



Cite this: DOI: 10.1039/d6cp00088f

# Generation and benchmarking of a diverse reaction database of quantum mechanical liquid-phase activation Gibbs free energies

 Lingfeng Gui,<sup>†a</sup> Alan Armstrong,<sup>b</sup> Claire S. Adjiman,<sup>a</sup> Fareed Bhasha Sayyed<sup>c</sup> and Amparo Galindo<sup>ib\*<sup>a</sup></sup>

Many chemical reactions occur in the liquid phase, making the accurate prediction of the liquid-phase activation Gibbs free energy,  $\Delta^\ddagger G^{\circ,L}$ , crucial for numerous applications. Quantum mechanical (QM) methods with implicit solvation models offer a valuable route to  $\Delta^\ddagger G^{\circ,L}$  prediction, although they are computationally demanding at high levels of theory and for larger systems. Data-driven surrogate models can address this issue but require extensive training and test datasets. We present here the liquid phase reaction energy database (LiPRED-2026), a QM reaction database containing 4513  $\Delta^\ddagger G^{\circ,L}$  values for 28 diverse chemical reactions computed in various solvents at 298.15 K. The reactions have been chosen for their sensitivity to solvent effects and the availability of experimental data. The SMD model is employed to calculate solvation contributions to  $\Delta^\ddagger G^{\circ,L}$  because it can be used to account for the effect of solvent on the geometries of the reactants and transition states and it is suitable for charged species. The database contains  $\Delta^\ddagger G^{\circ,L}$  obtained from seven calculation methods, including the thermodynamic cycle method, the direct method, and their variants. Using a subset of the database, a benchmarking study shows that the best methods achieve a mean absolute error of 2.89 kcal mol<sup>-1</sup> in absolute  $\Delta^\ddagger G^{\circ,L}$  and 1.00 kcal mol<sup>-1</sup> in relative  $\Delta^\ddagger G^{\circ,L}$ , respectively, with the lower error for the relative  $\Delta^\ddagger G^{\circ,L}$  being mainly attributable to error cancellation. The use of a higher level of theory to calculate  $\Delta^\ddagger G^{\circ,L}$  improves relative  $\Delta^\ddagger G^{\circ,L}$  values only, but not absolute ones. These results provide valuable insights into the choice of methods and levels of theory appropriate for calculating  $\Delta^\ddagger G^{\circ,L}$ , while the database can serve for training and testing surrogate models.

 Received 9th January 2026,  
 Accepted 30th April 2026

DOI: 10.1039/d6cp00088f

[rsc.li/pccp](http://rsc.li/pccp)

## 1 Introduction

Chemical reactions in the liquid phase hold particular significance in science and engineering. Most biological processes, from DNA replication, protein folding, to cellular respiration, occur in an aqueous environment.<sup>1</sup> The production of many chemical products, such as pharmaceuticals, pesticides, dyes and polymers, also involves liquid-phase chemical reactions, typically in organic solvents. In this context, furthering our

understanding of solvent effects on chemical reactions is of critical interest.<sup>2</sup>

In modelling liquid-phase reactions, explicit solvation models and implicit solvation models can be used. In the former case, solvent molecules are introduced explicitly. From this perspective, solvent molecules and solute molecules are essentially treated in the same way, though different approximation/parameterisation schemes may be applied to the solute and the solvent molecules. Explicit solvation models are commonly used in classical molecular dynamics and Monte Carlo simulations where the intermolecular interactions are relatively cheap to evaluate. For example, given the importance of water in biological systems, various explicit water models, such as SPC, TIP3P, TIP4P and their other variations, have been developed.<sup>3,4</sup> When it comes to more computationally expensive quantum mechanical (QM) modelling, which is particularly relevant for reaction kinetics, implicit solvation models are often adopted, where the solvent environment is represented as a continuum medium characterised by several

<sup>a</sup> Department of Chemical Engineering, The Sargent Centre for Process Systems Engineering, Imperial College London, London, SW7 2AZ, UK.  
 E-mail: a.galindo@imperial.ac.uk

<sup>b</sup> Department of Chemistry, Imperial College London, Molecular Sciences Research Hub, White City Campus, London, W12 0BZ, UK

<sup>c</sup> Synthetic Molecule Design and Development, Eli Lilly Services India Pvt Ltd, Devarabeesanahalli, Bengaluru 560103, India

<sup>†</sup> Present address: School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, UK.



parameters. The preference for implicit solvation models is often due to the fact that they avoid the introduction of extra degrees of freedom and provide an accurate description of the strong long-range electrostatic interactions.<sup>5</sup> Popular implicit solvation models include the polarisable continuum model (PCM),<sup>6</sup> the SMx series,<sup>7</sup> the COSMO<sup>8,9</sup> models, and their variations. There also exist hybrid methods of explicit and implicit solvation models,<sup>10,11</sup> wherein one or more layers of solvent molecules are treated explicitly and the remaining solvent molecules are treated implicitly as a continuum.

A significant challenge with using QM methods to calculate the liquid-phase activation Gibbs free energy  $\Delta^\ddagger G^{\circ,L}$  is their substantial computational expense, especially at higher levels of theory. As a result, the availability of accurate and chemically diverse reaction databases is highly valuable. Such databases, whether derived from QM calculations or experimental measurements, provide essential reference data for benchmarking and validating kinetic models. In particular, ensuring the accuracy and relevance of  $\Delta^\ddagger G^{\circ,L}$  data is crucial for reliable reaction modeling. Thus, the generation, curation, and benchmarking of diverse reaction databases is beneficial in advancing data-driven approach to reaction prediction.

Many existing reaction databases that catalogue activation energies and reaction kinetics focus on gas-phase reactions predominantly.<sup>12–19</sup> However, liquid-phase reactions are often of greater practical interest, which underscores the need for liquid-phase databases. Jorner *et al.*<sup>20</sup> curated a dataset of 449 experimental liquid-phase rate constants for  $S_NAr$  reactions from the literature, including a broad range of solvents, nucleophiles, and leaving groups. The corresponding  $\Delta^\ddagger G^{\circ,L}$  values, derived from the rate constants using the Eyring equation, range from 12.5 to 42.4 kcal mol<sup>-1</sup>. This dataset was used to train and validate a Gaussian process regression model, which achieved a MAE of 0.77 kcal mol<sup>-1</sup> on an independent test set. Chung and Green<sup>21</sup> presented an extensive dataset comprising over 8.3 million solvation free energies of activation,  $\Delta^\ddagger \Delta G^{\text{solv}}$ , and solvation enthalpies of activation,  $\Delta^\ddagger \Delta H^{\text{solv}}$ , computed at 298 K using COSMO-RS at the BP-TZVPD-FINE level for 28 318 neutral reactions and 295 solvents.  $\Delta^\ddagger \Delta G^{\text{solv}}$  and  $\Delta^\ddagger \Delta H^{\text{solv}}$  are defined as the differences in solvation free energy and solvation enthalpy, respectively, between the transition state and the reactant(s).  $\Delta^\ddagger \Delta G^{\text{solv}}$  is particularly useful for calculating the liquid-phase rate constant relative to that in the gas phase or in a different solvent. Among the reactions evaluated, 26 448 are unimolecular in both the forward and reverse directions, involving up to two products and generally characterised by high gas-phase energy barriers. Because most of these reactions are uncommon, they were only used to pre-train a machine learning model, which was subsequently fine-tuned using an additional set of 1870 more common reactions. This additional set was generated *via* the AutoTST framework<sup>22</sup> for H-abstraction, H-migration and R-addition reactions. The resulting machine learning surrogate model for  $\Delta^\ddagger \Delta G^{\text{solv}}$  and  $\Delta^\ddagger \Delta H^{\text{solv}}$  takes only the molecular structures of the reactants and solvents as input

and achieves a MAE of 0.68 kcal mol<sup>-1</sup> in predicting relative  $\Delta^\ddagger G^{\circ,L}$  values between two phases on an experimental test set. These relative  $\Delta^\ddagger G^{\circ,L}$  values were obtained from 165 experimental relative rate constants curated in the earlier work of Chung and Green<sup>23</sup> for 15 neutral closed-shell or free radical reactions and 49 solvents over a temperature range from 273 to 392 K.

The  $\Delta^\ddagger G^{\circ,L}$  can be calculated by subtracting the liquid-phase Gibbs free energy of each reactant from that of the transition state. The calculation of  $\Delta^\ddagger G^{\circ,L}$  *via* a thermodynamic cycle (TC) may, however, be preferred if the ideal gas-phase activation Gibbs free energy,  $\Delta^\ddagger G^{\circ,IG}$ , and solvation free energy,  $\Delta G^{\circ,\text{solv}}$ , can be evaluated more accurately in a separate manner,<sup>24</sup> although it has been argued that the TC method may not necessarily always be more accurate than the direct calculation of  $\Delta^\ddagger G^{\circ,L}$ .<sup>24,25</sup> Ho and Ertem<sup>24</sup> performed a benchmarking study over an extensive dataset comprised of 175 reaction Gibbs free energy changes and 83 activation free energies within liquid phase, which showed that the accuracies of the direct and TC methods are generally very similar to each other. For the activation Gibbs free energy,  $\Delta^\ddagger G^{\circ,L}$ , the mean absolute deviation (MAD) between the direct and the TC methods is about 0.24 kcal mol<sup>-1</sup> for reactions involving neutral reactants and about 0.96 kcal mol<sup>-1</sup> for reactions involving charged reactants, with the direct method generally agreeing better with experiments when charged reactants are involved. However, with the number of experimental  $\Delta^\ddagger G^{\circ,L}$  values (18 Diels Alder reactions and 13  $S_N2$  reactions) used for comparison in their work, it is difficult to conclude which method is better in general.

Regardless of whether the direct method or the TC method is used, the choice of the level(s) of theory used along with these two methods is key in terms of the accuracy of the  $\Delta^\ddagger G^{\circ,L}$  calculation. When the solvation free energy is calculated explicitly as in the TC method, it is most appropriately obtained using levels of theory that are consistent with the parameterisation scheme of the solvation model in use. For example, the SMD solvation model used in the current work was parameterised using an average of several relatively low levels of theory in the original work,<sup>26</sup> and it is unclear whether one of these levels of theory or the average would produce the best results for  $\Delta^\ddagger G^{\circ,L}$ . In addition, both the TC method and the direct method involve single-point calculations at a high level of theory, for which QM composite methods, such as G3MP2<sup>27</sup> and CBS-QB3,<sup>28,29</sup> are suitable,<sup>24</sup> but for which different options also exist. There is therefore a need to investigate further which method/level of theory combinations (MLOTC) are the most accurate in general and for a specific reaction type.

In the current work, we present the liquid phase reaction free energies database (LiPRED-2026). LiPRED-2026 contains both absolute and relative liquid-phase activation Gibbs free energies,  $\Delta^\ddagger G^{\circ,L}$ , for 28 diverse organic reactions computed in various solvents at 298.15 K using different MLOTCs. Instead of using reactions generated *in silico*, LiPRED-2026 encompasses reactions with well-documented solvent effects and mostly with



experimental data available. The SMD model<sup>26</sup> is chosen over COSMO-type approaches because it is parameterised extensively with experimental data to reproduce solvation Gibbs free energies of both neutral and charged species, and charged reactants are common in organic synthesis. Furthermore, SMD makes it possible to account for the effect of the solvent on the geometries of the reactants and transition state and therefore, on the reaction pathway. The Menschutkin reaction between pyridine and phenacyl bromide serves as the primary case for investigating solvent effects, and  $\Delta^\ddagger G^{\circ,L}$  is calculated in 531 solvents. A further 2812 calculations are carried out for different reaction-solvent combinations. A subset of reactions from the database is used to benchmark various MLoTCs against experimental  $\Delta^\ddagger G^{\circ,L}$  data from the literature, providing insights into the balance between computational cost and accuracy.

The remainder of this paper is organised as follows: in Section 2, the computational methods employed in this work are introduced, including the thermodynamic cycle method and its variations (Section 2.1), the direct method (Section 2.2) and the workflow for implementing these methods to calculate  $\Delta^\ddagger G^{\circ,L}$  (Section 2.3). For Section 3.1, an overview of the LiPRED-2026 database is provided, detailing the reactions (Section 3.1.1) and calculation schemes (Section 3.1.2) involved. In Section 3.2, the results of the benchmarking study of different MLoTCs are presented. Conclusions are presented in Section 4.

## 2 Methods

### 2.1 The thermodynamic cycle method

In the TC method, the liquid-phase activation Gibbs free energy  $\Delta^\ddagger G^{\circ,L}$  is decomposed according to the thermodynamic cycle in Fig. 1 and expressed as:<sup>24</sup>

$$\Delta^\ddagger G^{\circ,L} = \Delta^\ddagger G^{\circ,IG} + \Delta G_{TS}^{\circ,solv} + \sum_{r \in D} \nu_r \Delta G_r^{\circ,solv} + \left(1 + \sum_{r \in D} \nu_r\right) RT \ln \frac{RT}{P_0}, \quad (1)$$

where  $\Delta^\ddagger G^{\circ,IG}$  is the ideal gas activation Gibbs free energy,  $\Delta G_{TS}^{\circ,solv}$  is the solvation free energy of the transition state,  $\Delta G_r^{\circ,solv}$  is the solvation free energy of reactant  $r$ ,  $P_0$  is the reference pressure,  $D$  is the set of reactant(s) and  $\nu_r$  is the

stoichiometric coefficient of reactant  $r$ . The last term is the standard-state correction from the gas-phase standard state defined by  $T = 298.15$  K and  $P_0 = 1$  atm to the solution-phase standard state of  $1 \text{ mol L}^{-1}$ . Since all calculations in this work are performed using the Gaussian 16 software,<sup>30</sup> the remainder of the paper will express the terms in eqn (1) in terms of quantities that can be computed and output by Gaussian 16. Definitions of the thermodynamic quantities reported by Gaussian can be found in the documentation by Ochterski.<sup>31</sup>  $\Delta^\ddagger G^{\circ,IG}$  can be calculated as follows:

$$\Delta^\ddagger G^{\circ,IG} = G_{TS}^{\circ,IG} + \sum_{r \in D} \nu_r G_r^{\circ,IG} = \left(E_{TS}^{\text{el,IG}} + G_{TS}^{\text{therm,IG}}\right) + \sum_{r \in D} \nu_r \left(E_r^{\text{el,IG}} + G_r^{\text{therm,IG}}\right), \quad (2)$$

where  $G_{TS}^{\circ,IG}$  and  $G_r^{\circ,IG}$  are the ideal gas Gibbs free energies of the transition state and reactant  $r$ , respectively,  $E_{TS}^{\text{el,IG}}$  and  $E_r^{\text{el,IG}}$  are the electronic energies of the transition state and reactants  $r$ , respectively, and  $G_{TS}^{\text{therm,IG}}$  and  $G_r^{\text{therm,IG}}$  are the thermal corrections to the free energies of the transition state and species  $r$ , respectively.

The solvation free energies  $\Delta G_{TS}^{\circ,solv}$  and  $\Delta G_r^{\circ,solv}$  are calculated by applying the SMD solvation model<sup>26</sup> as:

$$\Delta G_{TS}^{\circ,solv} = E_{TS}^{\text{el,L}} - E_{TS}^{\text{el,IG}}, \quad (3)$$

$$\text{and } \Delta G_r^{\circ,solv} = E_r^{\text{el,L}} - E_r^{\text{el,IG}}, \quad r \in D,$$

where,  $E_r^{\text{el,L}}$  and  $E_r^{\text{el,IG}}$  are the electronic energies of reactant  $r$  calculated in the liquid phase (with the SMD solvation model) and in the ideal gas phase, respectively, and  $E_{TS}^{\text{el,L}}$  and  $E_{TS}^{\text{el,IG}}$  are the electronic energies of the transition state calculated at the same level of theory in the liquid phase and in the ideal gas phase, respectively. The liquid-phase energies,  $E_r^{\text{el,L}}$  and  $E_{TS}^{\text{el,L}}$ , are calculated at geometries optimised in the liquid phase. The expressions in eqn (3) include electrostatic and non-electrostatic contributions to the solvation free energy *via* the SMD model parameterisation.<sup>26</sup>

The benefit of using the TC method is that one can calculate  $\Delta^\ddagger G^{\circ,IG}$  with a high level of theory and  $\Delta G_i^{\circ,solv}$ ,  $i \in D \cup \{TS\}$  with a level of theory that is consistent with the parameterisation scheme of the solvation model separately.<sup>25</sup> Therefore,  $\Delta^\ddagger G^{\circ,L}$  can be expressed using the combination of different levels of theory as:

$$\Delta^\ddagger G_{TC}^{\circ,L} = \left(E_{TS,\text{high}}^{\text{el,IG}} + G_{TS,\text{geom}}^{\text{therm,IG}}\right) + \sum_{r \in D} \nu_r \left(E_{r,\text{high}}^{\text{el,IG}} + G_{r,\text{geom}}^{\text{therm,IG}}\right) + \left(E_{TS,\text{low}}^{\text{el,L}} - E_{TS,\text{low}}^{\text{el,IG}}\right) + \sum_{r \in D} \nu_r \left(E_{r,\text{low}}^{\text{el,L}} - E_{r,\text{low}}^{\text{el,IG}}\right) + \left(1 + \sum_{r \in D} \nu_r\right) RT \ln \frac{RT}{P_0}, \quad (4)$$

where the extra subscripts (“high”, “geom” and “low”) added to each term denote the level of theory at which the corresponding term is calculated. “geom” refers to the level of theory

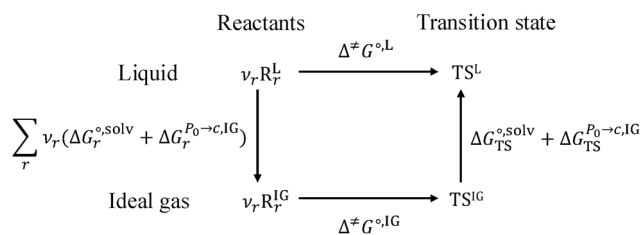


Fig. 1 Schematic representation of the thermodynamic cycle method for a multi-molecular reaction;  $R_r$  represents reactant  $r$  and TS is the transition state.



at which the geometry of the species is optimised. “high” refers to a high level of theory used for calculating the electronic energy of an optimised structure in the ideal gas phase; in this work, G3MP2<sup>27</sup> and CBS-QB3<sup>28,29</sup> are used, as these composite methods are designed to provide high thermochemical accuracy and are often discussed in the context of achieving “chemical accuracy” on standard thermochemical test sets, while maintaining a manageable computational cost for small- to medium-sized systems.<sup>32</sup> In principle, however, other levels of theory, such as g-xTB,<sup>33</sup> may also be used, depending on the requirements. “low” refers to a low level of theory at which the solvation free energy is calculated. Here M062X, B3LYP and HF in conjunction with the 6-31+G(d) basis set are used in order to be consistent with the work of Ho and Ertem.<sup>24</sup> In this work, we set “geom” = “low”.

In eqn (4), the use of gas-phase thermal corrections is considered appropriate because any solvation-induced changes may already be implicitly incorporated into the “electronic energy” term through the SMD model parameterisation. Using explicit solution-phase thermal corrections may risk double counting, and the use of an ideal-gas partition function is not formally appropriate for solution-phase geometries and frequencies.<sup>34</sup> However, Ribeiro *et al.* argued that, in the SMD parameterisation, the training set, which consists largely of small rigid molecules, was chosen such that solvation-induced structural changes are typically small compared with the intrinsic error of continuum models, rather than with the intention of folding these effects into the parameters.<sup>35</sup> Therefore, when solvation-induced changes in geometry and/or vibrational frequencies are expected to be significant, accounting for them explicitly through solution-phase thermodynamic corrections can be justified and may improve accuracy. In later work, Ho and Ertem also noted that solvation-induced changes in thermal motion can be significant, and that explicitly including these contributions can in principle be more accurate, although the benefit is system-dependent and does not always guarantee improved agreement.<sup>24</sup>

Therefore, when this vibrational change is significant, the vibrational correction term  $G_{i,\text{geom}}^{\text{therm,L}} - G_{i,\text{geom}}^{\text{therm,IG}}$  should be added to the solvation free energy components in eqn (4) to account explicitly for the solvation-induced vibrational change. As a result, the  $G_{i,\text{geom}}^{\text{therm,IG}}$  term in eqn (4) cancels, leaving only  $G_{i,\text{geom}}^{\text{therm,L}}$ . Equivalently,  $G_{i,\text{geom}}^{\text{therm,IG}}$  is replaced by  $G_{i,\text{geom}}^{\text{therm,L}}$ , such that:

$$\begin{aligned} \Delta^\ddagger G_{\text{TC-vib}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,IG}} + G_{\text{TS,geom}}^{\text{therm,L}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,IG}} + G_{r,\text{geom}}^{\text{therm,L}} \right) \\ &+ \left( E_{\text{TS,low}}^{\text{el,L}} - E_{\text{TS,low}}^{\text{el,IG}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{low}}^{\text{el,L}} - E_{r,\text{low}}^{\text{el,IG}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}. \end{aligned} \quad (5)$$

The TC method with the vibrational correction (eqn (5)) is denoted as “TC-vib” in the remainder of the paper.

Another variation that can be made to the TC method (eqn (4)) is to use a “high” level of theory for the solvation free energy calculations of the TS and reactants. This method is referred to as the “TC-high” method, which can be expressed as:

$$\begin{aligned} \Delta^\ddagger G_{\text{TC-high}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,L}} + G_{\text{TS,geom}}^{\text{therm,IG}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,L}} + G_{r,\text{geom}}^{\text{therm,IG}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}, \end{aligned} \quad (6)$$

since the electronic ideal gas terms cancel in this case. The difference between the TC-high method and the original TC method is referred to as the high level of theory (LoT) correction.

Given that the SMD model is parameterised using the following error function,<sup>26</sup>

$$\chi = \sum_{J=1}^{N_D} \left[ \Delta G_{\text{expt},J}^{\circ,\text{solv}} - \frac{1}{6} \sum_{j=1}^6 \Delta G_{j,J}^{\text{EP}} - G_J^{\text{CDS}} \right]^2, \quad (7)$$

where  $\Delta G_{\text{expt},J}^{\circ,\text{solv}}$  is the experimental solvation free energy for data point  $J$ ,  $\Delta G_{j,J}^{\text{EP}}$  is the electrostatic contribution to the solvation free energy calculated at level of theory  $j$  for the data point  $J$ ,  $G_J^{\text{CDS}}$  is the non-electrostatic contribution to the solvation free energy for the data point  $J$  and  $N_D$  is the number of data points ( $N_D = 2489$  in the work of Marenich *et al.*<sup>26</sup>). For maximum consistency with the model, the solvation free energy is calculated as the average of the results from the six levels of theory used for the SMD model parameterisation, *i.e.*, M05-2X/MIDI!6D, M052X/6-31G(d), M05-2X/6-31+G(d,p), M05-2X/cc-pVTZ, B3LYP/6-31G(d), and HF/6-31G(d), as this strictly follows the parameterisation scheme of the SMD model. The TC method corrected with this approach is referred to as “TC-SMD” and is expressed as:

$$\begin{aligned} \Delta^\ddagger G_{\text{TC-SMD}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,IG}} + G_{\text{TS,geom}}^{\text{therm,IG}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,IG}} + G_{r,\text{geom}}^{\text{therm,IG}} \right) \\ &+ \left( E_{\text{TS,SMD}}^{\text{el,L}} - E_{\text{TS,SMD}}^{\text{el,IG}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{SMD}}^{\text{el,L}} - E_{r,\text{SMD}}^{\text{el,IG}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}, \end{aligned} \quad (8)$$

where  $E_{i,\text{SMD}}^{\text{el,L}}$  is the average of the electronic energies calculated at the six low levels of theory mentioned above for species  $i$ .

A final modification considered is to use the quasi-rigid rotor harmonic oscillator (quasi-RRHO) model to calculate the thermal corrections such that errors caused by low-lying vibrational frequencies in the harmonic approximation can be corrected.<sup>36</sup> The method that uses quasi-RRHO is suffixed



by “quasi” and expressed as:

$$\begin{aligned} \Delta^\neq G_{\text{TC-quasi}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,IG}} + G_{\text{TS,geom}}^{\text{therm,IG,quasi}} \right) \\ &+ \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,IG}} + G_{r,\text{geom}}^{\text{therm,IG,quasi}} \right) \\ &+ \left( E_{\text{TS,low}}^{\text{el,L}} - E_{\text{TS,low}}^{\text{el,IG}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{low}}^{\text{el,L}} - E_{r,\text{low}}^{\text{el,IG}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}, \end{aligned} \quad (9)$$

where  $G_{i,\text{geom}}^{\text{therm,IG,quasi}}$  is the thermal correction to the Gibbs free energy for species  $i$  using the quasi-RRHO model.

All variants of the TC method used in this work are summarised in Table 1.

## 2.2 The direct method

The direct method is equivalent to calculating the Gibbs free energy of each species  $i$  directly within the continuum solvation model. It can be simply expressed as:

$$\begin{aligned} \Delta^\neq G_{\text{direct}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,L}} + G_{\text{TS,geom}}^{\text{therm,L}} \right) + \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,L}} + G_{r,\text{geom}}^{\text{therm,L}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}. \end{aligned} \quad (10)$$

The direct method is equivalent to augmenting the TC method result with both the vibrational correction and the high LoT correction for solvation free energy. The standard state correction is also needed if the thermal corrections are obtained using the gas-phase standard state. The direct method using the quasi-RRHO model is then expressed as:

$$\begin{aligned} \Delta^\neq G_{\text{direct-quasi}}^{\circ,\text{L}} &= \left( E_{\text{TS,high}}^{\text{el,L}} + G_{\text{TS,geom}}^{\text{therm,L,quasi}} \right) \\ &+ \sum_{r \in D} \nu_r \left( E_{r,\text{high}}^{\text{el,L}} + G_{r,\text{geom}}^{\text{therm,L,quasi}} \right) \\ &+ \left( 1 + \sum_{r \in D} \nu_r \right) RT \ln \frac{RT}{P_0}. \end{aligned} \quad (11)$$

It should be noted that although the direct method involves fewer constituent terms, it can be more computationally

expensive than the TC method as the electronic energies need to be evaluated with the same (potentially high) level of theory in both the gas and liquid phases. The variants of the direct method considered in the current work are summarised in Table 1.

In this study, the term “method” refers exclusively to the variants of the TC method and the direct method, whereas “level of theory” denotes the quantum-mechanical levels used for the high, low, and geom subscripts. The combination of a method with specific levels of theory is abbreviated as “MLoTC”.

The  $\Delta^\neq G^{\circ,\text{L}}$  values calculated using eqn (1)–(11) are referred to as absolute  $\Delta^\neq G^{\circ,\text{L}}$  values, as they represent the kinetics of a reaction in a specific solvent. In practice, however, relative values are often preferred, because systematic errors associated with the calculations or measurements tend to cancel. The relative  $\Delta^\neq G^{\circ,\text{L}}$  values, are determined by subtracting the absolute  $\Delta^\neq G^{\circ,\text{L}}$  value in a reference solvent,  $\Delta^\neq G^{\circ,\text{L,ref}}$ , from that in the solvent of interest,  $\Delta^\neq G^{\circ,\text{L}}$ , or by subtracting the absolute  $\Delta^\neq G^{\circ,\text{L}}$  value for a reference substrate from that for the substrate of interest when comparing substrate effects. In the current work, the reference solvent or reaction is either chosen to be consistent with the reported experimental values or corresponds to the reaction with the lowest free energy in the relevant reaction type series.

## 2.3 Workflow for calculating $\Delta^\neq G^{\circ,\text{L}}$

**2.3.1 Conformation treatment.** In the current work, a single conformer is used for each species (reactant or transition state) in all calculations. This choice is made because most reactions considered here involve small and largely rigid molecules that are either ring-constrained or lack long flexible chains. For reaction subsets where multiple conformers may be relevant, either published optimised geometries from previous studies are adopted or a conformer search is carried out using the MMFF94 force field,<sup>37</sup> with the lowest-energy conformer selected for further calculations.

Although Boltzmann averaging would offer a more rigorous treatment of conformational effects, its impact is expected to be limited for most reactions included in the present work, since the majority of species in the current version of the database are either rigid or only mildly flexible.<sup>38,39</sup> Furthermore, Wittmann *et al.*<sup>40</sup> reported in a benchmark study of solvation free energies and partition ratios for flexible solutes that Boltzmann-averaged results were very similar to those obtained using only the lowest-energy conformer, supporting the simplified approach adopted here. Further details on the conformer selection for each reaction are provided in the SI.

**2.3.2 Quantum mechanical calculations.** All the QM calculations for database generation and benchmarking study are performed with the Gaussian 16 software<sup>30</sup> with the workflow shown in Fig. 2. Initial structures are pre-optimised at the HF/3-21G level of theory in vacuum. The pre-optimised structures are then refined to local minima (or saddle points for transition-state structures) at a specified “low” level of theory in both the gas phase and liquid phase, with frequency calculations performed to obtain thermal corrections and confirm that true

**Table 1** Thermal correction terms and solvation free energy expressions used in each method

Method	Thermal correction	Solvation free energy
TC	$G_{i,\text{geom}}^{\text{therm,IG}}$	$E_{i,\text{low}}^{\text{el,L}} - E_{i,\text{low}}^{\text{el,IG}}$
TC-vib	$G_{i,\text{geom}}^{\text{therm,L}}$	$E_{i,\text{low}}^{\text{el,L}} - E_{i,\text{low}}^{\text{el,IG}}$
TC-SMD	$G_{i,\text{geom}}^{\text{therm,IG}}$	$E_{i,\text{SMD}}^{\text{el,L}} - E_{i,\text{SMD}}^{\text{el,IG}}$
TC-high	$G_{i,\text{geom}}^{\text{therm,IG}}$	$E_{i,\text{high}}^{\text{el,L}} - E_{i,\text{high}}^{\text{el,IG}}$
TC-quasi	$G_{i,\text{geom}}^{\text{therm,IG,quasi}}$	$E_{i,\text{low}}^{\text{el,L}} - E_{i,\text{low}}^{\text{el,IG}}$
Direct	$G_{i,\text{geom}}^{\text{therm,L}}$	$E_{i,\text{high}}^{\text{el,L}} - E_{i,\text{high}}^{\text{el,IG}}$
Direct-quasi	$G_{i,\text{geom}}^{\text{therm,L,quasi}}$	$E_{i,\text{high}}^{\text{el,L}} - E_{i,\text{high}}^{\text{el,IG}}$



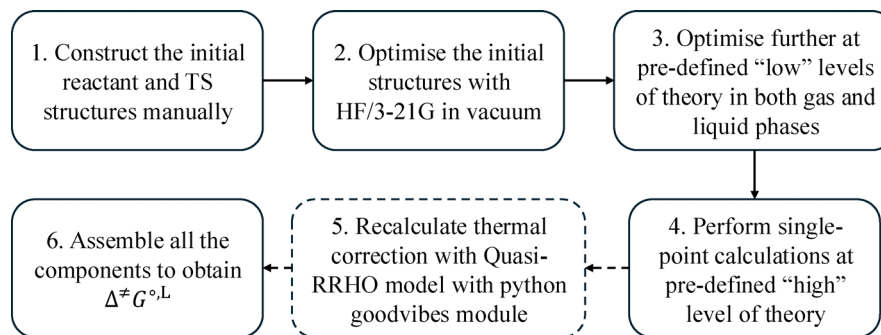


Fig. 2 Workflow for calculating  $\Delta^\ddagger G^{\circ,L}$ . The dashed lines indicate a step which is used in some methods only.

minima or saddle points have been found. Next, electronic calculations at the “high” level of theory are performed for the optimised geometry, only in the gas phase for the TC method and in both the gas and liquid phases for the direct method. Where desired, the thermal corrections are also calculated using the quasi-RRHO model with the Python goodvibes module.<sup>41</sup>

## 3 Results and discussion

### 3.1 Overview of the LiPRED database

The LiPRED database contains 4513  $\Delta^\ddagger G^{\circ,L}$  values calculated for 28 reactions using the combination of 7 methods, 3 low levels of theory and 2 high levels of theory. We characterise each reaction with a unique reactant or a unique pair of reactants. All of the reactions included in the database have previously been investigated experimentally and have well-documented solvent effects (see the references cited in Section 3.1.1). The solvents included in the database range from those commonly used in chemical laboratories with available experimental descriptors, such as those in the Minnesota solvent descriptor database,<sup>42</sup> to solvents constructed from predefined atom groups,<sup>43,44</sup> as well as hypothetical solvents<sup>45</sup> defined solely by descriptor values without corresponding molecular structures. The full list of reaction–solvent pairs considered is provided in the Excel sheet available in the Zenodo online repository.

**3.1.1 Overview of the reaction set.** The reaction set consists of 16  $S_N2$  reactions, 8 Diels–Alder (DA) reactions,<sup>46</sup> the aromatic nucleophilic substitution (ANS) reaction of 2,4-dinitrochlorobenzene and piperidine,<sup>47</sup> and 3 nucleophilic addition reactions<sup>48,49</sup> (Fig. 3 and 4). The 16  $S_N2$  reactions include 3 reactions with the cyanide ion as the nucleophile,<sup>24</sup> 9 identity reactions of various alkyl chloride substrates,<sup>24</sup> 2 Menschutkin reactions (*p*-nitrobenzyl chloride and trimethylamine<sup>50</sup> and phenacyl bromide and pyridine<sup>51</sup>), the Williamson ether synthesis reaction of sodium  $\beta$ -naphthoxide and benzyl bromide (O-alkylation), and its C-alkylation side reaction.<sup>52</sup> The 3 nucleophilic addition reactions include the activation of glycine using diisopropylcarbodiimide (DIC) and the intramolecular cyclisation of the *Z*- and *E*-configurational adducts of ethyl cyano(hydroxyimino)acetate (Oxyrna) and DIC.<sup>48,49</sup> The chemical equations of all the reactions as well as an overview of the solvent(s) used are summarised in Fig. 3 and 4. The complete list of the solvents used is provided in the

Excel sheet available in the Zenodo online repository. The Excel sheet also contains 64  $\Delta^\ddagger G^{\circ,L}$  values for Fmoc (9-fluorenylmethoxycarbonyl) deprotection reactions, calculated using M062X for all free energy components. These values are not comparable with the results from the other calculations and are therefore not discussed in this work. It is important to note that the SMD solvation model is parameterised using experimental properties for certain solvents. In cases where experimental data are unavailable for a solvent, property values calculated through group contribution methods are employed. Additionally, some calculations utilise hypothetical solvents, *i.e.*, a combination of property values that do not correspond to any known solvent. For example, the liquid-phase activation free energies of the Menschutkin reaction of phenacyl bromide and pyridine are calculated in 184 solvents using property values reported from experiments, most of which are tabulated in the Minnesota solvent descriptor database,<sup>42</sup> 338 solvents with property values computed by group contribution methods<sup>53</sup> and 9 hypothetical solvents.<sup>45</sup> The QM investigation of solvent effects on this Menschutkin reaction represents the most comprehensive analysis of any reaction within the database. It should also be noted that for the  $S_N2$  reactions in acetonitrile, *tert*-butyl chloride, allyl chloride, and benzyl chloride as substrates may favour the  $S_N1$  reaction pathway; according to Ho and Ertem<sup>24</sup> and the reference therein,<sup>54</sup> the reported values of  $\Delta^\ddagger G^{\circ,L}$  correspond to  $S_N2$  reactions. As the original experimental data or clarifications could not be located in the cited sources, one should exercise caution. In the current work, these reactions are treated as  $S_N2$  reactions in the calculations.

**3.1.2 Method/level of theory combinations (MLoTCs) used.** For the reactions in Fig. 3, all possible MLoTCs (42 MLoTCs = 6 LoT combinations  $\times$  7 methods) are used such that they can be compared to the experimental values. For most of the other reactions, only M062X/G3MP2 is used. The number of calculations performed for each type of reaction and the various MLoTCs are summarised in Fig. 5. In Fig. 6, the distribution of the  $\Delta^\ddagger G^{\circ,L}$  values obtained using the thermodynamic cycle method combined with the M062X/G3MP2 level of theory is shown. The distribution is positively skewed, with a peak in the range of 22.0 to 23.0 kcal mol<sup>−1</sup>, and a greater number of reactions exhibiting higher  $\Delta^\ddagger G^{\circ,L}$  values (>23 kcal mol<sup>−1</sup>) than those exhibiting lower values (<22 kcal mol<sup>−1</sup>). The peak



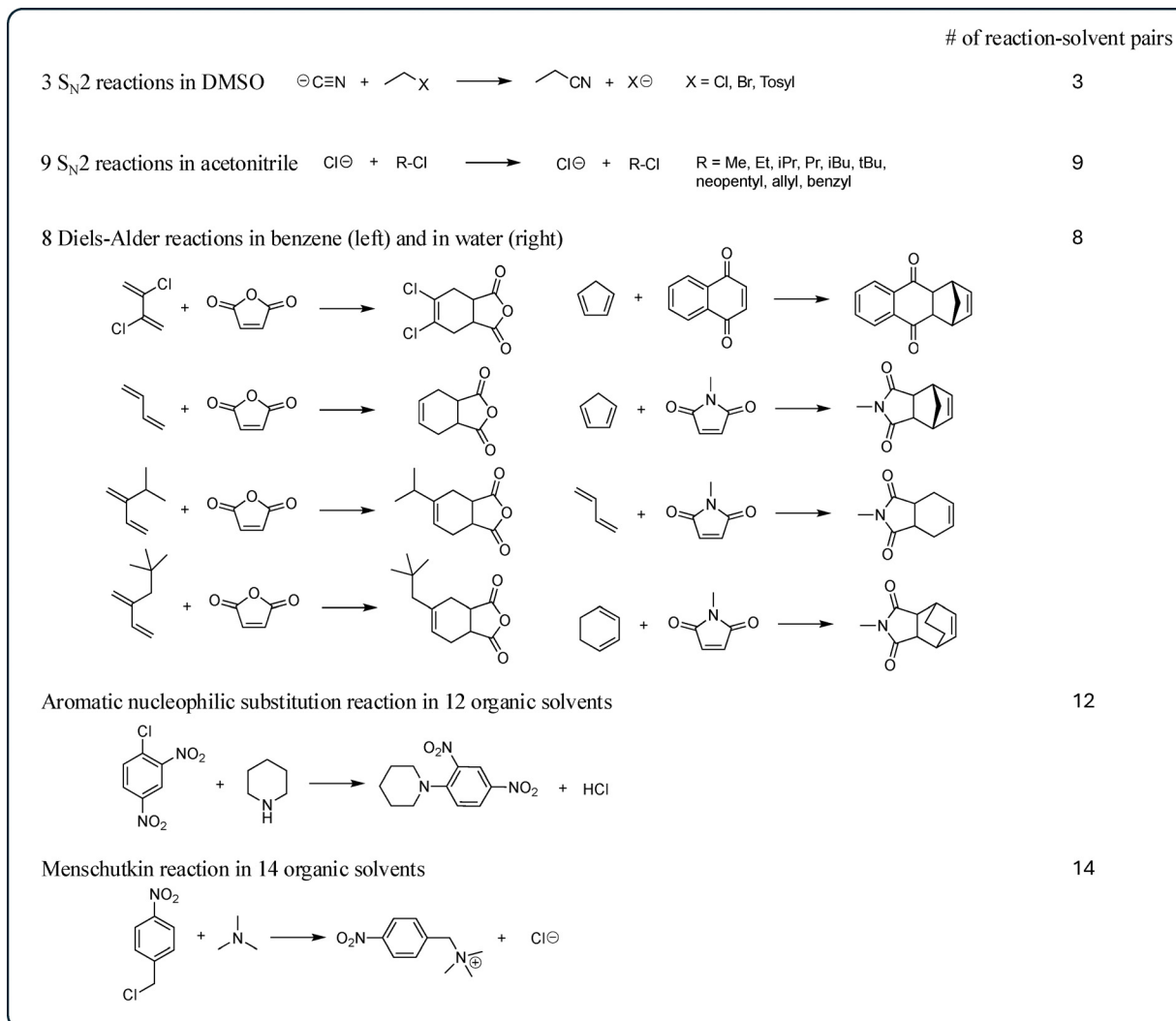


Fig. 3 Reactions in the benchmarking study from the LiPRED-2026 database used in the benchmarking study. 46 reaction-solvent combinations are studied. The Diels-Alder reactions are the only ones that involve exclusively neutral species, *i.e.*, their transition states do not exhibit a strong ionic character or charge separation.

centre value ( $22.5 \text{ kcal mol}^{-1}$ ), converted *via* the transition state theory, corresponds to a half-life of approximately 58 minutes for a unimolecular reaction at 298 K, suggesting that the majority of the reactions can be easily followed experimentally near room temperature, with only a small number being extremely slow or fast.

### 3.2 Benchmarking $\Delta^\ddagger G^{\circ,\text{L}}$ calculations

Experimental  $\Delta^\ddagger G^{\circ,\text{L}}$  of the 46 reaction-solvent pairs in Fig. 3 are collected from literature. Both absolute and relative  $\Delta^\ddagger G^{\circ,\text{L}}$  values are compared in the current work. Experimental absolute  $\Delta^\ddagger G^{\circ,\text{L}}$  values have been reported only for the 3 S<sub>N</sub>2 reactions involving cyanide as the nucleophile, the 8 DA reactions and the 12 ANS reactions. For the remaining reactions in the set, only relative  $\Delta^\ddagger G^{\circ,\text{L}}$  values have been reported. To assess relative  $\Delta^\ddagger G^{\circ,\text{L}}$  values, the experimental dataset used in the current work is constructed by including the relative  $\Delta^\ddagger G^{\circ,\text{L}}$  values reported in the literature (9 S<sub>N</sub>2 reactions with chloride

as the nucleophile and 14 Menschutkin reactions) and values that are calculated from absolute  $\Delta^\ddagger G^{\circ,\text{L}}$  values by subtracting the smallest experimental absolute  $\Delta^\ddagger G^{\circ,\text{L}}$  values as a reference value for each corresponding reaction type (3 S<sub>N</sub>2 reactions with cyanide as the nucleophile, 8 DA reactions and 12 ANS reactions).

The mean absolute error (MAE), is used as the performance indicator in the current work and is calculated for each MLoTC considered by comparing each computed liquid-phase activation free energy,  $\Delta^\ddagger G^{\circ,\text{L},\text{cal}}$ , to its corresponding experimental liquid-phase activation free energies,  $\Delta^\ddagger G^{\circ,\text{L},\text{exp}}$ :

$$\text{MAE} = \frac{\sum_{i=1}^N |\Delta^\ddagger G^{\circ,\text{L},\text{cal}} - \Delta^\ddagger G^{\circ,\text{L},\text{exp}}|}{N}, \quad (12)$$

where  $N$  is the total number of data points used. This expression is applied to both absolute and relative  $\Delta^\ddagger G^{\circ,\text{L}}$  values.



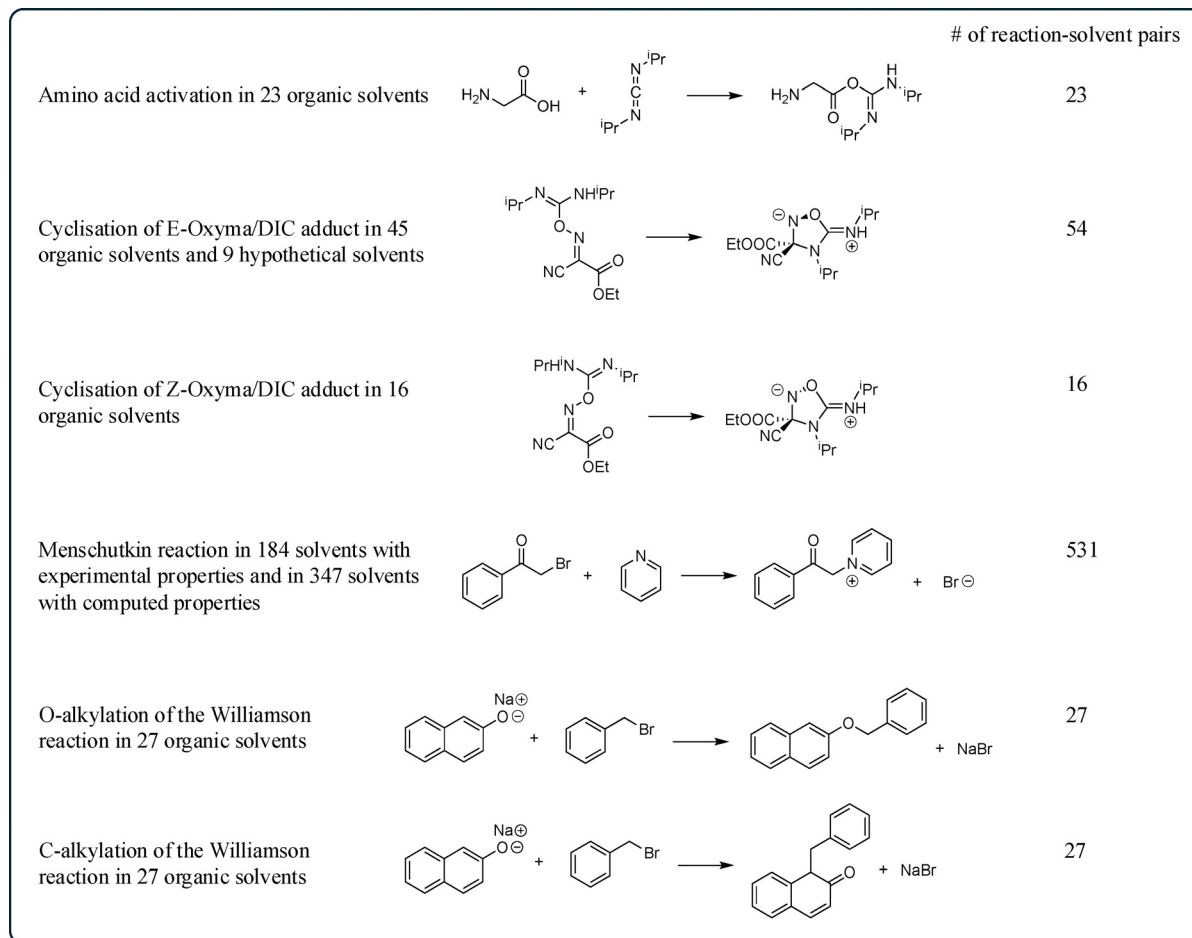


Fig. 4 Reactions included in the LiPRED-2026 database but not included in the benchmarking study.

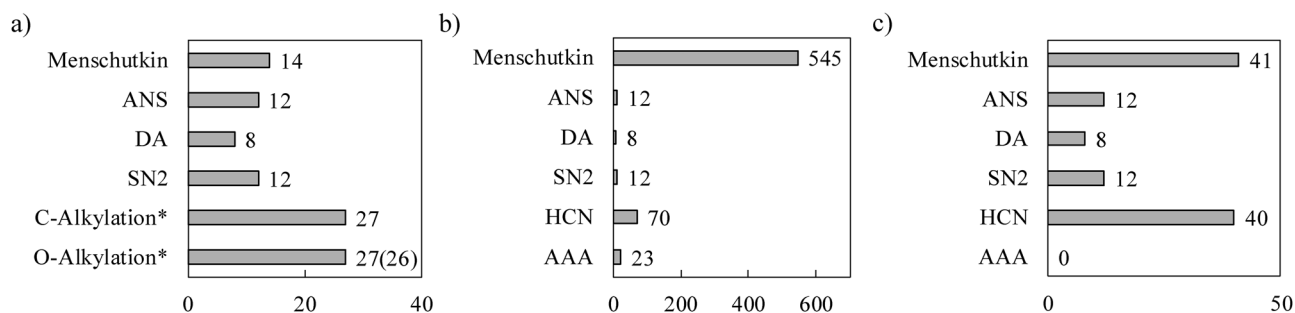


Fig. 5 Number of calculations performed for each type of reaction broken down in terms of the MLToTCs used: (a) all levels of theory except M062X/G3MP2 + all methods; (b) M062X/G3MP2 + (TC, TC-vib, TC-quasi) and (c) M062X/G3MP2 + (TC-SMD, TC-high, direct, direct-quasi). \*For C-/O-alkylation, only B3LYP/G3MP2 is used.

To ensure that the MAEs obtained in the benchmarking study are meaningful, we inspected (1) the uncertainties in the experimental data for the reactions in the benchmarking set, where available, and (2) the extent of solvent-induced variation in the experimental  $\Delta^\ddagger G^{\circ,L}$  values, so that any conclusions drawn from the benchmarking study are not limited by uncertainty or insufficient variation in the experimental data. For example, for the ANS reactions, the reported uncertainty in the

rate constants is less than 2 to 3% across all solvents, while the uncertainty in  $\Delta^\ddagger G^{\circ,L}$  is reported to be around 0.7 kcal mol<sup>-1</sup>.<sup>47</sup> For the DA reactions in water, the reported uncertainties in the rate constants are within 6%,<sup>55,56</sup> which corresponds, by the Eyring equation, to an estimated uncertainty of about 0.04 kcal mol<sup>-1</sup>. Thus, where experimental uncertainties are reported, they are clearly smaller than the MAEs obtained in the benchmarking study (see Fig. 7). Regarding the solvent-induced variation in the



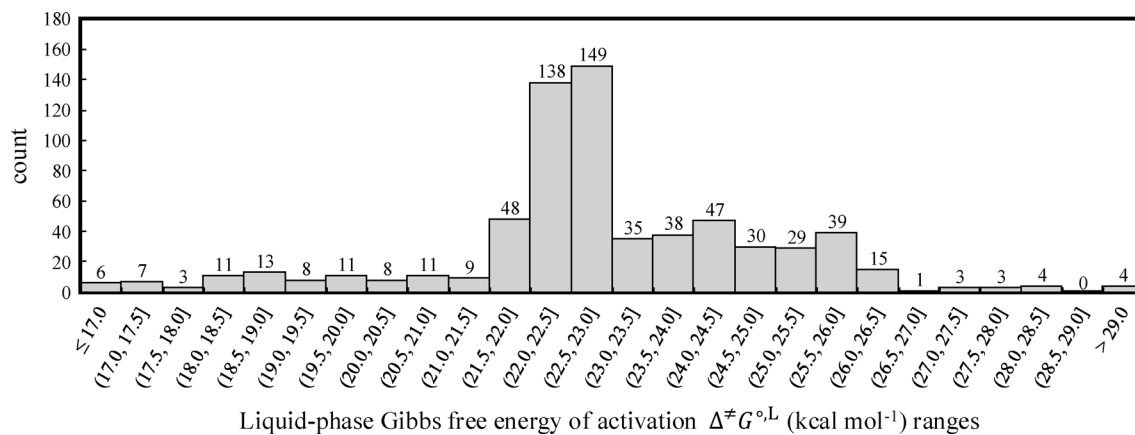


Fig. 6 Distribution of liquid-phase Gibbs free energies of activation computed using the combination of thermodynamic cycle method and the M062X/G3MP2 level of theory.

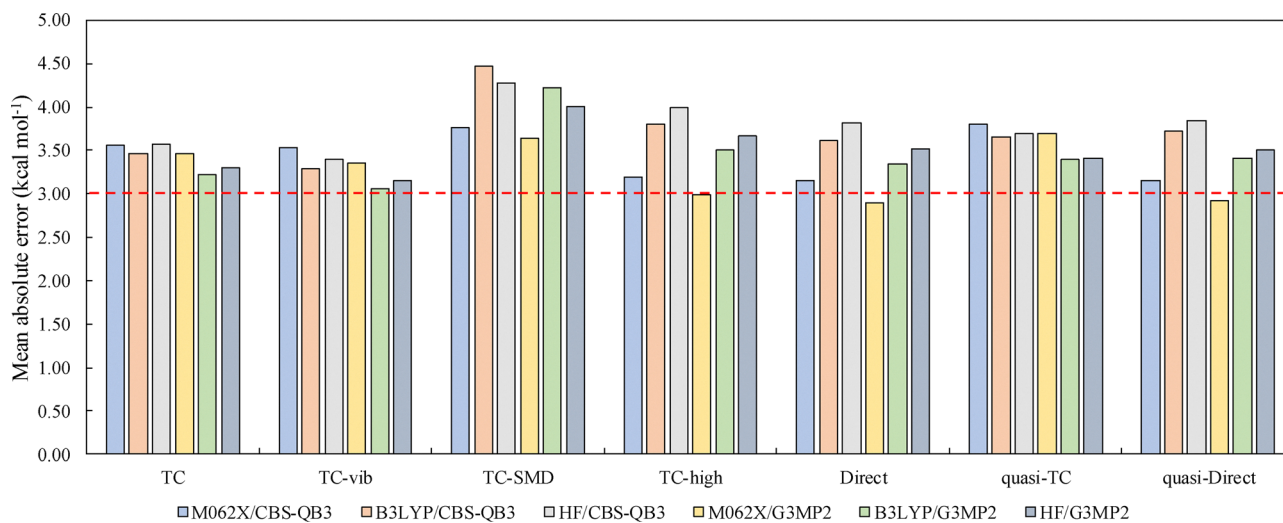


Fig. 7 Mean absolute error of each method/level of theory combination for 23 absolute  $\Delta^\ddagger G^{\circ,L}$  predictions (3  $S_N2$  reactions with cyanide as the nucleophile, 8 DA reactions and 12 ANS reaction–solvent pairs). The red dashed line corresponds to 3 kcal mol<sup>-1</sup>.

experimental  $\Delta^\ddagger G^{\circ,L}$  values, the relative activation free energies span approximately 0 to 9 kcal mol<sup>-1</sup> for the  $S_N2$  reactions, 0 to 11 kcal mol<sup>-1</sup> for the DA reactions, and about -2 to 4 kcal mol<sup>-1</sup> for the Menshutkin reactions. These ranges are substantially larger than the MAEs obtained for both the relative and absolute activation free energies, indicating that the solvent effect is significant compared with the prediction error. Only for the ANS reactions is the solvent-dependent range relatively small, at approximately 0 to 2 kcal mol<sup>-1</sup>. These solvent-dependent variations are investigated further in Fig. 9.

The results for absolute  $\Delta^\ddagger G^{\circ,L}$  values are summarised in Fig. 7. It can be seen from the comparison of the TC method and the TC-vib method that the addition of the vibrational correction reduces the MAE slightly for all the levels of theory considered, while the use of the TC-high method results in an improvement of the performance of the TC methods only when using the M062X functional. Combining the vibrational

correction with the TC-high method, *i.e.*, using the direct method, the use of the M062X functional further improves the MAE, while the direct methods with levels of theory other than M062X once again perform slightly worse than the original TC methods. Using the average of six levels of theory does not improve the performance even though this choice is most consistent with the parameterisation scheme of the SMD solvation model. Given that  $\Delta^\ddagger G^{\circ,L}$  is a much more complex quantity than the solvation free energies used to parameterise the SMD model, the performance of the TC-SMD model in predicting  $\Delta^\ddagger G^{\circ,L}$  does not depend only on being consistent with the SMD parameterisation. For example, the reactions considered here involve species, particularly transition states, that are not represented in the training set. Therefore, it is not surprising that the TC-SMD method, despite being the most consistent with the original SMD parameterisation, does not necessarily lead to improved predictions of  $\Delta^\ddagger G^{\circ,L}$ .



In general, using the quasi-RRHO model does not improve model performance, and slightly degrades the performance of the TC method, in which the use of the quasi-RRHO model affects only  $\Delta^\ddagger G^{\circ,L}$ . Given the presence of empirical components, *i.e.*, the solvation free energies, which are known to contribute a significant portion of the overall error, any benefit provided by quasi-RRHO may be diminished in this case. Further investigation of this aspect could include testing alternative entropy treatments such as 1D-HR<sup>57</sup> on a small subset, but is beyond the scope of the present benchmark study.

Focusing on the choice of the high level of theory, the use of G3MP2 consistently provides better performance than that of CBS-QB3. Overall, the direct method, in conjunction with the M062X/G3MP2 level of theory, yields the smallest MAE (2.89 kcal mol<sup>-1</sup>) for absolute  $\Delta^\ddagger G^{\circ,L}$  predictions. This is consistent with the conclusion of Ho and Ertem<sup>24</sup> that the direct method outperforms the TC method when the transition state is ionic, given that most of the reactions evaluated involve transition states that are either ionic or exhibit a strong ionic character (except for the DA reactions). For the DA reactions, the MAEs achieved using the TC method and the direct method are 1.58 and 1.68 kcal mol<sup>-1</sup>, respectively, when using M062X/G3MP2. This performance is also comparable to that reported by Chung and Green,<sup>23</sup> who evaluated only neutral closed-shell or free radical reactions and obtained an MAE of 1.30 kcal mol<sup>-1</sup> with their best level of theory combination:  $\omega$ B97XD/def2-TZVP for the gas-phase activation free energy and BP-TZVP-G16 for the COSMO-RS solvation evaluation.

At this point, it is useful to note that the MAEs from most of the MLoTCs are above 3 kcal mol<sup>-1</sup> and they may be subject to significant systematic error. If the systematic error exists, its impact can be alleviated by calculating relative  $\Delta^\ddagger G^{\circ,L}$  values, as is commonly done.<sup>21,23</sup>

The performance of each MLoTC for relative  $\Delta^\ddagger G^{\circ,L}$  prediction is shown in Fig. 8. Overall, it can be seen that the MAEs for the relative  $\Delta^\ddagger G^{\circ,L}$  values are 2 kcal mol<sup>-1</sup> smaller than that for

corresponding absolute  $\Delta^\ddagger G^{\circ,L}$ , indicating a significant portion of the systematic error is eliminated. In contrast to the case of absolute  $\Delta^\ddagger G^{\circ,L}$ , the inclusion of the vibrational correction leads to an increase in the MAE for all the levels of theory considered. Similarly, the use of the TC-high method either makes no difference or increases the MAE compared to the TC method. The direct method performs worse than the TC method for all the levels of theory. In principle, the addition of the vibrational correction is expected to be most useful for more flexible systems where solvation-induced changes in vibrational contributions are larger, whereas for predominantly rigid solutes it can lead to a deterioration of the predictions due to the potential double counting of thermal solvation effects already embedded in the parameterisation. In such cases, the standard TC approximation can be a more consistent choice given the underlying solvation model. In addition, similar to the case of absolute  $\Delta^\ddagger G^{\circ,L}$  predictions, using neither the average of six levels of theory nor the quasi-RRHO model lead to a decrease in the MAE. Overall, CBS-QB3 and G3MP2 provide similar performance. For relative  $\Delta^\ddagger G^{\circ,L}$ , the MLoTCs affording the smallest MAEs are the TC method + M062X/CBS-QB3 (1.00 kcal mol<sup>-1</sup>) and TC-high method + M062X/G3MP2 (1.01 kcal mol<sup>-1</sup>), followed by TC method + M062X/G3MP2 (1.02 kcal mol<sup>-1</sup>). The MAEs for relative  $\Delta^\ddagger G^{\circ,L}$  reported by Chung and Green<sup>23</sup> are similar across all the levels of theory combinations they evaluated, approximately 0.4 kcal mol<sup>-1</sup>, which is comparable to the best MAE we obtain here for the DA reaction using TC-vib + M062X/G3MP2 (0.44 kcal mol<sup>-1</sup>).

The performance of two well-performing MLoTCs, the TC method + M062X/G3MP2 and the direct method + M062X/G3MP2, is further evaluated *via* the parity plots shown in Fig. 9. In Fig. 9a and c, overall, the direct method yields a higher  $R^2$  value of 0.308, compared to 0.250 from the TC method, for the prediction of absolute  $\Delta^\ddagger G^{\circ,L}$ . Significant systematic deviations are observed in the predicted absolute  $\Delta^\ddagger G^{\circ,L}$  values for the S<sub>N</sub>2 reactions and the ANS reactions. The

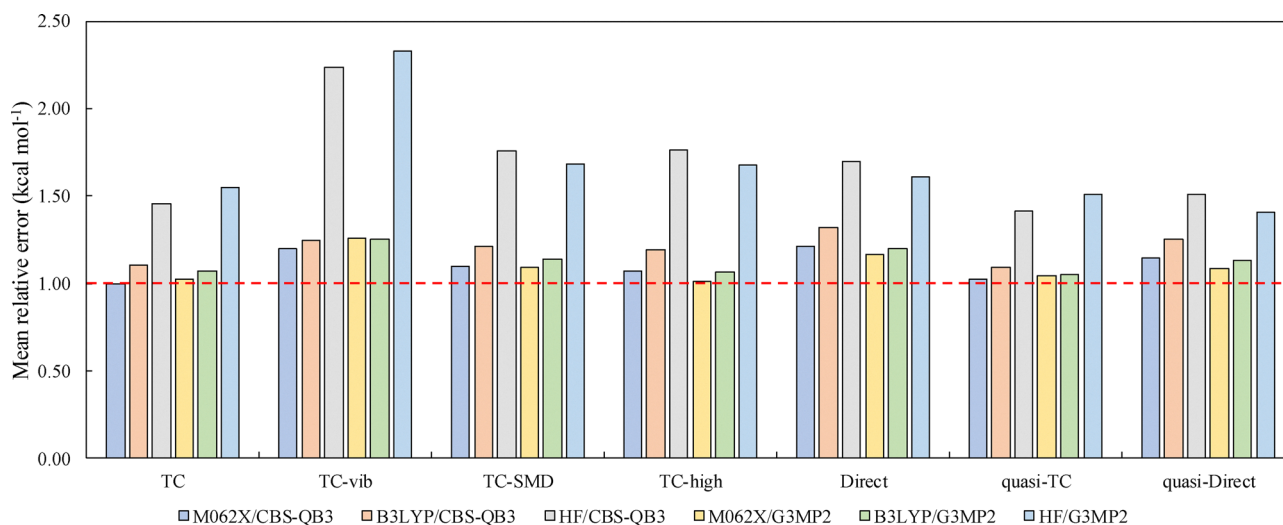
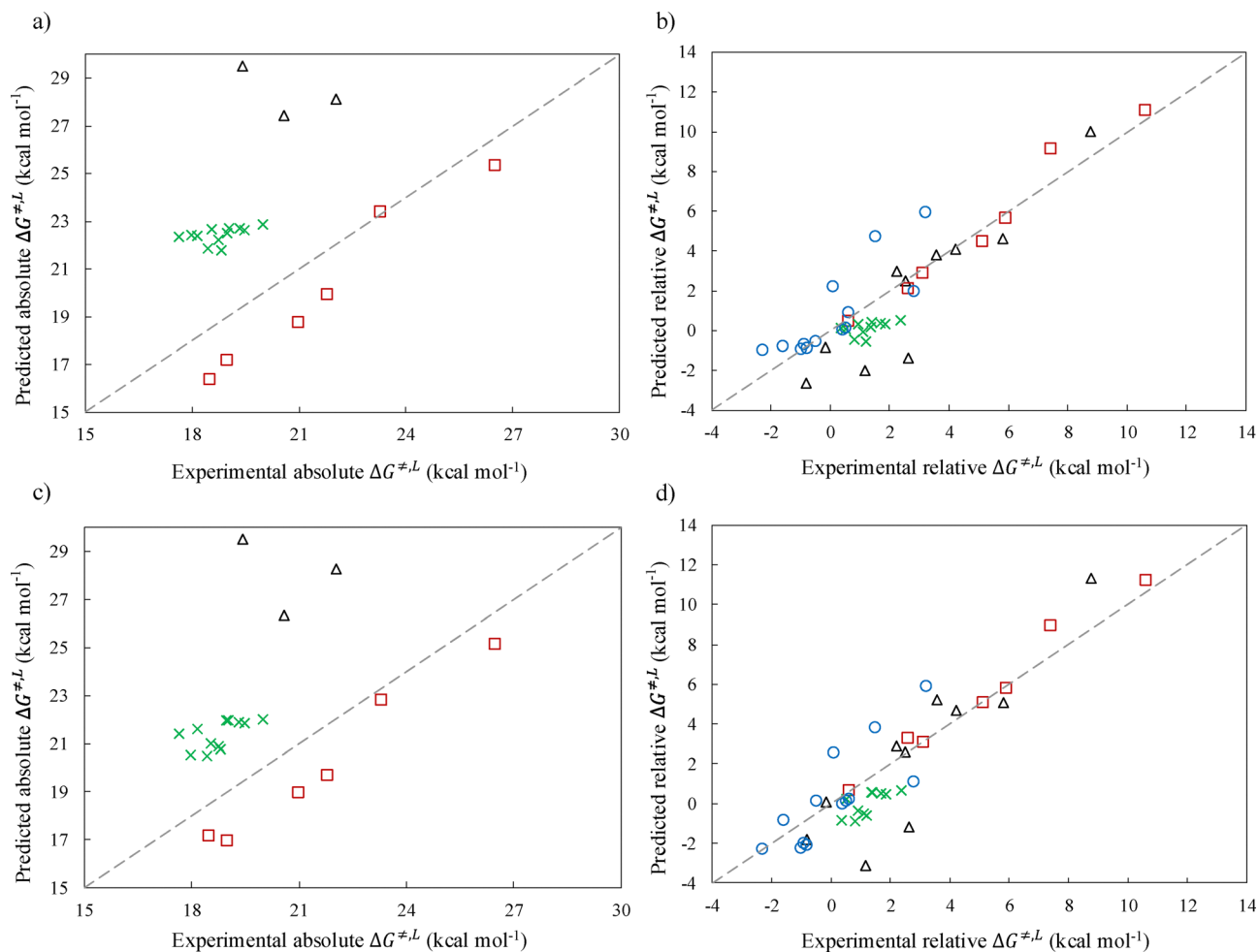


Fig. 8 Mean absolute error for each method/level of theory combination for 41 relative  $\Delta^\ddagger G^{\circ,L}$  predictions (all 46 reaction-solvent pairs with 5 reference pairs excluded). The red dashed line corresponds to 1 kcal mol<sup>-1</sup>.





**Fig. 9** Parity plots of predicted vs. experimental liquid-phase activation Gibbs free energy,  $\Delta^\ddagger G^{\circ,L}$ , using the TC method + M062X/G3MP2: (a) absolute and (b) relative, and the direct method + M062X/G3MP2: (c) absolute and (d) relative. Black triangles represent  $S_N2$  reactions; red squares represent Diels–Alder reactions; green crosses represent aromatic nucleophilic substitution reactions; blue circles represent Menschutkin reactions. The grey dashed line denotes the  $y = x$  reference.

calculated  $\Delta^\ddagger G^{\circ,L}$  values corresponding to the DA reactions are closest to the experiments. The errors of the TC method can be decomposed into contributions from the gas-phase activation free energies and the solvation free energies. The composite methods employed in this work, such as G3MP2 and CBS-QB3, have been shown in benchmark studies to be accurate for gas-phase reaction barriers.<sup>58,59</sup> Therefore, the dominant source of error is more likely to lie in the solvation models. This is supported by the work of Ho<sup>25</sup> on the calculation of  $pK_a$  values and reduction potentials using the TC method with the SMD model, which showed that increasing the accuracy of the electronic structure method does not necessarily improve agreement with experiment. As noted in the description of the SMD parameterisation, the SMD model was parameterised using a dataset that is dominated by neutral solutes, with a smaller fraction of ionic solvation free energies. This imbalance generally leads to better performance for neutral systems, with reported MAEs of 0.6 to 1.0 kcal mol<sup>-1</sup> and 4 kcal mol<sup>-1</sup> for the solvation free energies of neutral and ionic species, respectively.<sup>26</sup> Given the potential propagation of errors from

the various solvation free energy terms entering the calculation of  $\Delta^\ddagger G^{\circ,L}$ , deviations of up to 7 kcal mol<sup>-1</sup> (Fig. 10a) for the  $S_N2$  reactions, in which both the reactants and the TS involve ionic species, are not abnormal. The slightly better performance of the direct method may arise because species with strong ionic character are subject to larger solvation-induced geometrical and vibrational changes, which are more accurately described by the direct method. This reaction-type dependence indicates that applying SMD to reactions involving charged and highly polar species is more challenging, and leading to the larger errors observed in the absolute free energies of activation of ionic reactions.

Nevertheless, we note that SMD can provide reasonably good performance in the predictions of relative  $\Delta^\ddagger G^{\circ,L}$  values for ionic reactions, as shown in the parity plots in Fig. 9b and d. In this case, the impact of systematic errors is reduced for both the TC and direct methods (as indicated by the distribution of the data points along the  $y = x$  line). The overall trends are predicted well across different types of reactions and solvents, although we note that the  $R^2$  value for the direct method (0.830)



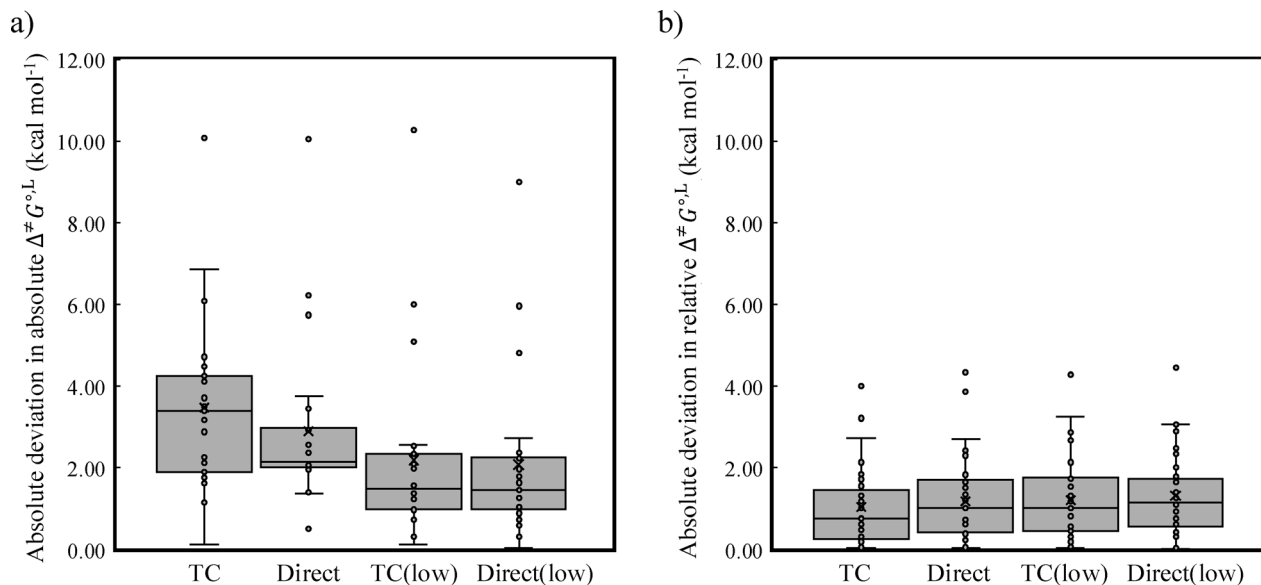


Fig. 10 Box plots of (a) absolute errors and (b) relative errors of  $\Delta^\ddagger G^{\circ,L}$  calculations using the TC method, the direct method and their counterparts using a low level of theory, M062X/6-31+G(d), for gas-phase activation free energy calculations.

is slight higher than that for the TC method (0.815), in contrast to the MAE values (1.17 kcal mol<sup>-1</sup> vs. 1.02 kcal mol<sup>-1</sup>), suggesting that both methods perform comparably.

We further compare the performance of these two MLoTCs against their counterparts that utilise a low level of theory, M062X/6-31+G(d), in place of the high level of theory calculations, *i.e.*, the gas-phase G3MP2 calculations used in the TC method and the gas- and liquid-phase G3MP2 calculations in the direct method (Fig. 10). Interestingly, employing a high level of theory results in a larger error in the absolute liquid phase activation Gibbs free energies (Fig. 10a) but a smaller error in the relative values (Fig. 10b). This suggests that high levels of theory introduce more systematic error in the resulting activation free energy. This could potentially be corrected through calibration against a large experimental dataset. However, given the marginal improvements observed when using high levels of theory with the SMD model, employing a low level of theory may offer better balance between computational cost and accuracy.

## 4 Conclusions

In summary, we have introduced the LiPRED-2026 database, comprising 4513 liquid-phase activation Gibbs free energy values calculated for 28 reactions in a range of solvents using the SMD model and seven computational schemes with a range of levels of theory. A subset of the database has been employed to benchmark the accuracy of various MLoTCs in the prediction of both absolute and relative activation free energies. All methods considered perform well for relative activation free energies (or reaction rates). The TC method, in which  $\Delta^\ddagger G^{\circ,IG}$  and solvation energy are calculated separately, paired with the M062X/CBS-QB3 level of theory provides the best performance

in terms of relative  $\Delta^\ddagger G^{\circ,L}$ , with an MAE of 1.00 kcal mol<sup>-1</sup>. The combination of the direct method, in which all the calculations are performed in a liquid phase, with the M062X/G3MP2 level of theory has been found to be most effective for predicting absolute activation Gibbs free energies, albeit with an MAE of 2.89 kcal mol<sup>-1</sup>. The smaller MAE of relative  $\Delta^\ddagger G^{\circ,L}$  is mainly due to the cancellation of systematic error in absolute  $\Delta^\ddagger G^{\circ,L}$ . The TC-high method in conjunction with M062X/G3MP2 strikes a good balance between absolute (MAE: 2.99 kcal mol<sup>-1</sup>) and relative (MAE: 1.01 kcal mol<sup>-1</sup>) Gibbs free energy predictions. We have also investigated the need to use a high level of theory in both the TC and direct methods. While this improves the accuracy of relative energy predictions, it tends to result in larger errors in absolute energies. Therefore, when comparisons between different solvents are not critical and computational resources are limited, a low level of theory may be sufficient for general applications.

The current version of LiPRED-2026 can provide valuable training and test data for developing data-driven models that do not require large amounts of data, such as a linear free energy relationship or a small-scale feedforward neural network model, to predict the kinetics of common, practically significant organic reactions. Such data-lean models have previously been found to be highly effective in solvent selection.<sup>45,51</sup> For complex machine learning architectures, particularly those based on deep learning, LiPRED-2026 can support fine-tuning<sup>21</sup> and benchmarking, similar to the roles of the fine-tuning and test sets in the work of Chung and Green,<sup>21</sup> where the amount of data required is much smaller than that used for pretraining the model.

However, we acknowledge several limitations of the current LiPRED-2026 database that will be addressed in future expansions. First, a single conformer has been considered for each species in a given chemical environment, either in vacuum or



in a specific solvent. Although solvation-induced geometric relaxation is captured through the SMD continuum solvation model in methods that include vibrational corrections, such as TC-vib and the direct method, potential solvation-induced changes in conformer populations have not been taken into account in the current work. This may lead to errors for systems with substantial conformational flexibility, for which Boltzmann averaging in both the gas and liquid phases would offer a more rigorous treatment.

Second, although LiPRED-2026 contains 4513  $\Delta^\ddagger G^{\circ,L}$  values, the reaction set remains relatively limited. In its current form, LiPRED-2026 is intended mainly as a benchmark and training resource for  $\Delta^\ddagger G^{\circ,L}$  in the reaction classes currently represented. The 28 reactions currently included do not cover several common reaction classes, such as elimination and radical reactions. Despite its relatively narrow focus, the outcomes of our benchmarking study are consistent with general expectations and provide useful insights on model performance. This motivates future work to extend the benchmarking to other solvation models and to expand substantially both the number and the diversity of reactions included in the database. This latter goal will require the generation of larger sets of reliable experimental data for both absolute and relative  $\Delta^\ddagger G^{\circ,L}$  values.

## Author contributions

Lingfeng Gui: conceptualisation, methodology, software, testing, investigation, data curation, writing – original draft, visualisation. Alan Armstrong: conceptualisation, methodology, writing – review & editing, supervision, funding acquisition. Claire S. Adjiman: conceptualisation, methodology, testing, writing – review & editing, supervision, project administration, funding acquisition. Fareed Bhasha Sayyed: conceptualisation, writing – review & editing, supervision. Amparo Galindo: conceptualisation, methodology, testing, writing – review & editing, supervision, project administration, funding acquisition.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

Data underlying this article and not available in the references cited are provided in the Zenodo repository (<https://doi.org/10.5281/zenodo.8396100>) and in the supplementary information (SI) under a CC BY licence. The Zenodo repository includes the complete quantum mechanical data generated for the liquid-phase activation free energy calculations reported in this work, provided as an Excel file in the “chapter5” folder of the repository. Supplementary information: details of the conformer selection procedure used in this work. See DOI: <https://doi.org/10.1039/d6cp00088f>.

## Acknowledgements

Funding from Eli Lilly and Company and the UK EPSRC, through the PharmaSEL-Prosperty Programme (EP/T005556/1), is gratefully acknowledged. AG acknowledges the funding of a Research Chair by the Royal Academy of Engineering and Eli Lilly and Company (RCSRF1819/7/33). The Imperial College Research Computing Service (DOI: <https://doi.org/10.14469/hpc/2232>) is also gratefully acknowledged for providing support and resources for the quantum mechanical calculations. The authors thank Dr S. P. Kolis (Eli Lilly and Company) for helpful discussions.

## Notes and references

- 1 J. L. Finney and A. K. Soper, *Chem. Soc. Rev.*, 1994, **23**, 1–10.
- 2 M. Orozco and F. J. Luque, *Chem. Rev.*, 2000, **100**, 4187–4226.
- 3 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
- 4 P. Mark and L. Nilsson, *J. Phys. Chem. A*, 2001, **105**, 9954–9960.
- 5 C. J. Cramer and D. G. Truhlar, *Chem. Rev.*, 1999, **99**, 2161–2200.
- 6 S. Miertus, E. Scrocco and J. Tomasi, *Chem. Phys.*, 1981, **55**, 117–129.
- 7 C. J. Cramer and D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 760–768.
- 8 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
- 9 A. Klamt, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 699–709.
- 10 G. Brancato, N. Rega and V. Barone, *J. Chem. Phys.*, 2008, **128**, 144501.
- 11 M. S. Lee, F. R. Salsbury Jr. and M. A. Olson, *J. Comput. Chem.*, 2004, **25**, 1967–1978.
- 12 J. A. Manion, R. E. Huie, R. D. Levin, D. R. Burgess Jr., V. L. Orkin, W. Tsang, W. S. McGivern, J. W. Hudgens, V. D. Knyazev, D. B. Atkinson, E. Chai, A. M. Tereza, C.-Y. Lin, T. C. Allison, W. G. Mallard, F. Westley, J. T. Herron, R. F. Hampson and D. H. Frizzell, NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2015.09, 2015, <https://kinetics.nist.gov>.
- 13 M. R. McGillen, W. P. L. Carter, A. Mellouki, J. J. Orlando, B. Picquet-Varrault and T. J. Wallington, *Earth Syst. Sci. Data*, 2020, **12**, 1203–1216.
- 14 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
- 15 G. F. von Rudorff, S. N. Heinen, M. Bragato and O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045026.
- 16 C. A. Grambow, L. Pattanaik and W. H. Green, *J. Phys. Chem. Lett.*, 2020, **11**, 2992–2997.



- 17 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
- 18 K. Spiekermann, L. Pattanaik and W. H. Green, *Sci. Data*, 2022, **9**, 417.
- 19 V. K. Prasad, Z. Pei, S. Edelmann, A. Otero-de-la Roza and G. A. DiLabio, *J. Chem. Theory Comput.*, 2022, **18**, 151–166.
- 20 K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.
- 21 Y. Chung and W. H. Green, *Chem. Sci.*, 2024, **15**, 2410–2424.
- 22 N. Harms, C. Underkoffler and R. West, *ChemRxiv*, 2020, preprint, DOI: [10.26434/chemrxiv.13277870.v2](https://doi.org/10.26434/chemrxiv.13277870.v2).
- 23 Y. Chung and W. H. Green, *J. Phys. Chem. A*, 2023, **127**, 5637–5651.
- 24 J. Ho and M. Z. Ertem, *J. Phys. Chem. B*, 2016, **120**, 1319–1329.
- 25 J. Ho, *Phys. Chem. Chem. Phys.*, 2015, **17**, 2859–2868.
- 26 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
- 27 L. A. Curtiss, P. C. Redfern, K. Raghavachari, V. Rassolov and J. A. Pople, *J. Chem. Phys.*, 1999, **110**, 4703–4709.
- 28 J. A. Montgomery, Jr., M. J. Frisch, J. W. Ochterski and G. A. Petersson, *J. Chem. Phys.*, 1999, **110**, 2822–2827.
- 29 J. A. Montgomery, Jr., M. J. Frisch, J. W. Ochterski and G. A. Petersson, *J. Chem. Phys.*, 2000, **112**, 6532–6542.
- 30 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian~16 Revision C.01*, Gaussian Inc., Wallingford CT, 2016.
- 31 J. W. Ochterski, *Thermochemistry in Gaussian*, Gaussian, inc. technical report, 2000.
- 32 K. A. Peterson, D. Feller and D. A. Dixon, *Theor. Chem. Acc.*, 2012, **131**, 1079.
- 33 T. Froitzheim, M. Müller, A. Hansen and S. Grimme, *ChemRxiv*, 2025, preprint, DOI: [10.26434/chemrxiv-2025-bjxvt](https://doi.org/10.26434/chemrxiv-2025-bjxvt).
- 34 J. Ho, A. Klamt and M. L. Coote, *J. Phys. Chem. A*, 2010, **114**, 13442–13444.
- 35 R. F. Ribeiro, A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2011, **115**, 14556–14562.
- 36 S. Grimme, *Chem. – Eur. J.*, 2012, **18**, 9955–9964.
- 37 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.
- 38 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A*, 2021, **125**, 4039–4054.
- 39 J. Gorges, S. Grimme, A. Hansen and P. Pracht, *Phys. Chem. Chem. Phys.*, 2022, **24**, 12249–12259.
- 40 L. Wittmann, C. E. Selzer and S. Grimme, *Chem. Sci.*, 2025, **16**, 22976–22995.
- 41 G. Luchini, J. Alegre-Requena, I. Funes-Ardoiz and R. Paton, *F1000Research*, 2020, **9**, 1–14.
- 42 P. Winget, D. M. Dolney, D. J. Giesen, C. J. Cramer and D. G. Truhlar, *Minnesota Solvent Descriptor Database*, 2021, <https://comp.chem.umn.edu/solvation/mnsddb.pdf>.
- 43 E. Grant, Y. Pan, J. Richardson, J. R. Martinelli, A. Armstrong, A. Galindo and C. S. Adjiman, *13th International Symposium on Process Systems Engineering (PSE 2018)*, Elsevier, 2018, vol. 44, pp. 2437–2442.
- 44 L. Gui, Y. Yu, T. O. Oliyide, E. Sioungkrou, A. Armstrong, A. Galindo, F. B. Sayyed, S. P. Kolis and C. S. Adjiman, *Comput. Chem. Eng.*, 2023, **177**, 108345.
- 45 L. Gui, A. Armstrong, A. Galindo, F. B. Sayyed, S. P. Kolis and C. S. Adjiman, *Mol. Syst. Des. Eng.*, 2024, **9**, 1254–1274.
- 46 S.-Y. Tang, J. Shi and Q.-X. Guo, *Org. Biomol. Chem.*, 2012, **10**, 2673–2682.
- 47 P. M. E. Mancini, R. D. Martinez, L. R. Vottero and N. S. Nudelman, *J. Chem. Soc., Perkin Trans. 2*, 1984, 1133–1138.
- 48 A. D. McFarland, J. Y. Buser, M. C. Embry, C. B. Held and S. P. Kolis, *Org. Process Res. Dev.*, 2019, **23**, 2099–2105.
- 49 L. Gui, C. S. Adjiman, A. Galindo, F. B. Sayyed, S. P. Kolis and A. Armstrong, *Ind. Eng. Chem. Res.*, 2023, **62**, 874–880.
- 50 M. H. Abraham, *J. Chem. Soc. D*, 1969, 1307–1308.
- 51 H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Sioungkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, *Nat. Chem.*, 2013, **5**, 952–957.
- 52 A. Diamanti, Z. Ganase, E. Grant, A. Armstrong, P. M. Piccione, A. M. Rea, J. Richardson, A. Galindo and C. S. Adjiman, *React. Chem. Eng.*, 2021, **6**, 1195–1211.
- 53 T. J. Sheldon, C. S. Adjiman and J. Cordiner, *Fluid Phase Equilib.*, 2005, **231**, 27–37.
- 54 P. R. Rablen, B. D. McLarney, B. J. Karlow and J. E. Schneider, *J. Org. Chem.*, 2014, **79**, 867–879.
- 55 A. Meijer, S. Otto and J. B. F. N. Engberts, *J. Org. Chem.*, 1998, **63**, 8989–8994.
- 56 T. Rispens and J. B. F. N. Engberts, *J. Org. Chem.*, 2002, **67**, 7369–7377.
- 57 J. Pfaendtner, X. Yu and L. J. Broadbelt, *Theor. Chem. Acc.*, 2007, **118**, 881–898.
- 58 Y. Lan, L. Zou, Y. Cao and K. N. Houk, *J. Phys. Chem. A*, 2011, **115**, 13906–13920.
- 59 L. A. Curtiss, P. C. Redfern and K. Raghavachari, *Chem. Phys. Lett.*, 2010, **499**, 168–172.

