Journal of Molecular Modeling (2025) 31:7
https://doi.org/10.1007/s00894-024-06225-3

ORIGINAL PAPER

On the prediction by density functional theory of entropies in solution within implicit solvation models

Victoria Castor-Villegas[1] · Vincent Tognetti[1] · Laurent Joubert[1]

Received: 31 August 2024 / Accepted: 14 November 2024 / Published online: 4 December 2024

Abstract

Context Entropies are fundamental contributions to Gibbs energies that carry important chemical information, in particular when investigating reaction mechanisms. However, evaluating them in solution is far from being straightforward. In this paper, we focus on its evaluation within the framework of implicit solvation models. To this aim, successive corrections (with increased complexity) involving only contributions available from any standard quantum chemistry code and macroscopic solvent properties are built and assessed by comparison to more than one hundred experimental entropy values measured in a liquid phase. It turns out that significant improvement with respect to the standard ideal gas approximation can be achieved at an almost negligible computational cost, affording a robust and transferable predictive model.

Methods DFT calculations with the ADF software at the PBE or PBE0/TZ2P level of theory with COSMO solvent model. Python scripts for regressions.

Keywords Entropy · Gibbs energy · Solvent phase · Continuum models · Density functional theory

Introduction

Quantum chemistry is often used to determine the possible mechanisms of a chemical reaction, often within the potential energy surface (PES) paradigm. For reactions performed at constant temperature T and pressure (this is the most common case in organic synthesis or in homogeneous catalysis), the relevant thermodynamic state function is the Gibbs energy G defined by

$$G = H - TS \tag{1}$$

where H and S respectively denote the enthalpy and entropy of the studied system. This last one thus plays a crucial role in thermodynamic calculations and in understanding the behavior of chemical systems, in particular when the number of molecules Δn varies along the reaction. Indeed, while the variations of Gibbs energies and enthalpies are quite close to each other for isomerization processes or conformational changes (that correspond to $\Delta n = 0$), chemical additions (inducing negative values for Δn) and eliminations ($\Delta n > 0$) can exhibit consequent entropy effects. In some cases, entropy can even reveal the main driving force or resistance of the reaction and may also induce kinetic selectivity. [1]

In fact, as recalled by [2], "a useful rule of thumb" is that "for any reaction step in which two molecules combine to form a single one, the standard Gibbs energy change at room temperature will be roughly $10\,\mathrm{kcal\,mol^{-1}}$ less favourable (or more unfavourable) than the electronic energy change," a value that should be compared to the $10\,\mathrm{kcal\,mol^{-1}}$ to $30\,\mathrm{kcal\,mol^{-1}}$ values typically encountered for activation barriers. Besides, this effect is even enhanced for multicomponent reactions involving more than two partners, a synthetic approach that is notably popular in drug design. We refer the interested reader to the enlightening didactical paper by [3] for more examples and an insightful analysis of important chemical trends.

In this study, we will focus on reactions occurring in solvents. The still most widespread approach is to approximate the entropy in solution by its gas phase expression (see more details in next section), while it has been recognized for many

✉ Vincent Tognetti
vincent.tognetti@univ-rouen.fr

✉ Laurent Joubert
laurent.joubert@univ-rouen.fr

Victoria Castor-Villegas
victoria.castor-villegas@univ-rouen.fr

[1] Normandy Univ., COBRA UMR 6014 & FR 3038, Université de Rouen, INSA Rouen, CNRS, 1 rue Tesnière, 76821 Mont St Aignan Cedex, France

decades that entropy values in solution and in the gas phase can significantly differ (see for instance the work by [4] based on experimental results). Several correcting schemes have been proposed in the last years (we emphasize that we do not consider to extensively review them in this paper), such as the seminal works by [5, 6] and [7], this last one being the basis for the Ariai and Gellrich's methodology [8]. Of particular significance is the remarkable automated approach by [9] and [10], which is well suited to accurately evaluate entropies of large molecules in solution [11].

It can be noticed that some of these methods are not implemented in standard quantum chemistry software or that they require complex computational pipelines. Crudely speaking, entropy can be estimated either using static or dynamic approaches. For instance, [12] have assessed the use of ab initio molecular dynamics applied to organometallic complexes, while [13] have shown how metadynamics can be efficiently used to evaluate entropies at all temperatures by a single simulation. Although these approaches are both physically grounded and chemically attractive, they can reveal computationally intractable if one wants to rely on a pure quantum chemistry level without resorting to any classical mechanics methods.

Besides, a second important factor that affects the calculation time is how the solvent is described. While explicit approaches in which a given number of solvent molecules are described at the atomic scale allow for a precise description of the interactions between the solute and the solvent, they are considerably more expensive than implicit ones, epitomized by the various flavors of the polarizable continuum model (PCM) [14]. These last ones disregard the atomistic structure of the environment since the molecule is placed into a cavity inside a continuum characterized by its relative permittivity ϵ_r, but the mutual polarization between the solute electron density and the infinite continuum dielectric medium is however taken into account by solving the appropriate Poisson electrostatic equation.

Owing to their simplicity, their versatility (all common solvents have been parameterized), and their high accuracy/time ratio, PCMs are ubiquitous in the modelling of chemical reactions and of physico-chemical properties of molecules in solution. In this contribution, we thus only consider static PCM approaches. Furthermore, still targeting at a minimal computational cost, only properties already computed by standard density functional theory (DFT) codes will be employed, so that this strategy can be straightforwardly implemented from the output of any DFT software.

Henceforth, this paper will be divided into the following four next sections: a theoretical one that presents the main equations to evaluate the various entropy components and entropy corrections, followed by the description of the experimental database on which our several models (each incorporating different mathematical complexities and physical considerations) will be trained and of the computational protocol, before reporting their performances and discussing their meaning and usefulness.

Theory

The mathematical expressions for entropy in the gas phase are based on standard statistical physics formulas (see for instance Ochterski's paper [15]) within the ideal gas (IG) hypothesis. Even if these expressions are standard, we decided to report them here, so that this paper remains self-explanatory. More specifically, the IG entropy (thus incorrect for molecules in solution) of a system can be divided into four contributions: vibrational (S_v), rotational (S_r), translational (S_t), and electronic entropies (S_e) according to the following:

$$S_{IG} = S_v + S_r + S_t + S_e \tag{2}$$

If we assume that the electronic excited states are energetically far in energy from the ground state, the electronic component is simply the following:

$$S_e = R \ln(d) \tag{3}$$

where d denotes the electronic degeneracy equal to the (electronic) spin multiplicity and R the molar gas constant. For non-radical species, this term exactly vanishes, so that we will not consider it any longer. The translational contribution is evaluated from the celebrated Sackur-Tetrode formula:

$$S_t = R(\ln(q_t) + 5/2) \tag{4}$$

where q_t represents the translational thermodynamic partition function

$$q_t = \left(\frac{2\pi m k_B T}{h^2}\right)^{3/2} V \tag{5}$$

and h and k_B are the Planck and Boltzmann's constants, respectively, and V is the volume. Similarly, the rotational entropy reads as follows (model of the quantum rigid rotor):

$$S_r = R(\ln(q_r) + 5/2) \tag{6}$$

where the rotational partition function q_r involves the three moments of inertia J_x, J_y, and J_z (along each Cartesian axis generically denoted α, which can be all three conveniently collected in the $\vec{J}$ vector) of the molecule. More precisely, once the three rotational temperatures are defined by $\Theta_\alpha^r = h^2/(8\pi^2 k_B J_\alpha)$, we have

$$q_r = \frac{\sqrt{\pi}}{\sigma_r}\left(\frac{T^{3/2}}{\sqrt{\Theta_x^r \Theta_y^r \Theta_z^r}}\right) \tag{7}$$

where σ_r is the rotation symmetry number (equal to 1 for non-symmetric molecules).

Finally, for each of the $3N-6$ vibrational modes (for non-linear molecules) of frequency ν_i, we define the characteristic vibrational temperature by the following:

$$\Theta_i^v = h\nu_i/k_B \tag{8}$$

The vibrational entropy can be evaluated by a more intricate, but closed formula, in the case we assumed that pure harmonic vibrations (i.e., the energy levels are those from the quantum oscillator model):

$$S_v^h = \sum_i \left(\frac{\Theta_i^v/T}{e^{\Theta_i^v/T}-1} - \ln\left(1-e^{\Theta_i^v/T}\right)\right) \tag{9}$$

Obviously, the harmonic approximation is sometimes too crude. Full anharmonic calculations (for instance using second-order perturbation theory (VPT2) or vibrational mean-field (VSCF) approaches) are available, but are in general feasible only for systems of small or moderate size. A popular cheap alternative is the simple scaling of harmonic frequencies, but it does not really cure the problems of the harmonic approximation, in particular for the low-lying vibrational frequency modes.

However, as shown by Eq. 9, these modes are those who contribute the most to the vibrational entropy (S_v^h actually diverges for $\nu_i \rightarrow 0$). In 2012, [16] proposed to correct this incorrect behavior (leading to the so-called rigid-rotor-harmonic-oscillator (RRHO) approximation) by building a smooth interpolation (using a damping function w_{HG} proposed by [17]) between a free rigid rotor whose effective inertia moment depends on the vibrational frequency (and is thus not the molecular one) and the harmonic oscillator at a negligible computational cost. More precisely, the entropy for each normal mode is computed by the following:

$$S_v^G(\nu_i) = w_{HG}(\nu_i)S_r(\nu_i) + (1-w_{HG}(\nu_i))S_v^h(\nu_i) \tag{10}$$

which is not diverging anymore for vanishing ν_i values. It should be emphasized that this RRHO approach should not be confounded with the more refined *hindered* rotor approach to model the low-frequency vibrations [18, 19] in which the potential entering the vibrational Schrödinger equation is fitted by a Fourier series (which thus required many points on the PES). Finally, the total vibrational entropy obtained by the simple (but efficient) Grimme's treatment, S_v^G, is obtained by summation on all frequencies, allowing us to define Grimme's correction according to the following:

$$\Delta S_v^G = S_v^G - S_v^h \tag{11}$$

Having recalled the main equation for gas phase entropies, we now come back to our main purpose that is to predict accurate values for entropies in solution. These should not be confused with solvation entropies that are defined by the following:

$$\begin{aligned}\Delta S^{solvation} = & S\left(\vec{J}^{solv}, \left\{\Theta_i^{r,solv}\right\}, ...\right) \\ & -S\left(\vec{J}^{gas}, \left\{\Theta_i^{r,gas}\right\}, ...\right)\end{aligned} \tag{12}$$

In this last equation, we have made explicitly some relevant variables in the two contributions in the right-hand side, which are the entropy in solution and the entropy in the gas phase. As geometries in the gas phase and in solvent often differ (for instance, some structures are more folded in the gas phase), the inertia moments will also have different values. Besides, as solvent and gas phase PESs are not the same, the vibrational temperatures $\Theta_i^{r,solv}$ and $\Theta_i^{r,gas}$ are not equal. So, even if we use the same formulas to evaluate S^{solv} and S^{gas}, values will differ if we use the molecular properties computed specifically in the two phases. This is in a way reminiscent of the fact that, in DFT, the Kohn-Sham exact exchange value differs from the Hartree-Fock exchange value since, even if the very same orbital functional is used, it is applied on different orbitals.

It can be added that using these gas phase formulas with gas phase quantities will only lead to approximate gas phase entropy values due to the various approximations we have already mentioned. However, it is possible that solvation entropies would be not so bad due to spurious error compensation in Eq. 12.

We now come back to our main target: building an accurate model of entropies in solution using the formulas of gas phase entropy applied on quantities evaluated in solution. To this aim, it has been well established that translational and rotational degrees of freedom are reduced in solution [20] (from a physical point of view, the available volume is decreased due to the solvent pressure, so that, in virtue of Eq. 5, S_t decreases).

As a consequence, [21] proposed to fully remove S_t and S_r in their calculations. In a less drastic way, [22] proposed to scale S_t and S_r by a two-third factor, a simple correction that has been regularly used, mainly in the context of organometallic catalysis [23]. From thermodynamic measurements, [24] proposed a very close factor value (0.65), also close to the one (0.60) we used [25] in a recent paper in organic synthesis. All these corrections correspond to the following one-parameter model (in a general way, the number of parameters in our various models will be indicated by an integer value in subscript position):

$$S_1^l = \alpha\,(S_t + S_r) + S_v^h \tag{13}$$

An easy generalization of this linear model (hence the l superscript) implies four parameters:

$$S_4^l = \alpha S_t + \beta S_r + \gamma S_v^h + \delta \Delta S_v^G \tag{14}$$

These two first models can be seen as a first-order Taylor expansion of the entropy in solution using the ideal gas-type entropy components as basic variables. Pushing toward second-order would involve S_t^2, S_r^2, $(S_v^h)^2$ squared terms, as well as $S_t S_r$, $S_t S_v$, and $S_r S_v$ crossed ones. Preliminary tests have shown that the inclusion of S_t^2 and S_r^2 squared terms do not significantly improve the model. As we also seek the most parsimonious models (in terms of parameters) in order to prevent it from too much overfitting, only $S_t S_v$, $S_r S_v$, and $(S_v^h)^2$ (this one becoming dominant when the molecular size increases) will be further considered.

However, as entropy increases with the molecular size, the mixed variables would become largely dominant for extended molecules and will bias the regression process. It is thus preferable to work with some "normalized" quantities, which would have all the same dimension (that of an entropy). Our own non-linear (*nl*, or second-order) extensions of S_1^l, where S_t and S_r are still grouped and where no action is made on S_v, is accordingly as follows:

$$S_2^{nl} = \alpha\,(S_t + S_r) + S_v^h + \beta\left(\frac{S_t S_v^h}{S_t + S_r + S_v^h} + \frac{S_r S_v^h}{S_t + S_r + S_v^h}\right) \tag{15}$$

involving two scaling factors. If we now allow to correct the vibrational part, the natural second-order extension of Eq. 14 includes seven parameters:

$$S_7^{nl} = \alpha S_t + \beta S_r + \gamma S_v^h + \delta \Delta S_v^G + \zeta \frac{S_t S_v^h}{S_t + S_r + S_v^h} + \eta \frac{S_r S_v^h}{S_t + S_r + S_v^h} + \theta \frac{(S_v^h)^2}{S_t + S_r + S_v^h} \tag{16}$$

It should also be noticed that using this normalization of the second-order terms, all corresponding parameters are *unitless*, so that their magnitude can be safely controlled.

A refinement of these schemes is to correct the first-order vibrational entropy using the main solvent properties, namely the relative permittivity ϵ_r and the radius R_{solv} used in implicit solvent model computations (see Computational details below). Still, in the quest for the simplest models, the last (experimental) parameter has been retained by us for the vibrational correction:

$$S_3^{nlv} = \alpha\,(S_t + S_r) + f_\gamma\,(R_{solv})\,S_v^h + \beta\left(\frac{S_t S_v^h}{S_t + S_r + S_v^h} + \frac{S_r S_v^h}{S_t + S_r + S_v^h}\right) \tag{17}$$

where γ is a parameter controlling the vibrational weight, which can be further extended by the following:

$$S_4^{nlv} = \alpha\,(S_t + S_r) + f_\gamma\,(R_{solv})\,S_v^h + \beta\left(\frac{S_t S_v^h}{S_t + S_r + S_v^h} + \frac{S_r S_v^h}{S_t + S_r + S_v^h}\right) + g_\delta\,(\epsilon_r) \tag{18}$$

In these two last equations, f_γ and g_δ are tailored functions whose expressions will be based (and justified) on the results obtained from the previous models (note that we do not include ΔS_v^G, and that S_r and S_r are grouped here for reasons that will be reported in the results section).

The molecular database

Experimental values for entropy in solution are extracted from the National Institute of Standards and Technology (NIST) Chemistry WebBook [26], labelled S_0^{liq} in the database, corresponding to the entropy of the pure liquid in standard thermodynamical conditions. When several values are available for a given chemical system, the most recent one was systematically chosen (unfortunately, the database does not provide error margins). All solvents reported in the online ADF [27] and [28] manuals were considered, excluding solvent mixtures and, obviously, solvents for which no experimental data were available, to which we added selected small molecules still from the NIST website.

In total, we collected 106 molecules of small or moderate size (from 3 to 56 atoms), including all common solvents used in organic synthesis, covering the full type spectrum (polar, apolar, protic, aprotic, with high or low dielectric constant), and the main chemical functions and families (alkanes, alkenes, alkynes, ethers, alcohols, amines, nitriles, carboxylic acids, esters, ketones, nitro and halogenated compounds, thiols...) in both aliphatic and aromatic series. The full list can be found in the supplementary material.

More precisely, Fig. 1 shows the distributions of values for the relative permittivity (left panel) and for the experimental standard entropies (right panel) along the whole dataset. It can be seen that ϵ_r values spread from almost 1 to 80 (water) with an important concentration in the [1,20] range where the distribution is quite uniform. The shape for the entropy distribution strongly differs, close to a Gaussian one with a major peak around 220 J mol^{-1} K^{-1}.

It should be noticed that the considered experimental range is restricted by the fact that, in our approach, the system should be in a pure liquid phase in standard conditions (for instance, for several interesting chemical systems, only entropies in the solid state have been measured). The entropy value distribution is a little bit biased toward small values: we thus conducted a full search in the NIST database for

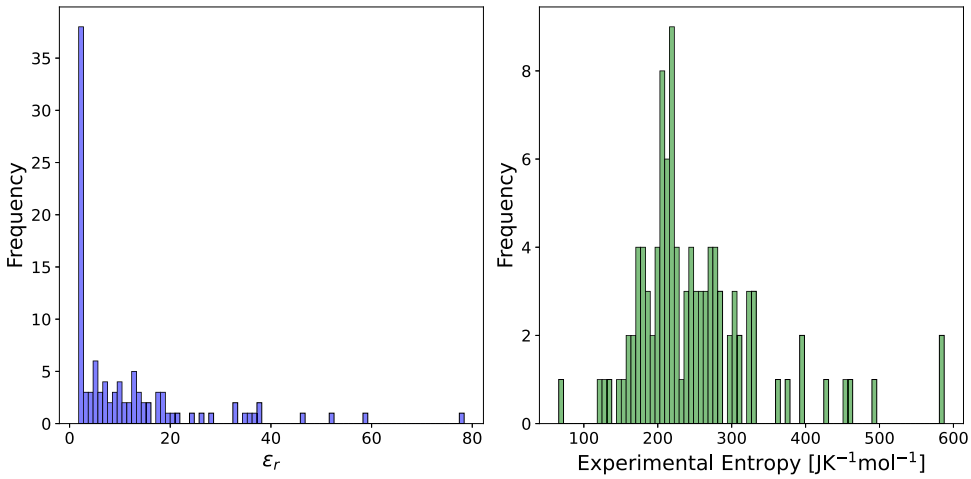

Fig. 1 Distribution of relative permittivity (left panel) and experimental standard entropy (right panel) values across the dataset

all molecules including from 14 to 19 carbon atoms, and we incorporated all those for which full experimental data was available (unfortunately, in some cases where liquid entropies values have been measured, we were not able to find the corresponding ϵ_r values, precluding the use of these points in our dataset).

Computational details

As mentioned in the previous section, experimental values correspond to the pure liquid at 298.15 K. This means that the solute and the solvent correspond to the same molecule. Within all this work, the COSMO implicit model [29], as implemented in the ADF software [30] (with default parameters), was exclusively considered (for instance, the experimental value of entropy for liquid dichloromethane (DCM) will be retrieved by performing a DFT calculation on one DCM molecule immersed inside a DCM continuum). Such a computation requires the knowledge of the relative permittivity and of the solvent radius. This last one was evaluated according to the following:

$$R_{solv} = 0.735 \left(\frac{M|_{\mathrm{g/mol}}}{d_l|_{\mathrm{g/mL}}} \right)^{1/3} \text{Å} \tag{19}$$

where M and d_l represent the molar mass and liquid density, respectively, whose values are taken either from the NIST or the [31] website.

All DFT calculations were performed with the ADF software [30] using either the PBE [32] or PBE0 [33] exchange-correlation functionals with Grimme's BJDAMP dispersion correction [34], this comparison between a pure gradient generalized approximation (GGA) and a hybrid functional allowing for discussing the possible transferability of the optimized models. The TZ2P triple-ζ basis set was used, with default numerical quality and integration grid settings. No relativistic effects were included, and no frozen core approximation was applied.

From a theoretical viewpoint, experimental values actually correspond to an average on all stable conformations. A conformational search was thus carried out for all systems by adapting an available recipe [35] in the Python Library for Automating Molecular Simulation (PLAMS) library, in which for each chemical system specified by its Simplified Molecular Input Line Entry System (SMILES) notation, the computational workflow automatically performs the following sequence of steps within the very same run:

i) The RDKit [36] generator is used to extract a three-dimensional structure from the SMILES code and to generate a set of randomized molecular geometries for extensively exploring the PES, in an almost instantaneous way.

ii) These geometries are then fully optimized at the Density-Functional based Tight-Binding (DFTB) level of theory that is a computationally cheap but reliable semi-empirical method.

iii) An energy threshold of 5.0 kcal mol^{-1} $\approx$ 20.9 kJ mol^{-1} and a structural root mean square deviation (RMSD) threshold equal to 3.0 Å are then applied to filter the most stable representative geometries.

iv) These last ones are fully reoptimized at the chosen DFT level (so that both energies and geometries are refined at this more time-consuming step).

v) A final filtering with threshold of 2.5 kcal mol^{-1} $\approx$ 10.5 kJ mol^{-1} (let us recall that RT is about 0.59 kcal mol^{-1} $\approx$ 2.48 kJ mol^{-1} at room temperature) and 1.5 Å for energy and RMSD are respectively used.

vi) Vibrational frequencies and associated physico-chemical properties are calculated on all of the retained structures (single point calculations). The absence of any imaginary values for frequencies was checked to ensure that genuine energy minima were obtained. Grimme's vibrational correction was computed using the default values as implemented in ADF.

A weight w_i is subsequently associated with each retained conformer according to Boltzmann's distribution, here based on the self-consistent-field (SCF) energy (thermal and entropy effects—which are not yet known since the models have not been built—are in general not important when dealing with conformational preferences, as recalled in the introduction):

$$w_i = Ae^{-\frac{E_i^{SCF}}{RT}} \tag{20}$$

where A is the normalization constant. For any property P, an average value is then simply obtained by (with obvious notations)

$$\overline{P} = \sum_i w_i P_i \tag{21}$$

In machine learning (ML), model parameters are those that minimize the so-called loss function, which is often built using a norm of the difference between the vector $(\overrightarrow{Y}_{ref})$ gathering the reference values and the one $(\overrightarrow{Y}_{pred})$ collecting the predicted values. Popular norms are those belonging to the p-norm family (here for a vector space of dimension n):

$$\left\| \overrightarrow{u} \right\|_p = \left(\sum_{i=1}^{n} u_i^p \right)^{1/p} \tag{22}$$

p=1 corresponds to the so-called taxicab norm associated with the mean absolute error (MAE) that is the most common performance metrics in theoretical chemistry, and p=2 to the Euclidean norm linked to the root mean square error, while p=∞ returns the maximal value. Whereas norms are equivalent in any space vector of finite dimension, the learnt models will differ from a norm to the other, since each norm emphasizes one given aspect of the value distribution. For instance, $\|.\|_2$ is more sensitive to high values, while it is insensitive to small ones. In order to reduce the learning bias induced by a particular norm, one can mix them. In this study, we consider the following interpolation for n data,

$$\left\| \overrightarrow{u} \right\|_{mix} = (1-\lambda) \left\| \overrightarrow{u} \right\|_\infty + \lambda \left\| \overrightarrow{u} \right\|_1 / n \tag{23}$$

which reduces to $\lambda \left\| \overrightarrow{u} \right\|_1 / n$ for λ=1 and to $\left\| \overrightarrow{u} \right\|_\infty$ for λ=0.

Assessing the performances of a ML model requires a proper splitting of data. Traditionally, they are split into three non-overlapping sets (to reduce overfitting, avoiding that final tests are applied on already seen data). The *training* set is made of data used to fit machine learning models under construction, the *validation* set is used to tune and control the model at this stage (for instance by determining the hyperparameters), and the *test* set is employed to provide an unbiased final evaluation of the model. Here, we will evaluate this final entropy prediction performance by assessing the mean absolute error (MAE) values according to the following:

$$MAE = \frac{1}{N_{mol}} \sum_{i=1}^{N_{mol}} \left| S_i^{exp} - S_i^{pred} \right| \tag{24}$$

All regression models were optimized using Python libraries. More specifically, ML parameters were determined using the *minimize* function available in the *scipy.optimize* Python library using the sequential least squares programming (SLSQP) algorithm. Initial guesses for the parameters to optimize were randomly generated using a uniform distribution of the [0.0,1.0] range. Similarly, bounds imposed for the optimized parameters also corresponded to the same range. This prevents to get too much high values that will have no physical interpretation, which could bring numerical instability, or that could be too much sensitive to data noise. Besides, in order to get rid of the initial random guess bias, this minimization protocol was repeated 500 times.

Such constraints actually bear some similarity with regularization procedures such as in Tikhonov (ridge) and lasso regressions in which penalty terms are added to the loss function to add parameter control. For instance, if we come back to Eq. 14, fitting without any constraint in the regression procedure may lead to negative values for α, β, or γ that are not physically motivated (the translation, rotational, and vibrational contributions should be positive), or to values that are above 1.0 for S_r and S_t, while we know that the rotational and vibrational entropies are reduced in the condensed phase.

All of the other mathematical manipulations are handled using standard *numpy* routines, while graphs were generated using *pyplot*.

Results and discussion

As explained in the previous sections, our feature set consists in the rotational, translational, and vibrational entropies from ADF calculations, taking into account the conformer populations at 298.15 K calculated using the Boltzmann distribution. This is illustrated by Fig. 2 in the case of butanone for which three conformers are retained, with populations equal to 72%, 15%, and 13% when modelled by immersion into the COSMO butanone implicit medium.

According to Eqs. 4 and 5, they feature as expected the same value for the translation entropy. Actually, they really differ from their vibrational entropy that ranges from 58.0 to 61.8 J mol^{-1} K^{-1}. In order to have more insight into the importance of a correct PES sampling, we define for each

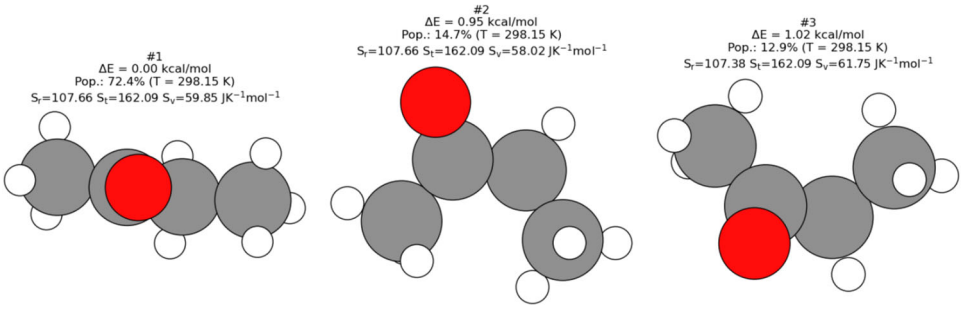

Fig. 2 Conformers selected based on potential energy surface exploration of butanone, with their respective Boltzmann weights at 298.15 K. These three conformers were identified and chosen using the method detailed in the "Computational details" section

chemical system the vibrational conformational span on all retained conformations by the following:

$$\Delta S_v^{conf} = \underset{\{conf\}}{max}\,(S_v^h) - \underset{\{conf\}}{min}\,(S_v^h) \tag{25}$$

We now complement these first results by having a quick look at the role of Grimme's anharmonic correction, as we show in Fig. 3. Three different behaviors can be identified. The first one is when $\overline{\Delta S_v^G}$ can be fully negligible. This is the case for molecules that do not exhibit low vibrational wavenumbers (typically lower than 100 cm^{-1}), as epitomized by the water molecule with its two a1 normal modes around 1600 cm^{-1} and 3700 cm^{-1}, and its b1 one at 3800 cm^{-1}. However, in the major cases, ΔS_v^G is negative, spanning the whole [-20, 0] $J\,K^{-1}\,mol^{-1}$ range, so that this correction tends to counter the entropy overestimation induced by the IG approximation, as expected. More surprisingly, positive values (between 10 and 15 $J\,K^{-1}\,mol^{-1}$) were found in three cases, namely, artifacts probably due to the simplicity of this approach. Thus, it turns out that, in absolute value, $T\overline{\Delta S_v^G}$ can reach up to 6 kJ mol^{-1}, even for moderate-size systems, a magnitude that is far from negligible when chemical accuracy is sought.

After having focused on the vibrational component, it is also valuable to compare it with the two other ones. Figure 3 shows the Boltzmann-averaged values, denoted $\overline{S}_t$, $\overline{S}_r$, $\overline{S}_v^h$, for the whole dataset.

It appears that the translational one is always higher than the rotational one and that both predominate for systems exhibiting a total entropy lower than 300 J mol^{-1} K^{-1}). Conversely, for the molecules with the highest entropy values, $\overline{S}_v^h$ clearly becomes the major contribution. It can be also noticed that the slope for $\overline{S}_v^h$ is much higher, suggesting that it is much more sensitive to the system size or composition.

Predictive models can now be built. However, the existence of various conformers makes the task not so straightforward. Indeed, as explained in "The molecular database," section the experimental value corresponds to an average value. As discussed before, the number of retained conformers may vary from one system to the other. Two approaches (note that this chemical complexity may be addressed by more refined techniques such as multi-instance machine learning [37] that are outside the scope of this paper) are possible: either all components in the model are Boltzmann-averaged and then these values are used for fitting or each component for each conformer are corrected, and then the Boltzmann weighting is performed. Mathematically, in the presence of non-linear terms, both approaches are actually not equivalent.

The first one is certainly the simplest to implement since there is exactly one average value for each component of a

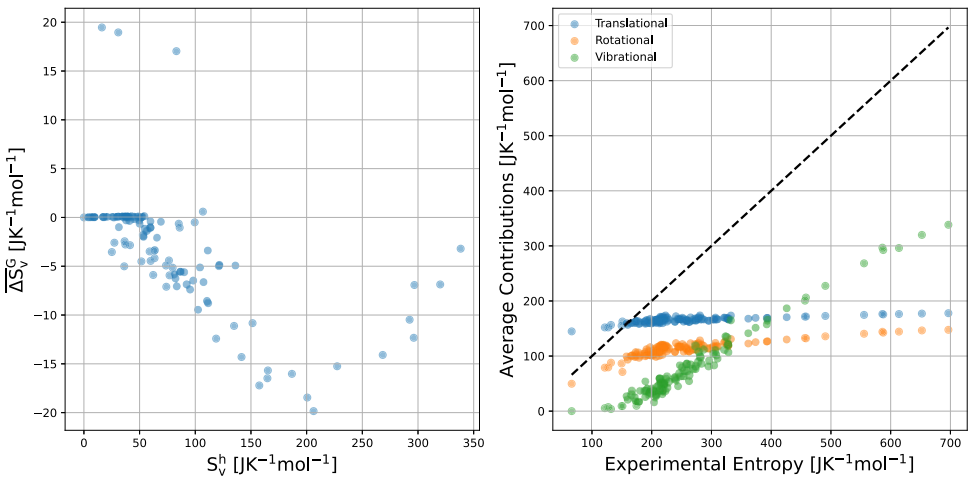

Fig. 3 *Left panel*: Comparison of harmonic contributions (x-axis) and anharmonic corrections (y-axis) in the system under study. *Right panel*: Plot of the average entropy contribution across conformers, with the experimental entropy on the x-axis and the Boltzmann-weighted average contribution on the y-axis

given system, but it would provide a model only made to evaluate average entropies. It is thus in principle not suited if one is interested in the entropy value for a specific conformation (which is useful when investigating, for instance, competitive reaction pathways). Moreover, its application requires that an exhaustive PES sampling has been performed, something that a user could not or should not want to carry out for more complex systems. The second approach, however, will not only learn how to correct entropy, but also include learning of weight correction, entangling both thermodynamical and electronic effects in an intricate way, making the task more complicated to achieve.

Obviously, both fitting methodologies reduce to the same one when only one conformer is retained, removing the ambiguity of the whole protocol. We thus decided to split our dataset into two parts. On the one hand, the *training* set will be made only of systems for which only one conformer was retained after the conformational search: it encompasses 54 molecules at the PBE level of theory. On the other hand, the remaining systems define the *validation* and *test* sets. The entropy of each conformation is then computed using the model optimized on the training set. Then, for a given system, the average entropy is calculated using the Boltzmann average and can then be compared to the experimental value.

It should added that, in a way, this splitting will give a kind of upper bound for the MAE values. Indeed, only one retained conformer usually corresponds to cases of small and rigid molecules, while multiple conformers are encountered for larger and flexible structures (not used at all for the full training). This means that the training and the test sets might be chemically quite different. Good performances on the test set for a given model would thus suggest its high versatility and would suggest that it can be trustfully applied on a large variety of systems.

Then, in order to split validation and test sets, we numbered (from 1 to 52) the systems with multiple retained conformers in increasing experimental entropy values. Systems with an odd label constitute the validation set, while the test set collects systems with even labels. By doing so, we ensure that both sets cover the full molecular range.

As explained in the "Computational details" section, training is performed by minimizing a given loss function. In the following, we will restrict our analysis to the special one expressed by Eq. 23. First, we trained the four first models, namely S_1^l (Eq. 13), S_4^l (Eq. 14), S_2^{nl} (Eq. 15), and S_7^{nl} (Eq. 16), for various values of the λ hyperparameter, by minimizing this loss function on the training set. All these models are then evaluated on the validation set by calculating the corresponding $MAE_{valid}(\lambda)$ value.

We have found that for any λ value within the [0.0,1.0] range, the optimal α parameter determined on the training set for S_1^l (Eq. 13, only one scaling coefficient) is almost constant, belonging to the [0.60, 0.62] range (a value that is close to the popular 2/3 correction), so that $MAE_{valid}(\lambda)$ is also almost constant for this one-parameter linear model, close to 37.3 J mol^{-1} K^{-1}.

The situation is much more contrasted for the non-linear extension, S_2^l that includes a normalized second-order term, still grouping translation and rotational contributions (Eq. 15). In the [0.0,0.4] λ range, the trained models converge to the same solution: α=0.578 and β=0.403, leading to a significant improvement with $MAE_{valid}(\lambda)$ equal to 19.5 J mol^{-1} K^{-1}. Conversely, for higher λ values, the optimal model changes. For instance, for λ=0.8, α=0.591 and β=0.111, deteriorating the performances on the validation set with MAE_{valid} equal to 30.8 J mol^{-1} K^{-1}.

At first sight, this could seem a counter-intuitive result. Indeed, as λ increases, the weight of $\|.\|_1$ in the loss function also increases. The optimal model on the training set is thus close to the one that minimizes the MAE on the training set. But, it appears that, concomitantly, it is less accurate, in terms of MAE, for the validation set. This can actually be rationalized by the fact that, as already mentioned, the training and the validation sets importantly differ. In a way, introducing the λ parameter allows for mitigating these differences and the data heterogeneity. It also underlines the importance of choosing relevant loss functions and the high impact on this choice in the final performances.

After reviewing these models (plotted in Fig. 4) with only one and two parameters, we can look at the four-parameter one represented by Eq. 14 (linear model in which the translational, rotational, and vibrational contributions are separately scaled). Strikingly, its performance of the validation set is disappointing with $MAE_{valid}(\lambda)$ values within the [33,41] J mol^{-1} K^{-1} range, being significantly higher than the one-parameter linear model. The best model corresponds to α=0.422, and $\beta = \gamma = \delta$=1.0, this last value being the maximal bound for optimization and leading to a suboptimal model. One could then wonder whether adding second-order terms and fully decoupling translational and rotational effects would be more efficient, leading to Eq. 16. We found that $MAE_{valid}(\lambda)$ then exhibits a high λ-dependence, varying from 23.7 J mol^{-1} K^{-1} (for $\lambda = 0$) to 47.2 J mol^{-1} K^{-1} ($\lambda = 1$), still being larger than the minimal value for S_2^l.

This last model, in spite of its simplicity (that has the advantage of preventing overfitting), can thus be considered the more satisfying one. This is the reason why we built Eq. 17 (S_3^{nlv} model) from Eq. 14 and not from Eq. 16. As shown in the left panel in Fig. 4, Eq. 14 leads to an underestimation of entropy for the highest values. As discussed earlier (see also Fig. 3), this corresponds to cases for which the vibrational entropy dominates, and also to the largest molecules. Intuitively, one thus would like to slightly increase the S_v^h weight with increasing R_{solv} values. Once more, we would like to do

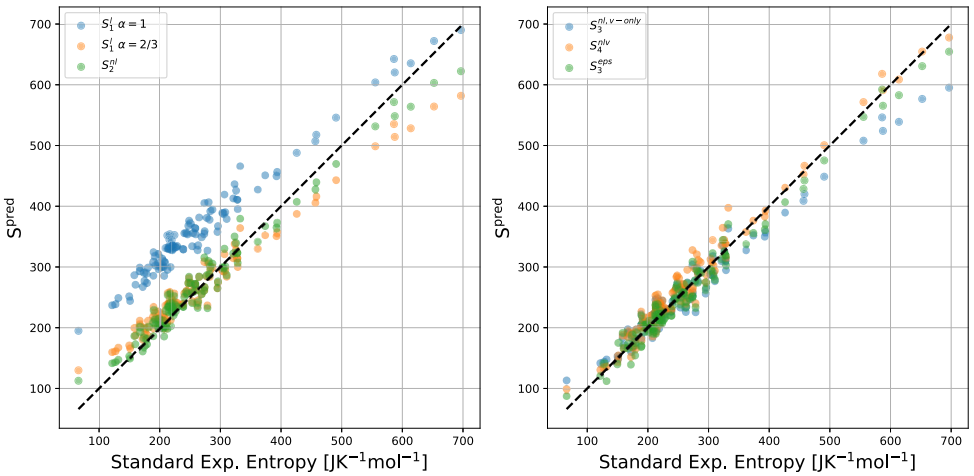

Fig. 4 Plots of the various models for entropies in solution at 298.15 K. x-axis, experimental values; y-axis, predicted values, all in J mol^{-1} K^{-1}. The black line represents a perfect correlation between experimental and predicted values

it in a controlled way and prevent the correction to take too much large values. Such a control is already at work for the mixed terms. Indeed, as all entropy components are positive, it is straightforward to show that $\frac{S_t S_v^h}{S_t+S_r+S_v^h}$, for instance, is already bounded by S_t, which becomes a minor contribution for the largest systems. Arbitrarily, we hence chose to bound the scaling f function in Eq. 17 to 20%:

$$f_\gamma(R_{solv}) = 1 + 0.2\widetilde{f}_\gamma(R_{solv}) \tag{26}$$

with

$$\begin{aligned} 0 \leq \gamma \leq 1 \\ \widetilde{f}_\gamma(0) = 0 \\ \lim_{R_{solv}\to\infty} \widetilde{f}_\gamma(R_{solv}) = 1 \end{aligned} \tag{27}$$

Simple functions obeying these conditions are sigmoids. We tested several standard flavors of them such as the logistic, arctan, and tanh functions. The best performance on the validation set was obtained using the Gudermannian function:

$$gd(x) = 2arctan(tanh(x)) \tag{28}$$

We thus optimized the following vibrational weight correction:

$$f_\gamma(R_{solv}) = 1 + 0.4\, arctan(tanh(\gamma R_{solv})) \tag{29}$$

with parameter γ constrained, as before, to belong to the [0.0,1.0] range, describing how "quick" the (unoptimized) asymptotic limit is reached. For λ=0, the optimized γ value was found equal to 0.0, so that this S_3^{nlv} model exactly reduces to S_2^{nl}, and this remains the case on the whole [0.0,0.50] λ range. Then, $MAE_{valid}(\lambda)$ decreases, reaching its minimum (17.9 J mol^{-1} K^{-1}) around λ=0.60 (with α=0.582, β=0.238, γ=0.131), before significantly increasing for higher λ values. The best model represents in fact a noticeable improvement with respect to the S_2^{nl} model. The right panel of Fig. 4 shows the obtained model, with an overall MAE of 16.6 J mol^{-1} K^{-1}.

Nevertheless, it appears that the highest values remain underestimated, while, on the opposite, the water molecule (lowest experimental value) is importantly overestimated. The first family gathers molecules with relative permittivity values around 2.0, while water distinguishes itself by its very high ϵ_r value (78.5). In order to cure these two deficiencies of the S_3^{nlv} model, an additional function of ϵ_r can be introduced (see Eq. 16). Ideally, it should give a positive value near 2.0, a negative one for water, and should be almost vanishing for the other systems. We thus proposed the following ansatz:

$$g_{delta}(\epsilon_r) = \delta ln(A/\epsilon_r) \tag{30}$$

The A value determines the transition between negative (for high ϵ_r) to positive (low ϵ_r) corrections. A simple choice is to choose it as the median ϵ_r value, that is to say, 5.82. Only δ is then still to be determined. Here, we decided to fix its value (it must be noticed that we did not reoptimize the other parameters) so that the standard deviation between MAE_{pred}, MAE_{valid}, and MAE_{test} is the smallest possible. This was reached with $\delta = 5.4$. It turned out that the corresponding S_3^{nlv} model improves over S_2^{nlv}, with a MAE on the full database equal to 15.4 J mol^{-1} K^{-1}. This corresponds to an energy at room temperature (multiplying by 298.15 K) equal to 4.6 kJ mol^{-1} ≈1.1 kcal mol^{-1}. In other words, this four-parameter non-linear model reaches the so-called "chemical accuracy."

Continuing to focus on the idea that the main contribution to the entropy trend comes from the vibrational component, we propose the following three-parameter model:

$$S_3^{nl,v-only} = (S_v^h)^{1+\alpha} - \beta\epsilon_r + \gamma \tag{31}$$

We constrained β to the range $0.0 \leq \beta \leq 1.0$ and γ to the range $0.0 \leq \gamma \leq 200.0$. Minimizing the mean absolute error

Table 1 Mean absolute error (MAE) values for the various trained models with respect to experimental entropy values (in J mol^{-1} K^{-1})

Model	MAE_{train}	MAE_{valid}	MAE_{pred}	MAE_{tot}
S_1^l, α=1.0	111.0	82.6	82.2	97.0
S_1^l, α=2/3	21.7	27.9	27.5	24.6
S_2^{nl}	16.0	19.4	19.3	17.7
S_3^{nlv}	15.2	17.9	18.3	16.6
S_4^{nlv}	15.4	15.3	15.5	15.4
$S_3^{nl,v-only}$	15.0	17.3	16.5	15.9

(MAE_{valid}) yielded $\lambda = 0.5$ as the optimal value, with $\alpha = 0.045$, $\beta = 0.904$, and $\gamma = 166$. The performance of this model is shown in the right panel of Fig. 4 and summarized in Table 1.

Despite its promising accuracy, the physical interpretation of the $S_3^{nl,v-only}$ model remains unclear due to several factors.

First, the model does not explicitly account for the rotational or translational contributions to entropy. While these may not be numerically significant, they are important in understanding the overall entropy behavior. Second, the application of a fractional exponent to the vibrational entropy introduces concerns regarding its physical justification. Though this approach works well for the systems in our dataset where vibrational contributions are predominant and avoids underestimation, it carries the risk of a significant "acceleration" in entropy for systems with extremely large vibrational effects, potentially leading to overestimation. Finally, the constant term γ implies a non-zero entropy even in the absence of any elements, which prompts further questions: does γ represent an intrinsic entropy of solvation, or is it capturing an average contribution from rotational and translational entropy? This leaves open questions about broader applicability and scalability.

Finally, we briefly discuss the transferability of our models by applying them to data obtained using the PBE0 exchange-correlation hybrid functional, instead of GGA PBE. For this purpose, all geometries were fully reoptimized with PBE0, followed by frequency calculations. We then applied our final S_4^{nlv} (Eq. 18) and $S_3^{nl,v-only}$ (Eq. 31) models using the parameter values initially derived with PBE (i.e., no refitting). Pleasingly, the respective MAEs for the entire dataset remained very close to the previous ones: 15.7 and 16.8 J mol^{-1} K^{-1}, respectively. These results increase our confidence that the models can be reliably used with any exchange-correlation functional.

Conclusions

In this paper, we developed and evaluated various models for entropies in solution within implicit solvation approaches with increasing complexity and physical considerations, involving only a few parameters and only features that are available from any quantum chemical calculations. Our models were subsequently trained on publicly available experimental values using different metrics, careful parameter control, and data splitting. The best-performing models were found to be transferable and close to the chemical accuracy, at a negligible computational cost.

Supplementary information Data used to build, validate and test the predictive models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00894-024-06225-3.

Acknowledgements The Centre Régional Informatique et d'Applications Numériques de Normandie (CRIANN) is acknowledged for providing access to computational resources.

Author contribution All authors have contributed to the conceptualization, scientific research, writing, and review of the paper.

Funding This work has been partially supported by University of Rouen Normandy, INSA Rouen Normandy, the "Centre National de la Recherche Scientifique" (CNRS), the European Regional Development Fund (ERDF), Labex SynOrg (ANR-11-LABX-0029), Carnot Institut I2C, the graduate school for research XL-Chem (ANR-18-EURE-0020 XL CHEM), and the "Région Normandie."

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Dedication It is a pleasure to dedicate this paper to our colleague and friend Prof. Alejandro Toro-Labbé, in recognition of his inspiring works, among others, on chemical reactivity and conceptual DFT.

References

1. Tantillo DJ (2022) Portable models for entropy effects on kinetic selectivity. J Am Chem Soc 144(31):13996–14004. https://doi.org/10.1021/jacs.2c04683
2. Harvey JN, Himo F, Maseras F et al (2019) Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. ACS Catal 9(8):6803–6813. https://doi.org/10.1021/acscatal.9b01537
3. Watson L, Eisenstein O (2002) Entropy explained: the origin of some simple trends. J Chem Educ 79(10):1269. https://doi.org/10.1021/ed079p1269
4. Abraham MH (1981) Relationship between solution entropies and gas phase entropies of nonelectrolytes. J Am Chem Soc 103(22):6742–6744. https://doi.org/10.1021/ja00412a036
5. Leung BO, Reid DL, Armstrong DA et al (2004) Entropies in solution from entropies in the gas phase. J Phys Chem A 108(14):2720–2725. https://doi.org/10.1021/jp030265a

6. Liu SC, Zhu XR, Liu DY et al (2023) DFT calculations in solution systems: solvation energy, dispersion energy and entropy. Phys Chem Chem Phys 25(2):913–931. https://doi.org/10.1039/d2cp04720a
7. Garza AJ (2019) Solvation entropy made simple. J Chem Theory Comput 15(5):3204–3214. https://doi.org/10.1021/acs.jctc.9b00214
8. Ariai J, Gellrich U (2023) The entropic penalty for associative reactions and their physical treatment during routine computations. Phys Chem Chem Phys 25(20):14005–14015. https://doi.org/10.1039/d3cp00970j
9. Pracht P, Grimme S (2021) Calculation of absolute molecular entropies and heat capacities made simple. Chem Sci 12(19):6551–6568. https://doi.org/10.1039/d1sc00621e
10. Gorges J, Grimme S, Hansen A et al (2022) Towards understanding solvation effects on the conformational entropy of non-rigid molecules. Phys Chem Chem Phys 24(20):12249–12259. https://doi.org/10.1039/d1cp05805c
11. Conquest OJ, Roman T, Marianov A et al (2021) Calculating entropies of large molecules in aqueous phase. J Chem Theory Comput 17(12):7753–7771. https://doi.org/10.1021/acs.jctc.1c00848
12. Besora M, Vidossich P, Lledós A et al (2018) Calculation of reaction free energies in solution: a comparison of current approaches. J Phys Chem A 122(5):1392–1399. https://doi.org/10.1021/acs.jpca.7b11580
13. Michel C, Laio A, Milet A (2009) Tracing the entropy along a reactive pathway: the energy as a generalized reaction coordinate. J Chem Theory Comput 5(9):2193–2196. https://doi.org/10.1021/ct900177h
14. Lipparini F, Mennucci B (2016) Perspective: polarizable continuum models for quantum-mechanical descriptions. J Chem Phys 144(16). https://doi.org/10.1063/1.4947236
15. Gaussian Inc. (2024) Thermochemistry in Gaussian. Accessed 06 Aug 2024
16. Grimme S (2012) Supramolecular binding thermodynamics by dispersion-corrected density functional theory. Chem Eur J 18(32):9955–9964. https://doi.org/10.1002/chem.201200497
17. Chai JD, Head-Gordon M (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. Phys Chem Chem Phys 10(44):6615. https://doi.org/10.1039/b810189b
18. Dzib E, Merino G (2021) The hindered rotor theory: a review. Wiley Interdiscip Rev Comput Mol Sci 12(3). https://doi.org/10.1002/wcms.1583
19. Vansteenkiste P, Van Speybroeck V, Marin GB et al (2003) Ab initio calculation of entropy and heat capacity of gas-phase n-alkanes using internal rotations. J Phys Chem A 107(17):3139–3145. https://doi.org/10.1021/jp027132u
20. Ardura D, López R, Sordo TL (2005) Relative Gibbs energies in solution through continuum models: effect of the loss of translational degrees of freedom in bimolecular reactions on Gibbs energy barriers. J Phys Chem B 109(49):23618–23623. https://doi.org/10.1021/jp0540499
21. Sumimoto M, Iwane N, Takahama T et al (2004) Theoretical study of trans-metalation process in palladium-catalyzed borylation of iodobenzene with diboron. J Am Chem Soc 126(33):10457–10471. https://doi.org/10.1021/ja040020r
22. Cooper J, Ziegler T (2002) A density functional study of SN2 substitution at square-planar platinum(II) complexes. Inorg Chem 41(25):6614–6622. https://doi.org/10.1021/ic020294k
23. Tobisch S (2005) Organolanthanide-mediated ring-opening Ziegler polymerization (ROZP) of methylenecycloalkanes: a theoretical mechanistic investigation of alternative mechanisms for chain initiation of the samarocene-promoted rozp of 2-phenyl-1-methylenecyclopropane. Chem Eur J 11(10):3113–3126. https://doi.org/10.1002/chem.200401102
24. Chen P, Dougan BA, Zhang X et al (2013) Reactions of a tungsten alkylidyne complex with mono-dentate phosphines: thermodynamic and theoretical studies. Polyhedron 58:30–38. https://doi.org/10.1016/j.poly.2012.07.042
25. Jouanno LA, Di Mascio V, Tognetti V et al (2014) Metal-free decarboxylative hetero-diels-alder synthesis of 3-hydroxypyridines: a rapid access ton-fused bicyclic hydroxypiperidine scaffolds. J Org Chem 79(3):1303–1319. https://doi.org/10.1021/jo402729a
26. Linstrom P (1997) NIST Chemistry Webbook, NIST Standard Reference Database 69. https://doi.org/10.18434/T4D303
27. SCM TC (2023) COSMO: conductor-like screening model. https://www.scm.com/doc/ADF/Input/COSMO.html, software for Chemistry & Materials, Amsterdam, The Netherlands
28. Gaussian I (2023) SCRF: self-consistent reaction field. https://gaussian.com/scrf/, gaussian, Inc., Wallingford CT
29. Klamt A (1995) Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. J Phys Chem 99(7):2224–2235. https://doi.org/10.1021/j100007a062
30. te Velde G, Bickelhaupt FM, Baerends EJ et al (2001) Chemistry with ADF. J Comput Chem 22(9):931–967. https://doi.org/10.1002/jcc.1056
31. Stenutz R (2023) Rolf Stenutz's chemistry pages. https://www.stenutz.eu/chem/
32. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868. https://doi.org/10.1103/physrevlett.77.3865
33. Adamo C, Barone V (1999) Toward reliable density functional methods without adjustable parameters: the PBE0 model. J Chem Phys 110(13):6158–6170. https://doi.org/10.1063/1.478522
34. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32(7):1456–1465. https://doi.org/10.1002/jcc.21759
35. SCM TC (2023) PLAMS interfaces: conformers. https://www.scm.com/doc/plams/interfaces/conformers.html, software for Chemistry & Materials, Amsterdam, The Netherlands
36. SCM TC (2023) PLAMS components: RDKit. https://www.scm.com/doc/plams/components/mol_rdkit.html, software for Chemistry & Materials, Amsterdam, The Netherlands
37. Zankov D, Madzhidov T, Varnek A et al (2023) Chemical complexity challenge: is multi-instance machine learning a solution? Wiley Interdiscip Rev Comput Mol Sci 14(1). https://doi.org/10.1002/wcms.1698