Front. Chem. Sci. Eng. 2022, 16(2): 152–167
https://doi.org/10.1007/s11705-021-2060-z

RESEARCH ARTICLE

A computational toolbox for molecular property prediction based on quantum mechanics and quantitative structure-property relationship

Qilei Liu, Yinke Jiang, Lei Zhang (✉), Jian Du

Institute of Chemical Process Systems Engineering, School of Chemical Engineering, Dalian University of Technology, Dalian 116024, China

Abstract Chemical industry is always seeking opportunities to efficiently and economically convert raw materials to commodity chemicals and higher value-added chemical-based products. The life cycles of chemical products involve the procedures of conceptual product designs, experimental investigations, sustainable manufactures through appropriate chemical processes and waste disposals. During these periods, one of the most important keys is the molecular property prediction models associating molecular structures with product properties. In this paper, a framework combining quantum mechanics and quantitative structure-property relationship is established for fast molecular property predictions, such as activity coefficient, and so forth. The workflow of framework consists of three steps. In the first step, a database is created for collections of basic molecular information; in the second step, quantum mechanics-based calculations are performed to predict quantum mechanics-based/derived molecular properties (pseudo experimental data), which are stored in a database and further provided for the developments of quantitative structure-property relationship methods for fast predictions of properties in the third step. The whole framework has been carried out within a molecular property prediction toolbox. Two case studies highlighting different aspects of the toolbox involving the predictions of heats of reaction and solid-liquid phase equilibriums are presented.

Keywords molecular property, quantum mechanics, quantitative structure-property relationship, heat of reaction, solid-liquid phase equilibrium

Received November 20, 2020; accepted March 16, 2021

E-mail: keleiz@dlut.edu.cn

1 Introduction

The molecular property plays an important role in life cycles of chemical products, for example, the conceptual designs of chemical-based products and their manufacturing processes. How to fast and accurately obtain the property data of interested molecules is a persistent research topic. The most often reliable solution is to measure the molecular properties by experiments. However, it is practically infeasible to perform thousands of costly and time-consuming experiments for all compounds, as the chemical space is still an unexplored galaxy with more than 10^{60} molecules [1].

An alternative solution is to develop property prediction models for reasonable and fast predictions of molecular properties through quantitative structure-property relationship (QSPR) methods [2] based on data and/or knowledge. QSPR has been developed and used for nearly 140 years since Mills probably first developed QSPR methods for predictions of melting and boiling points in 1884 [3]. Dearden et al. presented a guidance on how to develop reasonable QSPR methods, where 21 types of common errors are discussed with examples [4]. In general, a complete procedure for QSPR developments consists of four steps: (1) prepare samples with the target properties (dependent variables); (2) select a number of appropriate descriptors (independent variables); (3) determine a specific algorithm (e.g., least square algorithm) to establish a linear or nonlinear model associating the target properties with descriptors; (4) using evaluation criteria and/or validation samples to evaluate the developed QSPR methods. Although QSPR methods cannot completely replace the experiments, they are able to quickly identify promising molecular candidates which could then be verified through focused experiments.

Up to now, a number of toolboxes has been developed for QSPR (or quantitative structure-activity relationship

(QSAR)) methods, as shown in Table 1. Kim and Cho proposed an integrated standalone python package (PyQSAR) that combines all QSAR modeling process in one workbench [5]. PyQSAR is devoted to developing reliable QSAR models on a single platform with an easy-to-follow workflow. Enciso et al. provided an easy-to-use open-source software platform (BioPPSy) for QSPR/QSAR modelling [6]. Three key biochemical properties used in drug discovery are presented to demonstrate the program capabilities. Pirhadi et al. [7] had reviewed a few open-source toolboxes for QSPR/QSAR modelling, such as AZOrange [8], Bioalerts [9], camb [10], eTOXlab [11] and Open3DQSAR [12]. Although above toolboxes are open-source and powerful, the parameters in their property prediction models are all fitted by external supports of (pseudo) experimental data. The limited number or even lack of (pseudo) experimental data will result in unreasonable or failed predictions for molecular properties. Dimitrov et al. [13] developed an OECD QSAR toolbox for governments, where missing (pseudo) experimental data are supported through a read-across method. For molecules without property data, the read-across method is able to fill the data gap through searching available property data of similar molecules belonging to the same category. However, the reliability of data provided by the read-across method may be low.

With the developments of computational chemistry, molecular simulation techniques are receiving increasing attention among the community of molecular property prediction. Compared with experiments and QSPR methods, these calculation approaches focus on the microscopic information of molecular systems (e.g., electron, atom, bond length, bond angle, dihedral angle, etc.) based on highly recognized theories including molecular mechanics (MM), quantum mechanics (QM), and so forth. Therefore, these methods are not strongly dependent on (pseudo) experimental data.

MM is the simplest and fastest way to evaluate molecular systems. Based on a set of parameterized empirical potential energy functions, it allows efficient simulation of large biological systems or material assemblies with thousands of atoms [14]. As one of the most popular MM-based molecular simulation techniques, molecular dynamics (MD) is generally used for modeling the time-dependent motions (trajectories) of macromolecules (biological proteins, polymers, crystals, and so forth) [15]. The simulation of the motion is achieved by the numerical solution of the classical Newtonian dynamic equations [16]. There are several free program packages for MD simulation, for example, Amber [17], CHARMM [18], GROMACS [19], LAMMPS [20], NAMD [21], and so forth. Note that MM-based methods (e.g., MD methods) may fail to make predictions of properties with acceptable accuracy if appropriate force field parameters are missing in the simulation systems.

Compared with MM-based methods, QM-based methods including rigorous QM methods (e.g., Hartree-Fock (HF), density functional theory (DFT), etc.) and semi-empirical QM methods (e.g., PM7, etc.) predict molecular properties by solving the Schrödinger equations without any force field parameter. The prediction accuracy of QM methods, especially for rigorous QM, for molecular properties is superior to MM-based methods due its consideration of electronic impacts, not simply nucleic impacts, on molecular systems. However, the computational cost of QM methods, especially for rigorous QM, is more intensive compared with MM-based methods and increases exponentially with the system size of molecules. Macromolecules are hard to be modeled by rigorous QM methods. If super classical computers and linear scaling algorithms (exist minor accuracy loss) are supported [22], it may be possible to perform some QM tasks like single point energy for macromolecules. QM has also been applied in some commonly used software tools, for example, Gaussian [23], ORCA [24], GAMESS [25], MOPAC [26], and so forth. Gaussian is the most popular commercial software in computational chemistry area. It has comprehensive functions for property predictions and enjoys great reputations among the community of molecular simulation. ORCA is a free software which has experienced rapid developments these years. With the chain-of-spheres exchange algorithm [27], ORCA has greatly reduced the computational expense of rigorous QM methods at a cost of minor accuracy loss for calculations of weak interactions for large molecular systems. GAMESS has received great attentions for its open-source character-

Table 1 Toolboxes developed for QSPR/QSAR methods

Name	Open-source	Toolbox provides (pseudo) experimental data by itself	Applicability domain
PyQSAR [5]	Yes	No	Biochemistry and chemical engineering
BioPPSy [6]	Yes	No	Biochemistry
AZOrange [8]	Yes	No	Biochemistry, chemical engineering and pharmacy
Bioalerts [9]	Yes	No	Biochemistry and pharmacy
Camb [10]	Yes	No	Biochemistry
eTOXlab [11]	Yes	No	Biochemistry
Open3DQSAR [12]	Yes	No	Pharmacy
OECD QSAR Toolbox [13]	No	Yes (read-across method)	Biochemistry, chemical engineering and pharmacy

istic. MOPAC is a free semiempirical QM software with efficient calculation speed for large molecules, even for macromolecules, at the expense of precision. Semiempirical QM methods are faster than linear scaling algorithm-based rigorous QM methods, while giving more accuracy loss.

To sum up, on one hand, although QSPR toolboxes are able to develop property prediction models and make fast predictions for molecular properties, they need reliable methods to provide (pseudo) experimental data for QSPR developments; on the other hand, QM methods are free from dependence on (pseudo) experimental data and force field parameters compared with QSPR and MM-based methods, respectively. They are capable of making accurate property predictions. However, QM methods cost extremely high computational expense, which does not meet the requirements of high-throughput property predictions (seconds per molecule). Thus, it is desirable to develop shortcut methods to accelerate QM calculations with minor accuracy loss. In this work, a QM-QSPR framework combining QM calculations and QSPR methods is established for fast predictions of molecular properties. In section 2, the framework is established through three steps including database establishment, QM calculation and QSPR development, where QSPR methods are used to accelerate QM calculations through linear/nonlinear models, while QM calculations are used to provide accurate pseudo experimental data for QSPR developments. Note that the established QM-QSPR framework is more recommendable for molecular properties that suffer from a limited experimental data. In section 3, the whole QSPR framework is carried out within a molecular property prediction toolbox called "QM-QSPR". The software architecture of QM-QSPR with its dataflow and workflow is systematically introduced. In section 4, the application of the QM-QSPR framework and toolbox is highlighted through two case studies involving the predictions of heats of reaction and solid-liquid phase equilibriums.

2 The QM-QSPR framework

The diagrammatic sketch of QM-QSPR framework is shown in Fig. 1, which contains three steps including database establishment, QM calculation and QSPR development. More details about this framework are given in the following text.

2.1 Database establishment

It is necessary to establish a database before QM calculations and QSPR developments. Many molecular databases are available online, such as PubChem, DrugBank, ProCAPD database, NIST Chemistry webbook, ZINC, ChEMBL, and so forth. When focusing on different categories of molecules, different databases are required, for examples, solvents, drugs, refrigerants, ionic liquids, adsorbents, and many more. Thus, the molecules with their basic molecular information (e.g., chemical abstracts service (CAS) number, simplified molecular-input line-entry system (SMILES), etc.) should be carefully selected in the database to develop a balanced tailor-made QSPR method between generalization and accuracy for one or multiple fixed categories of molecules. Note that macromolecules are not considered in this paper due to the computational cost. Conformational isomers are not taken into account due to the lack of conformation search algorithm in the current QM-QSPR framework.

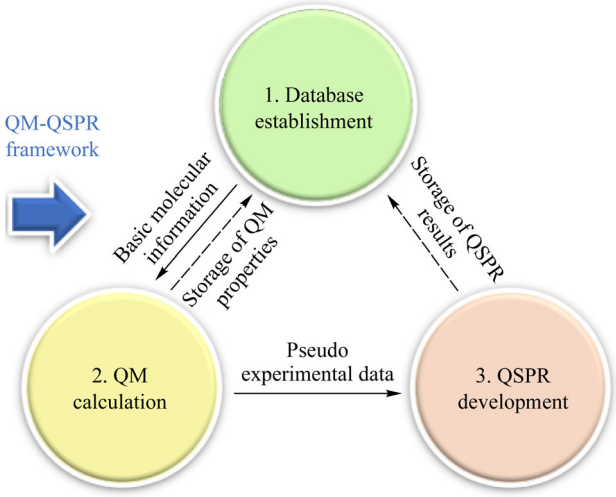

Fig. 1 The diagrammatic sketch of QM-QSPR framework.

2.2 QM calculation

After the database is established, the QM calculation step is carried out to provide pseudo experimental data for molecular properties. Starting from the basic molecular information in the database, the stereoscopic representations of molecules (e.g., cartesian coordinates) are obtained through either on-line (e.g., application program interfaces (APIs) in website database using CAS number) or off-line tools (e.g., OpenBabel software [28] using SMILES). Before predicting properties through QM calculations, geometry optimizations need to be performed first for molecules to find their stable structures. Note that the prediction accuracy and computational cost with regard to the geometry optimizations and other QM functional calculations (e.g., single point energy calculation, frequency analysis, etc.) are greatly dependent on the QM methods. Therefore, it is essential to select appropriate QM methods for specific problems. Three main QM methods are introduced here [29]:

(1) HF is one of the main QM methods to obtain the

numerical solution of the Schrödinger equations through wave functions [30]. Modern post-HF methods are able to provide the highest accuracy with the largest computational expense [31]. For example, using the coupled cluster method CCSD(T) and large basis sets (often called the "gold standard" in QM area [32]) for energy predictions, mean unsigned errors of about 2 $kJ \cdot mol^{-1}$ are identified for small molecules with 1 or 2 non-hydrogen atoms [31].

(2) DFT is another popular QM method to solve the Schrödinger equations using the distribution of electron density [33]. Compared with post-HF methods, it has a faster prediction speed, though lower precision [34]. For example, mean unsigned errors of about 8–24 $kJ \cdot mol^{-1}$ are identified in a typical DFT method of B3LYP, which is, however, 10^3 to 10^4 faster than CCSD(T) [31].

(3) Semiempirical QM methods, for example, PM7, PM6, and forth, significantly reduce accuracy, but greatly improve efficiency compared with post-HF and DFT methods. For instance, unsigned errors of 120 $kJ \cdot mol^{-1}$ are identified in semiempirical methods while their expense is found lower by factors around 10^8 compared with CCSD (T) [31].

Next, for QM predicted properties, they are discussed in terms of two types as in the following text, including the QM-based property and QM-derived property.

2.2.1 QM-based property

The QM-based property is defined as the molecular property being directly predicted with the basic molecular information through QM calculations. For examples, enthalpy H, free energy G, dipole moment μ, energy of HOMO ε_{HOMO}, energy of LUMO ε_{LUMO}, etc., are regarded as QM-based properties. Three commonly used thermodynamic properties are discussed in detail as examples in the following text, including standard enthalpy of formation at ideal gas state $\Delta_f H_g^\theta$, standard Gibbs free energy of formation at ideal gas state $\Delta_f G_g^\theta$ and standard entropy at ideal gas state S_g^θ.

2.2.1.1 Standard enthalpy of formation at ideal gas state

$\Delta_f H_g^\theta$ is one of the key properties in thermochemistry. The $\Delta_f H_g^\theta$ of a compound is defined as the enthalpy change from its elements in their most stable states at a pressure of 1 bar and the specific temperature of 298.15 K [35]. Thus, the general way to obtain $\Delta_f H_g^\theta$ is to calculate H_g^θ for the target molecule and its elements first, and then to add the experimental standard enthalpy of vaporization ΔH_{vap}^θ or sublimation ΔH_{sub}^θ for the elements that are liquids or solids [35], respectively. Here is an example of benzaldehyde for calculating its $\Delta_f H_g^\theta$ The generation route is shown as follows.

$$7C(s,graphite) + 3H_2(g) + \frac{1}{2}O_2(g) \xrightarrow{7\Delta H_{sub}^\theta(graphite)} 7C(g)$$

$$+ 3H_2(g) + \frac{1}{2}O_2(g) \rightarrow C_7H_6O_2(g)$$

Based on the generation route, $\Delta_f H_g^\theta(C_7H_6O_2) = H_g^\theta(C_7H_6O_2) - 3H_g^\theta(H_2) - \frac{1}{2}H_g^\theta(O_2) - 7H_g^\theta(C) + 7\Delta H_{sub}^\theta$ (graphite). The H_g^θ for each compound is obtained through QM calculations using appropriate QM methods. Some commonly used QM methods for predictions of thermodynamic properties are listed in Table A1 (cf. Electronic Supplementary Material, ESM). The experimental ΔH_{vap}^θ and ΔH_{sub}^θ for liquid and solid elements (being in their most stable states at 1 bar and 298.15 K) are collected from database (in this paper, the database refers to the Lange's handbook of chemistry [36] and NIST Chemistry Webbook by default). Normally, the spin multiplicity of compounds is set to 1, while the elements need selecting appropriate values of spin multiplicity when performing QM calculations for their H_g^θ. The experimental ΔH_{vap}^θ, ΔH_{sub}^θ and recommended spin multiplicity for elements are listed in Table A2 (cf. ESM).

2.2.1.2 Standard Gibbs free energy of formation at ideal gas state

Similar to $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ is obtained through Gibbs free energy change from its elements. The experimental ΔG_{vap}^θ and ΔG_{sub}^θ for elements are calculated through $\Delta G_{vap}^\theta = \Delta H_{vap}^\theta - T\Delta S_{vap}^\theta$ and $\Delta G_{sub}^\theta = \Delta H_{sub}^\theta - T\Delta S_{sub}^\theta$, respectively, where ΔS_{vap}^θ and ΔS_{sub}^θ are experimental data collected from database. The experimental ΔG_{vap}^θ, ΔG_{sub}^θ, ΔS_{vap}^θ and ΔS_{sub}^θ for liquid and solid elements are listed in Table A2 in Appendix A (cf. ESM).

2.2.1.3 Standard entropy at ideal gas state

S_g^θ is calculated through Eq. (1),

$$S_g^\theta = \frac{H_g^\theta - G_g^\theta}{T}, \quad (1)$$

where H_g^θ and G_g^θ are predicted by QM calculations, T represents temperature.

2.2.2 QM-derived property

In contrast to the QM-based property, the QM-derived property is defined as the molecular property being predicted with the QM-based property through quantitative property-property relationship (QPPR) models (the model

correlating QM-based property and QM-derived property). Note that the prediction accuracy of QM-derived property is strongly dependent on the QM methods, as the model parameters in QPPR are fitted in advance using several samples with a fixed QM method. Two key thermodynamic properties, standard enthalpy of vaporization $\Delta H^{\theta}_{\mathrm{vap}}$ and activity coefficient γ, are discussed as examples in the following text.

2.2.2.1 Standard enthalpy of vaporization

Standard enthalpy of formation at liquid state $\Delta_{\mathrm{f}}H^{\theta}_{1}$ is often needed in practice and it is obtained through Eq. (2),

$$\Delta_{\mathrm{f}}H^{\theta}_{1} = \Delta_{\mathrm{f}}H^{\theta}_{\mathrm{g}} - \Delta H^{\theta}_{\mathrm{vap}}, \tag{2}$$

where $\Delta H^{\theta}_{\mathrm{vap}}$ is the standard enthalpy of vaporization and its experimental data may not be available for the compound of interest. Based on the knowledge that $\Delta H^{\theta}_{\mathrm{vap}}$ is dependent on noncovalent interactions and predicted with good accuracy from quantitative features of the computed potentials on the molecular surfaces [35], Politzer et al. [35] developed a QPPR model to predict $\Delta H^{\theta}_{\mathrm{vap}}$ at 298.15 K with the characteristics of QM-based properties of surface potential $V_{\mathrm{S}}(\mathbf{r})$ and molecular surface area A_{S} The formula of the developed QPPR model with parameters regressed from experimental data is shown in Eq. (3) [35],

$$\Delta H^{\theta}_{\mathrm{vap}}(298.15\ \mathrm{K}) = 1.3556A_{\mathrm{S}}^{0.5} + 1.1760(v\sigma^{2}_{\mathrm{tot}})^{0.5} - 10.4331, \tag{3}$$

in which $\sigma^{2}_{\mathrm{tot}}$ represents the total variance of $V_{\mathrm{S}}(\mathbf{r})$ and v is the electrostatic balance of $V_{\mathrm{S}}(\mathbf{r})$. To guarantee the prediction accuracy of the QM-derived property, the QM method "B3PW91/6-31G(d,p)" is selected for geometry optimizations and single point energy calculations to obtain the necessary $\sigma^{2}_{\mathrm{tot}}$, v and A_{S}. More details about the formula derivations, calculation procedures, selected samples and model performances are given in Politzer et al. [35].

2.2.2.2 Activity coefficient

Activity coefficient γ is a key mixture property in separation and reaction unit operations [37]. UNIFAC models are often employed to predict γ. However, a large set of measured data are required for the regression of group parameters for UNIFAC, which hinders the applications of UNIFAC models to molecules with missing group parameters. Here, a QPPR model named conductor like screening model for segment activity coefficient (COSMO-SAC) [38] is employed to predict γ with the characteristics of QM-based properties of surface charge q and molecular cavity volume V_{C}. The general formula of the developed QPPR (COSMO-SAC) model with a small set of group-independent parameters regressed from experimental data is described as the following Eq. (4),

$$\gamma = f(p(\sigma), V_{\mathrm{C}}), \tag{4}$$

where $p(\sigma)$ is the surface charge density profiles derived from q. To guarantee the prediction accuracy of the QM-derived property, the QM method "B3LYP/6-31G(d,p)" is selected for geometry optimizations and COSMO calculations to obtain the necessary $p(\sigma)$ and V_{C}. More details about the formula derivations, calculation procedures, model parameters and model performance are given in Chen et al. [39].

2.3 QSPR development

The QM calculation is a reliable alternative to the experiment for predictions of molecular properties with acceptable accuracy. However, the costly QM calculations prevent the high-throughput design/screen of promising molecules with regard to target properties. Thus, it is desirable to develop QSPR methods (shortcut methods) with a number of pseudo experimental data generated from QM to accelerate the QM calculations. QSPR combines computer science and mathematical method to investigate the correlations between the physical, chemical, biological properties of compounds and their molecular structures. It assumes that the properties of compounds are expressed as a linear or nonlinear function of chemical structures. In this way, the QM computational cost is significantly shortened to fast identify molecules with target properties. In this paper, QSPR methods are divided into two types: linear QSPR method and nonlinear QSPR method.

2.3.1 Linear QSPR method

Linear QSPR method is generally performed through the least square method. One of the most popular linear QSPR methods in chemical areas is the group contribution (GC) method, which assumes that the functional groups share the same property contributions among all compounds and the property value of each compound is a summation of all GCs involved in this molecule [40]. The general model formula of the GC method is described in Eq. (5),

$$f(X) = \sum\nolimits_{i} N_i C_i, \tag{5}$$

where $f(X)$ is a function of property X and it may contain additional adjustable model parameters (universal constants) depending on the property involved, C_i is the contribution of the group of type-i that occurs N_i times [41]. Note that the GC method is essentially a knowledge-

based QSPR method (or in other words, semi-empirical QSPR method). Generally, external validation is an option in such a knowledge-based QSPR method as the knowledge ensures its extrapolation. Also, it is possible to improve the predictive capability and application range of the GC-based property model through including all of the available experimental data of the property in the regression [42].

2.3.2 Nonlinear QSPR method

One of the most popular nonlinear QSPR methods is the back propagation-artificial neural network (BP-ANN) method [43], the model of which could be built/trained by following Eqs. (6) and (7),

$$\min f_{\text{loss}}(\mathbf{p}^{\text{pre}},\mathbf{p}^{\text{tar}}), \tag{6}$$

$$\mathbf{p}^{\text{pre}} = F_{\text{BP-ANN}}(\mathbf{D},\mathbf{P}), \tag{7}$$

where f_{loss} is the loss function to measure the differences between the target outputs $\mathbf{p}^{\text{tar}}$ and the prediction outputs $\mathbf{p}^{\text{pre}}$, $\mathbf{P}$ is the set of parameters (e.g., weights and biases) and hyperparameters (e.g., hidden layers), $\mathbf{D}$ is the input dataset, and $F_{\text{BP-ANN}}$ is the optimization function that helps the BP-ANN model to correlate the inputs and outputs. Note that an external validation is compulsory in the BP-ANN method because it is a data-based QSPR method (or in other words, empirical QSPR method). In such method, external validation must be given to avoid overfitting issues.

3 QM-QSPR: software architecture

The workflow of the QM-QSPR framework with its associated dataflow and computer aided tools is integrated into the "QM-QSPR" toolbox, which provides the additional option of QSPR method for molecular design in OptCAMD software [44]. The software architecture of QM-QSPR is shown in Fig. 2. Note that the current QM-QSPR toolbox needs the paid prerequisite software (Gaussian software [23]), while other prerequisite tools (e.g., OpenBabel [28], Multiwfn [45,46], etc.) are all free.

The QM calculation section in the QM-QSPR architecture consists of four steps: (a) user interface, (b) pre-processing, (c) execution and (d) downstream applications. In the user interface step, a set of molecules represented as CAS or SMILES from the established database are given by users and input through user interface, which is written in python [47]. In the pre-processing step, the website API script or OpenBabel [28] software is employed to transform the CAS or SMILES to three-dimensional cartesian coordinates (".gjf" input files). Then, in the execution step, based on the stereoscopic molecular information, geometry optimizations are performed to find the stable molecular structures through appropriate QM methods in Gaussian software [23] prior to other QM functional calculations (e.g., single point energy calculation, frequency analysis, etc.). During this period, users can select additional instructions (e.g., solvation model, pseudopotential, etc.) for the execution step. The execution results are stored in the ".log" output files. In the downstream application step, on one hand, QM-based properties are directly obtained from the output files; on the other hand, QM-derived properties are predicted with the intermediate data (σ^2_{tot}, $p(\sigma)$, etc.) post-processed from the QM-based properties (e.g., $V_{\text{S}}(\mathbf{r})$, q, etc.) through additional tools (e.g., Multiwfn [45,46], etc.) or directly with the QM-based properties according to the QPPR models.

All properties calculated from the QM calculation section (QM-based/derived properties) are stored in database and further used to develop QSPR methods, the results of which are also stored in database and provide high-throughput property predictions for other applications (e.g., OptCAMD for molecular design [44]).

4 Case studies

Two case studies highlighting the use of QM-QSPR computational toolbox, in which one involving the predictions of heats of reaction and another for the predictions of solid-liquid phase equilibriums, are presented below.

4.1 Prediction of heat of reaction

The goal of case study 1 is to predict heat of reaction. The heat of reaction (also known as enthalpy of reaction) $\Delta_{\text{r}}H^{\theta}_{\text{state}}(T)$ is the change in the enthalpy of a chemical reaction that occurs at a constant pressure. $\Delta_{\text{r}}H^{\theta}_{\text{state}}(T)$ is an important criterion for reactions and is described as Eq. (8) [48]:

$$\begin{aligned}\Delta_{\text{r}}H^{\theta}_{\text{state}}(T) &\approx \Delta_{\text{r}}H^{\theta}_{\text{state}}(298.15\ \text{K}) \\ &= \sum_j \nu_j \Delta_{\text{f}}H^{\theta}_{\text{state},j}(298.15\ \text{K}),\end{aligned} \tag{8}$$

where $\Delta_{\text{r}}H^{\theta}_{\text{state}}(T)$ approximates to $\Delta_{\text{r}}H^{\theta}_{\text{state}}(298.15\ \text{K})$ over a modest range of temperature T [49], $\Delta_{\text{f}}H^{\theta}_{\text{state},j}(298.15\ \text{K})$ represents the standard enthalpy of formation at 298.15 K for compound j, ν_j is the stoichiometric coefficient of compound j, state represents gas (g) or liquid (l) state, θ represents the standard state. In the following, the subscript j will be simplified among properties if unnecessary.

To calculate heat of reaction, predictions for $\Delta_{\text{f}}H^{\theta}_{\text{state}}(298.15\ \text{K})$ are needed. As introduced in section 2, the QM-based property $\Delta_{\text{f}}H^{\theta}_{\text{g}}(298.15\ \text{K})$ is directly predicted by QM calculations. For $\Delta_{\text{f}}H^{\theta}_{\text{l}}(298.15\ \text{K})$, it is calculated

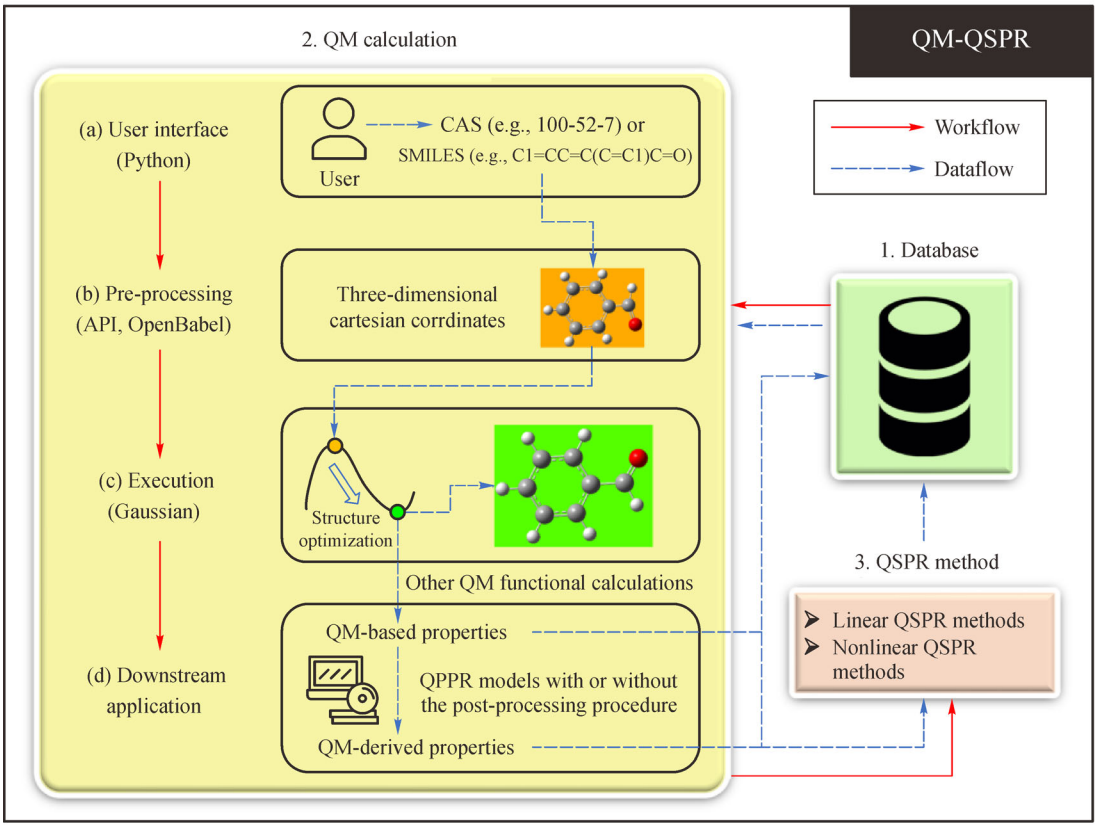

Fig. 2 The software architecture of QM-QSPR with its dataflow and workflow.

through Eq. (2), where the QM-derived property $\Delta H^{\theta}_{\mathrm{vap}}(298.15\ \mathrm{K})$ is calculated through Eq. (3).

Besides $\Delta_{\mathrm{f}}H^{\theta}_{\mathrm{g}}$, $\Delta_{\mathrm{f}}H^{\theta}_{\mathrm{l}}$ and $\Delta H^{\theta}_{\mathrm{vap}}$ at 298.15 K are considered, other thermodynamic properties $\Delta_{\mathrm{f}}G^{\theta}_{\mathrm{g}}$, $\Delta_{\mathrm{f}}G^{\theta}_{\mathrm{l}}$, S^{θ}_{g}, S^{θ}_{l}, $\Delta G^{\theta}_{\mathrm{vap}}$ and $\Delta S^{\theta}_{\mathrm{vap}}$ at 298.15 K are also taken into account in this case study because another important reaction property, reaction equilibrium constant $K^{\theta}_{\mathrm{state}}(T)$, is often needed in some problems and calculated through $K^{\theta}_{\mathrm{state}}(T) = \exp(-\Delta_{\mathrm{r}}G^{\theta}_{\mathrm{state}}(T)/\mathrm{R}T)$ and $\Delta_{\mathrm{r}}G^{\theta}_{\mathrm{state}}(298.15\,\mathrm{K}) = \sum_j \nu_j \Delta_{\mathrm{f}}G^{\theta}_{\mathrm{state},j}(298.15\ \mathrm{K})$ (or $\Delta_{\mathrm{r}}G^{\theta}_{\mathrm{state}}(T \neq 298.15\ \mathrm{K}) \approx \sum_j \nu_j \Delta_{\mathrm{f}}H^{\theta}_{\mathrm{state},j}(298.15\ \mathrm{K}) - T\sum_j \nu_j S^{\theta}_{\mathrm{state},j}(298.15\ \mathrm{K})$) [48], where $\Delta_{\mathrm{f}}G^{\theta}_{\mathrm{g}}$, $\Delta_{\mathrm{f}}G^{\theta}_{\mathrm{l}}$, S^{θ}_{g}, S^{θ}_{l}, $\Delta G^{\theta}_{\mathrm{vap}}$ and $\Delta S^{\theta}_{\mathrm{vap}}$ at 298.15 K are indispensable for calculations of $K^{\theta}_{\mathrm{state}}(T)$.

However, the QM calculations for the above properties are computationally cost. The GC methods are usually used to fast predict thermodynamic properties for compounds [42]. To accelerate the QM calculations, this case study employs the QM-QSPR toolbox to develop new GC methods regressed from the QM calculated data for fast predictions of $\Delta_{\mathrm{f}}H^{\theta}_{\mathrm{state}}$, $\Delta_{\mathrm{f}}G^{\theta}_{\mathrm{state}}$, $S^{\theta}_{\mathrm{state}}$, $\Delta H^{\theta}_{\mathrm{vap}}$ and $\Delta G^{\theta}_{\mathrm{vap}}$ at 298.15 K based on drug and solvent database. The workflow of case study 1 is shown in Fig. 3.

4.1.1 Database establishment

A database is created consisting of 2859 neutral molecules with CAS numbers and Canonical SMILES, where 1956 drug molecules are screened from the DrugBank database using the Lipinski's "Rule of Five" and 903 solvent molecules not repeated in drug molecules are selected from the Virginia Tech database [50]. Compared with the solvent molecules in Virginia Tech database, the drug molecules in DrugBank database are more structurally complex and their molecular weights are generally larger. Selecting drug database is able to enlarge the applicability domain of new developed GC methods since the group values of existing GC methods to thermodynamic properties regressed from experimental data are usually blank among structurally complex molecules (e.g., drugs) [42].

4.1.2 QM calculation

Before predicting properties, the stereoscopic representations of all 2859 molecules with cartesian coordinates are obtained from PubChem using CAS numbers through the website API script in the QM-QSPR toolbox.

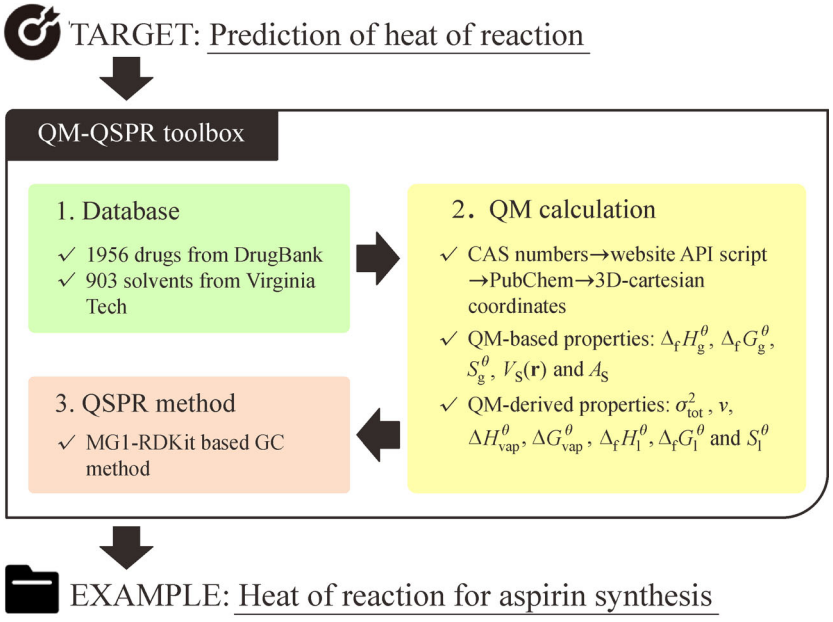

Fig. 3 The workflow of case study 1.

4.1.2.1 Standard thermodynamic property at ideal gas state

The unpackaged hybrid QM method in Table A1 in Appendix A (cf. ESM) is employed to predict $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ and S_g^θ at 298.15 K with the Gaussian software [23], where geometry optimization and frequency analysis are performed using "B3LYP/6-31G(d)" and single point energy calculation is carried out with "M062X/def2TZVP em = GD3". In this way, a balance between prediction accuracy and computational cost is achieved, and large molecules (e.g., drugs) are able to be handled by QM. Before performing the hybrid QM method for 2859 molecules, 22 representative compounds including the types of alkane, alkene, alkyne, halogen, alcohol, ether, aldehyde, ketone, acid, etc. are taken for examples to confirm the prediction accuracy through comparisons between the predicted values and experimental data, the results of which are shown in Fig. 4 and the corresponding raw data are listed in Table B1 in Appendix B (cf. ESM).

From Fig. 4, it is found that the calculated values of $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, S_g^θ for most compounds are in good agreements with their experimental data. The mean absolute error (MAE) of comparison results between calculated values and experimental data for $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, S_g^θ of 22 examples are 8.7, 12.2 kJ·mol^{-1}, and 10.7 J·mol^{-1}·K^{-1}, respectively, which verifies the feasibility and effectiveness of the QM calculation for predictions of thermodynamic properties $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, S_g^θ at 298.15 K. Using the hybrid QM method, $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, S_g^θ at 298.15 K for 2859 molecules are successfully predicted and prepared for further QSPR developments.

4.1.2.2 Standard thermodynamic property of vaporization

Considering that the thermodynamic properties at liquid state are also important in process industry, the QM-derived properties, standard thermodynamic properties of vaporization, are calculated in this section.

For ΔH_{vap}^θ, the fixed QM method of "B3PW91/6-31G(d, p)" [35] are employed to optimize the structures of 22 example molecules and calculate their single point energies. Afterwards, a powerful wave function analysis program, Multiwfn [45,46], is employed to make surface quantitative analysis for compounds to obtain σ_{tot}^2 and ν, and subsequently ΔH_{vap}^θ through Eq. (3). The prediction results are compared with experimental data and listed in Table 2.

To our best knowledge, there is no appropriate model for predictions of ΔS_{vap}^θ. Therefore, the Trouton's rule [51] is applied to obtain the standard entropy of vaporization of organic compound at boiling point ($\Delta S_{vap}^\theta (T_b) = = 88$ J·mol^{-1}·K^{-1}), which is supported by the fact that the gas entropy is significantly larger than the liquid entropy and thus the latter can be ignored. In this paper, an additional assumption is given that ΔS_{vap}^θ at 298.15 K is approximate to $\Delta S_{vap}^\theta (T_b)$, which makes it possible to calculate ΔG_{vap}^θ at 298.15 K through the equation of $\Delta G_{vap}^\theta = \Delta H_{vap}^\theta - T \times \Delta S_{vap}^\theta$. The calculated values of ΔH_{vap}^θ, ΔG_{vap}^θ, ΔS_{vap}^θ at 298.15 K of 22 examples are compared with the corresponding experimental data and listed in Table 2. Also, using the appropriate QM method, ΔH_{vap}^θ, ΔG_{vap}^θ, ΔS_{vap}^θ at 298.15 K for 2859 molecules are

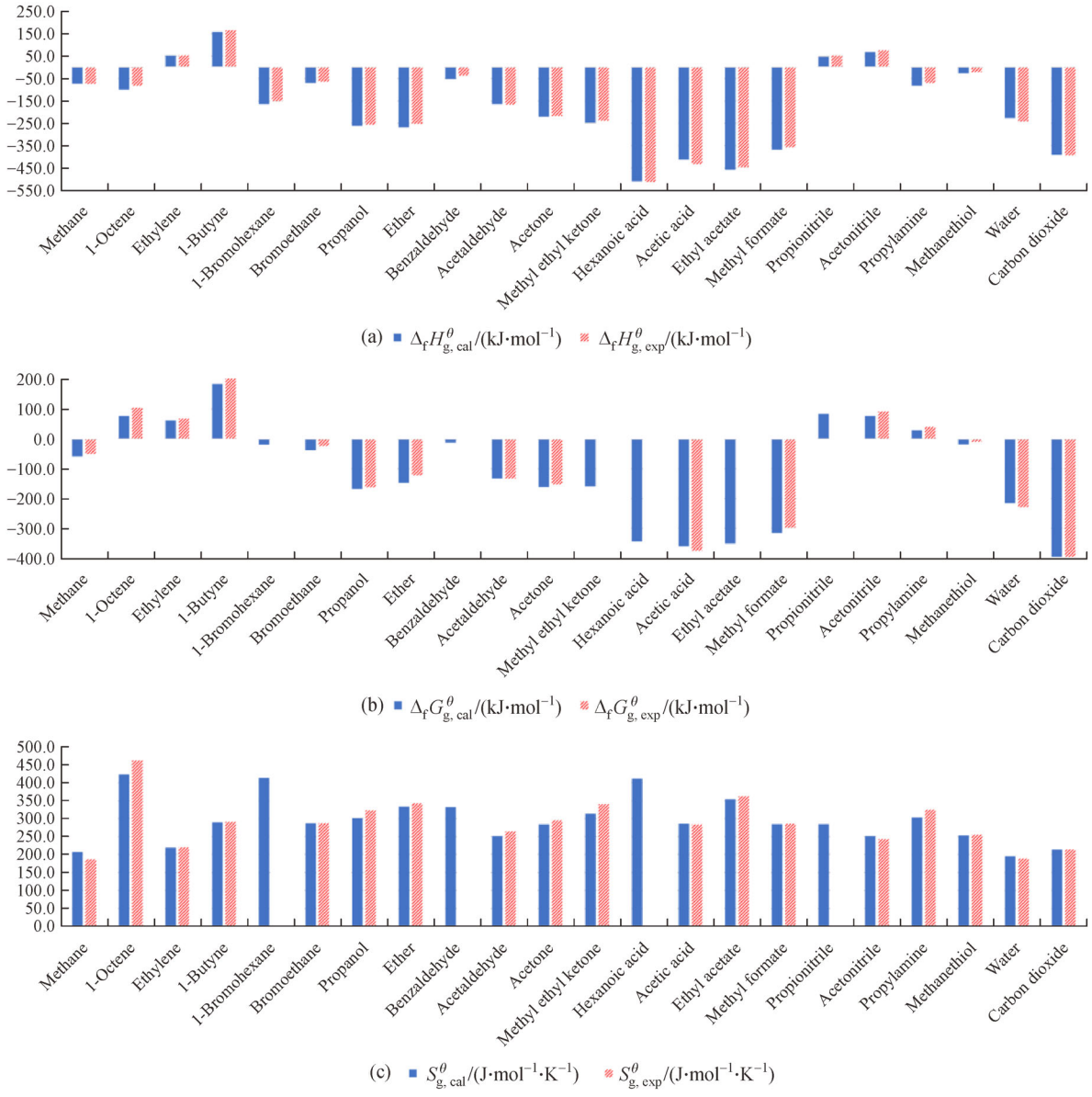

Fig. 4 Comparison results between QM calculated values and experimental data for (a) $\Delta_f H^{\theta}_g$, (b) $\Delta_f G^{\theta}_g$ and (c) S^{θ}_g at 298.15 K of 22 representative compounds (blank means experimental data are unavailable in database).

successfully predicted and prepared for further QSPR developments.

Some calculated values of ΔG^{θ}_{vap} have large errors compared with the experiment data, which are mainly caused by the assumption of $\Delta S^{\theta}_{vap}(T_b) \approx \Delta S^{\theta}_{vap}(298.15\ K)$. Besides, the coefficients for the ΔH^{θ}_{vap} expression (Eq. (3)) are regressed by fitting to only 30 experimental data [35], which may limit the application scope of Eq. (3) (e.g., No. 12 compound in Table 2). Therefore, it is desirable to develop reliable models for predictions of $\Delta S^{\theta}_{vap}(298.15\ K)$ in the future. Also, it is desirable to employ more experimental data for fitting the model coefficients of Eq. (3) or even revise Eq. (3) based on knowledge if the fitting result of original Eq. (3) with enough experimental data is not acceptable.

4.1.2.3 Standard thermodynamic property at liquid state

Based on the obtained gas thermodynamic properties and vaporization properties at 298.15 K, liquid thermodynamic properties at 298.15 K are obtained through $\Delta_f H^{\theta}_l = \Delta_f H^{\theta}_g - \Delta H^{\theta}_{vap}$, $\Delta_f G^{\theta}_l = \Delta_f G^{\theta}_g - \Delta G^{\theta}_{vap}$ and $S^{\theta}_l = S^{\theta}_g - \Delta S^{\theta}_{vap}$. The calculated values of $\Delta_f H^{\theta}_l$, $\Delta_f G^{\theta}_l$, S^{θ}_l at 298.15 K of 22 examples are compared with the corresponding experimental data and listed in Table 3. Also, $\Delta_f H^{\theta}_l$, $\Delta_f G^{\theta}_l$, S^{θ}_l at 298.15 K for 2859 molecules are

Table 2 The calculated values of ΔH^{θ}_{vap}, ΔG^{θ}_{vap}, ΔS^{θ}_{vap} at 298.15 K of 22 examples and their corresponding experimental data [a)]

No.	Compound	CAS number	Calculated ΔH^{θ}_{vap} /(kJ·mol^{-1})	Experimental ΔH^{θ}_{vap} /(kJ·mol^{-1})	Calculated ΔG^{θ}_{vap} /(kJ·mol^{-1})	Experimental ΔG^{θ}_{vap} /(kJ·mol^{-1})	Calculated ΔS^{θ}_{vap} /(J·mol^{-1}·K^{-1})	Experimental ΔS^{θ}_{vap} /(J·mol^{-1}·K^{-1})
1	Methane	74-82-8	4.4		–21.8		88.0	
2	1-Octene	111-66-0	46.9	40.4	20.7		88.0	
3	Ethylene	74-85-1	18.7		–7.5		88.0	
4	1-Butyne	107-00-6	34.9		8.6		88.0	
5	1-Bromohexane	111-25-1	50.3		24.1		88.0	
6	Bromoethane	74-96-4	31.3		5.0		88.0	
7	Propanol	71-23-8	48.5	47.5	22.2	8.8	88.0	129.1
8	Ether	60-29-7	34.6	27.4	8.4	–5.6	88.0	170.3
9	Benzaldehyde	100-52-7	63.2		37.0		88.0	
10	Acetaldehyde	75-07-0	27.6	26.1	1.3	–5.4	88.0	103.4
11	Acetone	67-64-1	35.1	31.3	8.8		88.0	96.5
12	Methyl ethyl ketone	78-93-3	41.3	19.6	15.0		88.0	100.8
13	Hexanoic acid	142-62-1	72.8		46.5		88.0	
14	Acetic acid	64-19-7	53.3	52.2	27.1	16.0	88.0	123.6
15	Ethyl acetate	141-78-6	38.5	35.7	12.3	5.3	88.0	106.1
16	Methyl formate	107-31-3	38.0	28.7	11.8		88.0	
17	Propionitrile	107-12-0	38.3		12.0		88.0	
18	Acetonitrile	75-05-8	35.0	39.6	8.8	5.4	88.0	93.7
19	Propylamine	107-10-8	44.4	31.3	18.2		88.0	
20	Methanethiol	74-93-1	27.2		1.0		88.0	
21	Water	7732-18-5	35.1	44.0	8.9	8.5	88.0	118.9
22	Carbon dioxide	124-38-9	18.1	19.8	–8.1	–8.4	88.0	94.4

a) Blank means experimental data which are unavailable in database.

calculated and prepared for further QSPR developments. Some calculated values have large errors compared with the experiment data, which are mainly caused by the prediction error of ΔH^{θ}_{vap}, ΔG^{θ}_{vap} and ΔS^{θ}_{vap}.

To sum up, the QM calculation method is capable of predicting $\Delta_f H^{\theta}_g$, $\Delta_f G^{\theta}_g$ and S^{θ}_g with acceptable prediction accuracy. However, some large errors are identified in the predicted vaporization and liquid thermodynamic properties. Models for predictions of ΔS^{θ}_{vap} and ΔH^{θ}_{vap} need further improvements in the future. In spite of this, the GC methods are still developed for vaporization and liquid properties in the next section just in case they are needed for some problems without any experimental data support. Besides, the regression results of GC method may also provide some inspirations for future QSPR developments of vaporization and liquid properties.

4.1.3 QSPR development

With the obtained pseudo experimental data, the GC method is developed for fast predictions of thermodynamic properties in this case study. Here, for convenient interaction of computational tools, a group division script (convert molecular SMILES to pre-defined group sets) is written in python [47] using the RDKit package, where 209 group sets (named as MG1-RDKit) are defined in SMILES arbitrary target specification. Most of these newly defined group sets refer to the commonly used first-order Marrero and Gani (MG1) group sets [42] as they are able to cover a large variety of organic compounds. Besides, a certain number of small compounds (e.g., H_2O, O_2, etc.) are also included in MG1-RDKit group sets. The whole group sets are able to be expanded and/or revised by users flexibly. Then, the least square method is employed to regress the MG1-RDKit GCs to the thermodynamic properties with 2859 molecules (samples) through Eq. (5) ($f(X) = X - UC = \sum_i N_i C_i$, UC represents a universal constant). The fitting results (R^2)/error criterion (MAE) of $\Delta_f H^{\theta}_g$, $\Delta_f G^{\theta}_g$, S^{θ}_g, ΔH^{θ}_{vap}, ΔG^{θ}_{vap}, $\Delta_f H^{\theta}_l$, $\Delta_f G^{\theta}_l$ and S^{θ}_l are listed in Table 4.

The results in Table 4 indicate that the developed GC method with MG1-RDKit group sets has ability in fast

Table 3 The calculated values of $\Delta_f H_1^\theta$, $\Delta_f G_1^\theta$, S_1^θ at 298.15 K of 22 examples and their corresponding experimental data [a)]

No.	Compound	CAS number	Calculated $\Delta_f H_1^\theta$ /(kJ·mol^{-1})	Experimental $\Delta_f H_1^\theta$ /(kJ·mol^{-1})	Calculated $\Delta_f G_1^\theta$ /(kJ·mol^{-1})	Experimental $\Delta_f G_1^\theta$ /(kJ·mol^{-1})	Calculated S_1^θ /(J·mol^{-1}·K^{-1})	Experimental S_1^θ /(J·mol^{-1}·K^{-1})
1	Methane	74-82-8	–78.1		–36.9		118.7	
2	1-Octene	111-66-0	–146.2	–121.8	56.3		335.0	
3	Ethylene	74-85-1	32.8		69.4		130.9	
4	1-Butyne	107-00-6	122.4		174.7		201.2	
5	1-Bromohexane	111-25-1	–215.0		–43.3		325.9	
6	Bromoethane	74-96-4	–101.2		–42.7		198.4	
7	Propanol	71-23-8	–309.5	–303.5	–190.1	–170.6	212.8	193.6
8	Ether	60-29-7	–303.2	–279.4	–155.8	–116.7	245.4	172.4
9	Benzaldehyde	100-52-7	–115.7		–48.8		244.5	
10	Acetaldehyde	75-07-0	–192.0	–192.2	–134.6	–127.6	163.5	160.4
11	Acetone	67-64-1	–255.6	–248.3	–170.0		195.8	198.8
12	Methyl ethyl ketone	78-93-3	–288.5	–258.2	–174.0		225.7	239.1
13	Hexanoic acid	142-62-1	–581.9		–390.5		323.4	
14	Acetic acid	64-19-7	–465.1	–484.4	–387.3	–390.2	197.6	159.9
15	Ethyl acetate	141-78-6	–496.0	–481.1	–362.9		265.7	256.7
16	Methyl formate	107-31-3	–406.1	–386.1	–327.9		196.4	
17	Propionitrile	107-12-0	8.1		72.1		196.4	
18	Acetonitrile	75-05-8	32.2	34.4	68.2	86.5	163.4	149.7
19	Propylamine	107-10-8	–125.7	–101.3	10.4		215.1	
20	Methanethiol	74-93-1	–54.8		–19.9		165.2	
21	Water	7732-18-5	–262.1	–285.8	–224.5	–237.1	106.6	69.9
22	Carbon dioxide	124-38-9	–408.7	–413.3	–386.2	–386.0	126.1	119.4

a) Blank means experimental data which are unavailable in database.

Table 4 The fitting results (R^2)/error criterion (MAE) of $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, S_g^θ, $\Delta_{vap} H^\theta$, $\Delta_{vap} G^\theta$, $\Delta_f H_1^\theta$, $\Delta_f G_1^\theta$ and S_1^θ between GC predictions and QM predictions

Property	R^2	MAE
$\Delta_f H_g^\theta$	0.990	28.1 kJ·mol^{-1}
$\Delta_f G_g^\theta$	0.988	27.7 kJ·mol^{-1}
S_g^θ	0.993	6.7 J·mol^{-1}·K^{-1}
ΔH_{vap}^θ	0.925	4.0 kJ·mol^{-1}
ΔG_{vap}^θ	0.925	4.0 kJ·mol^{-1}
$\Delta_f H_1^\theta$	0.990	28.0 kJ·mol^{-1}
$\Delta_f G_1^\theta$	0.988	27.7 kJ·mol^{-1}
S_1^θ	0.993	6.7 J·mol^{-1}·K^{-1}

predicting these thermodynamic properties with acceptable accuracy. Note that the MG1-RDKit based GC method may lead to poor predictions for $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$, $\Delta_f H_1^\theta$, $\Delta_f G_1^\theta$ in some cases as the MAEs of these properties are not small, which is caused by the fact that the interactions among first-order group sets cannot be ignored for the structurally complex molecules (e.g., drugs). Therefore, higher-order group sets are needed to further improve the prediction accuracy of MG1-RDKit based GC method in the future.

The MG1 group sets regressed from experimental $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ and S_g^θ values [42,48] are also compared with the MG1-RDKit group sets. First, the number of group sets with $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ and S_g^θ values are counted and compared between these two GC methods, the results of which are shown in Fig. 5. It is found that the numbers of group sets with $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ and S_g^θ values in the MG1-RDKit based GC method have increased by 44, 64 and 103, respectively, compared with those in the MG1 based GC method, which demonstrates that the new developed GC method is able to predict properties of molecules in larger chemical space. In addition, an evaluation criterion of $SR_{database}^{GC}$ is employed to calculate the ratios of the molecules for which their group sets have values of $\Delta_f H_g^\theta$ and S_g^θ to all molecules in a database. It is found that $SR_{VirginiaTech}^{MG1} =$

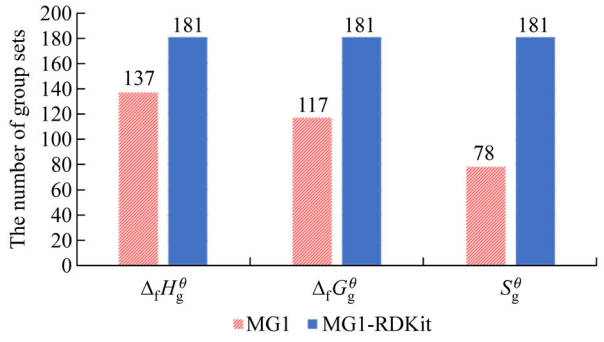

Fig. 5 The number of group sets with $\Delta_f H_g^\theta$, $\Delta_f G_g^\theta$ and S_g^θ values in the MG1 based GC method and MG1-RDKit based GC method.

$\frac{669}{903} \times 100\% = 74\%$, $SR_{\text{DrugBank}}^{\text{MG1}} = \frac{660}{1956} \times 100\% = 34\%$, $SR_{\text{VirginiaTech}}^{\text{MG1-RDKit}} = \frac{903}{903} \times 100\% = 100\%$ and $SR_{\text{DrugBank}}^{\text{MG1-RDKit}} = \frac{1956}{1956} \times 100\% = 100\%$, which indicates that the MG1-RDKit group sets are able to handle more structurally complex molecules (e.g., drugs) compared with the MG1 group sets due to the superiority of the QM-QSPR framework in providing more pseudo experimental data for GC regressions.

4.1.4 Heat of reaction for aspirin synthesis

Aspirin is used to treat pain and reduce fever or inflammation. It is generally produced by salicylic acid and acetic anhydride. The diagrammatic sketch of synthesis pathway for aspirin is shown in Fig. 6.

The properties, $\Delta_f H_g^\theta(298.15\ \text{K})$, for each compound in aspirin synthesis pathway have been obtained from the experiment, unpackaged hybrid QM method, MG1-RDKit based GC method and MG1 based GC method, which are listed in Table 5. Considering that the QM-predicted $\Delta_f H_g^\theta(298.15\ \text{K})$ are close to the experimental $\Delta_f H_g^\theta(298.15\ \text{K})$ in Fig. 4(a) and Table 5, it is reliable to adopt the prediction result of $\Delta_f H_g^\theta(298.15\ \text{K})$ for compound aspirin for calculations of $\Delta_r H_g^\theta(298.15\ \text{K})$. The results of $\Delta_r H_g^\theta(298.15\ \text{K})$ for unpackaged hybrid QM method and MG1-RDKit based GC method are –49.5 and –37.6 kJ·mol^{-1}, respectively, which confirms that prediction accuracy of MG1-RDKit based GC method is acceptable. Besides, it is found that the MG1 based GC method fails to predict $\Delta_f H_g^\theta(298.15\ \text{K})$ of aspirin while the MG1-RDKit based GC method succeeds, which indicates that the MG1-RDKit based GC method has a wider application range.

4.2 Prediction of solid-liquid phase equilibrium

This case study mainly refers to Liu et al. [52]. Here, four new crystallization solvents are investigated through the QM-QSPR toolbox. Solid-liquid phase equilibrium plays an important role in chemical engineering, e.g., crystallization process in pharmaceutical industry. It is described through Eq. (9),

$$\ln x_i^{\text{Sat}} - \frac{\Delta H_{\text{fus},i}}{\text{R}T_{\text{m},i}}\left(1 - \frac{T_{\text{m},i}}{T}\right) + \ln\gamma_i^{\text{Sat}} = 0, \tag{9}$$

where x_i^{Sat}, $\Delta H_{\text{fus},i}$, $T_{\text{m},i}$ and γ_i^{Sat} are the saturated mole

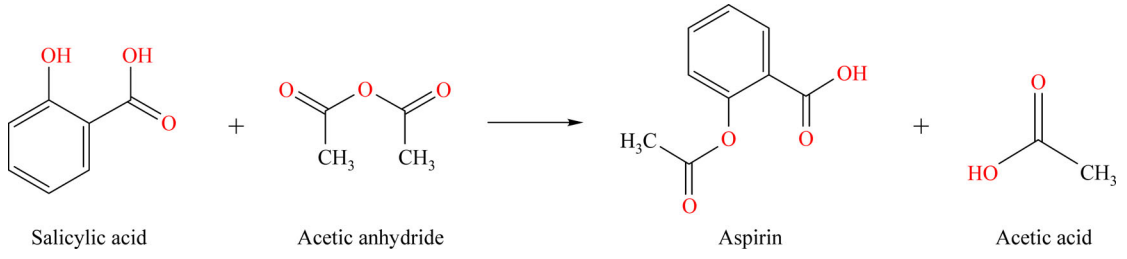

Fig. 6 The diagrammatic sketch of synthesis pathway for aspirin.

Table 5 The properties, $\Delta_f H_g^\theta(298.15\ \text{K})$, for each compound in aspirin synthesis pathway obtained from different methods (the experimental data refers to ICAS software [42]; blank means data which are unavailable)

Molecule	CAS number	Experiment $\Delta_f H_g^\theta$/(kJ·mol^{-1})	Unpackaged hybrid QM $\Delta_f H_g^\theta$/(kJ·mol^{-1})	MG1-RDKit $\Delta_f H_g^\theta$/(kJ·mol^{-1})	MG1 $\Delta_f H_g^\theta$/(kJ·mol^{-1})
Salicylic acid	69-72-7	–494.8	–473.5	–458.4	–469.4
Acetic anhydride	108-24-7	–572.5	–588.3	–587.0	–575.2
Aspirin	50-78-2		–677.1	–667.1	
Acetic acid	64-19-7	–432.8	–434.3	–415.9	–432.6

fraction, enthalpy of melting, melting point and saturated activity coefficient of solute i, respectively. T represents the crystallization temperature.

The solvent-related property, γ_i^{Sat}, is predicted by the COSMO-SAC model [38], where the $p(\sigma)$ of solvents are key model inputs for predictions of γ_i^{Sat}. However, the computational cost of generating $p(\sigma)$ through QM calculations are time-consuming, which hinders the high-throughput predictions of γ_i^{Sat} through the COSMO-SAC model. In this case study, the QM-QSPR toolbox is employed to develop a machine learning-based atom contribution (MLAC) method for fast predictions of $p(\sigma)$. The MLAC method is developed based on solvent database and QM calculations. The workflow of case study 2 is shown in Fig. 7.

4.2.1 Database establishment

A database is created by screening 1120 solvents containing H, C, N and O elements with CAS numbers and Isomeric SMILES from the Virginia Tech database [50]. Here, only H, C, N and O elements are considered as these elements can be found in most of the commonly used organic solvents.

4.2.2 QM calculation

Before predicting properties, the stereoscopic representations of all 1120 solvents with cartesian coordinates are obtained from Isomeric SMILES by OpenBabel [28] in the QM-QSPR toolbox.

The molecular $p(\sigma)$ is obtained by the summation of all atomic $p(\sigma)(p_{\mathrm{atom}}(\sigma))$ contributions based on the assumption of the COSMO-SAC model [38]. The $p_{\mathrm{atom}}(\sigma)$ of all solvents are calculated with the Gaussian software [23] and the COSMO-SAC model [38]. In Gaussian, the QM method "B3LYP/6-31G(d,p)" is selected for solvent geometry optimizations and COSMO calculations. In the COSMO-SAC model, the group-independent model parameters refer to Chen et al. [39] (listed in Table C1 in Appendix C (cf. ESM)), which have been specially reparametrized for "B3LYP/6-31G(d,p)" and verified to have good performance for the predictions of solvent properties (e.g., γ). More detailed information about the formula derivations of $p_{\mathrm{atom}}(\sigma)$ can be found in Chen et al. [38]. Finally, a database of $p_{\mathrm{atom}}(\sigma)$ is established as the outputs for the development of MLAC method, where the number of atomic samples for H, C, N and O elements is 15535, 9108, 305 and 1215, respectively.

4.2.3 QSPR development

With the obtained pseudo experimental data, the MLAC method is developed for fast predictions of molecular $p(\sigma)$ in this case study. The MLAC method starts from the molecular SMILES, which is first converted to the stereoscopic representation with cartesian coordinates through OpenBabel [28]. Then, a script is written in python [47] to transform the stereoscopic representation to the three-dimensional atomic descriptors, weighted atom-centered symmetry functions (wACSFs) [53]. More detailed information about the wACSFs can be found in Gastegger et al. [53]. Four separate element-based (H, C,

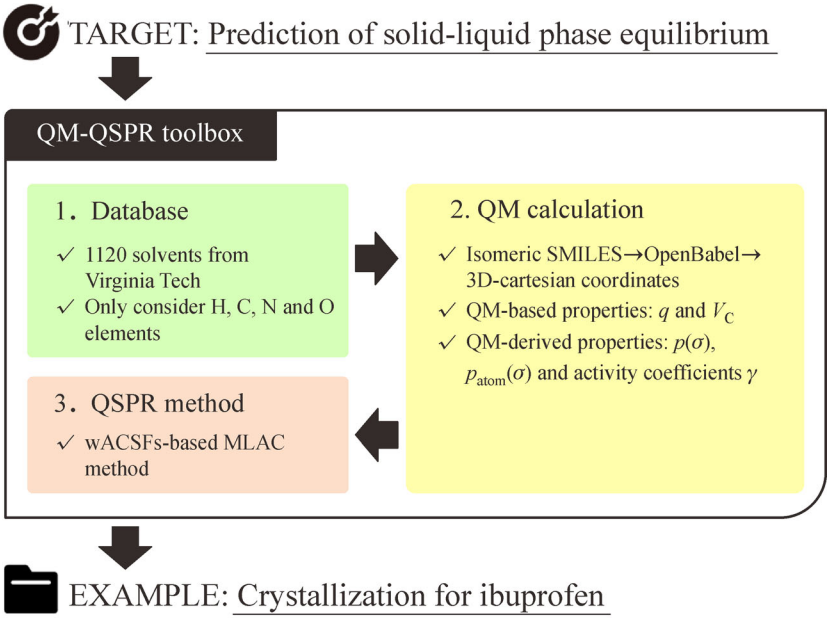

Fig. 7 The workflow of case study 2.

N, O) BP-ANNs are apriori built/trained through Eqs. (6) and (7) with atomic samples of 15535, 9108, 305, 1215 wACSFs (input data) and $p_{\text{atom}}(\sigma)$ (output data) for H, C, N, O elements, respectively, before its use in the MLAC method. With the established BP-ANNs, $p_{\text{atom}}(\sigma)$ are predicted with the wACSFs and the MLAC method ends with $p(\sigma)$ summed by $p_{\text{atom}}(\sigma)$. More detailed information related to the MLAC method (e.g., the architectures of BP-ANNs, the performance of MLAC method, the weakness of MLAC method that poor predictions may be identified in solvents containing N element, etc.) can be found in Liu et al. [52].

4.2.4 Crystallization for ibuprofen

The cooling crystallization process for solute ibuprofen with four solvents (1-propanol, 2-methyl-1-propanol, 3-methyl-1-butanol and ethyl acetate) is taken as examples to highlight the QM-QSPR toolbox. The experimental data of solid-liquid phase equilibrium are obtained from Wang et al. [54], which are compared with those predicted by Eq. (9) based on the QM calculation and MLAC method. The necessary crystallization parameters of $\Delta H_{\text{fus,ibuprofen}} = 27.94\ \text{kJ}\cdot\text{mol}^{-1}$ and $T_{\text{m,ibuprofen}} = 347.6\ \text{K}$ are taken from Hong et al. [55] and the crystallization temperature range is set to $260\ \text{K} \leqslant T \leqslant 320\ \text{K}$. The results of QM calculation and MLAC method are shown in Fig. 8. It is found that the solid-liquid phase equilibrium curves of MLAC predictions for four solvents (see Figs. 8(a–d)) are all close to the QM calculations, which indicates the reliability of the MLAC method in fast predictions of $p(\sigma)$ and γ for solvents with H, C and O elements. Although minor prediction errors are identified in solid-liquid phase equilibrium curves of solvents 3-methyl-1-butanol and ethyl acetate between the QM calculations and experiments (see Figs. 8(c) and 8(d)), the QM calculation (specifically, the COSMO-SAC model) is still worth considering for predictions of solid-liquid phase equilibrium curves of structurally complex molecules if their group parameters are missing in UNIFAC models.

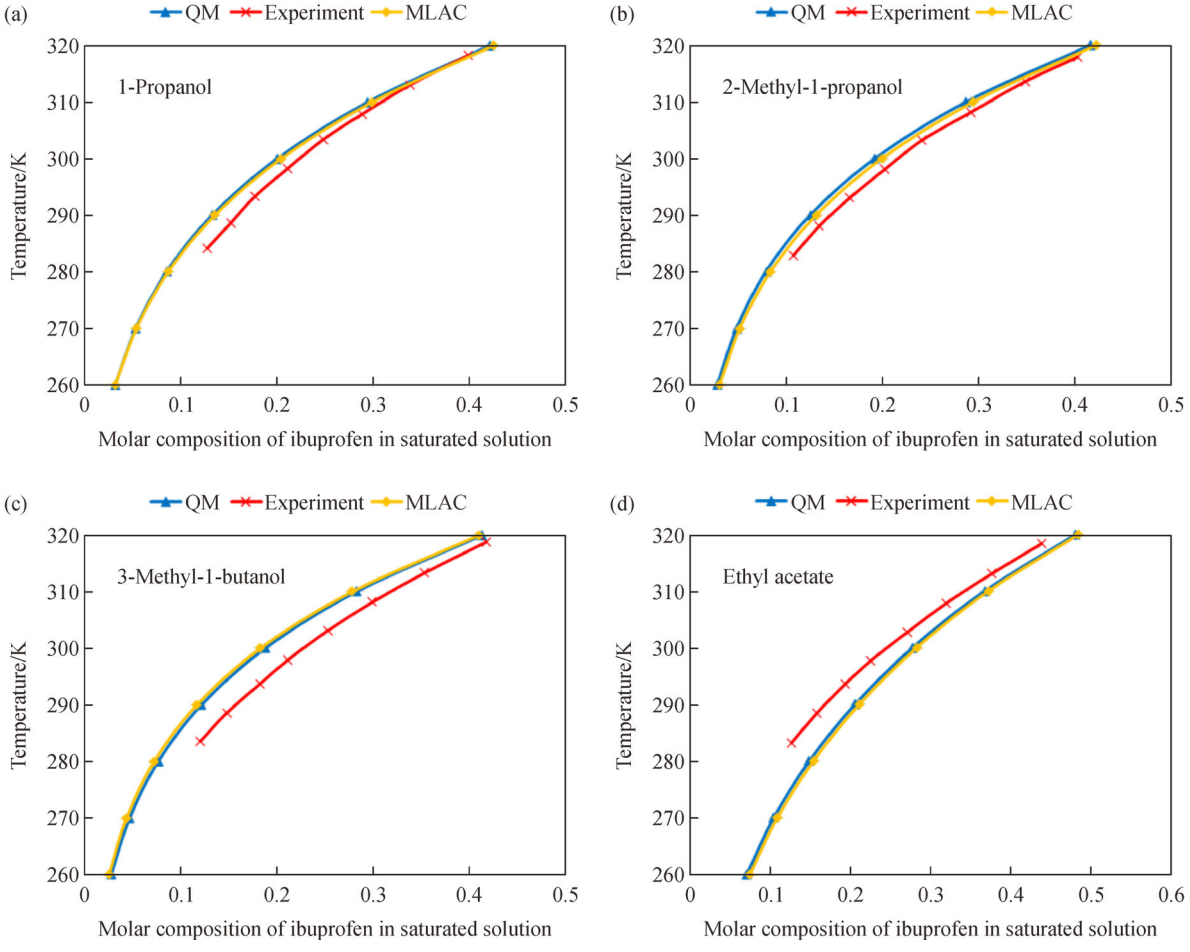

Fig. 8 Comparisons of solid-liquid phase equilibrium curves of solute ibuprofen and four solvents ((a) 1-propanol, (b) 2-methyl-1-propanol, (c) 3-methyl-1-butanol and (d) ethyl acetate) obtained from the experiment, QM calculation and MLAC method.

5 Conclusions

In this paper, a QM-QSPR framework is established for molecular property prediction, involving three steps of database establishment, QM calculation and QSPR development. In the QM calculation step, molecular properties are classified in terms of QM-based property and QM-derived property, for which appropriate QM methods are recommended considering prediction accuracy and computational efficiency. With the sufficient pseudo experimental data from the QM calculation step, QSPR methods are developed for fast predictions of molecular properties in terms of linear QSPR methods (e.g., GC method) and nonlinear QSPR methods (e.g., MLAC method). Two case studies involving the prediction of heat of reaction and solid-liquid phase equilibrium are presented to confirm the feasibility and effectiveness of the QM-QSPR framework. The targets of developing the MG1-RDKit based GC method with a wide application scope and the MLAC method with a high prediction accuracy for $p(\sigma)$ are both achieved in two case studies, respectively, through the developed computational toolbox of QM-QSPR. In the future, more QM-based/derived properties with corresponding QSPR methods will be incorporated into the QM-QSPR toolbox for convenient molecular property predictions and applications to other chemical-related software (e.g., OptCAMD [44] for molecular design).

Acknowledgements The authors are grateful for the financial supports of the National Natural Science Foundation of China (Grant Nos. 22078041 and 21808025) and the Fundamental Research Funds for the Central Universities (Grant No. DUT20JC41).

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://dx.doi.org/10.1007/s11705-021-2060-z and is accessible for authorized users.

References

1. Kirkpatrick P, Ellis C. Chemical space. Nature, 2004, 432(7019): 823
2. Katritzky A R, Lobanov V S, Karelson M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews, 1995, 24(4): 279–287
3. Mills E J. On melting point and boiling point as related to composition. Philosophical Magazine, 1884, 17(5): 173–187
4. Dearden J C, Cronin M T D, Kaiser K L E. How not to develop a quantitative structureactivity or structureproperty relationship (QSAR/QSPR). SAR and QSAR in Environmental Research, 2009, 20(3-4): 241–266
5. Kim S, Cho K H. PyQSAR: a fast QSAR modeling platform using machine learning and jupyter notebook. Bulletin of the Korean Chemical Society, 2019, 40(1): 39–44
6. Enciso M, Meftahi N, Walker M L, Smith B J. BioPPSy: an open-source platform for QSAR/QSPR analysis. PLoS One, 2016, 11(11): e0166298
7. Pirhadi S, Sunseri J, Koes D R. Open source molecular modeling. Journal of Molecular Graphics & Modelling, 2016, 69: 127–143
8. Stålring J C, Carlsson L A, Almeida P, Boyer S. AZOrange—high performance open source machine learning for QSAR modeling in a graphical programming environment. Journal of Cheminformatics, 2011, 3(1): 28
9. Cortes-Ciriano I. Bioalerts: a python library for the derivation of structural alerts from bioactivity and toxicity data sets. Journal of Cheminformatics, 2016, 8(1): 13
10. Murrell D S, Cortes-Ciriano I, van Westen G J P, Stott I P, Bender A, Malliavin T E, Glen R C. Chemically aware model builder (camb): an R package for property and bioactivity modelling of small molecules. Journal of Cheminformatics, 2015, 7(1): 45
11. Carrió P, López O, Sanz F, Pastor M. eTOXlab, an open source modeling framework for implementing predictive models in production environments. Journal of Cheminformatics, 2015, 7(1): 8
12. Tosco P, Balle T. Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. Journal of Molecular Modeling, 2011, 17(1): 201–208
13. Dimitrov S D, Diderich R, Sobanski T, Pavlov T S, Chankov G V, Chapkanov A S, Karakolev Y H, Temelkov S G, Vasilev R A, Gerova K D, et al. QSAR Toolbox—workflow and major functionalities. SAR and QSAR in Environmental Research, 2016, 27(3): 203–219
14. Kostal J. Advances in Molecular Toxicology. 1st ed. Cambridge: Elsevier, 2016, 139–186
15. Krokhotin A, Dokholyan N V. Methods in Enzymology. 1st ed. Waltham: Elsevier, 2015, 65–89
16. Polanski J. Comprehensive Chemometrics. 1st ed. Oxford: Elsevier, 2009, 459–506
17. Salomon-Ferrer R, Case D A, Walker R C. An overview of the Amber biomolecular simulation package. WIREs Computational Molecular Science, 2013, 3(2): 198–210
18. Jo S, Kim T, Iyer V G, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. Journal of Computational Chemistry, 2008, 29(11): 1859–1865
19. Berendsen H J C, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. Computer Physics Communications, 1995, 91(1): 43–56
20. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. Journal of Computational Physics, 1995, 117(1): 1–19
21. Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kalé L, Schulten K. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry, 2005, 26(16): 1781–1802
22. Li W, Chen C, Zhao D, Li S. LSQC: low scaling quantum chemistry program. International Journal of Quantum Chemistry, 2015, 115 (10): 641–646
23. Gaussian 16. Revision A.03. Wallingford, CT: Gaussian, Inc., 2016.
24. Neese F. The ORCA program system. WIREs Computational Molecular Science, 2012, 2(1): 73–78
25. Schmidt M W, Baldridge K K, Boatz J A, Elbert S T, Gordon M S, Jensen J H, Koseki S, Matsunaga N, Nguyen K A, Su S, et al. General atomic and molecular electronic structure system. Journal of Computational Chemistry, 1993, 14(11): 1347–1363

26. Stewart James J P. MOPAC: a semiempirical molecular orbital program. Journal of Computer-Aided Molecular Design, 1990, 4(1): 1–103
27. Neese F, Wennmohs F, Hansen A, Becker U. Efficient, approximate and parallel hartreefock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the hartreefock exchange. Chemical Physics, 2009, 356(1): 98–109
28. O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T, Hutchison G R. Open Babel: an open chemical toolbox. Journal of Cheminformatics, 2011, 3(1): 33
29. Mata R A, Suhm M A. Benchmarking quantum chemical methods: are we heading in the right direction? Angewandte Chemie International Edition, 2017, 56(37): 11011–11018
30. Vereecken L, Glowacki D R, Pilling M J. Theoretical chemical kinetics in tropospheric chemistry: methodologies and applications. Chemical Reviews, 2015, 115(10): 4063–4114
31. Zheng J, Zhao Y, Truhlar D G. The DBH24/08 database and its use to assess electronic structure model chemistries for chemical reaction barrier heights. Journal of Chemical Theory and Computation, 2009, 5(4): 808–821
32. Řezáč J, Hobza P. Describing noncovalent interactions beyond the common approximations: how accurate is the "gold standard," CCSD(T) at the complete basis set limit? Journal of Chemical Theory and Computation, 2013, 9(5): 2151–2155
33. Sun J, Furness J W, Zhang Y. Mathematical Physics in Theoretical Chemistry. 1st ed. Amsterdam: Elsevier, 2019, 119–159
34. Goerigk L, Hansen A, Bauer C, Ehrlich S, Najibi A, Grimme S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. Physical Chemistry Chemical Physics, 2017, 19(48): 32184–32215
35. Politzer P, Ma Y, Lane P, Concha M C. Computational prediction of standard gas, liquid, and solid-phase heats of formation and heats of vaporization and sublimation. International Journal of Quantum Chemistry, 2005, 105(4): 341–347
36. Speight J G. Book Lange's Handbook of Chemistry. 16th ed. New York: McGraw-Hill, 2005, 515–560.
37. Liu Q, Zhang L, Liu L, Du J, Meng Q, Gani R. Computer-aided reaction solvent design based on transition statetheory and COSMO-SAC. Chemical Engineering Science, 2019, 202: 300–317
38. Hsieh C M, Sandler S I, Lin S T. Improvements of COSMO-SAC for vaporliquid and liquidliquid equilibrium predictions. Fluid Phase Equilibria, 2010, 297(1): 90–97
39. Chen W L, Hsieh C M, Yang L, Hsu C C, Lin S T. A critical evaluation on the performance of COSMO-SAC models for vaporliquid and liquidliquid equilibrium predictions based on different quantum chemical calculations. Industrial & Engineering Chemistry Research, 2016, 55(34): 9312–9322
40. Gani R. Group contribution-based property estimation methods: advances and perspectives. Current Opinion in Chemical Engineering, 2019, 23: 184–196
41. Mattei M, Kontogeorgis G M, Gani R. Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method. Industrial & Engineering Chemistry Research, 2013, 52(34): 12236–12246
42. Hukkerikar A S, Sarup B, Ten Kate A, Abildskov J, Sin G, Gani R. Group-contribution^{+} (GC^{+}) based estimation of properties of pure components: improved property estimation and uncertainty analysis. Fluid Phase Equilibria, 2012, 321: 25–43
43. Goh A T C. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering, 1995, 9(3): 143–151
44. Liu Q, Zhang L, Liu L, Du J, Tula A K, Eden M, Gani R. OptCAMD: an optimization-based framework and tool for molecular and mixture product design. Computers & Chemical Engineering, 2019, 124: 285–301
45. Lu T, Chen F. Multiwfn: a multifunctional wavefunction analyzer. Journal of Computational Chemistry, 2012, 33(5): 580–592
46. Lu T, Chen F. Quantitative analysis of molecular surface based on improved marching tetrahedra algorithm. Journal of Molecular Graphics & Modelling, 2012, 38: 314–323
47. Oliphant T E. Python for scientific computing. Computing in Science & Engineering, 2007, 9(3): 10–20
48. Liu Q, Zhang L, Tang K, Feng Y, Zhang J, Zhuang Y, Liu L, Du J. Computer-aided reaction solvent design considering inertness using group contribution-based reaction thermodynamic model. Chemical Engineering Research & Design, 2019, 152: 123–133
49. Oxtoby D W, Gillis H P, Campion A, Helal H H, Gaither K P. Book Principles of Modern Chemistry. 7th ed. Belmont: CENGAGE Learning, 2011, 596
50. Mullins E, Oldland R, Liu Y A, Wang S, Sandler S I, Chen C C, Zwolak M, Seavey K C. Sigma-profile database for using COSMO-based thermodynamic methods. Industrial & Engineering Chemistry Research, 2006, 45(12): 4389–4415
51. Rooney J J. Trouton's rule. Nature, 1990, 348(6300): 398–398
52. Liu Q, Zhang L, Tang K, Liu L, Du J, Meng Q, Gani R. Machine learning-based atom contribution method for the prediction of surface charge density profiles and solvent design. AIChE Journal. American Institute of Chemical Engineers, 2021, 67(2): e17110
53. Gastegger M, Schwiedrzik L, Bittermann M, Berzsenyi F, Marquetand P. WACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials. Journal of Chemical Physics, 2018, 148(24): 241709
54. Wang S, Song Z, Wang J, Dong Y, Wu M. Solubilities of ibuprofen in different pure solvents. Journal of Chemical & Engineering Data, 2010, 55(11): 5283–5285
55. Hong J, Hua D, Wang X, Wang H, Li J. Solidliquidgas equilibrium of the ternaries ibuprofen + myristic acid + CO_2 and ibuprofen + tripalmitin + CO_2. Journal of Chemical & Engineering Data, 2010, 55(1): 297–302