Dossier

This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 70, No. 3, pp. 395-519 and available online here

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 70, n°2, pp. 395-519 et téléchargeable ici

DOSSIER Edited by/Sous la direction de : **V. Santos-Moreau**

IFP Energies nouvelles International Conference / Les Rencontres Scientifiques d'IFP Energies nouvelles

NEXTLAB 2014 - Advances in Innovative Experimental Methodology or Simulation Tools used to Create, Test, Control and Analyse Systems, Materials and Molecules

NEXTLAB 2014 - Innover dans le domaine de la méthodologie expérimentale et des outils de simulation pour créer, tester, contrôler et analyser des systèmes, matériaux et molécules

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 70 (2015), No. 3, pp. 395-519

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 70 (2015), No. 3, pp. 405-417

DOI: 10.2516/ogst/2014041

IFP Energies nouvelles International Conference
Rencontres Scientifiques d'IFP Energies nouvelles

NEXTLAB 2014 - Advances in Innovative Experimental Methodology or Simulation Tools used to Create, Test, Control and Analyse Systems, Materials and Molecules
NEXTLAB 2014 - Innover dans le domaine de la méthodologie expérimentale et des outils de simulation pour créer, tester, contrôler et analyser des systèmes, matériaux et molécules

Automatic and Systematic Atomistic Simulations in the MedeA® Software Environment: Application to EU-REACH

Xavier Rozanska[1]*, Philippe Ungerer[1], Benoit Leblanc[1], Paul Saxe[2] and Erich Wimmer[1]

[1] *Materials Design SARL, 18 rue de Saisset, 92120 Montrouge - France*
[2] *Materials Design, Inc., 6 First National Place, Angel Fire, NM 87710 - USA*
e-mail: xrozanska@materialsdesign.com - pungerer@materialsdesign.com - bleblanc@materialsdesign.com - psaxe@materialsdesign.com - ewimmer@materialsdesign.com

* Corresponding author

Abstract *— This work demonstrates the systematic prediction of thermodynamic properties for batches of thousands of molecules using automated procedures. This is accomplished with newly developed tools and functions within the Material Exploration and Design Analysis (MedeA®) software environment, which handles the automatic execution of sequences of tasks for large numbers of molecules including the creation of 3D molecular models from 1D representations, systematic exploration of possible conformers for each molecule, the creation and submission of computational tasks for property calculations on parallel computers, and the post-processing for comparison with available experimental properties. After the description of the different MedeA® functionalities and methods that make it easy to perform such large number of computations, we illustrate the strength and power of the approach with selected examples from molecular mechanics and quantum chemical simulations. Specifically, comparisons of thermochemical data with quantum-based heat capacities and standard energies of formation have been obtained for more than 2 000 compounds, yielding average deviations with experiments of less than 4% with the Design Institute for Physical PRoperties (DIPPR) database. The automatic calculation of the density of molecular fluids is demonstrated for 192 systems. The relaxation to minimum-energy structures and the calculation of vibrational frequencies of 5 869 molecules are evaluated automatically using a semi-empirical quantum mechanical approach with a success rate of 99.9%. The present approach is scalable to large number of molecules, thus opening exciting possibilities with the advent of exascale computing.*

Résumé — Simulations atomistiques automatiques et systématiques dans l'environnement logiciel de MedeA® : Application à EU-REACH — Ce travail démontre notre capacité à prédire systématiquement les propriétés thermodynamiques par lot de plusieurs milliers de molécules en utilisant des procédures automatiques. Ceci est accompli à l'aide de nouveaux outils et fonctions intégrés dans l'environnement logiciel de MedeA® (*Material Exploration and Design Analysis*), qui manipule l'exécution automatique de séquences de tâches sur de grands nombres de molécules comprenant la création de modèles moléculaires 3D à partir de représentations 1D, l'exploration systématique des conformations possibles pour chaque molécules, la création et la soumission de tâches informatiques pour calculer des propriétés sur des ordinateurs parallèles, et le post-traitement par comparaison avec les propriétés experimentales disponibles. Après la description des différentes fonctionnalités et méthodes de MedeA® qui facilitent grandement le traitement de si grand nombre

de calculs, nous illustrons la puissance et la force de l'approche avec des exemples sélectionnés à partir de simulations de mécanique moléculaire et chimie quantique. En particulier, la comparaison des données thermochimiques moléculaires obtenues par chimie quantique, notamment les énergies de formation et les chaleurs spécifiques moléculaires, avec les valeurs expérimentales a été obtenue pour plus de 2 000 composés, conduisant à des déviations entre des valeurs expérimentales tirées des bases de données de la DIPPR (*Design Institute for Physical PRoperties*) entre autres et les valeurs calculées de moins de 4 %. Le calcul automatique de la densité de fluides moléculaires est démontré pour 192 systèmes. Les relaxations dans leurs structures d'énergies minimales et le calcul des fréquences vibratoires de 5 869 molécules sont évalués automatiquement en utilisant une approche de mécanique quantique semi-empirique avec un taux de succès de 99,9 %. La présente approche peut être étendue à de grand nombre de molécules, ouvrant ainsi des possibilités excitantes en vue de l'avènement du calcul à l'échelle exascale.

INTRODUCTION

Within the context of the European Union Registration, Evaluation, Authorization and restriction of CHemicals protocol (EU-REACH) [1], a daunting number of chemical compounds has to be characterized in the form of at least 17 physico-chemical properties for each compound. Furthermore, the properties need to be determined not only for pure compounds, but also for any formulations of the compounds under which they are commercialized and transported. Around 17 000 compounds have been pre-registered already in the REACH database [2], and all compounds are required to be fully registered by the end of 2018. This is a large number of systems, yet it represents only a fraction of the 'known' existing chemical compounds: for instance, more than 88 million chemical substances were registered in the Chemical Abstract Service (CAS) database of the American Chemical Society as of May 2014 [3]. Only a small fraction of the properties of these compounds has been investigated by experimental means. The determination of the physico-chemical properties for such a large set of compounds is unpractical or even impossible without standardized and automatic protocols. When the means to characterize the properties of the chemical compounds were defined in the elaboration of the EU-REACH protocol, this was well-understood and carefully balanced. Hence, all properties do not need to be determined anew when they have already been obtained by well-established and documented protocols. When they are not known, they can be obtained experimentally, but also computationally. It is this last point that is of main interest to us here, and which served as basis for the research project PREDIMOL (*PREDIction des propriétés physico-chimiques des produits par modélisation MOLéculaire*) [4]. In fact, most of the software developments and results described here originate from this project. Several objectives were defined in the PREDIMOL project. First of all, we wanted to establish the minimum input information needed to compute the desired physico-chemical properties. Ideally only the chemical structure and formulation in case of mixture should be necessary. The second important set of tasks, which was distributed among the different partners of the PREDIMOL project, was to determine the error associated with the evaluation of the properties with respect to experimental reference data using different computational methods. This information also provides valuable guidelines for the range of applicability of various computational methods. To be as complete as possible, this second task included both a comprehensive literature review and an applicative program to extend the scope of the evaluations done in the literature. We describe some of the key results obtained in this applicative program here. In addition, methodological developments were also performed to extend the domain of applicability of Quantitative Structure-Property Relationship (QSPR) methods [5, 6] and atomistic simulations [7, 8] to organic peroxide and amine classes of molecules. Finally, an essential point, particularly within the context of the EU-REACH protocol, was to establish the validity of the simulation methods and protocols according to the regulatory criteria [9]. For instance, the validation of Quantitative Structure Activity Relationship (QSAR) methods [10] are recommended to agree with the five principles of the Organization for Economic Co-operation and Development (OECD) [11], namely:

- a defined end-point,
- an unambiguous algorithm,
- a defined domain of applicability,
- appropriate measure of goodness-of-fit, robustness, and predictive power,
- a mechanistic interpretation whenever possible.

As stated earlier, one of our contributions to the PREDIMOL research project was to evaluate the accuracy of particular simulation methods. This was not our main and only contribution. The most important tasks were the design and development of software capabilities to prepare, submit, and analyze very large number of atomistic simulations, typically on the order of 1 000, possibly in a single

batch, and to verify with real case applications the validity and robustness of the tools. When one handles such a large number of atomistic simulations, it becomes nearly impossible to manually prepare the different jobs. Hence, an effort of automation of the 'routine' atomistic simulation preparation procedure was done. The next important contribution was to identify default simulation parameters that result in stable and successful simulations. After one submits a large batch of calculations, it is not acceptable to manually intervene to correct problems that occurred for hundreds of them because of ill-defined input simulation parameters. This task of optimization of the simulation protocols and parameters is therefore very important and should be completely transparent to the end-user once successfully implemented.

In the following, we will first provide an overview of the MedeA® software environment before we explain the software functionalities and features that were added to achieve the objective of high-throughput atomistic simulations as specified in global projects like EU-REACH. To illustrate the capacities of the software and its environment, we describe selected applications of direct relevance for EU-REACH.

1 DATA, MODELS, AND METHODS

1.1 Selection of Molecules and Property Data

For organic molecules Simplified Molecular Input Line Entry Specification (SMILES) formulas were extracted from the DIPPR database [12]. This was done for 880 molecules representing 15 classes of organic molecules (Tab. 1).

The search criteria in the DIPPR database included all molecules of the different classes mentioned in Table 1 containing from 1 to 9 carbon atoms. The liquid densities at 298 K and 1 bar of these molecules and their standard heats of formation were also collected from the DIPPR. In the DIPPR database, the data and properties are not always experimental. They can also be predicted, extrapolated, and interpolated. The liquid densities are obtained from empirical relationships with an accuracy of 1% with respect to experimental data within the range of temperature applicability defined for each molecule. The temperature range of the molecular liquid density is typically bracketed by the melting and boiling temperatures at 1 bar. Out of the 880 molecular systems we computed the density for a subset of 192 systems. We checked that all empirical molecular liquid densities are in the temperature range of applicability for a temperature of 298 K. For the heat of formation of the molecules, we restricted the set to those systems where the experimental data have an accuracy of better than 5%. The experimental frequencies of vibration of the molecules were collected from the National Institute of Science and Technology (NIST) database [13]. The ideal gas heat capacities at constant pressure of the organic molecules were obtained from tabulated empirical values in Poling *et al.* [14]. These empirical values are reported with a precision of 1% with respect to experimental values. The molecular properties obtained from all databases were unambiguously attributed to those from the simulation using their molecular CAS-number identifier.

TABLE 1
Selection of organic molecules

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

The inorganic gas molecules list was taken from Knacke *et al.* [15], where thermodynamic values of crystals and molecules are reported. We identified inorganic gases from this work and built them using the molecular builder in MedeA® [16].

1.2 Computational Methods

The equilibrium structure of all selected molecules was computed using Molecular Orbital PACkage (MOPAC) [17] with the PM7 semi-empirical method [18] as implemented in MedeA®. The geometry optimizations were confirmed by vibrational frequency calculations in the rigid molecule harmonic approximation: molecules are considered to be in their ground state when there is no imaginary frequency of vibration. The absence of imaginary frequencies signifies a local minimum. The corresponding molecular structures are assumed to represent the ground state. It is understood that this is an approximation. Density Functional Theory (DFT) geometry optimization were also performed using Turbomole [19] with the BP86 [20, 21] and B3LYP [20, 22, 23] functionals with the Triple Zeta Valence plus Polarization (TZVP) basis set [24] followed by frequency calculations on all molecules. Again, the quality of the geometry optimization was monitored through the absence of imaginary modes of vibration.

The liquid density calculations were done using the molecular mechanics and molecular dynamics code Large-scale Atomistic/Molecular Massively Parallel Simulator (LAMMPS) [25], as integrated in MedeA®. For all simulations, the Polymer Consistent ForceField-enhanced version (PCFF+) was used [26]. This forcefield is derived from PCFF [27] and is enhanced to reproduce with high precision the properties and geometries of organic molecules in addition to those of the polymers. To compute the densities, cells containing 256 molecules were built. They were equilibrated in the isothermal-isobaric (NPT) ensemble for 1 ns with an integration time step of 0.5 fs. The sampling time to collect the density values was 2 ns. The non-bond interactions were described using Ewald summation [28], van der Waals tail corrections [29], and a distance radius cutoff of 9.5 Å.

1.3 The MedeA® Software Environment

MedeA® is characterized by a three-tier architecture (Fig. 1).

The first level in the architecture is a Graphical User Interface (GUI), which enables the query and retrieval of atomistic structural information from integrated crystallographic databases [30-32], the import from external files in various formats (*e.g.*, Crystalloagraphic Information File (CIF) and Protein Data Bank (PDB)), or the direct construction of periodic or molecular atomistic models, the definition of atomistic simulation protocols (flowcharts) *via* graphical modular construction, and finally the graphical analysis of the simulation results. It is supported for different computer Operating Systems (OS), namely, Windows and Linux OS.

After a simulation flowchart is created *via* the GUI, it is transmitted to the second level of MedeA®, namely the MedeA-JobServer. The purpose of the JobServer is to control the processing of the simulation protocols, which may consist of a number of computational tasks. The JobServer runs independently of the MedeA GUI. It can be connected and disconnected from the MedeA GUI and it can be accessed by a number of different users. The JobServer allows to monitor and control the simulations during the task processing and to analyze and collect the results after task completion. Very conveniently, it is also an efficient way to keep record and easy access to all simulations that were performed earlier. The JobServer automatically processes the jobs that are submitted by the user(s): it verifies the availability of the computer resources and submits the computer tasks accordingly. The JobServer can be installed locally and/or remotely and multiple JobServers can exist (associated with a given computer resource and duplicated on a same computer resource) as desired and configured by the user(s). The user interface of the JobServer(s) is accessed *via* any web browser interface (including tablets and cell phones) through its HyperText Transfer Protocol (http) address, optionally also in secure (https) mode.

The third level in the MedeA® architecture consists of the TaskServers. This level encapsulates all numerically intensive simulations and interacts directly with the compute-servers. The TaskServer and the associated executables of the simulation programs are based on a light-weight software architecture so that they can be readily ported to high-performance computing platforms. The TaskServers receive their instructions from the JobServer and transmit the results back to the JobServer, which stores the relevant information. The users have access to all these data *via* the JobServer interface or *via* direct login in the computers themselves. Direct access to the simulation files is generally unnecessary because the data post-treatment and summary of the main simulation results can be accessed *via* MedeA®'s GUI and the JobsServer's web interface. A range of atomistic simulation packages are integrated in the MedeA®'s simulation environment, namely, Vienna Ab initio Simulation Package (VASP) [33-35], LAMMPS [25], Gibbs [36], and MOPAC [17].

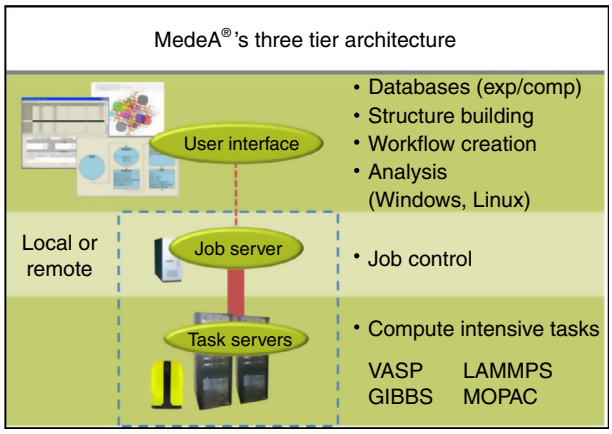

Figure 1

The three tier architecture of the MedeA® atomistic simulation software environment.

1.4 Automation of the Preparation, Processing, and Analysis of Simulations

The main additions to the MedeA® environment to permit preparing and submitting large number of atomistic simulations are first the possibility of defining and regrouping list of atomistic structures. A large versatility and flexibility is given to the user as to which structures can be added to this list. The atomistic structures can be imported from different atomistic file formats, from SMILES formulas [37], from the MedeA® canvas (and hence from the crystallographic databases that can be accessed from MedeA®). Additionally, the list of structures can also be generated directly from a simulation itself and used as soon as created inside the sim-

ulation protocol. All structure lists can be exported to other file formats if so desired by a user.

After the list of structures is created, it can be parsed in a flowchart when the user adds a module that will iterate over each structure. The operations that are performed on the structures are defined in a sub-flowchart that is equivalent to a flowchart that is defined for a single structure. The module looping over structure list is fully integrated in the MedeA®'s flowchart paradigm and can be used in connection with other task modules, in particular those that manipulate the structures *via* translation of atoms, supercell building, amorphous phase building, random atomistic substitution and atomistic simulation. From a given list of structures, this allows constructing a much larger number of structures that include defects, etc. In addition to the module looping over structure lists, other 'loop' modules exist and permit looping over simulation variables and parameters such as temperature and pressure, leading to complex and powerful simulation flowcharts that may include nested loops.

Finally, modules to create, define, and print values in user-defined table(s) are introduced to permit the user to collect and report only specific data among all available properties from all atomistic simulation calculations.

The necessary input information, graphically defined atomistic simulation task, and output values gathered in a table are summarized in Figure 2. In the example that is shown in Figure 2, the definition of the molecules is done *via* their names and SMILES formulas. The atomic coordinates for all molecules are automatically generated and regrouped in a file before they are used as input data in semi-empirical geometry optimizations. At the end of the simulations, we decided here to print in a table the name, chemical formula, total electronic energy, and dipole moment of each molecule contained in the list.

In the next sections, we will now focus on specific applications to better evaluate the accuracy of the different methods available in MedeA®.

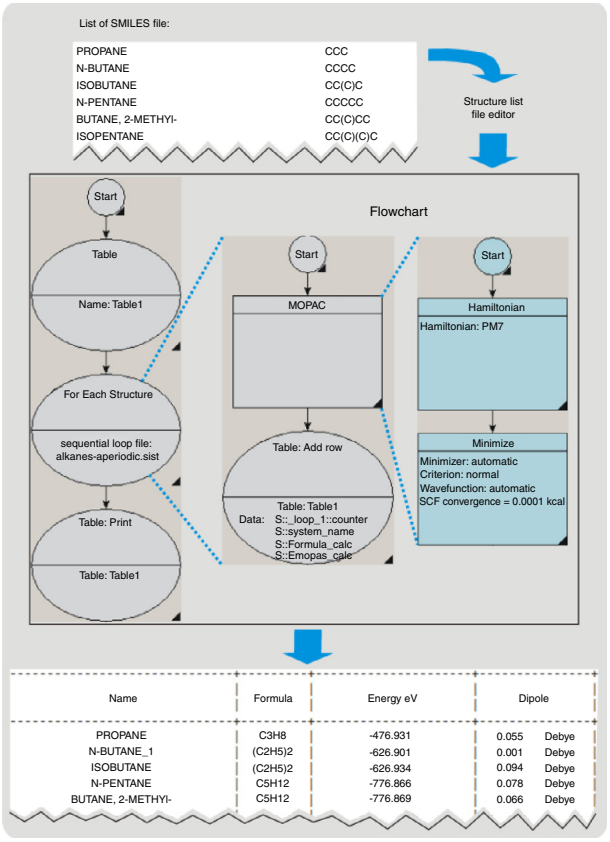

Figure 2

Example of input and output data with the corresponding and MedeA®-Flowchart from.

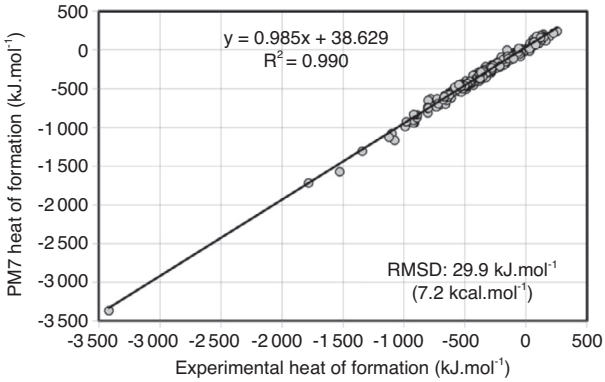

Figure 3

Comparison of the heat of formation of 428 organic molecules.

2 APPLICATIONS

2.1 Thermodynamic Properties of Organic Molecules

Recently, we performed a full systematic evaluation and comparison of semi-empirical PM7 method [38]. Here, we give selected results that were mentioned in this study, namely, the comparison of the frequencies of vibration and the ideal heat capacities, while we also report values that are new, in particular the comparison between the experimental and computed molecules heats of formation.

From our initial set of 880 molecules, we can unambiguously identify 428 experimental values for the heat of formation with accuracy better than 5% [12]. The comparison between the computed and experimental heats of formation is shown in Figure 3.

A linear fit of the values is an appropriate mathematical model as indicated by the regression coefficient R^2 in Figure 3. To better estimate the quality of the computed data,

we computed the dispersion of differences between the experimental and computed values with respect to the average difference. The Root-Mean-Square Deviation (RMSD) is 29.9 kJ.mol^{-1}. This value is in line with the more extensive comparison previously reported by Stewart [18], who reported on Averaged Unsigned Error (AUE) of 17 kJ mol^{-1} for a set of 1 366 organic compounds. Because the semi-empirical PM7 method parameters are fitted to reproduce the experimental heat of formation, the computed values are of better quality than most of other quantum chemical methods [18].

Several other thermochemical properties of the compounds, like the entropy, enthalpy, Gibbs free energy, and heat capacity, are obtained from the equations of statistical mechanics using the geometries and frequencies of vibration obtained from *ab initio* calculations [39]. Very demanding and precise quantum chemical methods, *e.g.* Coupled Cluster with the full explicit treatment of Singles and Doubles contributions and estimate of the Triples contributions *via* perturbation theory in the Complete Basis Set limit extrapolation (CCSD(T)/CBS) [40], can provide reference data that have precision equivalent to the experimental values. However, such methods are rather impractical for use on a routine basis because of their extremely high demand for computing time and memory. Alternatively, density functional theory methods [41] can be used although with a lower precision and computational times that are one or more orders of magnitude faster than CCSD(T) methods and a time scaling with respect to the size of the molecules (*i.e.* the number of electrons) that is much more favorable: CCSD(T) and DFT scale as N^7 and N^{1-3}, respectively, where N is the number of electrons in the molecule. Semi-empirical methods are even more interesting for intensive high-throughput calculations because they are typically two orders of magnitude faster than DFT methods [42] and a Central Processing Unit (CPU) time that scales like $N^{1.7}$. In the case of PM7 method, we mentioned that values for the molecular heat of formation are obtained with an accuracy relative to experimental data that is equivalent or even better than that obtained with DFT methods. This is due to the fact that PM7 is fitted to experimental data. Now the question is if this accuracy of PM7 is achieved also for other thermochemical properties. This aspect is addressed in the following.

Within our set of 880 organic molecules, we identified a subset of 52 molecules for which the experimental frequencies of vibration can be unambiguously assigned [13]. For these molecules, we collected the frequencies of vibration obtained from PM7 calculations, exported the geometries to Cartesian coordinates format, and further optimized the geometries with B3LYP/TZVP and BP86/TZVP DFT methods before computing the frequencies of vibration. All results are reported in Figure 4.

For vibrational frequencies below 2 500 cm^{-1}, all methods give results that are equivalent with respect to the experimental values. Above 2 500 cm^{-1}, good agreement remains for the DFT methods while PM7 results in bond force constants that are too weak, in particular the bonds where H atom is present, which gives an underestimation of the stretching frequencies of vibration of X-H, where X is any element. The underestimation can be corrected *via* linear rescaling of the frequencies of vibration above 2 500 cm^{-1}. We extended the frequency calculations to our entire set of organic molecules. We removed several molecules from this set because a one-to-one frequency assignment could not be done directly due to a different treatment of the symmetry operations in the DFT and semi-empirical codes. We rescaled the PM7 frequencies of vibration based on the values in Figure 4, and report the frequencies of vibration obtained with B3LYP/TZVP and PM7 as a function of those of BP86/TZVP in Figure 5. The semi-empirical and B3LYP values are close to those of BP86 as indicated by the linear fit parameters. The dispersion of the PM7 is clearly larger than that of B3LYP with respect to BP86.

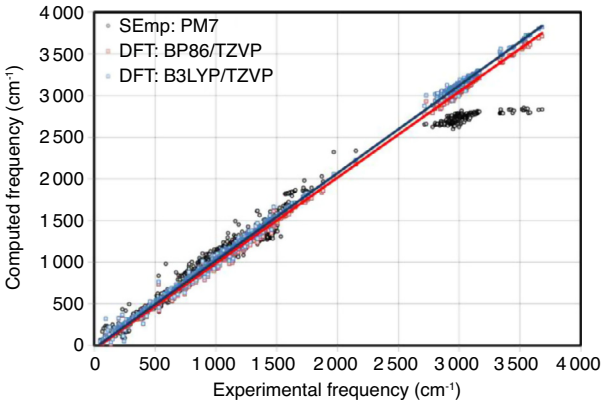

Figure 4

Comparison of the computed and experimental frequencies of vibrations for 52 organic molecules.

To better estimate the effect of the larger dispersion of the semi-empirical values, we compute from the frequencies of vibration the ideal gas heat capacity at constant pressure (C_p) as a function of temperature: this quantity is obtained directly solely from the geometry of the molecules as obtained from the calculations and the computed frequencies of vibration. We collected experimentally fitted data for 160 molecules that belong to our set of 880 molecules [14]. The average deviation and RMSD as a function of temperature between the computed and the experimentally fitted values are reported in Figure 6. There is essentially no difference between the Average Relative Deviations (ARD) of the computed C_p with respect to the experimental values for temperature between 300 and 1 000 K. The RMSD of PM7 is also close to that of BP86: over a large

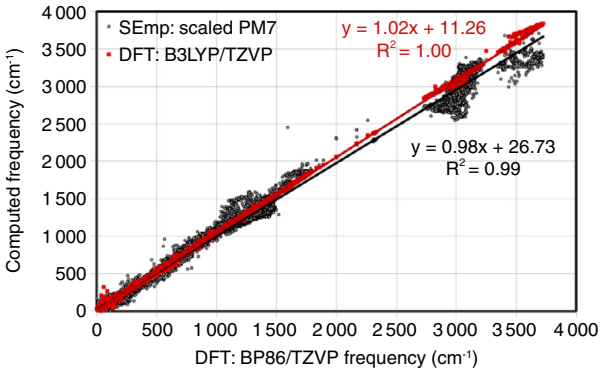

Figure 5

Comparison of the computed (B3LYP/TZVP in blue and PM7 in black) vibrational frequencies with respect to BP86/TZVP values of 795 organic molecules (37 113 frequencies) (adapted with permission from DOI: 10.1021/je500201y [38]. Copyright 2014, American Chemical Society).

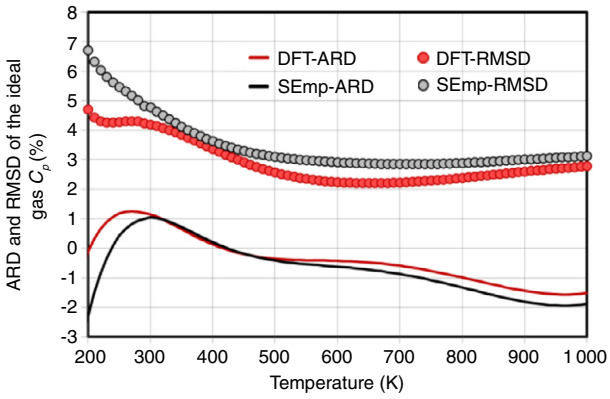

Figure 6

ARD and RMSD between the experimental and computed (BP86/TZVP in black and PM7 in blue) ideal gas constant pressure heat capacity as a function of temperature of 160 organic molecules.

range of temperature they are smaller than 4% and become as small as 2 and 3% at 600 K for BP86 and PM7, respectively. It is reported in the literature [41, 42] that it is not possible to reduce the accuracy of DFT-computed C_p values below 1% without changing the approximations in the statistical thermodynamic equations, and in particular going beyond the approximation of rigid molecules with harmonic vibrations. Below 300 K, the situation differs: relative errors increase up to 5 ~ 7%. For a large part, this relative increase stems from the fact that the C_p values are themselves much smaller at low temperature than at high temperature, which mathematically increases the relative errors.

The accuracy of PM7 in predicting thermochemical properties compared to experimental data is particularly satisfactory. This is not *a priori* granted because PM7 parameterization relies mainly on reproducing with very good precision the energies of the molecules [18] and as we observed, shows weaknesses in reproducing with good precision the experimental frequencies of vibration above 2 500 cm^{-1}.

PM7 parameterization has been found to reproduce with good precision the properties of organic molecules [18]. Data for inorganic molecules have also been used to parameterize PM7 but to a much lesser extent than the organic molecules for which reference experimental data are numerous. To verify the accuracy of PM7 for inorganic compounds, we identified from Knacke *et al.* [15] a set of 515 compounds that exist as gas phase molecules: the small sizes of the molecules allow using both DFT and semi-empirical methods to compute and compare properties for this large set of molecules. The ideal gas heat capacity RMSD are reported in Figure 7 for a temperature of 298 K. To illustrate the gas molecules that are typically found, we give also the list of 18 molecules containing Al (Fig. 7, top). On the full set of 515 molecules, the RMSD on the relative differences between PM7 and BP86 C_p values is 4.8% at 298 K. At the same temperature, the RMSD is 3.5% for the set of 795 organic molecules.

Although the statistical analysis is performed on more limited populations in Figure 7, we can notice that the RMSD are clearly not uniform: for several elements, especially Mn and Co, but also Li, the RMSD are significantly larger than 15%. Overall, the inorganic compounds show RMSD that are 1 to 3 times larger than those obtained for organic molecules. This is indeed a rather good result given complexity of the electronic structure of metal-containing compounds. One could have feared unphysical results by falling outside the applicability domain of the method. This failure is not seen and instead PM7 is revealed to be a robust semi-empirical method that covers with reasonable accuracy most of the periodic table.

In the first application, namely the calculation of thermochemical properties of gas phase molecules, we illustrated the fact that we can compute properties over lists of numerous molecules (1 395 organic and inorganic molecules). In the next application we illustrate the capacity of MedeA®-Flowcharts to perform and organize computations of elaborate simulation protocols with the determination of the fluid density of a set of organic pure compounds.

2.2 Fluid Density

The automated calculation of the density of molecular fluids is demonstrated here for 192 molecules. This was

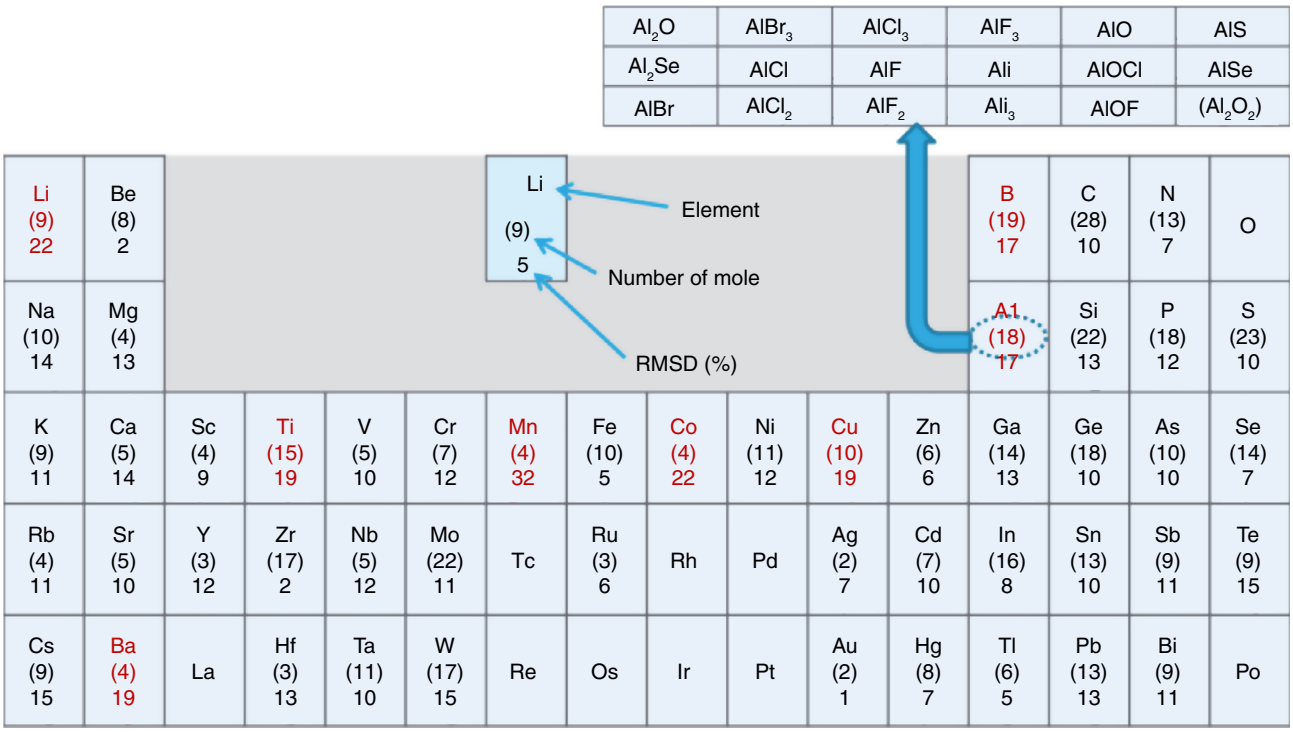

Figure 7

RMSD of the ARD between the PM7 and BP86/TZVP computed ideal gas heat capacities at $T = 298$ K for inorganic gases.

accomplished as follows. The starting point was a set of molecules defined by a file containing a one-dimensional representation of each molecule in the SMILES format. MedeA® converted this one-dimensional information into three-dimensional molecular structures under periodic boundary conditions. The initial cell parameters were handled automatically. The conversion of the SMILES formula list by MedeA®'s structure list editor resulted in a concatenated list of periodic cells with each one molecule inside. To perform the atomistic simulation to determine the density of a fluid, a cell containing at least 1 000 atoms should be built in order:

– to get reasonable sampling over temperature and configurations;
– to avoid most of the periodic artefacts due to the finite size of the cell [43].

Additionally, pre-conditioning and equilibration of the periodic cells containing the molecules are needed before sufficient sampling of the systems is achieved. The duration of this sampling is of course dependent on the characteristic time of the properties that is being considered in the simulation [43]. Most of the preparation tasks before the sampling takes place are actually repetitive, although it is important to keep control over them as they can be different depending on the nature and composition of the systems, *e.g.*, if the system is gaseous, liquid, or solid, if the systems is composed of small, medium, large molecules, or polymers (cross-linked or not). The definition of the tasks as done in the MedeA's software environment *via* graphical flowchart is particularly well-suited here:

– because the modular construction of the simulation permits easily to standardize protocols;
– because the flowcharts are electronic files that can be saved locally or collected from the GUI from the JobServer after the jobs ran by any of the users accessing the JobServer.

This infrastructure constitutes an efficient way to conserve the knowledge and expertise of past campaigns of simulations (Fig. 1). A full flowchart for computing the liquid density of organic compound is shown in Figure 8. This figure is obtained from actual snapshots from MedeA's GUI.

Similarly to density determination, analogous flowcharts can be constructed to compute other properties, *e.g.* viscosity [44] or self-diffusion coefficients. In the flowchart depicted in Figure 8, the forcefield assignment, the construction of the cell containing hundreds of molecules, initialization, equilibration, and data sampling before reporting the properties in a table, are all the tasks that are systematically and automatically performed on all structures contained in the list of molecules initially defined by the users. This protocol

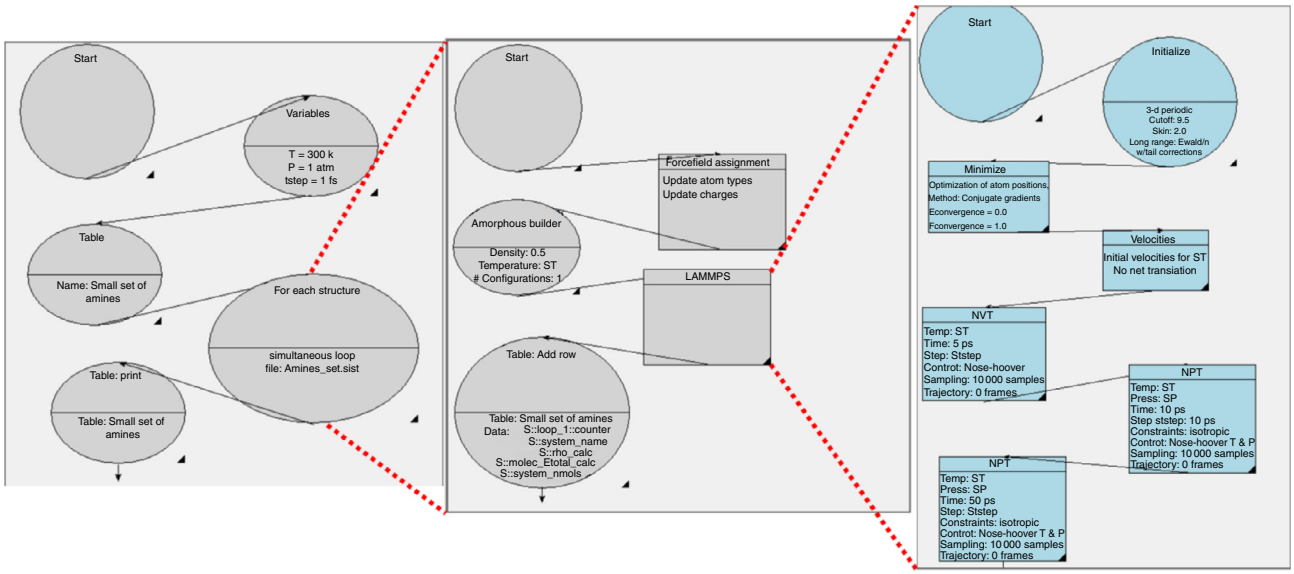

Figure 8

MedeA®'s Flowchart used to compute density of organic compounds in liquid state.

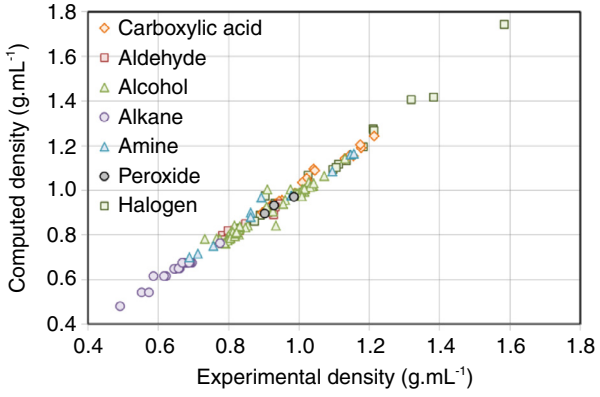

Figure 9

Computed densities of molecular fluids *vs* experimental values for a set of 192 organic molecules using the PCFF + forcefield.

TABLE 2

Statistical analysis of the computed *vs* experimental densities of organic fluids at 298 K and 1 atm

Org. molc. type	Pop.	Average Deviation ($g.mL^{-1}$)	RMSD ($g.mL^{-1}$)	ARD (%)	RMSD (%)
Acid, carb.	29	−0.010	0.015	−1.0	1.4
Aldehyde	41	0.001	0.010	0.1	1.1
Alcohol	71	0.009	0.021	0.9	2.4
Alkane	20	0.005	0.012	0.8	2.0
Amine	10	−0.015	0.024	−1.7	2.6
Peroxide	3	0.004	0.008	0.5	0.8
Halogen.	18	−0.028	0.044	−2.2	3.4
All	192	−0.001	0.024	0.0	2.4

is slightly different, in particular with respect to simulation times mentioned in Figure 8 compared to those that we used (Sect. 1.2).

The computed densities of a set of 192 molecules as a function of the experimental densities values are illustrated in Figure 9, and the statistical analysis of these data is given in Table 2. As seen in Figure 9, we cover a rather large range of liquid densities between 0.4 to 1.8 $g.mL^{-1}$ with different classes of organic molecules. The accuracy of these simulations on 192 demonstrates convincingly the generality and robustness of this approach. As much as possible we tried with the 192 molecules to sample representative molecules in the initial set of 880 molecules. In total, computing all the densities represents a simulation time of 576 ns. LAMMPS is a molecular mechanic simulation software that is particularly well-scalable as a function of the computational resources, and the density calculations can be easily distributed over numerous processors. Benchmarks concerning the scalability of LAMMPS as a function of the platform are detailed at [45].

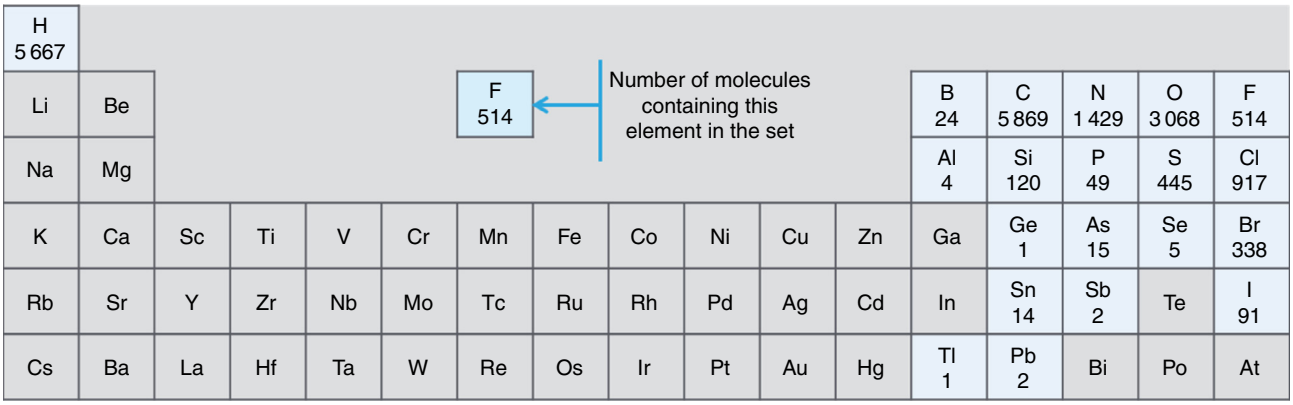

Figure 10

Chemical composition of the 5 869 organic molecules used in the present work.

The details of the statistical analysis in Table 2 reveal that the different classes of organic molecules are equally described with the PCFF+ forcefield: the average deviations and RMSD are uniform and equivalent. Overall, on the full set of molecules, the computed density is obtained with an RMSD with respect to experimental data of 0.024 g.mL^{-1} (2.4%).

2.3 High-Throughput Calculations and Beyond

In the applications discussed above we have computed the properties of organic and inorganic compounds using atomistic simulation methods. We showed that the preparation of the simulations is made as easy and simple as possible: creating molecular models and setting up the simulation parameters can be largely automated. Furthermore, using standardized flowcharts *via* insertion of task modules enables the construction of simple as well as elaborate simulation protocols. These, once created, can be re-used and applied to other chemical compounds, which leads to a standardization of the property calculations. We have determined the accuracy of the numerical models employed here. To some extent, it is possible to go beyond the level of approximation of the current numerical model to increase further the accuracy of the computer properties with respect to experimental reference data. The trade-off for improved accuracy is an increase in computational time. To lower the wall-clock time to results for large datasets, one can use Highly-Parallel Computers (HPC) that can currently perform up to 10^{15} FLoating point Operations Per Second (petaFLOPS) while the exaFLOPS limit is estimated to be crossed in about 10 years from now. In the present work, we have successfully computed properties of up to 880 molecules. This raises the question if the current software is able to handle efficiently and with good stability an order of magnitude more molecules? This is a critical question since there is a dire shortage of experimental physico-chemical data. For instance, Fink and Reymond [46] demonstrated that a set of 26.4 million chemically possible molecules (110.9 million stereoisomers) can be generated from a pool of up to 11 atoms of C, N, O, and/or F. The possibility of formation of the molecules included simple valency, chemical stability, and synthetic feasibility criteria. They analyzed that only 63 857 compounds of up to 11 atoms can be found in public databases. Systematic simulations on large datasets, *e.g.* on molecules with all possible isomeric forms and ranges of combinatorial substitutions, constitutes a most valuable source of comprehensive and consistent molecular and materials property data.

In this context, we performed atomistic simulations on 5 869 molecules. As described above, the test started from SMILES formulas that were collected from accessible lists [47, 48]. We restricted this list to molecules that possess a CAS name and number. The composition of the molecules in this list is indicated in Figure 10. All molecules are carbon-atom containing molecules, such as alkanes, olefins, aromatics, halogenated hydrocarbons, oxygenated hydrocarbons, thiols, sulfides, as well as hydrocarbons with less common elements, namely, B, Al, Si, P, Ge, As, Se, Sn, Sb, Tl, and Pb.

The next steps consisted in building a MedeA® structure list from the raw SMILES formula list and performed an automatic assignment of PCFF+ forcefield parameters to each molecule. As expected such an automatic forcefield assignment encounters some limitations since not all combinations of the different bond, angles, and dihedral angles as found in a 'blind' list of 5 869 molecules are parameterized. In the present case, for 99.15% of this arbitrarily chosen test

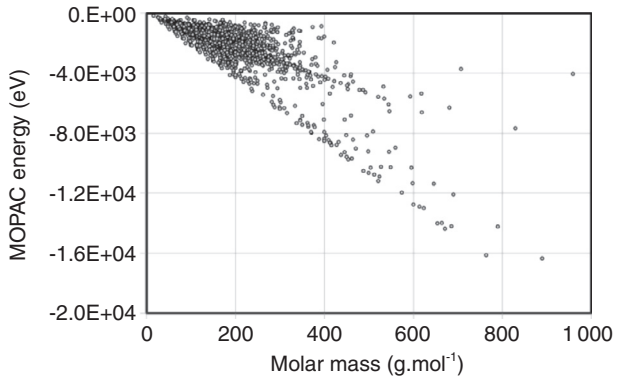

Figure 11

Correlation between molecular mass and electronic energy (PM7) for 5 869 organic molecules.

set the automatic assignment succeeded. The unassigned atom types occurred in molecules containing B, As, and Se plus 6 other multifunctional molecules (although without heteroatoms). This illustrates that the broad reach of PCFF+ but it also points out the need to continually expand this type general high-quality forcefields. However, this limitation inherent in valence forcefields is not present in quantum mechanical methods, as illustrated in the following.

We took the same list of molecules as above and carried out semi-empirical calculations on all of the molecules in the list including geometry optimization and frequency calculations. We printed all names, energies, and some other properties, and plotted the energies as a function of the molar masses in Figure 11. For all molecules but 7, the geometry optimizations and frequency calculations were processed successfully. The reason for the 7 unsuccessful optimizations was related to difficulty to localize the potential energy minimum, and could be resolved manually by changing parameters in the minimization method: they were not optimized using the default optimizer in MOPAC but a more demanding one, namely, with offering possibility in the minimization to increase energy and using a gradient-follower minimization follower with explicit calculation of the gradient each three geometry optimization cycles. In other words, a more conservative choice of computational parameters in this quantum chemical method led to a remarkable success rate.

The full list of molecules was processed in about 24 hours when the tasks are submitted simultaneously by batch of two using one CPU each. It was unnecessary to use more than one CPU to optimize the geometries and compute the frequencies of vibration with MOPAC of a molecule. This would be, however, different for other atomistic simulations. In the determination of the density using LAMMPS, the simulations were submitted by batches of 6 molecules at once each using 8 processors: other combinations are possible and depend on the available computer resources, and the intrinsic efficiency of the atomistic simulation software with respect to the size and nature of the systems to be simulated and hardware resources themselves [45].

CONCLUSIONS

We have presented the atomistic and molecular simulation software environment MedeA® and the new enhancements that have been incorporated to address the case of large number of simulations as will be needed for the systematic determination of molecular and materials properties, for instance in fulfillment of the requirements of EU-REACH. The new capabilities of MedeA® make the determination of properties for lists of several thousand molecules readily doable using atomistic simulation methods. The approach can be extended to larger datasets due to its inherent parallel characteristics and robustness, thus setting the stage for fully exploiting the growth in computer power. We have shown that the preparation of the simulations can be made very straightforward and simple. The tasks of creating molecular models and setting up the simulation parameters are largely automated. Using standardized flowcharts *via* insertion of task modules permits to construct simple as well as elaborate simulation protocols. These, once created, can be re-used whenever needed by chemical engineers.

Our objective was to demonstrate that of the order of 1000 simulations can be easily prepared, submitted, and their data analyzed. This was illustrated with different applications, namely:

- the calculation of thermochemical properties of about 1 400 organic and inorganic molecules and comparison with experimental and computational values to establish the validity and accuracy of the simulation method;
- the determination of the fluid density for a set of nearly 200 organic compounds;
- the determination of thermochemical properties of a set of nearly 6 000 organic molecules

The present results demonstrate that we can compute the heat of formation with an accuracy of about 5%, ideal gas heat capacity with an accuracy 3.5% for organic and 4.8% for inorganic compounds at 298 K. These values are obtained with fast (compared to DFT) semi-empirical method. We also analyzed with great detail the evolution of the thermochemical values as a function of temperature, in particular the organic molecule ideal gas heat capacities. The volumetric masses of 192 organic fluids were also investigated and validated with an accuracy of 2.4% with respect to experimental data. In summary, the MedeA® software environment with the recent enhancements set the stage

for the large-scale deployment of quantum mechanical and forcefield-based simulations as an extremely efficient source of molecular and materials properties.

ACKNOWLEDGMENTS

This work was carried out within the project PREDIMOL funded by the French Research Agency under contract ANR-2010-CD2I-007-05 in partnership with *ARKEMA, IFP Energies nouvelles, INERIS, ENSCP ParisTech, University Paris Sud 11*, and *CNRS* to provide software solutions to industry within the context of the EU-REACH protocol.

REFERENCES

1 Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH).

2 European Chemicals Agency (http://echa.europa.eu). For the list of pre-registered substances: http://echa.europa.eu/en/information-on-chemicals/pre-registered-substances.

3 http://www.cas.org/content/chemical-substances.

4 http://www.ineris.fr/predimol/.

5 Prana V., Fayet G., Rotureau P., Adamo C. (2012) Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds, *J. Hazard. Mater.* **235-236**, 169-177.

6 Prana V. (2013) Approches structure-propriété pour la prédiction des propriétés physico-chimiques des substances chimiques, *Doctoral dissertation*, Paris 6.

7 Orozco G.A., Lachet V., Nieto-Draghi C., Mackie A.D. (2013) A transferable force field for primary, secondary, and tertiary alcanolamines, *J. Chem. Theo. Comput.* **9**, 2097-2103.

8 Rotureau P., Fayet G., Prana V., Nieto-Draghi C., Adamo C., Rousseau B., Leblanc B., Rozanska X., Ungerer P., André D. (2012) Prediction of physico-chemical properties in the context of the French PREDIMOL project, in *15th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR 2012)*, Tallinn, Estonie, 18 June.

9 Guidance on information requirements and chemical safety assessment Chapter R.7a: Endpoint specific guidance. These guidance documents can be obtained *via* the website of the European Chemicals Agency Agency (http://echa.europa.eu/about/reach_en.asp).

10 Fayet G., Rotureau P., Adamo C. (2013) On the development of QSPR models for regulatory frameworks: The heat of decomposition of nitroaromatics as a test case, *J. Loss Prevention Process Ind.* **26**, 6, 1100-1105.

11 Organisation for Economic Co-operation and Development (2007) Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models, http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2.

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

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

14 Poling B.E., Prausnitz J.M., O'Connell J.P. (eds) (2007) The properties of gases and liquids, fifth International Edition, McGraw-Hill, Boston, pp. A.35-A.46.

15 Knacke O., Kubaschewski O., Hesselmann K. (1991) *Thermodynamic properties of inorganic substances*, Springer-Verlag, Berlin.

16 MedeA®: Materials Exploration and Design Analysis Copyright © 1998-2014 Materials Design, Inc. Version 2.14.6. http://www.materialsdesign.com.

17 Stewart J.J.P. (2012) *MOPAC2012*, *Stewart Computational Chemistry*, Colorado Springs, CO, USA, http://OpenMOPAC.net.

18 Stewart J.J.P. (2013) Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, *J. Mol. Model.* **19**, 1, 1-32.

19 Ahlrichs R., Bär M., Häser M., Horn H., Kölmel C. (1989) Electronic structure calculations on workstation computers: The program system turbomole, *Chem. Phys. Lett.* **162**, 3, 165-169.

20 Becke A.D. (1988) Density-functional exchange-energy approximation with correct asymptotic behavior, *Phys. Rev. A* **38**, 6, 3098.

21 Perdew J.P. (1986) Density-functional approximation for the correlation energy of the inhomogeneous electron gas, *Phys. Rev. B* **33**, 12, 8822.

22 Lee C., Yang W., Parr R.G. (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B* **37**, 2, 785.

23 Becke A.D. (1993) Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* **98**, 7, 5648-5652.

24 Schäfer A., Huber C., Ahlrichs R. (1994) Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr, *J. Chem. Phys.* **100**, 8, 5829-5835.

25 Plimpton S. (1995) Fast parallel algorithms for short-range molecular dynamics, *J. Comput. Phys.* **117**, 1, 1-19.

26 Sun H., Mumby S.J., Maple J.R., Hagler A.T. (1994) An *ab initio* CFF93 all-atom force field for polycarbonates, *J. Am. Chem. Soc.* **116**, 7, 2978-2987.

27 PCFF is distributed along with Materials Design© software package MedeA®.

28 Toukmaji A.Y., Board J.A. Jr (1996) Ewald summation techniques in perspective: a survey, *Comput. Phys. Comm.* **95**, 2, 73-92.

29 Frenkel D., Smit B. (2001) *Understanding molecular simulation: from algorithms to applications*, Vol. **1**, Academic press.

30 Inorganic Crystal Structure Database, FIZ Karlsruhe – Leibneiz-Institut fuer Informationsinfrastruktur GmbH, http://www.fiz-karlsruhe.de/icsd_home.html.

31 Mighell A.D., Karen V.K. (1996) NIST Crystallographic Databases for Research and Analysis, *J. Res.-Nat. Inst. Std. Tech.* **101**, 273-280.

32 Villars P., Cenzual K. (2009) Pearson's Crystal Data, *Crystal Structure Database for Inorganic Compounds, Release 2012, 13*.

33 Kresse G., Hafner J. (1993) *Ab initio* molecular dynamics for liquid metals, *Phys. Rev. B* **47**, 1, 558.

34 Kresse G., Furthmüller J. (1996) Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* **6**, 1, 15-50.

35 Kresse G., Furthmüller J. (1996) Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, *Phys. Rev. B* **54**, 16, 11169.

36 MedeA-Gibbs: Gibbs Licence IFPEN-CNRS-Université Paris-Sud

37 Weininger D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* **28**, 1, 31-36.

38 Rozanska X., Stewart J.P.P., Ungerer P., Leblanc B., Freeman C., Saxe P., Wimmer E. (2014) High-throughput calculations of molecular properties in the MedeA environment: Accuracy of PM7 in predicting vibrational frequencies, ideal gas entropies, heat capacities, and Gibbs free energies of organic molecules, *J. Chem. Eng. Data*. **59**, 10, 3136-3143.

39 Van Santen R.A., Niemantsverdriet J.W. (1995) (eds) Chemical kinetics and catalysis, Springer.

40 Raghavachari K., Trucks G.W., Pople J.A., Head-Gordon M. (1989) A fifth-order perturbation comparison of electron correlation theories, *Chem. Phys. Lett.* **157**, 6, 479-483.

41 Parr R.G., Yang W. (1989) *Density-functional theory of atoms and molecules*, Vol. **16**, Oxford University Press.

42 Prigogine J., Rice S.A. (eds) (1996) *Advances in chemical physics, new methods in computational quantum mechanics*, John Wiley & Sons, New York, pp. 713-714.

43 Allen M.P., Tildesley D.J., (eds) (1989) *Computer simulation of liquids*, Oxford University Press, pp. 24-29.

44 Hoover W.G., Moran B. (1980) Lennard-Jones triple-point bulk and shear viscosities. Green-Kubo theory, Hamiltonian mechanics, and nonequilibrium molecular dynamics, *Phys. Rev. A* **22**, 4, 1690-1697.

45 LAMMPS benchmarks and scalability efficiency evaluations on different platforms are accessible at http://lammps.sandia.gov/bench.html.

46 Fink T., Reymond J.-L. (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compounds classes, and drug discovery, *J. Chem. Inf. Model.* **47**, 342-353.

47 Training set of the Estimation Program Interface (EPI) Suite, EPA's Office of Pollution Prevention Toxics and Syracuse Research Corporation (SRC), Copyright 2000-2012 U.S. Environmental Protection Agency.

48 Computer-Aided Drug Design Group's Chemoinformatics Tools and User Service. Downloadable structure files of the National Cancer Institute open database compounds. http://cactus.nci.nih.gov/download/nci/.

Manuscript submitted in May 2014
Manuscript accepted in September 2014
Published online in December 2014

Cite this article as: X. Rozanska, P. Ungerer, B. Leblanc, P. Saxe and E. Wimmer (2015). Automatic and Systematic Atomistic Simulations in the MedeA® Software Environment: Application to EU-REACH, *Oil Gas Sci. Technol* **70**, 3, 405-417.