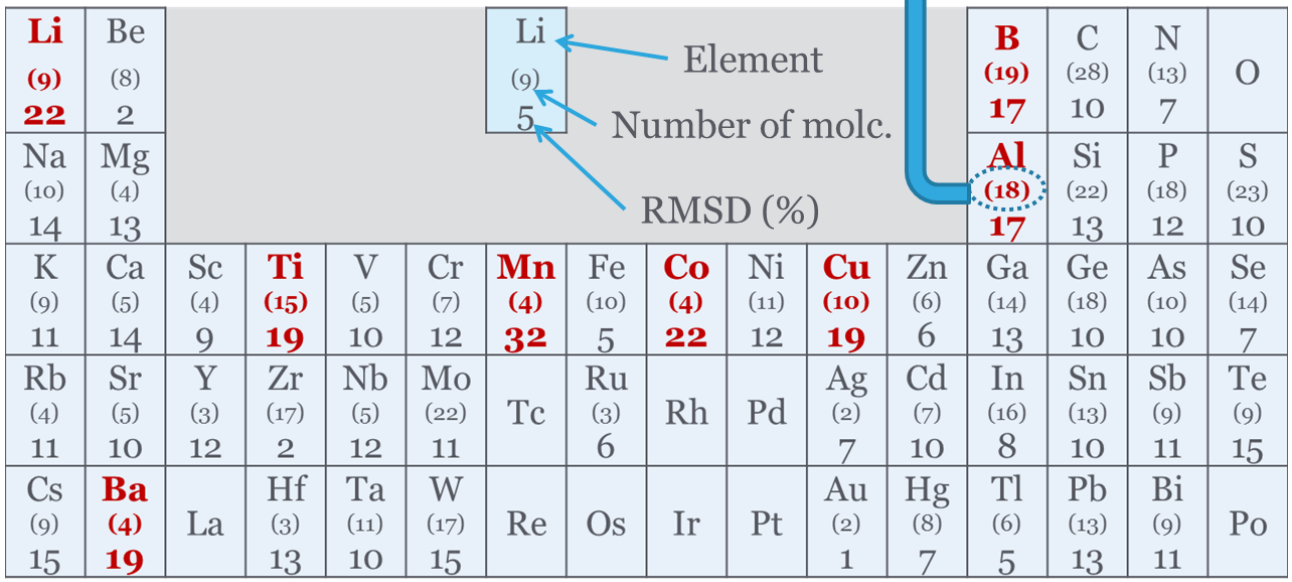
- Comparison of the unscaled semi-empirical frequencies of vibration with the DFT values for all organic and inorganic molecules : 1395 molecules, more than 42,000 freq.



Inorganic molecules

- Ideal gas C_P at T=298 K : RMSD of the average relative difference between SEmp and DFT

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	$(Al_2O)_2$



SEmp vs. DFT

515 Inorganic molc.
ΔC_p :
T=300 K, RMSD=4.8%
$\Delta S°$:
T=300 K, RMSD=5.0%

795 Organic molc.
ΔC_p :
T=300 K, RMSD=4.0%
$\Delta S°$:
T=300 K, RMSD=6.5%

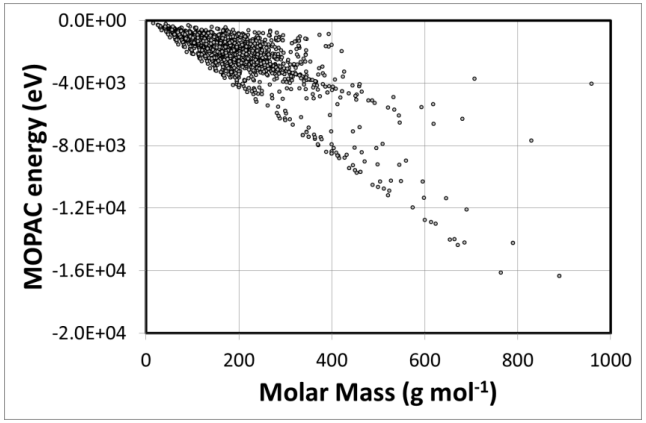
Source for the 515 inorganic molecules : Knacke *et al*. *Thermochemical properties of inorganic substances*, Springer-Verlag, Berlin, 1991

Stability and efficiency

- SMILES formula list of 5869 molecules (EPISuite, NCI)

H 5667																
Li	Be											B 24	C 5869	N 1429	O 3068	F 514
Na	Mg											Al 4	Si 120	P 49	S 445	Cl 917
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge 1	As 15	Se 5	Br 338
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn 14	Sb 2	Te	I 91
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl 1	Pb 2	Bi	Po	At

F 514 ← Number of molecules containing this element in the set

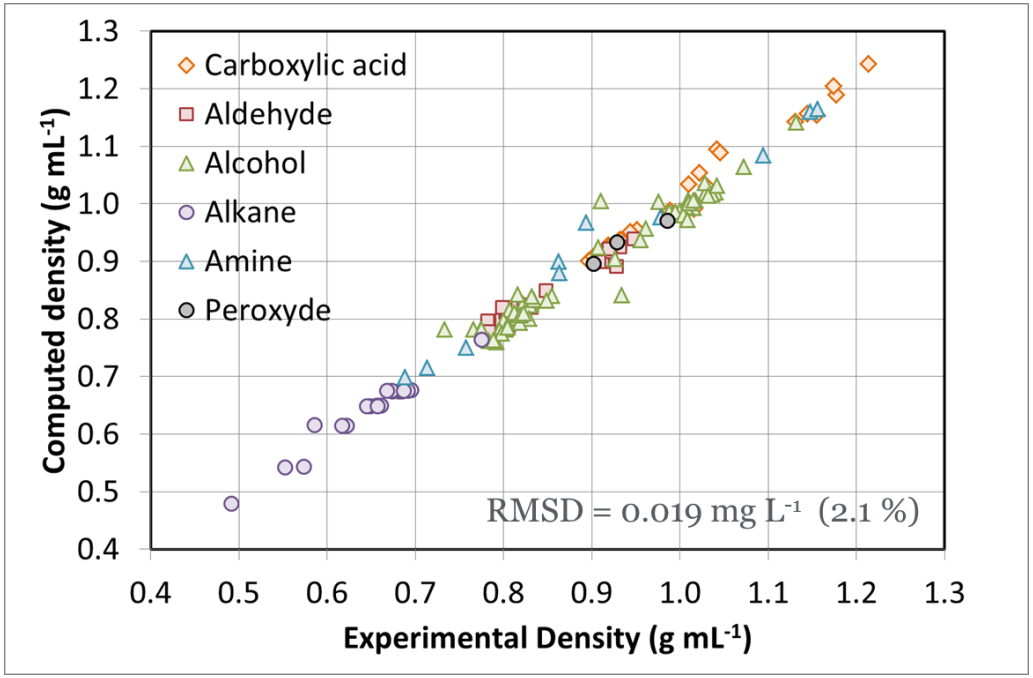


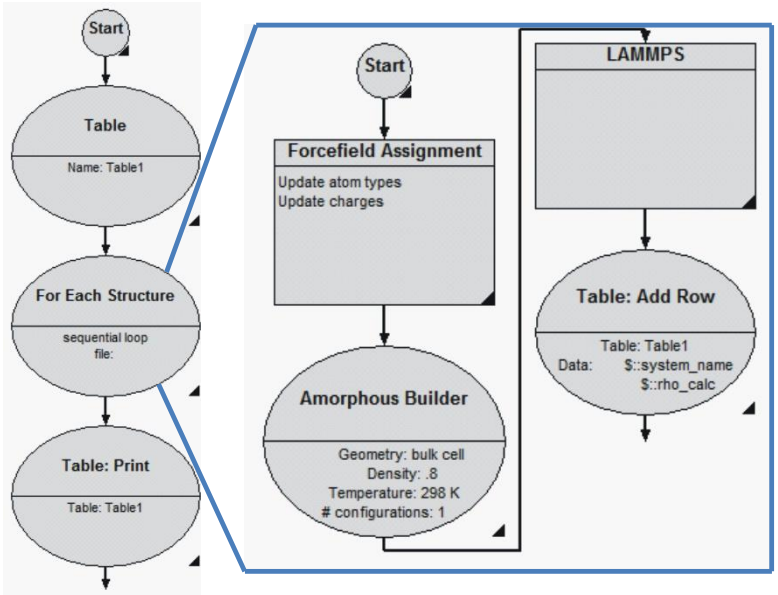
- MOPAC geometry optimization and frequency calculations on all structures

5. Other applications

Robustness of the MedeA's flowchart environment

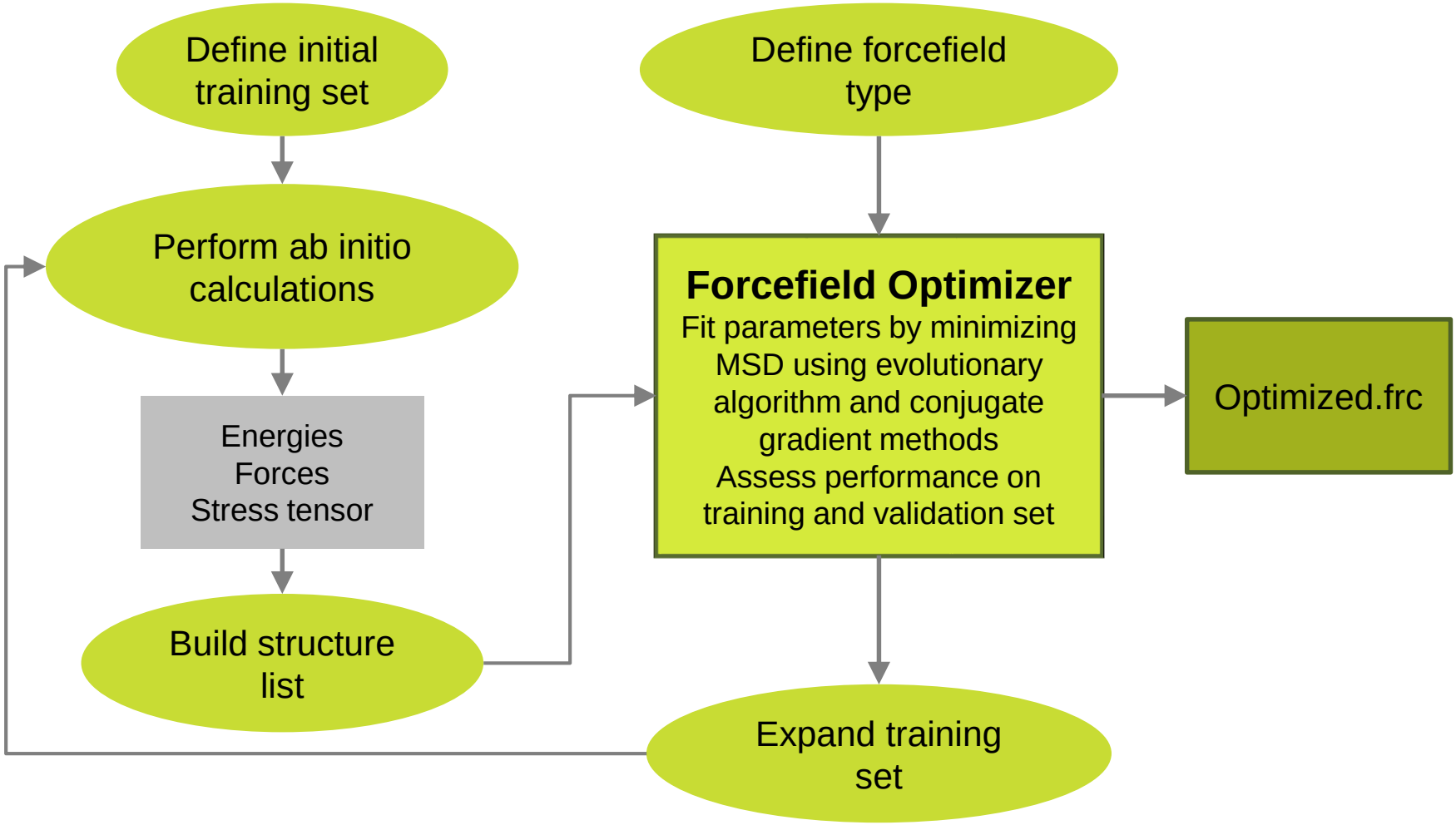
- The software automations are compatible with other softwares : e.g. LAMMPS and VASP
- Comparison of liquid density at *P*=1 bar and *T*=298 K for 174 molecules



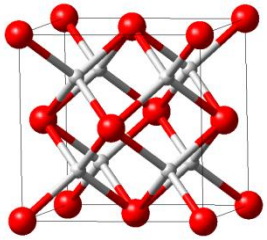


Source experimental data : DIADEM: The DIPPR Information and Data Evaluation Manager for the Design Institute for Physical Properties, Version 6.0.0, Database 2011

MedeA®-Forcefield Optimizer



Ionic Forcefield for Li_2O

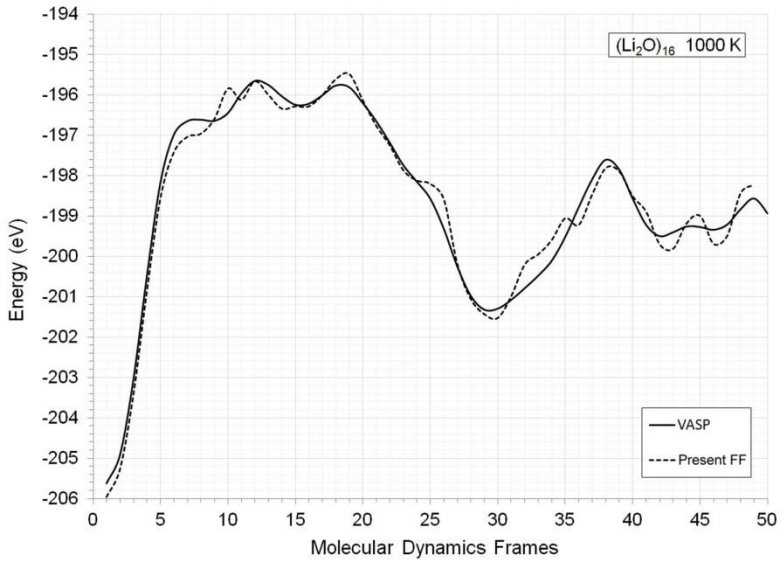


$$E = \sum_{j>i} \left(\frac{q_i q_j}{r_{ij}} + A_{ij} e^{-r_{ij}/\rho_{ij}} - C_{ij} r_{ij}^{-6} \right)$$

MedeA®-VASP Forcefield Optimizer

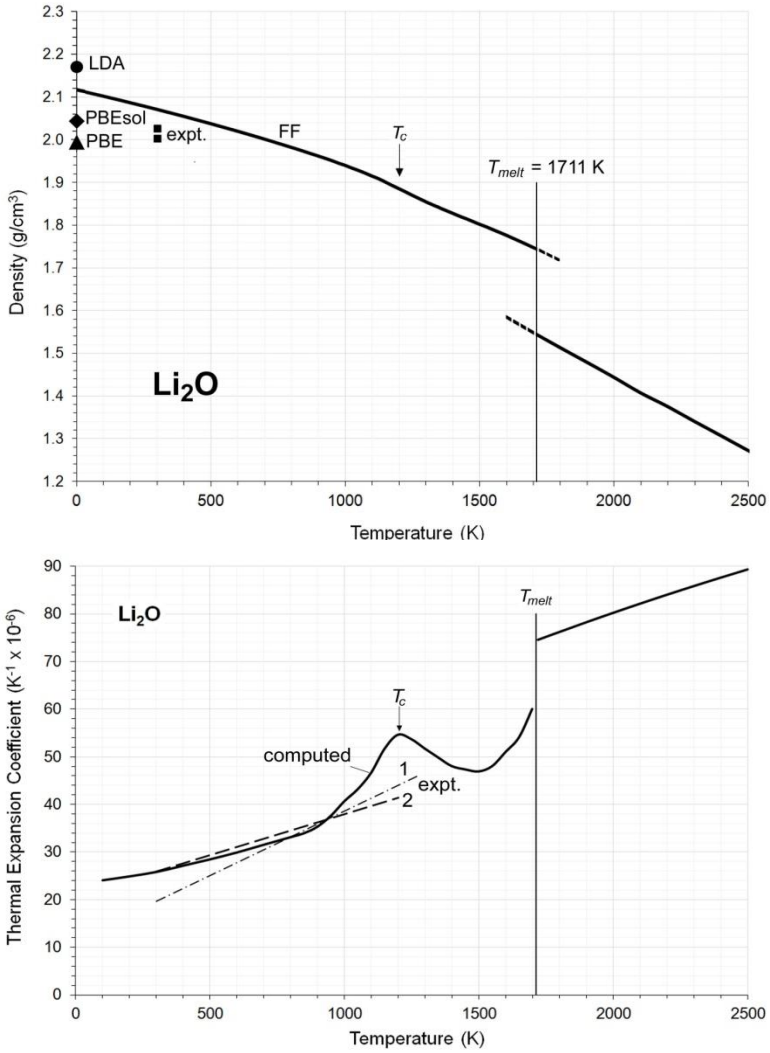
q_{Li}=0.79091, $A_{Li\text{-}O}$=1425.5, $\rho_{Li\text{-}O}$=0.23630

- Fit to ab initio (VASP) molecular dynamics trajectory
  - charges used in fitting
- Calibration with melting temperature
- Validation on structural and mechanical properties
- Application to thermal expansion and diffusion

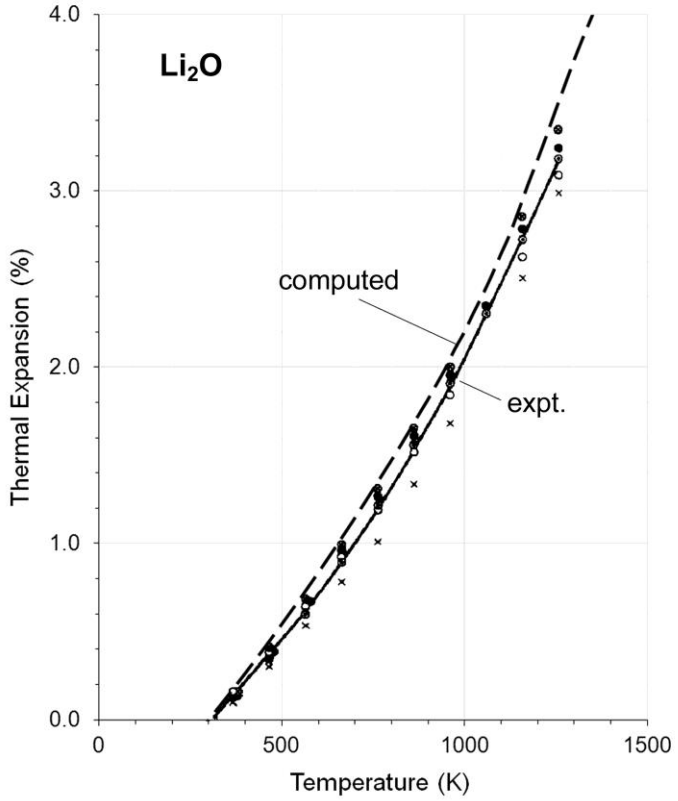


Asahi, R., Freeman, C. M., Saxe, P., & Wimmer, E. (2014). Thermal expansion, diffusion and melting of Li_2O using a compact forcefield derived from ab initio molecular dynamics. *Modelling and Simulation in Materials Science and Engineering*, *22*(7), 075009.

Density and Thermal Expansion



MedeA®-LAMMPS



Asahi et al.(2014)

5. Summary

Summary

- MedeA software environment
  - Structures list editor (SMILES formulas, conformer search, import/export)
  - Loop over structures list for LAMMPS, MOPAC, and VASP
    - fully integrated in the MedeA flowchart
      - manipulation of structures
      - manipulation of simulation parameters
    - cover a much wider range of properties than shown here
- Thermochemistry
  - Evaluation and comparison of PM7 SEmp method vs. Exp. and DFT
    - SEmp about 100 faster than DFT
    - Errors (RMSD) of SEmp and DFT *vs*. Exp are less than 1 % different
    - Errors on inorganic and organic compounds are the same
- High-throughput calculations: stability and efficiency
  - Automation of the simulation preparation, submission, processing, and collection of data tested on set of up to ~6000 molecules
  - Big Data generation

Acknowledgements

Support by the *Agence Nationale de la Recherche* (Research project PREDIMOL ANR-2010-CD2I-09)

PREDIMOL partners