Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

A Systematic Method to Estimate and Validate Enthalpies of Formation Using Error-Cancelling Balanced Reactions

Philipp Buerger[1], Jethro Akroyd[1],
Sebastian Mosbach[1], and Markus Kraft[1,2]

released: 03 February 2017

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge, CB2 3RA
United Kingdom
E-mail: mk306@cam.ac.uk

[2] School of Chemical and
Biomedical Engineering,
Nanyang Technological University,
62 Nanyang Drive,
Singapore 637459

Preprint No. 179

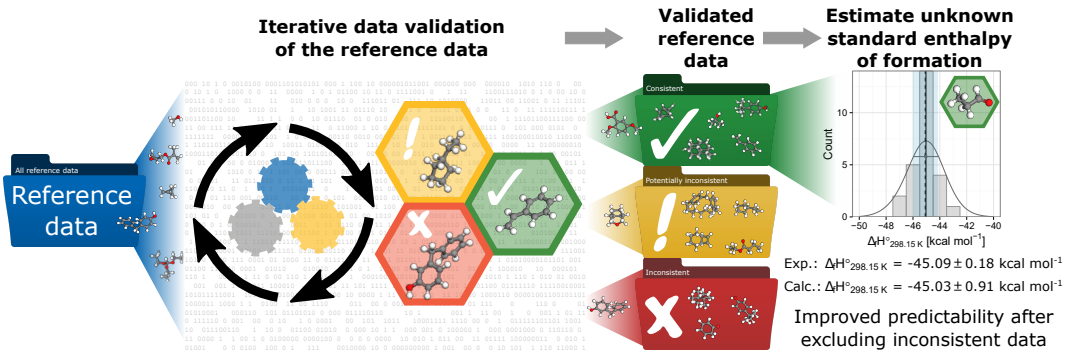
Keywords: enthalpy of formation, heat of formation, error-cancelling balanced reactions, validation, thermochemical data, big data, algorithm, methodology, data consistency

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
United Kingdom

Fax: + 44 (0)1223 334796
E-Mail: c4e@cam.ac.uk
World Wide Web: http://como.ceb.cam.ac.uk/

Abstract

This paper presents an automated framework that uses overlapping subsets of reference data to systematically derive an informed estimate of the standard enthalpy of formation of chemical species and assess the consistency of the reference data. The theory of error-cancelling balanced reactions (EBRs) is used to calculate estimates of the standard enthalpy of formation. Individual EBRs are identified using linear programming. The first part of the framework recursively identifies multiple EBRs for specified target species. A distribution of estimates can then be determined for each species from which an informed estimate of the enthalpy is derived. The second part of the framework iteratively isolates inconsistent reference data and improves the prediction accuracy by excluding such data. The application of the framework is demonstrated for test cases from organic and inorganic chemistry, including transition metal complexes. Its application to a set of 920 carbon, hydrogen and oxygen containing species resulted in a rapid decrease of the mean absolute error for estimates of the enthalpy of formation of each species due to the identification and exclusion of inconsistent reference data. Its application to titanium-containing species identified that the available reference values of TiOCl and $TiO(OH)_2$ are inconsistent and need further attention. Revised values are calculated for both species. A comparison with popular high-level quantum chemistry methods shows that the framework is able to deliver highly accurate estimates of the standard enthalpy of formation, comparable to high-level quantum chemistry methods for both hydrocarbons and transition metal complexes.



Highlights:

- Systematic method to obtain informed estimates of the standard enthalpy of formation.
- Linear programming used to identify error-cancelling balanced reactions.
- Assessment of the consistency of provided reference data.
- Method is applied to test cases from organic and inorganic chemistry, including transition metal complexes.
- Revised reference values of the standard enthalpy of formation for TiOCl and $TiO(OH)_2$.

Contents

1 Introduction

The development of automated procedures, for analysing chemical species and mechanisms [9, 10, 30, 38, 67, 68], facilitate the investigation of progressively complex reaction systems. The availability of large sets of consistent chemical data is of key importance. Many data sets used by such tools are collated literature data [54, 63] that are held in repositories [2, 62, 72, 75, 80] and are used for benchmarking computational methods [13, 15, 18–20, 103]. Alongside these opportunities there remain challenges. One question is concerned with the consistency of chemical data [see for example 25–28, 84, 85]. Using accurate single-point level calculations to validate one species at a time is computationally demanding and quickly becomes intractable for large systems. This paper therefore presents a solution to this problem and provides a framework to systematically evaluate the consistency of thermochemical data for chemical species.

Consistent and accurate thermochemical data, such as species enthalpies, heat capacities, and Gibbs free energies, are an essential part of any detailed chemical model. The standard enthalpy of formation is a fundamental parameter required to calculate accurate values of the enthalpy and Gibbs free energy changes of the reactions in a chemical model. Inconsistencies in the standard enthalpy of formation could lead to significant errors affecting the accuracy, predictive performance and quality of any model using such data.

In the past, different methods have been introduced to estimate the enthalpies of formation. The simplest are *additive or group contribution methods* [6, 16, 48, 104]. They rely on the regularity of molecular and structural groups. They are computationally cheap and predictive in nature. To achieve qualitatively accurate results, their application is limited to well studied systems with precisely and accurately defined functional groups [82, 94]. *Molecular mechanics methods* are computationally less demanding than other methods but are not universally applicable because they rely on empirical parameters and correction terms [82]. Electronic structure calculations at a high level of theory are used by *quantum chemistry methods* to estimate the enthalpies of formation. This type of single-point calculation is computationally demanding. The errors scale with the size of the molecule [98, 105], and the calculations become intractable for large molecules [49, 77]. In addition, care must be taken to choose the right level of theory [49, 89, 97] and various correction terms needed to achieve consistent and accurate estimates [90].

Fortunately, the errors incurred in electronic structure calculations are systematic. Different methods have been developed to reduce and cancel the impact of these errors on estimates of the enthalpy of formation. Among these are the bond additivity correction (BAC) [3, 43, 59, 60] and the atom additivity correction (AAC) [86, 94]. Both rely on predefined parameters associated with the level of theory used in the calculation.

Error-cancelling balanced reactions (EBRs) exploit structural and electronic similarities between the species in a reaction to reduce the impact of the inherited systematic errors. The standard enthalpy of formation from an EBR is calculated based on the application of Hess's Law to the reaction. This method has been applied to a variety of different systems [see for example 9, 71, 76, 87, 93, 99]. The absence of any empirical parameters makes this method suitable for automation. The total electronic energies for all species

in the reaction and the enthalpies of formation need to be known (experimentally or theoretically) for all except one species, for which the unknown enthalpy of formation can be estimated. The method requires the identification of suitable EBRs fulfilling a set of constraints defined by the type of EBR.

Since the introduction of EBRs by Pople and co-workers [41, 73], several types of EBRs have been proposed [33–35, 41, 73, 74, 76, 101, 102]. For example, isogyric, isodesmic, hypohomodesmotic, homodesmotic and hyperhomodesmotic reactions. The use of EBRs has been shown to enable the calculation of accurate estimates of the enthalpy of formation on the back of affordable electronic structure calculations. Generally, the more structural and electronic similarity that is preserved by the reaction, the more accurate the resulting estimate of the enthalpy of formation.

In our previous work [11] a high-level description was presented of an abstract and systematic framework to validate thermochemical data for chemical species and recommend what future experiments or calculations would be required to improve the data. The purpose of the current work is to give a detailed description of the algorithms used by the framework to identify error-cancelling balanced reactions (EBRs) and to calculate informed estimates of the standard enthalpies of formation. The framework facilitates the assessment of the consistency of chemical data, in this case for the standard enthalpy of formation in this work. This is achieved by iteratively isolating potentially inconsistent reference data. By excluding such data, an improved prediction accuracy is achieved. The performance of the framework is demonstrated using four different reaction classes and test cases from organic and inorganic chemistry, including transition metal complexes. Calculated estimates of the standard enthalpy of formation for hydrocarbons and titanium-containing species are compared against calculated values using popular high-level quantum chemistry methods.

2 Methodology

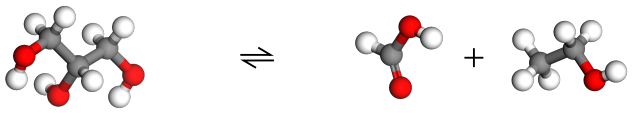
This section presents a detailed description of the algorithms used to identify a set of EBRs, calculate informed estimates of the standard enthalpy of formation for a species and assess the consistency of the required reference data. The algorithms are implemented as part of an automated and systematic framework to estimate the standard enthalpy of formation for a set of target species. A set of reference species, each consisting of the total electronic energy, the molecular connectivity, the spin multiplicity and a known enthalpy of formation, is required.

The definitions of the EBRs considered in this work are given in Section 2.1, followed by a description of the electronic structure calculations in Section 2.2. The reference data are described in Section 2.3, followed by a detailed algorithmic description of the framework in Section 2.4.

2.1 Types of error-cancelling balanced reactions

Many different reaction classes of EBRs have been proposed [33–35, 41, 73, 74, 76, 101, 102]. The following types of EBRs are used in this work.

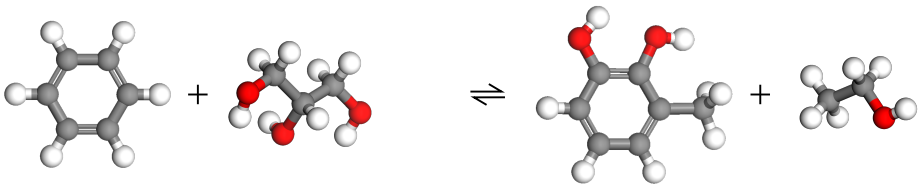
Isogyric reactions (reaction class RC1) are the least restrictive and only conserve the number of spin pairs on either side of the reaction. An example for an isogyric reaction is given in Figure 1.



Constraints	Reactants	Reactants Σ	Products Σ	Products	
Paired e⁻	50	50	50	24	26
Unpaired e⁻	0	0	0	0	0

Figure 1: *Example reaction for reaction class RC1 (isogyric reactions). The number of spin pairs is conserved on either side of the reaction.*

Isodesmic reactions (reaction class RC2) conserve the number of each type of bond on either side of the reaction. No constraint is placed on the chemical environment near each bond. An example for an isodesmic reaction is given in Figure 2.

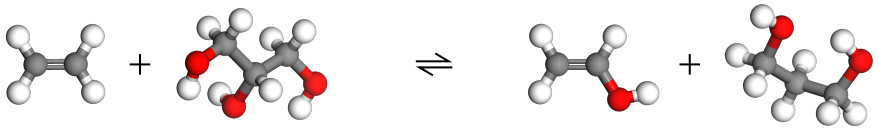


Constraints	Reactants		Reactants Σ	Products Σ	Products	
	3	2	5	5	4	1
	3	0	3	3	3	0
	6	5	11	11	6	5
	0	3	3	3	2	1
	0	3	3	3	2	1

Figure 2: *Example reaction for reaction class RC2 (isodesmic reactions). The number of each type of bond is conserved on either side of the reaction.*

Reaction class RC3 extends the concept of isodesmic reactions. The total bond order and identity of each atom on either side of the bond is conserved in addition to the number of each type of bond on either side of the reaction. Figure 3 presents an example of such a reaction.

Reaction class RC4 extends the concept of RC3. The bond type and identity of each neighbouring atom including the total bond order of the neighbouring atoms is conserved



Constraints	Reactants		Reactants Σ	Products Σ	Products	
3 — 3	0	2	2	2	0	2
2 = 2	1	0	1	1	1	0
3 — 0	4	5	9	9	3	6
3 — 1	0	3	3	3	1	2
1 — 0	0	3	3	3	1	2

Figure 3: *Example reaction for reaction class RC3. The number of each type of bond including the total bond order and identity of each atom on either side of the bond is conserved on either side of the reaction. The numbers next to the atoms in each bond show their total bond order.*

in addition to the constraints imposed by RC3. Possible constraints for propylene glycol ($C_3H_8O_2$) are presented in Figure 4 and an example reaction is given in Figure 5.

In addition, all EBR reaction classes conserve the atom-mass-balance of the reaction. The above reaction classes are presented in order of increasing restictiveness. Isogyric reactions (RC1) are the least restrictive. Reaction class RC4 is the most restrictive.

2.2 Electronic structure calculations

Ground state geometries and vibrational frequencies for all species used in this work were calculated using DFT at the B97-1/6-311+G(d,p) level of theory, as per previous works [9, 11, 71, 99]. This functional has shown to be accurate [8, 40] and well suited for transition metal complexes [31, 47, 91]. For comparison purposes, ground state geometries and vibrational frequencies for all gas-phase species containing carbon, hydrogen and oxygen, were additionally calculated using the B3LYP/6-311+G(d,p) level of theory.

To compensate for overestimated vibrational frequencies, scaling factors were used for both functionals as proposed by Merrick et al. [61]. A simple rigid-rotor harmonic-oscillator approximation was assumed [58]. This presents the worst case scenario with respect to the accuracy of the total energy calculation and gives an idea about the predictive power of the method.

All electronic structure calculations were performed using the Gaussian09 software package [29], running on Intel Xeon CPU X5472@3GHz/8GB nodes with 8 cores per node.

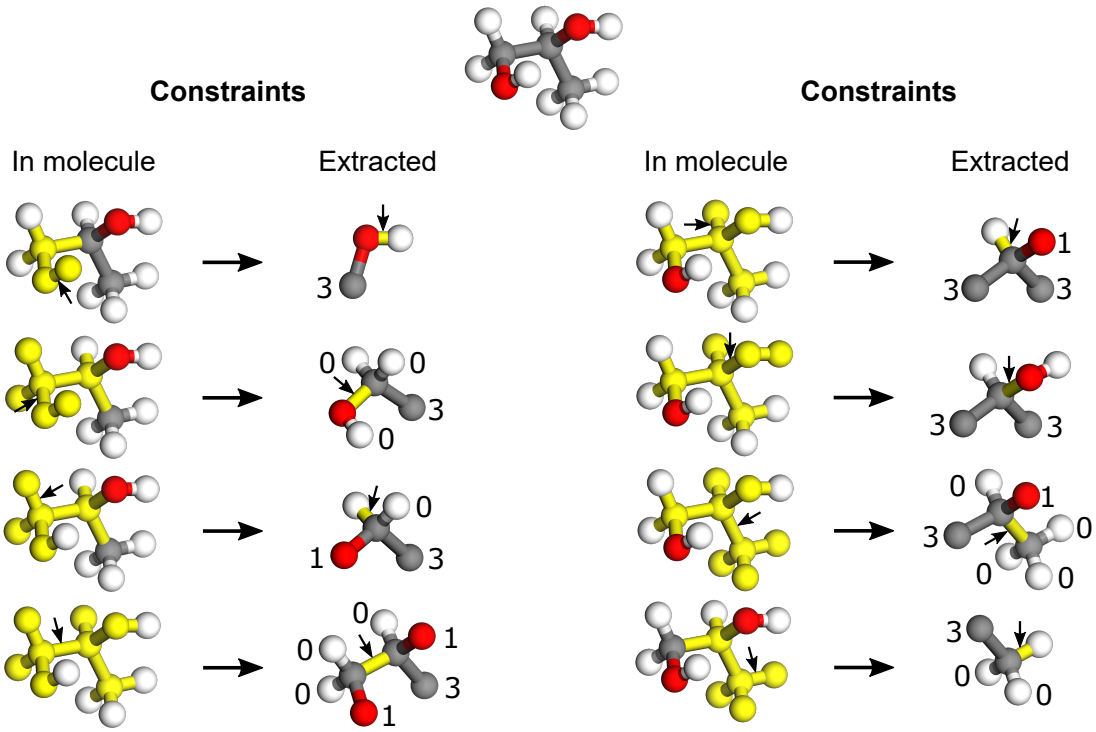


Figure 4: *Conserved structural groups of propylene glycol ($C_3H_8O_2$) using reaction class RC4. The labelled structural group is extracted and defines a constraint for the conserved bond. In each example, the conserved bond is labelled with a small arrow. The reaction class imposes that the number of each type of bond including the bond type, the identity of each neighbouring atom and the total bond order of the neighbouring atoms is conserved. The numbers next to the atoms show their total bond order.*

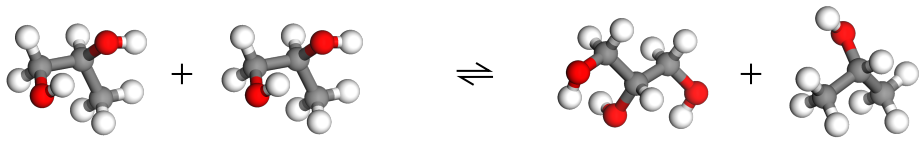
2.3 Reference data

2.3.1 Carbon-Hydrogen-Oxygen species

Reference data for 920 gas-phase species containing carbon, hydrogen and oxygen with known enthalpies of formation were retrieved from the NIST Chemistry WebBook [54]. This set includes open- and closed-shell species. The largest species is composed of 32 carbon and 66 hydrogen atoms. For each species, the reported 3D geometry was taken from the NIST Chemistry WebBook [54] as an initial guess of the geometry for the electronic structure calculations. A full list of the species is given in the Supplementary Material provided by Buerger et al. [11].

2.3.2 Other species

Reference data for other species, including titanium- and chlorine-containing species, were retrieved from various sources. The final set of reference values are presented in Table 1. The additional chlorine-containing species (listed under the heading *Other species*



Constraints	Reactants		Reactants $\sum$	Products $\sum$	Products	
3	2	2	4	4	3	1
0 0 3 0	1	1	2	2	2	0
0 1 3	2	2	4	4	4	0
0 0 1 1 3 0	1	1	2	2	2	0
1 3 3	1	1	2	2	1	1
3 3	1	1	2	2	1	1
0 1 3 0 0 0	1	1	2	2	0	2
3 0 0	3	3	6	6	0	6

Figure 5: *Example reaction for reaction class RC4. The number of each type of bond including the bond type and identity of each neighbouring atom and the total bond order of these neighbouring atoms is conserved on either side of the reaction. The numbers next to the atoms define their total bond order.*

in Table 1) are required for the validation of the data for the chlorine-containing titanium species. The reference values for TiOCl and $TiO(OH)_2$ have been revised as part of this work and are discussed in Sections 3.2.3 and 3.2.4. Ground state geometries for each titania species were taken from previous works [9, 99].

Table 1: *Reference data for standard enthalpies of formation for relevant species.*

species	$\Delta_f H^\circ_{298.15\,K}$ [kcal mol^{-1}]	species	$\Delta_f H^\circ_{298.15\,K}$ [kcal mol^{-1}]
Ti−Cl species		*Other species*	
$TiCl_4$	−182.40 [14, 54]	ClO	24.29 [1, 5, 7, 14, 54]
$TiCl_3$	−121.50 [42]	ClO_2	23.42 [14, 54]
$TiCl_2$	−49.00 [42]	ClO_3	48.04 [83]
TiCl	40.90 [42]	ClO_4	54.80 [88]
		OClO	22.6 [7, 14, 22, 54, 65]
Ti−O−Cl species		Cl_2	0.00 [14, 54]
$TiOCl_2$	−141.80 [96]	ClOCl	19.79 [1]
TiOCl	−68.42 [a]	Cl_2O	21.51 [14, 54]
		Cl_2O_2	36.86 [14, 51, 52, 54]
Ti−O species		Cl_2O_3	32.74 [1]
TiO_2	−73.00 [14, 54]	Cl_2O_4	44.48 [53]
TiO	13.00 [14, 54]	Cl_2O_5	61.74 [53]
		Cl_2O_6	66.56 [53]
Ti−O−H species		Cl_2O_7	76.79 [53]
$Ti(OH)_4$	−303.20 [96]	ClOClO	41.95 [14, 52, 54]
$TiO(OH)_2$	−200.65 [a]	ClOOCl	31.79 [1]
		O_2	0.00 [14, 54]
Ti−O−C−H species			
$Ti(OC_3H_7)_4$	−360.40 ± 2.20 [14, 54]		
$Ti(OC_2H_5)_4$	−324.60 ± 2.40 [14, 54]		
Other Ti species			
TiH	116.4 ± 2.3 [81]		

[a] this work

2.4 Algorithms

2.4.1 Overview

Figure 6 gives an overview of the method used to calculate informed estimates of the enthalpy of formation from a set of reference data. A global cross-validation was used to assess the consistency of the reference data. Data which gave cause for concern were identified and excluded. Multiple EBRs were identified and used to construct a distribution of estimates for the enthalpy of formation of each target species. The resulting distributions were post-processed to derive an informed estimate for each target species.

The algorithm used to identify individual EBRs is introduced in Section 2.4.2. This is extended to identify multiple EBRs in Section 2.4.3. The global cross-validation is described in Section 2.4.4. A modified version of the cross-validation that can be applied to a predefined set of species is introduced in Section 2.4.5. Pseudocode listings for the algorithms are provided as Supplementary Material.

2.4.2 Identification of an individual error-cancelling balanced reaction

Linear programming Linear programming is a constrained optimisation technique [21, 36, 95]. It was used in this work to identify possible EBRs fulfilling the constraints defined

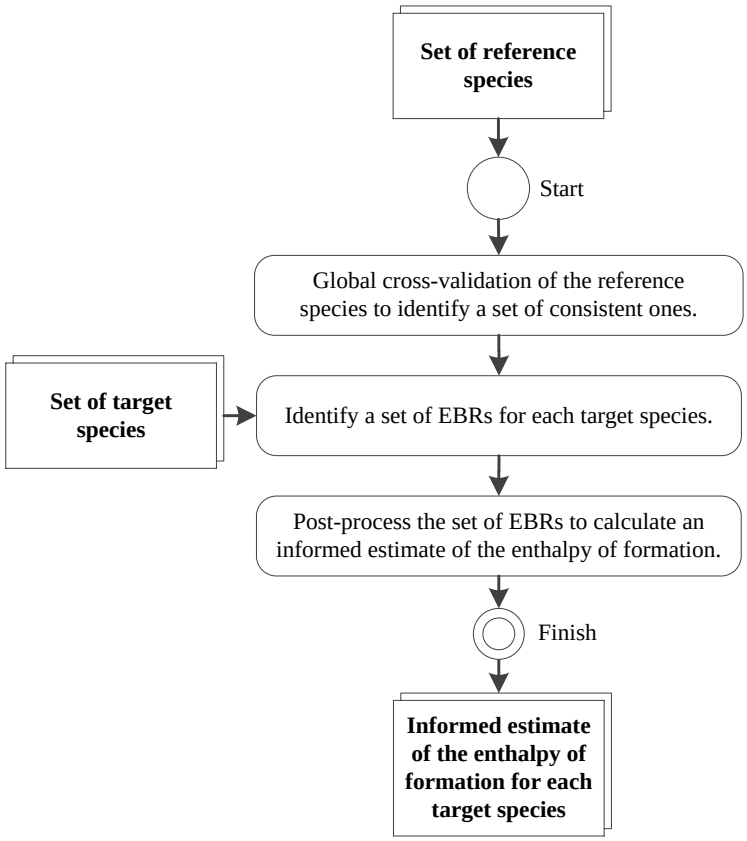


Figure 6: *Overview of the method used to estimate the enthalpy of formation.*

by the chosen reaction class. Each EBR is defined by a combination of reactants and products. The species must be chosen so as to conserve electronic and structural properties on either side of the reaction as required by the reaction class. This problem can be expressed in form of an objective function which can be solved using linear programming.

Problem The problem of identifying a possible EBR can be defined by applying the general linear objective function [21, 36, 95]. In this work it is defined by,

$$f(\nu) = \sum_{i=1}^{N_{\text{Species}}} |\nu_i| \sum_{j=1}^{N_{\text{Constraints}}} c_{ij} \tag{1}$$

where N_{Species} is the number of species, $N_{\text{Constraints}}$ is the number of distinct constraints, ν_i are the stoichiometric coefficients and c_{ij} are the constraints, defined by the selected reaction class. The objective function f is minimised with respect to ν, subject to the constraints

$$\sum_{i=1}^{N_{\text{Species}}} \nu_i c_{ij} = 0, \quad \forall j \in \{1, \ldots, N_{\text{Constraints}}\}, \tag{2}$$

such that the required quantities are conserved (see Section 2.1 for the definition of the reaction classes).

Solver Many linear programming solvers [12, 17, 37, 39, 55] are available. The GNU Linear Programming Kit (GLPK) [37] and lp_solve [55] linear programming solver were used in preliminary tests. It was observed that lp_solve did not always manage to solve the defined problem. This issue was also encountered in the work of Gearhart et al. [32]. No issues were encountered with GLPK and it was therefore used for the remainder of this work.

2.4.3 Identification of multiple error-cancelling balanced reactions

Manual identification of EBRs can be time-consuming and error-prone. An automated procedure that can be applied systematically to identify multiple EBRs offers many advantages, including the ability to calculate a distribution of values for the enthalpy of formation of any given species.

Figure 7 presents an overview of the algorithm used to identify multiple EBRs. Details about the algorithm are given below. In addition, a detailed description of the individual steps as well as two examples are provided as Supplementary Material.

Input This algorithm requires: (i) A set of reference data. (ii) A list of the target species for which the standard enthalpy of formation should be estimated. (iii) A hierarchy of reaction classes which are ordered from most to least restrictive. (iv) The minimum number of required EBRs (n). (v) The maximum number of identification attempts. (vi) The maximum search depth.

Output A set of identified EBRs for each target species. The EBRs may be subject to additional user defined constraints. For example to prevent a one-to-one mapping between a reactant and a product in cases where different values of the enthalpy of formation are available for a given species and are being evaluated to assess the accuracy of the reference data.

Initialisation The hierarchy of reaction classes is iteratively searched to find an initial EBR. The same reaction class is recommended to be used for the subsequent recursive identification of multiple EBRs.

Identification of EBRs The set of species that are included in the reference set is systematically manipulated in order to identify multiple EBRs. The manipulation is implemented using a recursive algorithm.

The initial step after entering the recursion is to check whether the required number of EBRs have been found, or whether the maximum search depth or maximum number of search attempts have been reached. If so, the algorithm is completed. If not, the algorithm recursively excludes from the reference set each combination of species (ignoring the target species) that exist in the current set of identified EBRs. At each recursion, the algorithm attempts to identify an EBR. If a distinct EBR is found and it adheres to any additional user defined constraints, it is added to the set of identified EBRs.

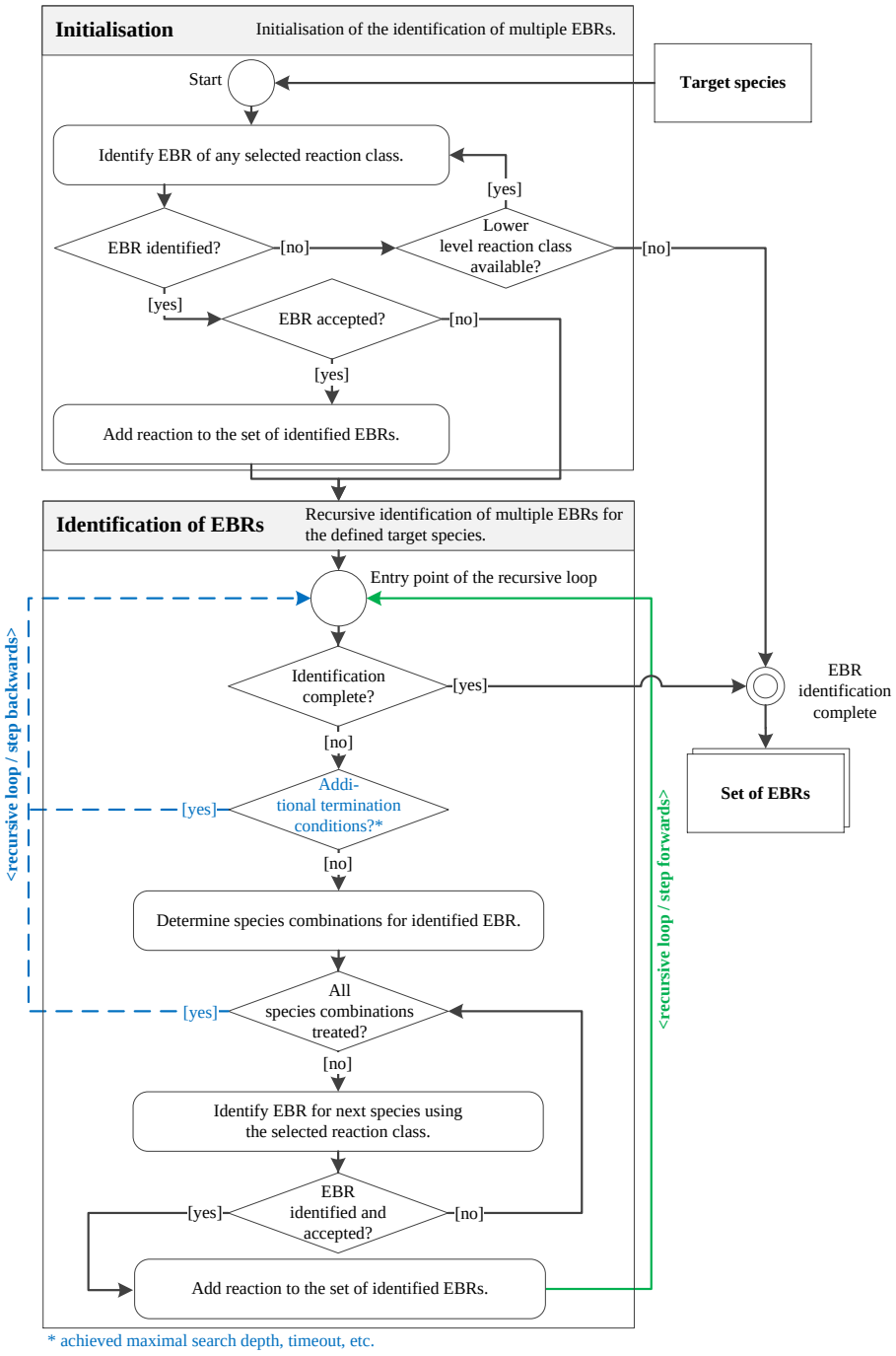


Figure 7: *Recursive algorithm used to identify a set of error-cancelling balanced reactions (EBRs). The input includes a set of reference data and a list of the target species for which the enthalpy of formation is to be estimated. A set of EBRs, which can be used to estimate the standard enthalpy of formation of the target species, are recursively identified and returned. The forward step in the recursion is indicated by the green solid line and the backward steps by the blue dashed lines.*

The algorithm terminates when it reaches the required number of EBRs, or when all combinations of species that exist in the set of identified EBRs have been excluded and analysed, or when it reaches a maximum number of search attempts.

This type of recursive algorithm is commonly used for processing an abstract data type known as a tree. A well-known example from the chemical literature is the mechanism reduction algorithm introduced by Lu and Law [56] and its subsequent developments [57, 66, 69, 70].

2.4.4 Global cross-validation

It is important to allow for the possibility that some of the reference data are potentially inconsistent. Simply excluding data for which the absolute difference between the calculated enthalpy of formation and the reference value exceeds a predefined error threshold could lead to the exclusion of accurate and consistent data, while inaccurate and inconsistent data remain due to the dependence on the order of processing the reference data. A method is required to:

- Assess the consistency of the reference data independently of the order of processing the data.
- Choose the reference data for a species where multiple conflicting choices of data exist.

Evaluating every possible combination of data would identify potentially inconsistent species. However, this is intractable for large reference sets. An alternative cross-validation algorithm is proposed to solve this issue. Figure 8 gives a simplified illustration of the algorithm.

Cross-validation of data sets is widely used in the field of data mining and statistical analysis [45, 79]. It assesses the predictive power of a model by separating the given data into complementary test and training sets. The training set is used to train the model which then attempts to predict results from the test set. The algorithm presented in this work has been developed based on this concept. It relates the resulting error from the cross-validation to each species in the reference data set. Based on the calculated error contributions, potentially inconsistent species are isolated and iteratively excluded from the reference set. The cross-validation is continued until defined convergence criteria are achieved or no further changes are observed.

Overview The algorithm used for the global cross-validation can be organised into three distinct modules:

- The *data pre-processing* is concerned with an initial evaluation of the reference data. It calculates an initial error contribution for each species. This provides the basis for further analysis.

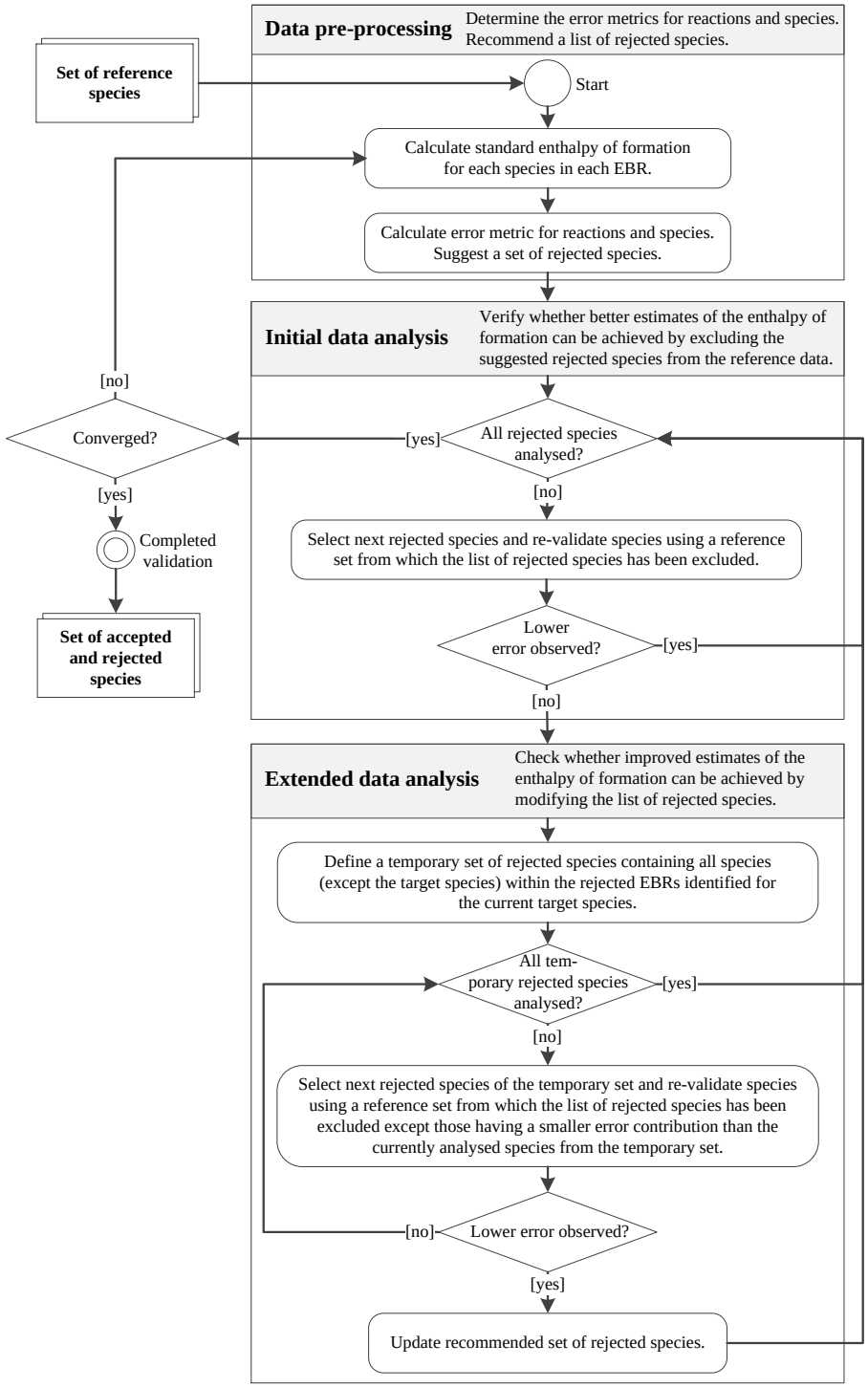


Figure 8: *Global cross-validation algorithm. A set of reference data is needed as input. Data which are likely to introduce inaccuracies are isolated by analysing error-cancelling balanced reactions (EBRs). The output is a set of consistent (accepted) and a set of inconsistent (rejected) reference data.*

- Based on the species error contribution, the *initial data analysis* attempts to identify and exclude the species from which the errors originate.
- If the initial data analysis is inconclusive, an *extended data analysis* is conducted to investigate problematic species in more detail.

In the following, the modules of the algorithm are explained in more detail. Where required, the algorithms presented in Sections 2.4.2 and 2.4.3 are used to identify a set of EBRs and estimate the standard enthalpy of formation for a species. An algorithmic description of the individual steps is provided as Supplementary Material.

Input This algorithm requires: (i) The full list of reference data to be evaluated. (ii) A hierarchy of reaction classes which are ordered from most to least restrictive class. (iii) The minimum number of required EBRs (n). (iv) The maximum number of identification attempts. (v) The maximum search depth. (vi) The magnitude of maximal acceptable error for each species. (vii) An upper limit of the number of iterations. (viii) A choice of how to calculate the error due to each species.

Output Two lists of reference data: (i) A list of reference data found to be consistent when used to estimate the enthalpies of formation. (ii) A list of potentially inconsistent reference data which were found to introduce inaccuracies when used to estimate enthalpies of formation.

Data pre-processing The pre-processing is used to initially identify a set of EBRs to calculate the standard enthalpy of formation for each species. Isolated species for which no EBRs are found are identified and excluded. The validation uses a pre-defined hierarchy of reaction classes. Results from the reaction class, highest in the hierarchy, leading to a successful termination are used and collected in a reaction set, R. The set of species participating in reaction $r \in R$ is denoted by,

$$S^{\mathrm{ref}}(r) := \{s \in S^{\mathrm{ref}} | s \text{ is involved in } r\}, \tag{3}$$

where S^{ref} is the full reference set containing all species. An error metric for each combination of reaction $r \in R$ and species $s \in S^{\mathrm{ref}}$ is calculated. The absolute difference between the reference value, $\Delta_{\mathrm{f}}^{\mathrm{ref}} H^{\circ}_{298.15\ K}(s)$, and the calculated enthalpy of formation, $\Delta_{\mathrm{f}} H^{\circ}_{298.15\ K}(r,s)$, is calculated,

$$\varepsilon_{\mathrm{r}}(r,s) = \begin{cases} |\Delta_{\mathrm{f}}^{\mathrm{ref}} H^{\circ}_{298.15\ K}(s) - \Delta_{\mathrm{f}} H^{\circ}_{298.15\ K}(r,s)| & \text{if } s \in S^{\mathrm{ref}}(r) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where reaction r is used to estimate the standard enthalpy of formation for species s.

A reaction $r \in R$ is labelled to be accepted if $\varepsilon_{\mathrm{r}}(r,s)$ is smaller than a defined upper limit $\varepsilon_{\mathrm{r}}^{\max}$ for all species s,

$$\varepsilon_{\mathrm{r}}(r,s) < \varepsilon_{\mathrm{r}}^{\max} \qquad \forall s \in S^{\mathrm{ref}}, \tag{5}$$

and otherwise rejected. The set of rejected reactions for a species $s \in S^{\text{ref}}$, for which the standard enthalpy of formation is to be determined, is defined by,

$$R^{\text{rej}}(s) := \{r \in R | \varepsilon_{\text{r}}(r,s) \geq \varepsilon_{\text{r}}^{\text{max}}\}. \tag{6}$$

The full set of all rejected reactions is then defined by,

$$R^{\text{rej}} := \bigcup_{s \in S^{\text{ref}}} R^{\text{rej}}(s), \tag{7}$$

and that of accepted reactions by,

$$R^{\text{acc}} := \{r \in R | \varepsilon_{\text{r}}(r,s) < \varepsilon_{\text{r}}^{\text{max}}\}, \qquad \forall s \in S^{\text{ref}} \tag{8}$$

so that by definition the sets of accepted and rejected reactions are complements of each other within R, *i.e.* $R^{\text{acc}} \cap R^{\text{rej}} = \emptyset$ and $R^{\text{acc}} \cup R^{\text{rej}} = R$.

The average error of a set of rejected reactions associated with a species $s \in S^{\text{ref}}$ is defined by,

$$\overline{\varepsilon}_{\text{r}}(s) = \frac{\sum_{r \in R^{\text{rej}}(s)} \varepsilon_{\text{r}}(r,s)}{|R^{\text{rej}}(s)|}, \tag{9}$$

where the vertical bar notation denotes the number of elements within a set. Although not used here, the median error could be employed. The contribution to ε_{r} can be calculated for each species s as a weighted contribution,

$$\varepsilon_{\text{s}}(r,s) = \frac{\nu(r,s)\,\varepsilon_{\text{r}}(r,s)}{\nu(r)} \qquad \forall r \in R, s \in S^{\text{ref}}, \tag{10}$$

where $\nu(r,s)$ are the weights of s in r. For example, the stoichiometry, number of atoms or the product of both. The sum of weights over $S^{\text{ref}}(r)$ is defined by,

$$\nu(r) = \sum_{s \in S^{\text{ref}}(r)} \nu(r,s) \qquad \forall r \in R. \tag{11}$$

The average error contribution for s is calculated from $\varepsilon_s(r,s)$,

$$\overline{\varepsilon}_{\text{s}}(s) = \frac{\sum_{r \in R^{\text{rej}}(s)} \varepsilon_{\text{s}}(r,s)}{|R^{\text{rej}}(s)|} \qquad \forall s \in S^{\text{ref}}. \tag{12}$$

Instead of the mean error contribution for s, the median error contribution could be used. In the remainder of the work, only results using Equation (12) are reported. The set of species assumed to be consistent is defined by,

$$S := \bigcup_{r \in R^{\text{acc}}} S^{\text{ref}}(r), \tag{13}$$

where all species appearing in identified reactions with an error lower than $\varepsilon_{\text{r}}^{\text{max}}$ are assumed to be consistent.

The full set of rejected species is then simply defined by,

$$S^{\text{rej}} := S^{\text{ref}} \setminus S. \tag{14}$$

At this stage it is unclear which species are the cause of the error $\varepsilon_{\text{r}}(r,s) \geq \varepsilon_{\text{r}}^{\text{max}}$ in the rejected reactions R^{rej}. It is assumed that the error could originate from any of the species.

Initial data analysis The initial data analysis checks whether improved estimates of the enthalpy of formation can be achieved by re-analysing each species in S^{rej} using the current subset of accepted reference species S, initially defined by Equation (13). The species in S^{rej} are analysed in order, based on the size of the species error contribution, largest first.

A working set of rejected species $\hat{S}^{\text{rej}}$ is defined. This is initially equal to the set of rejected species,

$$\hat{S}^{\text{rej}} \leftarrow S^{\text{rej}}. \tag{15}$$

The species with the largest error contribution $\overline{\varepsilon}_{\text{s}}$,

$$s^{\max} := \underset{y \in \hat{S}^{\text{rej}}}{\operatorname{argmax}}\, \bar{\varepsilon}_{\text{s}}(y), \tag{16}$$

is selected and excluded from $\hat{S}^{\text{rej}}$,

$$\hat{S}^{\text{rej}} \leftarrow \hat{S}^{\text{rej}} \setminus \{s^{\max}\}. \tag{17}$$

A new set of EBRs $\hat{R}^{\text{new}}(s^{\max})$ is determined for $s^{\max}$ using reference set S. $\hat{R}^{\text{new}}(s^{\max})$ is validated against the previous set of EBRs $R(s^{\max})$ for the same target species.

Two different validation methods are used: (i) The first validation method assumes that the alternative set of EBRs $\hat{R}^{\text{new}}(s^{\max})$ is an improvement over the previous set of EBRs $R(s^{\max})$ if the number of rejected reactions is smaller or, in cases with the same number of rejected reactions, the average error, as defined by Equation (9), is smaller than previously. (ii) The second validation method calculates and compares the average error, given by Equation (9), between the two sets.

In cases where the error is reduced, $s^{\max}$ is added to S. Otherwise, we enter the extended data analysis which attempts to modify the reference set S in order to reduce the error. After completion of the extended analysis the next iteration is performed.

Following a complete iteration through the set of rejected species, it is checked whether convergence has been achieved. The convergence criteria are whether or not the sets of accepted and rejected species are unchanged compared to the previous iteration or whether a maximum number of iterations has been achieved. If this is the case, the final sets of accepted and rejected species are returned. Otherwise, the current set of accepted species defines the new set of reference species, $S^{\text{ref}} \leftarrow S$, and is re-analysed via another iteration of the cross-validation.

Extended data analysis The purpose of the extended analysis is to improve the predictive performance for species s. This is accomplished by checking whether better results can be achieved by iteratively extending the set of accepted reference species S with species previously found to be present in rejected reactions identified for s.

The algorithm starts by identifying the set of species S^{rej} that were present in rejected reactions identified for s,

$$S^{\text{rej}}(s) := \bigcup_{r \in R^{\text{rej}}(s)} S^{\text{ref}}(r). \tag{18}$$

The species under investigation s is excluded from this set,

$$\tilde{S}^{\text{rej}}(s) := S^{\text{rej}}(s) \setminus \{s\}. \tag{19}$$

The species in $\tilde{S}^{\text{rej}}(s)$ with the smallest error contribution $\overline{\varepsilon}_{\text{s}}(s)$,

$$s^{\min}(s) := \underset{y \in \tilde{S}^{\text{rej}}(s)}{\operatorname{argmin}} \bar{\varepsilon}_{\text{s}}(y), \tag{20}$$

is excluded from $\tilde{S}^{\text{rej}}(s)$,

$$\tilde{S}^{\text{rej}}(s) \leftarrow \tilde{S}^{\text{rej}}(s) \setminus \{s^{\min}(s)\}, \tag{21}$$

and is used to define the set of species that have smaller average error contributions. This set is defined as a subset of the full set of rejected species S^{rej},

$$\tilde{S}^{\text{err}}(s) := \{x \in S^{\text{rej}} | \bar{\varepsilon}_{\text{s}}(x) \leq \bar{\varepsilon}_{\text{s}}(s^{\min}(s))\}. \tag{22}$$

If there is more than one species with the same species error contribution $\overline{\varepsilon}_{\text{s}}(s)$, a random selection is conducted.

Combining $\tilde{S}^{\text{err}}(s)$ and S,

$$S \leftarrow S \cup \tilde{S}^{\text{err}}(s), \tag{23}$$

results in an updated set of reference data. All species within $\tilde{S}^{\text{err}}(s)$ have a lower average species error contribution than $s^{\min}(s)$ and it is assumed that by adding them to the reference set S, the resulting error for the identified alternative EBRs will be smaller than by adding species with larger average error contributions. The revised species set S is used to identify distinct EBRs and a new set of reactions $\tilde{R}^{\text{new}}(s)$ for species s.

If the error, based on the selected validation method as discussed for the initial data analysis, is reduced by the use of $\tilde{R}^{\text{new}}(s)$ instead of $R(s)$, the set of species required for the definition of the new reaction set,

$$S^{\text{new}}(s) := \bigcup_{r \in \tilde{R}^{\text{new}}(s)} S^{\text{ref}}(r). \tag{24}$$

is used to update the list of recommended rejected species,

$$S^{\text{rej}} \leftarrow S^{\text{rej}} \setminus S^{\text{new}}(s), \tag{25}$$

so that species required to achieve the improvement are excluded from the set of rejected species but the others remain in S^{rej}, even if they were temporarily in the reference set during the extended data analysis. Using the refined set of rejected species, S is then updated according to,

$$S \leftarrow S^{\text{ref}} \setminus S^{\text{rej}}. \tag{26}$$

If the error has increased or no results could be obtained, the species with the next larger $\overline{\varepsilon}_{\text{s}}(s)$ in $\tilde{S}^{\text{rej}}(s)$ is used. This is repeated until better results are achieved or all species in $\tilde{S}^{\text{rej}}(s)$ have been treated.

2.4.5 Modified global cross-validation

In some cases it is desired to only validate a subset of species rather than the full reference set. This is required for the validation of the titanium-containing species. The global cross-validation method described in the previous section is modified by adding an additional input to specify a set of target species. Instead of validating every species within the reference set, only the specified target species are validated. The validation of all other species is skipped. It is assumed that they are consistent.

3 Results

The performance of the framework is demonstrated using different test cases. These include species from organic and inorganic chemistry, including transition metal complexes, and are organised as follows:

- The first test case consists of 920 species containing carbon, hydrogen and oxygen [11]. This is a large reference set and is used to demonstrate the functionality of the global cross-validation using a representative set of reference data that are widely available in the literature. Validation of these species is required to validate the titanium-containing species.
- The second test case consists of oxychloride species listed in Table 1. This demonstrates the application of the framework to a small reference set where there is considerable variation in the literature values of the reference data. Validation of these species data is essential to validate the titanium-chlorine-containing species.
- The final test case consists of titanium-containing species. The species are classified as Titanium-Oxygen-Chlorine (*i.e.* Ti−Cl, Ti−O−Cl and Ti−O species) and Titanium-Oxygen-Carbon-Hydrogen species (*i.e.* Ti−O, Ti−O−H, Ti−O−C−H and TiH species). The considered species are listed in Table 1. The estimation of standard enthalpies of formation for transition metal complexes is challenging [23, 24, 46, 47, 50] and several quantum chemistry methods have been validated for such transition metal complexes. These species were chosen to demonstrate the applicability of the framework to such systems.

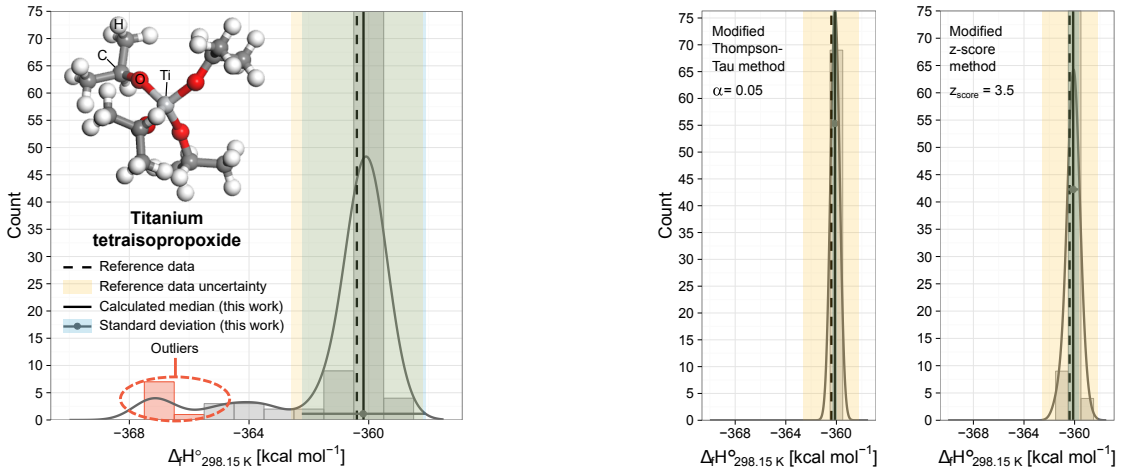
The following sections show that the framework delivers significant benefits that derive from the consideration of multiple EBRs, outlier detections, and the global cross-validation of the reference data. The framework is used to calculate and recommend new reference values of the standard enthalpy of formation for TiOCl and $TiO(OH)_2$. These are important species in titanium oxygen systems [9, 99, 100].

3.1 Benefits of multiple error-cancelling balanced reactions

The use of multiple error-cancelling balanced reactions results in significant statistical benefits compared to using a single reaction. This includes a systematic and standardised

way of analysing the results and the calculation of an informed estimate. The algorithms described in Section 2.4.2 and 2.4.3 were used for the identification of multiple reactions for a given target species.

Figure 9 presents example results for titanium tetraisopropoxide (TTIP, $Ti(OC_3H_7)_4$) using reaction class RC2 (isodesmic reactions), calculated from multiple EBRs. The use of multiple EBRs enables the construction of a histogram of calculated estimates of the standard enthalpy of formation for each sample.



(a) *Full set of identified EBRs without the exclusion of outliers.*

(b) *Application of two common outlier detection methods: The modified Thompson-Tau method [4] and the modified z-score method [44].*

Figure 9: *Histogram of the estimated values of the standard enthalpy of formation for titanium tetraisopropoxide (TTIP, $Ti(OC_3H_7)_4$). 106 distinct isodesmic reactions (RC2) were identified. The kernel density was estimated from post-processing the histogram. The reference value of -360.40 kcal mol^{-1} (dashed line) for TTIP was taken from the NIST Chemistry WebBook [54]. Outliers giving particularly poor estimates are highlighted.*

The histogram derived from the full set of identified EBRs is shown in Figure 9a. The use of a central measure leads overall to a better estimate of the enthalpy of formation compared to an estimate that relies on a single reaction. The width of the distribution can be used to gain information about the statistical uncertainty of the calculation. The distribution in Figure 9a results in an estimate of -360.18 ± 2.06 kcal mol^{-1}, which compares well to a reference value of -360.40 ± 2.15 kcal mol^{-1} [54]. The difference is significantly less than 3.00 kcal mol^{-1}, which is the chemical accuracy of any transition metal complex [23, 24], and is within the reported statistical uncertainty of 2.15 kcal mol^{-1}.

The exclusion of outliers provides additional benefit. Figure 9b shows two post-processed distributions that were determined by automatically identifying and excluding outliers from the full distribution (Figure 9a). A modified Thompson-Tau method [4] with an α value of 0.05 and a modified z-score method [44] with a z_{score} value of 3.5 were used. Important information about potentially inconsistent reference data is included in outliers.

By systematically analysing this information, inconsistent data can be identified and excluded from the reference set. This information is exploited by the global cross-validation.

Similar results were observed for the test case consisting of species containing carbon, hydrogen and oxygen. This is discussed in detail elsewhere [11].

3.2 Identification of potentially inconsistent reference data

3.2.1 Carbon-Hydrogen-Oxygen-containing species

Figure 10 (top panel) shows the decrease in the mean absolute error that was achieved by iteratively excluding inconsistent species from the set of carbon, hydrogen and oxygen containing species. The bottom panel shows the number of species which have been excluded for each iteration.

Different configurations of the parameters in the global cross-validation were evaluated. The first parameter was the rejection threshold ε_r^{max} as per Equation (5). It defines the magnitude of the maximum acceptable error for each species as defined by Equation (4). The second parameter defines the weighting of the species error contribution within a reaction as used in Equation (10). The weights in Equation (10) were defined as either the stoichiometry (S1), the number of atoms (S2) or the product of both (S3). The final parameter defines the validation method used to compare the two sets of EBRs. Method V1 assumed that the set with the smaller number of rejected reactions was the preferred set of EBRs. In cases with the same number of rejected reactions, the average error was used for comparison. Method V2 always used the average error.

A rapid asymptotic decrease in the observed mean absolute error was observed for all parameter configurations. The results were grouped by the value of ε_r^{max}. This implies that ε_r^{max} was the most influential and sensitive parameter. The choice of how to weigh the species error contribution and the validation method had less impact. The lower the rejection threshold ε_r^{max}, the more stringent and uncompromising the identification of consistent reference data. A mean absolute error lower than ε_r^{max} was observed for a threshold $\varepsilon_r^{max} \geq 2.0$ kcal mol^{-1}. On the other hand, for $\varepsilon_r^{max} \leq 1.0$ kcal mol^{-1} a mean absolute error just above 1.0 kcal mol^{-1} was observed. Reasons for this could include the loss of the statistical benefits from the selection of multiple EBRs due to the reduction in the size of the set of reference data, the choice of reaction class or the selected level of theory. In this work, errors in this range are to be expected from the use of the B97-1/6-311+G(d,p) level of theory. Below this point it is difficult to define whether the resulting discrepancies stem from the electronic structure calculations or inconsistencies in the reference data. Therefore, a trade-off must be made when choosing the level of theory, the configuration of the global cross-validation and the reaction class. Repeated analyses using the same configuration only showed small differences in the calculated mean absolute errors.

3.2.2 Oxychloride species

The reported literature values of the standard enthalpies of formation vary significantly for some of the oxychloride species. For example, differences of up to 22.56 kcal mol^{-1}

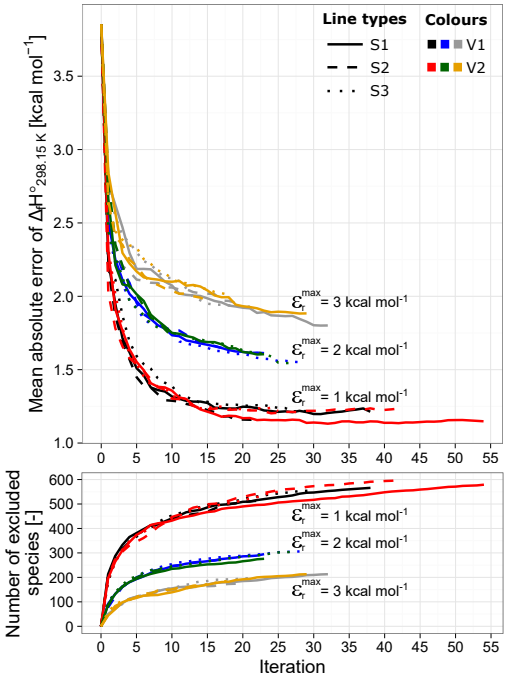


Figure 10: *The mean absolute error (top panel) and the number of excluded species (bottom panel) for different configurations of the parameters in the global cross-validation. The rejection threshold parameter ε_r^{max} defines the magnitude of the maximum acceptable error for each species. The weighting of the species error contribution within a reaction as used in Equation (10) is either calculated using the stoichiometry (S1), the number of atoms (S2) or the product of both (S3). Two different validation methods are used to compare sets of error-cancelling balanced reactions (EBRs). The first method (V1) recommends the set with the smaller number of rejected EBRs. In cases with the same number of rejected reactions, the average error contribution is used for comparison. The second method (V2) always uses the average error contribution.*

can be observed between the reported literature values for Cl_2O_7. Choosing the wrong reference value inevitably leads to the propagation of the error to any estimates of the enthalpy of formation that use this data. The global cross-validation was used to find the most suitable set of reference data to estimate the standard enthalpies of formation for species with large uncertainties in the literature data.

The combination of a small reference set with the large uncertainties in reference data impacted the performance of the global cross-validation algorithm. Therefore, configurations using $\varepsilon_r^{max} \geq 3.0$ kcal mol^{-1} were applied. Smaller values of ε_r^{max} were found to be unsuitable due to the exclusion of too many species.

Multiple validation executions identified three potentially inconsistent reference species ($\varepsilon_r^{max} = 3.0$ kcal mol^{-1}): ClOCl, ClOClO and Cl_2O_7. Using the full reference set led to a mean absolute error of 3.70 kcal mol^{-1}. The effect of excluding permutations of the above three potentially inconsistent reference species was evaluated. The largest im-

provement was achieved by excluding all three species from the reference set. However, the level of improvement was marginal. In some situations, the statistical benefit gained by considering more reference species outweighs the benefits of excluding species.

3.2.3 Titanium-Oxygen-Chlorine species

The number of titanium-oxygen-chlorine species for which reference data exist is very limited. As discussed in the previous section, the reference set for the oxychloride species is already small. This significantly affects the identification of multiple EBRs for the titanium-chlorides and titanium-oxychlorides, and consequently the accuracy of the estimated values of the standard enthalpy of formation. The modified global cross-validation (see Section 2.4.5) was applied to assess the quality of the reference data for $TiCl_4$, $TiCl_3$, $TiCl_2$, TiCl, TiOCl, $TiOCl_2$, TiO and TiO_2 in addition to the oxychloride species.

The first execution of the global-cross validation was performed using a rejection threshold of 3.0 kcal mol^{-1}, which is the assumed chemical accuracy for transition metal complexes [23, 24]. It was found that the reported NIST Chemistry WebBook reference values of the standard enthalpy of formation of TiOCl (with a value of -58.38 kcal mol^{-1}) and $TiOCl_2$ (with a value of -130.39 kcal mol^{-1}) were potentially inconsistent. When these values were replaced with reference values for TiOCl and $TiOCl_2$ as estimated by West et al. [99] or for $TiOCl_2$ by Wang et al. [96], the cross-validation found all the titanium-containing species to be consistent.

Upon decreasing the rejection threshold to 2.0 kcal mol^{-1}, the reference value for TiOCl was found to be potentially inconsistent. Excluding TiOCl and using the reference value for $TiOCl_2$ reported by Wang et al. [96] (and recommended by the cross-validation) led to a mean absolute error of 0.78 kcal mol^{-1}. A new estimate of -68.42 ± 0.54 kcal mol^{-1} (see Table 1) was calculated for TiOCl using reaction class RC2. This value differs significantly from that reported in the NIST Chemistry WebBook [54] and to a lesser extent from the value reported by West et al. [99].

3.2.4 Titanium-Oxygen-Carbon-Hydrogen species

The consistency of the following Ti−O−C−H species were validated using the modified global cross-validation: $Ti(OC_3H_7)_4$, $Ti(OC_2H_5)_4$, $Ti(OH)_4$, $TiO(OH)_2$, TiO and TiO_2. TiH was excluded from the analysis because of the absence of another species containing a Ti−H bond. The chemical accuracy of transition metal complexes [23, 24] was used for the rejection threshold parameter ($\varepsilon_r^{max} = 3.0$ kcal mol^{-1}). The reference set included the validated ($\varepsilon_r^{max} = 1.0$ kcal mol^{-1}) set of carbon, hydrogen and oxygen containing species.

$TiO(OH)_2$ and TiO_2 were found to be potentially inconsistent. This result was unexpected for TiO_2. Manually inspecting the identified EBRs revealed that both species, $TiO(OH)_2$ and TiO_2, always appeared as a pair. As a result of the limited number of titanium-containing reference data, the algorithm was not able to clearly identify the species from which the error originated.

TiO_2 was validated in the previous section for the Ti−O−Cl test case. Therefore, a new reference set consisting of the oxychloride species, all titanium-containing species and

the validated hydrocarbon bases species ($\varepsilon_r^{max} = 1.0$ kcal mol^{-1}) was defined. Using this set, a global cross-validation for the Ti−O−C−H species was performed which led to a slightly different conclusion. In this case only $TiO(OH)_2$ was identified to be potentially inconsistent. This was a consequence of the larger reference set which enabled the global cross-validation to identify the origin of the error. Excluding $TiO(OH)_2$ led to a mean absolute error of 0.35 kcal mol^{-1} for the Ti−O−C−H test case.

Based on the global cross-validation it was assumed that the reference values for TiO_2 and $Ti(OH)_4$ are accurate. Using reaction,

$$Ti(OH)_4 + TiO_2 \longleftrightarrow 2\,TiO(OH)_2, \tag{27}$$

gave an estimate of -200.65 kcal mol^{-1} compared to the reference value of -195.85 ± 11.83 kcal mol^{-1} [64] for $TiO(OH)_2$. The value of -200.65 ± 3.18 kcal mol^{-1} is recommended and listed in Table 1.

A further global cross-validation using reaction class RC1 (isogyric reactions) allowed the calculation of the standard enthalpy of formation of TiH as 110.81 ± 9.42 kcal mol^{-1}. The error is significantly worse than for the estimates where it was possible to use RC2. This is not entirely unexpected due to the less restrictive reaction class. Similar errors were observed for the other titanium-containing species when using RC1 instead of RC2.

3.3 Effect of reaction classes and excluding inconsistent reference data

The availability of sufficient reference data affects whether or not it is possible to apply higher-level reaction classes, for example RC4 and RC3. This is not an issue for the large set of species containing carbon, hydrogen and oxygen, but is a limitation for the other test cases. This section investigates the extent to which the choice of reaction class affects the results for the test case consisting of species containing carbon, hydrogen and oxygen.

Figure 11 shows the effect of the number of EBRs and the reaction class on the estimated enthalpies of formation using different sets of reference data. The lines represent the mean error calculated over 50 independent executions. The shaded areas represent the standard deviations. The full reference set (RS1) contains all 920 species taken from the NIST Chemistry WebBook [54]. RS2 and RS3 are reduced versions of RS1, determined by applying the global cross-validation with a rejection threshold of $\varepsilon_r^{max} = 1.0$ kcal mol^{-1}. RS2 was a randomly selected reference set out of 50 independent executions of the global cross-validation on the full reference set of 920 species containing carbon, hydrogen and oxygen. RS3 has been created by combining all species found to be consistent at least once in any of these independent runs.

The mean absolute errors converged rapidly to an asymptotic value for each configuration. The results were shown to be repeatable between independent runs, where the statistical uncertainty decreased as the number of EBRs increased. However, considerable differences between the asymptotic errors were observed for the different reference sets. As expected, the lowest mean absolute errors were reported using the reduced reference set RS2. Although the mean absolute error using RS3 was significantly reduced compared to

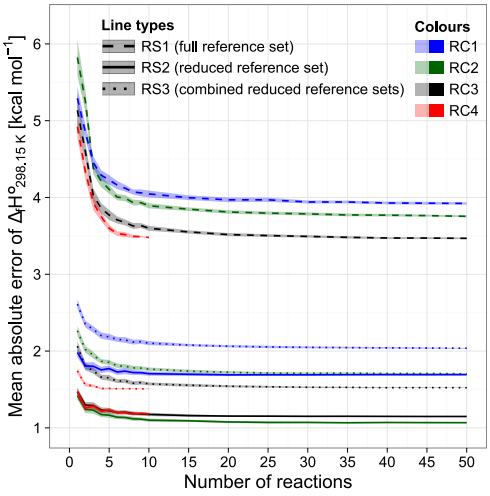


Figure 11: *The mean absolute error in the estimated values of the standard enthalpies of formation for the carbon, hydrogen and oxygen containing test species set as a function of the number of considered error-cancelling balanced reactions (EBRs) for different reference sets and reaction classes. The lines represent the calculated mean and the shaded areas show the standard deviation over 50 independent runs. Reference set RS1 consists of the full set of 920 hydrocarbon based species retrieved from the NIST Chemistry WebBook [54], RS2 is a single randomly selected reference set out of 50 independent executions and RS3 is a combination of all identified consistent reference species over 50 independent global-cross validation runs using the same configuration.*

the full reference set RS1, the error was larger than for RS2. This is a consequence of not always identifying the same EBRs from the space of possible solutions during the global cross-validation. Therefore, potentially less consistent species are present in reference set RS3.

Generally, the mean absolute errors followed the rigorousness of the reaction classes. The more rigorous the chosen reaction class, the more accurate the resulting estimate of the standard enthalpy of formation. Nevertheless, reaction class RC4 delivered less good results than RC2 for reference set RS2. This can be attributed to a loss of the statistical benefit due to a reduction in the numbers of EBRs that could be found for the RC4 reaction class. However, the effect of the reduction in the number of EBRs that could be found for RC4 did not carry through to the statistical uncertainties. The calculated statistical uncertainties were reduced by selecting a more rigorous reaction class. Table 2 reports the mean standard deviations over all reference species when estimating the standard enthalpies of formation. Applying outlier detection methods, such as the modified Thompson-Tau [92] and modified z-score method [44] reduced the statistical uncertainty. The most significant effect was observed for the full reference set RS1. This effect was reduced for the two validated reference sets (RS2 and RS3) due to the exclusion of inconsistent reference data.

Table 2: *Effect of reference sets, reaction classes and outlier detection methods on the mean standard deviations calculated over all reference species.*

reaction type	mean standard deviation [kcal mol^{-1}]								
	full reaction set			revised reaction set excluding outliers					
	RS1	RS2	RS3	RS1 [a]	RS2 [a]	RS3 [a]	RS1 [b]	RS2 [b]	RS3 [b]
RC1	4.08	1.45	2.03	2.85	1.43	1.86	1.72	1.21	1.37
RC2	5.21	1.22	1.79	2.69	1.20	1.69	1.43	1.03	1.16
RC3	4.41	1.24	1.78	2.50	1.22	1.70	1.45	0.92	1.19
RC4	2.98	1.14	1.33	1.78	1.06	1.11	1.66	0.88	1.08

[a] Modified z-score method with $z_{\text{score}} = 3.50$
[b] Modified Thompson-Tau method with $\alpha = 0.05$

For validation purposes, all calculations were additionally performed at the B3LYP/6-311+G(d,p) level of theory for species containing carbon, hydrogen and oxygen. Similar results were obtained, although the B97-1/6-311+G(d,p) led to slightly lower mean absolute errors. The other test cases were only evaluated using the reaction classes RC1 and RC2. In all cases, it was found that the reported mean absolute error using RC1 was significantly larger than by using RC2.

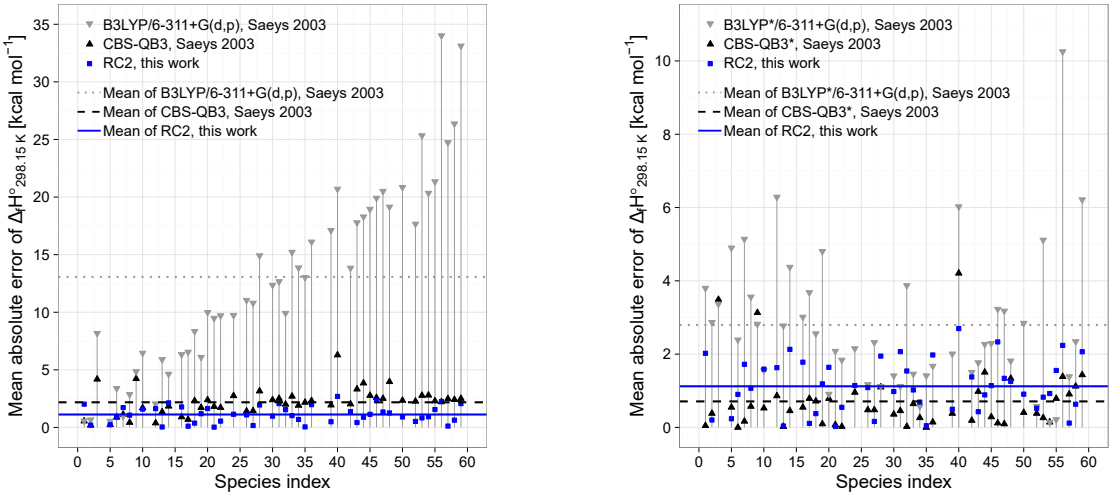
There is a trade-off between the reaction class, the number of EBRs, the choice of reference set and the expected accuracy of the resulting estimates. Given sufficient accurate reference data, an increased number of EBRs and a more restrictive reaction class improves the accuracy of the estimate. However, the use of an overly restrictive reaction class can result in accurate estimates for individual samples at the expense of losing the statistical benefits of multiple EBRs, resulting in an overall loss of accuracy. Loose constraints, such as the conservation of spin states with isogyric reactions (RC1), result in poorer estimates even if a large number of EBRs and a validated reference set is considered. Isodesmic reactions (RC2) give a good compromise, providing that the reference set is sufficiently large.

3.4 Comparison to other methods

3.4.1 Carbon-Hydrogen-Oxygen-containing species

Figure 12 compares the calculated mean absolute errors for species overlapping between this work and that of Saeys et al. [86]. Although calculations were conducted using all four reaction classes in combination with reference set RS2, only results using reaction class RC2 are shown for ease of presentation. No additional outlier detection method was applied. An accuracy comparison was conducted against popular quantum chemistry methods without (Figure 12a) and with additional AAC (Figure 12b). The species in Figure 12 are ordered by size and agree with the order defined by Saeys et al. [86].

The most accurate estimates reported in Figure 12a were achieved by applying the method presented in this work and have a mean absolute error of 1.12 kcal mol^{-1}. The use of



(a) *No atom-additivity-correction* **(b)** *With atom-additivity-correction*

Figure 12: *Comparison of calculated absolute errors against quantum chemistry methods with and without consideration of atom-additivity-correction (AAC). The data were reported by Saeys et al. [86].*

CBS-QB3 led to a mean absolute error of 2.18 kcal mol^{-1}, which is nearly twice the error obtained by using RC2 with RS2. The DFT method resulted in a significant mean absolute error of 13.07 kcal mol^{-1}. A clear size dependence of the error was observed for DFT. A smaller size dependence was detected for the CBS-QB3 method. The errors calculated by the method in this work did not depend on the size of the molecule. It must be noted that Redfern et al. [78] observed a dependence of the error on the size of the molecule using single isodesmic reactions (RC2) for *n*-alkane species. These calculations were repeated using the method presented in this work with RC2 and RS2 and did not show any dependence on size.

The errors using the quantum chemistry methods were significantly reduced by additionally considering AAC as shown in Figure 12b. The error dependence of the quantum chemistry methods on the size of the molecule was no longer observed. Despite significant improvement, using DFT with AAC still led to a mean absolute error of 2.80 kcal mol^{-1}. CBS-QB3 with AAC led to a mean absolute error of 0.71 kcal mol^{-1}. Even though the use of AAC significantly reduced the mean absolute errors, it is noted that AAC is not generally applicable to any system nor is it applicable for all levels of theory [86]. Replacing RC2 with more rigorous reaction classes reduced the error to 1.06 kcal mol^{-1} for reaction class RC3 and to 0.70 kcal mol^{-1} for reaction class RC4. Similar results were obtained using the B3LYP functional. The methodology presented in this work is capable of delivering estimates of the standard enthalpy of formation comparable in accuracy with computationally demanding quantum chemistry methods and without a dependence of the error on the size of the molecule.

3.4.2 Titanium-containing species

The standard enthalpies of formation for various titanium-containing species considered in this work were also calculated by Wang et al. [96] using the computationally demanding coupled-cluster method [CCSD(T)] with complete basis set extrapolation. This is also known as the *"gold standard"* of quantum chemistry [98].

Table 3: *Comparison of calculated standard enthalpies of formation for titanium-containing species using isodesmic reactions against selected reference values and CCSD(T)/CBS estimates as reported by Wang et al. [96]. The total energies were calculated at the B97-1/6-311+G(d,p) level of theory.*

species	this work[a] [kcal mol^{-1}]	CCSD(T)/CBS[b] [kcal mol^{-1}]	literature value[c] [kcal mol^{-1}]
$TiCl_4$	-181.61 ± 2.63	-181.5	-182.4
$TiOCl_2$	-141.26 ± 0.65	-141.8	-141.8[d]
TiO_2	-73.60 ± 0.68	-67.8	-73.0
$Ti(OH)_4$	-302.87 ± 0.22	-303.2	-303.2[d]

[a] Calculated using isodesmic reactions using the reference values defined in Table 1 and the set of species containing carbon, hydrogen and oxygen [11]. The B97-1/6-311+G(d,p) level of theory was used.
[b] Calculated by Wang et al. [96].
[c] Reference values as selected in this work (Table 1).
[d] Reference value taken from Wang et al. [96].

Table 3 presents estimates of the standard enthalpy of formation for titanium-containing species calculated using the methodology presented in this work, compared to high-level quantum chemistry calculations at the CCSD(T)/CBS level of theory as reported by Wang et al. [96], and selected literature values as listed in Table 1. Excellent agreement was observed between our estimates and those obtained by CCSD(T)/CBS for $TiCl_4$, $TiOCl_2$ and $Ti(OH)_4$. There was excellent agreement with the literature value for TiO_2 but less so with the calculated value by Wang et al. [96]. The proposed method was able to predict highly accurate standard enthalpies of formation comparable to computationally demanding quantum chemistry methods.

4 Conclusions

This paper presents an automated framework that uses overlapping subsets of reference data to systematically derive an informed estimate of the standard enthalpy of formation of a species using error-cancelling balanced reactions (EBRs). A distribution of estimates is derived for each species using multiple EBRs from which an informed estimate can be determined. Overall, this is a more accurate estimate than can be obtained from a

single reaction. A global cross-validation is used to assesses the consistency of the reference data. The EBRs are used to calculate the error contribution from each species in the reference data set, enabling potentially inconsistent reference data to be isolated and excluded.

The functionalities of the framework were demonstrated using test cases from organic and inorganic chemistry, including transition metal complexes. The cases included 920 species containing carbon, hydrogen and oxygen retrieved from the NIST Chemistry WebBook [54], titanium-containing species and oxychloride species from various sources. Electronic structure calculations were performed using DFT at the B97-1/6-311+G(d,p) level of theory for all species considered in this work.

Constrained optimisation, in the form of linear programming, was used to identify individual EBRs. This was combined with a recursive algorithm to systematically identify multiple distinct EBRs. It was found that using multiple EBRs to calculate a distribution of the standard enthalpy of formation resulted in significantly better estimates than from a single reaction. An estimate of the expected statistical uncertainty resulting from the calculation could be derived from the distribution of possible enthalpy values.

A global-cross validation was developed to assess the consistency of the reference data. Different parametrisations were evaluated. For all parameter configurations, the mean absolute error decreased asymptotically as a result of excluding potentially inconsistent reference species. The results of the cross-validation were found to be most sensitive to the rejection threshold parameter. Aspects such as the choice of the level of theory, the reaction class, the statistical benefit from the consideration of multiple EBRs and uncertainties in the reference data require consideration when choosing the rejection threshold. In general, the lower the rejection threshold, the smaller the expected mean absolute error and therefore the higher the accuracy.

Applying the global cross-validation to the test case with carbon, hydrogen and oxygen containing species showed that excluding potentially inconsistent species from the reference data set resulted in a significant reduction in the error in the estimated standard enthalpies of formation. A considerably reduced effect was observed for the smaller oxychloride reference set. In cases where the error reduction is significantly limited, a trade-off has to be made as to whether the statistical benefit from a larger reference set outweighs the exclusion of the potentially inconsistent species. As long as the given reference set is sufficiently large, using isodesmic reactions or an even more restrictive reaction class, should be considered for the global cross-validation. It is then further suggested to choose the same or a more restrictive reaction class for the estimation of the enthalpy of formation of the target species.

The choice of reaction class had an impact on the accuracy of the estimates for all of the reference sets investigated in the hydrocarbon based test case. Generally, the more rigorous the reaction class, the more accurate the expected estimate. It is noted that there is a trade-off between the reaction class, the number of EBRs and the reference set. The statistical benefit of using multiple EBRs can outweigh the choice of a more rigorous reaction class. Overall, it was found that isogyric reactions should be avoided while isodesmic reactions offered a good compromise.

Applying outlier detection methods decreased the statistical uncertainty. The decrease

was significantly larger in cases where the full reference set was used, compared to cases that used reduced sets of reference data that excluded potentially inconsistent species identified by the global cross-validation.

Comparison of the estimates of the enthalpy of formation obtained in this work for hydrocarbon species *versus* those calculated by Saeys et al. [86] showed that the framework is able to predict highly accurate standard enthalpies of formation, comparable to high-level quantum chemistry methods. This was further supported by the comparison of the estimates for the transition metal complexes to values obtained by the *"gold standard"* coupled-cluster calculation method with complete basis set extrapolation, for which excellent agreement for $TiCl_4$, $TiOCl_2$ and $Ti(OH)_4$ was achieved. The estimate for TiO_2 was in excellent agreement with reported NIST-JANAF reference values [14, 54] but in slightly less good agreement with the results from the coupled-cluster method.

Two potentially inconsistent transition metal complexes were found using the global cross-validation: $TiO(OH)_2$ and TiOCl. Revised standard enthalpies of formation for both complexes were proposed. The reference data for all other considered transition metal complexes were found to be consistent.

The application to the titanium-containing species demonstrates that the framework is able to calculate accurate enthalpies of formation for systems where only a few reference values are available. This allows for a systematic and automated investigation of increasingly complex reaction systems. The global cross-validation is useful in identifying inconsistent reference data which can then be targeted to improve the quality of model.

Acknowledgements

This project is partly funded by the National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The authors thank Huntsman Pigments and Additives for financial support.

Supplementary material

Two examples of the identification of multiple error-cancelling balanced reactions (EBRs) are provided. A detailed description of the individual steps, in the form of pseudocode listings, are included for the recursive identification of multiple error-cancelling balanced reactions (EBRs) and for the global cross-validation. The optimised molecular geometries can be provided on request.

5 Appendix

This document presents further information on the developed automated framework that uses overlapping subsets of reference data to systematically validate large sets of thermochemical data for chemical species. The first section provides two examples of the identification of multiple error-cancelling balanced reactions (EBRs). This is followed by detailed algorithmic descriptions in the form of pseudocode listings for the identification of multiple EBRs and the global cross-validation algorithm used to assess the consistency of the reference data.

5.1 Examples of the identification of multiple error-cancelling balanced reactions

Figure 13 and 14 show examples of the identification of multiple error-cancelling balanced reactions (EBRs).

Each node in the graphs defines an attempt to identify an EBR using a given reference set. The full reference set is only used in the root node. The nodes marked with a *cross* represent reference sets for which no EBR can be identified, whereas the shaped nodes represent identified EBRs. Each node is defined by a unique label N_x ($1 \leq x \leq N$), where N is the total number of nodes. For ease of presentation the number of nodes shown is kept to a minimum. Each edge represents a step used to systematically modify the current reference set. The label of the edge defines the species excluded from the previous reference set used by the nodes higher up in the hierarchy.

The table underneath each figure shows the state of the reference set at each node. The full reference set consists of the listed reference species plus the target species. Each species that is included in the reference set at a certain node is indicated by a *tick*, whereas the excluded species are indicated by a *cross*. If a reaction is identified using the reference set for a node, the reaction label is listed in the table. Reference sets which would be analysed multiple times are labelled as duplicates. Only the first instance of each duplicate is analysed.

5.1.1 Example 1

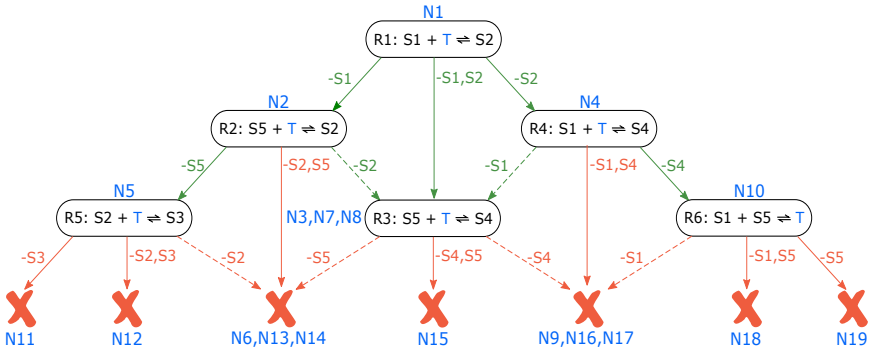
Figure 13 is a hypothetical example. The dashed edges indicate steps that are skipped to avoid analysing duplicate sets of references data. In this example, EBRs were identified for the target species T. The full reference set consisted of the species $\{S1, S2, S3, S4, S5, S6\}$.

A total of six distinct reactions were identified. The root N1 identifies the first reaction R1. Three possible combinations, excluding the empty set, could be derived from the species set $\{S1, S2\}$. Both of these species are required within R1 to determine an estimate of the standard enthalpy of formation for T. Excluding $\{S1\}$ from the full reference set used in N1 results in the identification of R2. R3 (N3) and R4 (N4) are identified by excluding the species combinations $\{S1, S2\}$ and $\{S2\}$.

Using the reference set of N2 and combinations of the reference species $\{S5, S2\}$ that

were involved in the reaction R2 that was found at node N2 enables the identification of a new EBR. This is achieved by further excluding $\{S5\}$ which results in the identification of R5. Another possible combination is $\{S2\}$. This is not treated because the species combination $\{S1, S2\}$ has already been considered going from N1 to N3.

No further EBRs are found from N5. Therefore, a backwards step within the tree is taken.



Duplicate*	Node label	Reference species S1	S2	S3	S4	S5	S6	Identified Reaction	
	N1	✓	✓	✓	✓	✓	✓	R1	✓
	N2	✗	✓	✓	✓	✓	✓	R2	✓
!	N3	✗	✗	✓	✓	✓	✓	R3	✓
	N4	✓	✗	✓	✓	✓	✓	R4	✓
	N5	✗	✓	✓	✓	✗	✓	R5	✓
!	N6	✗	✗	✓	✓	✗	✓	-	✗
!	N7	✗	✗	✓	✓	✓	✓	R3	✓
!	N8	✗	✗	✓	✓	✓	✓	R3	✓
!	N9	✗	✗	✓	✗	✓	✓	-	✗
	N10	✓	✗	✓	✗	✓	✓	R6	✓

Duplicate*	Node label	Reference species S1	S2	S3	S4	S5	S6	Identified Reaction	
!	N11	✗	✗	✓	✓	✗	✓	-	✗
	N12	✗	✗	✗	✓	✗	✓	-	✗
	N13	✗	✓	✗	✓	✗	✓	-	✗
!	N14	✗	✗	✓	✓	✗	✓	-	✗
	N15	✗	✓	✓	✗	✗	✓	-	✗
!	N16	✗	✗	✓	✗	✓	✓	-	✗
!	N17	✗	✗	✓	✗	✓	✓	-	✗
	N18	✗	✗	✓	✗	✗	✓	-	✗
	N19	✓	✗	✓	✗	✗	✓	-	✗

* Duplicates are labelled with !. Only one of the duplicates was treated while the others were skipped to reduce computational demand. A total of 14 steps, out of 64 possible combination of the reference species, were executed.

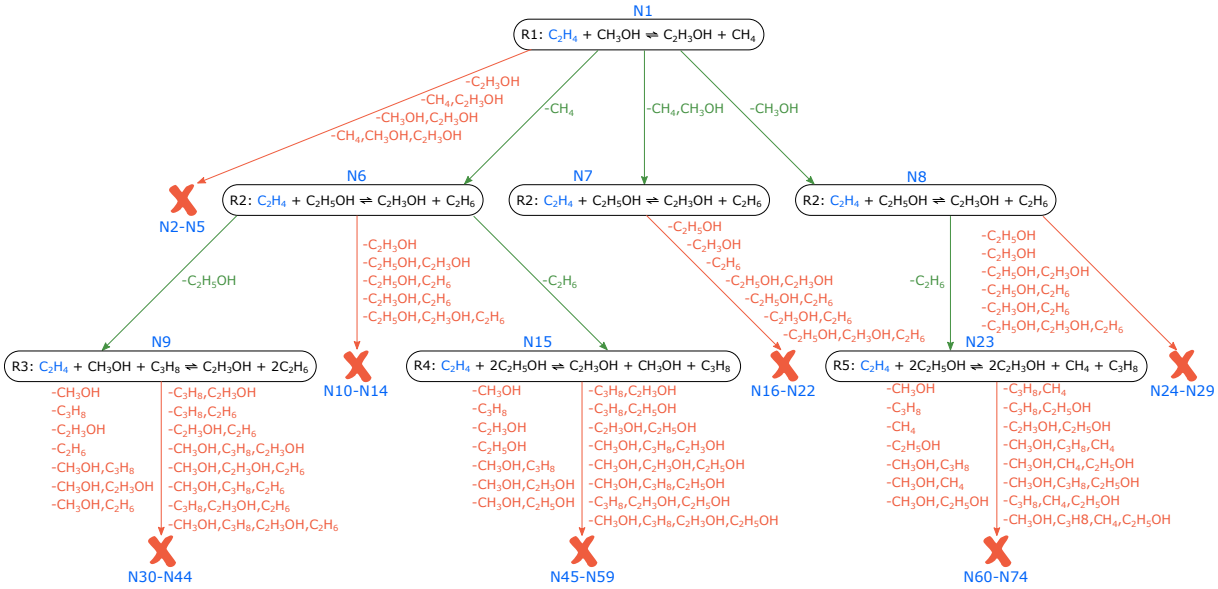
Figure 13: *Hypothetical example to illustrate the recursive identification of multiple error-cancelling balanced reactions. Each node defines an attempt to identify an error-cancelling balanced reaction (EBR) given a modified reference set. Species next to edges are recursively excluded from the reference set, forcing the algorithm to identify distinct EBRs. Species combinations that have already been considered are ignored (dashed edges).*

In this hypothetical example, a total of six reference species would lead to 64 possible combinations. This is a manageable size and all possible combinations could be analysed to identify all possible EBRs. However, the application of the algorithm reduced this to a total of just 14 steps to identify multiple possible EBRs. This is critical when analysing larger real-world systems.

5.1.2 Example 2

Figure 14 shows an example of identifying multiple EBRs for the target species ethylene (C_2H_4). The isodesmic reaction class (RC2) was used. The full reference set consisted of

$\{CH_4, C_2H_6, C_3H_8, CH_3OH, C_2H_3OH, C_2H_5OH\}$.



Duplicate*	Node label	Reference species CH_4	C_2H_6	C_3H_8	CH_3OH	C_2H_3OH	C_2H_5OH	Identified Reaction	
	N1	✓	✓	✓	✓	✓	✓	R1	✓
	N2	✓	✓	✓	✓	✗	✓	-	✗
!	N3	✗	✓	✓	✓	✗	✓	-	✗
!	N4	✓	✓	✓	✗	✗	✓	-	✗
!	N5	✗	✓	✓	✗	✗	✓	-	✗
	N6	✗	✓	✓	✓	✓	✓	R2	✓
	N7	✗	✓	✓	✗	✓	✓	R2	✓
	N8	✓	✓	✓	✗	✓	✓	R2	✓
	N9	✗	✓	✓	✓	✓	✗	R3	✓
!	N10	✗	✓	✓	✓	✗	✓	-	✗
!	N11	✗	✓	✓	✓	✗	✗	-	✗
!	N12	✗	✗	✓	✓	✓	✗	-	✗
!	N13	✗	✗	✓	✓	✗	✓	-	✗
!	N14	✗	✗	✓	✓	✗	✗	-	✗
	N15	✗	✗	✓	✓	✓	✓	R4	✓
!	N16	✗	✓	✓	✗	✓	✗	-	✗
!	N17	✗	✓	✓	✗	✗	✓	-	✗
!	N18	✗	✗	✓	✗	✓	✓	-	✗
!	N19	✗	✓	✓	✗	✗	✗	-	✗
!	N20	✗	✗	✓	✗	✓	✗	-	✗
!	N21	✗	✗	✓	✗	✗	✓	-	✗
!	N22	✗	✗	✓	✗	✗	✗	-	✗
	N23	✓	✗	✓	✗	✓	✓	R5	✓
!	N24	✓	✓	✓	✗	✓	✗	-	✗
!	N25	✓	✓	✓	✗	✗	✓	-	✗
!	N26	✓	✓	✓	✗	✗	✗	-	✗
!	N27	✓	✗	✓	✗	✓	✗	-	✗
!	N28	✓	✗	✓	✗	✗	✓	-	✗
!	N29	✓	✗	✓	✗	✗	✗	-	✗
!	N30	✗	✗	✓	✓	✓	✗	-	✗
	N31	✗	✓	✗	✓	✓	✗	-	✗
!	N32	✗	✓	✓	✗	✓	✗	-	✗
!	N33	✗	✓	✓	✓	✗	✗	-	✗
!	N34	✗	✗	✗	✓	✓	✗	-	✗
!	N35	✗	✗	✓	✗	✓	✗	-	✗
!	N36	✗	✗	✓	✓	✗	✗	-	✗
	N37	✗	✓	✗	✗	✓	✗	-	✗
	N38	✗	✓	✗	✓	✗	✗	-	✗
!	N39	✗	✓	✓	✗	✗	✗	-	✗
!	N40	✗	✗	✗	✗	✓	✗	-	✗
!	N41	✗	✗	✓	✗	✗	✗	-	✗
!	N42	✗	✗	✗	✓	✗	✗	-	✗
	N43	✗	✓	✗	✗	✗	✗	-	✗
!	N44	✗	✗	✗	✗	✗	✗	-	✗
	N45	✗	✗	✗	✓	✓	✓	-	✗
!	N46	✗	✗	✓	✗	✓	✓	-	✗
!	N47	✗	✗	✓	✓	✗	✓	-	✗
!	N48	✗	✗	✓	✓	✓	✗	-	✗
!	N49	✗	✗	✗	✗	✓	✓	-	✗
	N50	✗	✗	✗	✓	✗	✓	-	✗
!	N51	✗	✗	✗	✓	✓	✗	-	✗
!	N52	✗	✗	✓	✗	✗	✓	-	✗
!	N53	✗	✗	✓	✗	✓	✗	-	✗
!	N54	✗	✗	✓	✓	✗	✗	-	✗
!	N55	✗	✗	✗	✗	✗	✓	-	✗
!	N56	✗	✗	✗	✗	✓	✗	-	✗
!	N57	✗	✗	✗	✓	✗	✗	-	✗
!	N58	✗	✗	✓	✗	✗	✗	-	✗
!	N59	✗	✗	✗	✗	✗	✗	-	✗
!	N60	✗	✗	✓	✗	✓	✓	-	✗
	N61	✓	✗	✗	✗	✓	✓	-	✗
!	N62	✓	✗	✓	✗	✗	✓	-	✗
!	N63	✓	✗	✓	✗	✓	✗	-	✗
!	N64	✗	✗	✗	✗	✓	✓	-	✗
!	N65	✗	✗	✓	✗	✗	✓	-	✗
!	N66	✗	✗	✓	✗	✓	✗	-	✗
!	N67	✓	✗	✗	✗	✗	✓	-	✗
!	N68	✓	✗	✗	✗	✓	✗	-	✗
!	N69	✓	✗	✓	✗	✗	✗	-	✗
!	N70	✗	✗	✗	✗	✗	✓	-	✗
!	N71	✗	✗	✓	✗	✗	✗	-	✗
!	N72	✗	✗	✗	✗	✓	✗	-	✗
	N73	✓	✗	✗	✗	✗	✗	-	✗
!	N74	✗	✗	✗	✗	✗	✗	-	✗

* Duplicates are labelled with !. Only one of the duplicates was treated while the others were skipped to reduce computational demand. A total of 43 steps, out of 64 possible combination of the reference species, were executed.

Figure 14: *Example to the recursive identification of multiple error-cancelling balanced reactions for ethylene (C_2H_4). Each node defines an attempt to identify an error-cancelling balanced reaction (EBR) given a modified reference set. Species next to edges are recursively excluded from the reference set, forcing the algorithm to identify distinct EBRs. Species combinations that have already been considered are ignored.*

A total of five distinct reactions were identified in this example (Figure 14). The first reaction R1 is identified at the root of Figure 14. Seven combinations, excluding the empty set, of the three reference species $\{CH_4, CH_3OH, C_2H_3OH\}$ present in R1 could be determined. Only the combinations $\{CH_4\}$, $\{CH_3OH\}$ and $\{C_2H_3OH\}$ led to the identification

of new reactions. In this case, the same reaction R2 was found for each of the three manipulated reference sets defined by the nodes N6, N7 and N8. Seven species combinations were derived from the three additional species in R2.

The additional exclusion of $\{C_2H_5OH\}$ from the reference set of N6 led to the identification of R3. Reactions R4 and R5 were found by the exclusion of $\{C_2H_6\}$ from the reference sets of N6 and N8. No other combinations led to the identification of any EBR and none of the seven combinations led to an EBR from N7. The reactions R3, R4 and R5 were the last reactions found in the recursive search.

The full reference set consisted of six distinct species. In this case, 43 steps were required to identify these reactions, which is less than the 64 possible combinations.

5.2 Pseudocode listings

In the following sections, pseudocode listings are presented for the two main algorithms presented in this paper. Section 5.2.1 presents the pseudocode listings for the recursive identification of multiple error-cancelling balanced reactions (EBRs). The global cross-validation algorithm is presented in section 5.2.2.

Note that for ease of presentation, procedural pseudocode listings are given. They do not represent the most efficient and effective way of implementing the algorithms.

5.2.1 Identification of Multiple Error-Cancelling Balanced Reactions

Algorithm 1: Initialisation of the identification of multiple EBRs.

Input : S^{ref}, all reference data
s, the species of interest for which the enthalpy of formation is unknown
n, number of required EBRs
C, an ordered list of EBRs classes

Output : R, a set of EBRs to estimate the enthalpy of formation for species s

Purpose : Initialise the identification of n EBRs for species s.

```
function identifyEBRs(S^ref, s, n, C)
    // empty reaction class selection
    c^r ← ∅;
    // empty set
    R ← ∅;
    // iteration through EBR class hierarchy
    foreach c ∈ C do
        // identify single EBR of class c
        r ← identifyEBR (S^ref, s, c);
        // check whether we want to keep this reaction
        if validateEBR (r) then
            // select EBR class
            c^r ← c;
            // add valid EBR to R
            R ← {r};
            // exit loop
            break;
        end
    end
    // check whether an EBR class has been selected
    if c^r == ∅ then
        // no class selected, exit loop
        return ∅;
    end
    // recursion control variable
    d ← 0;
    // identification of the remaining EBRs (see Algorithm 2)
    R ← R ∪ identifyEBRs (r, S^ref, s, c^r, ∅, d, n − 1);
    // set of identified EBRs
    return R;
```

Algorithm 2: Recursive identification of multiple EBRs.

Input : r, previously identified EBR
S^{ref}, all reference species
s, the species of interest for which the enthalpy of formation is unknown
c^{r}, selected EBR class
S^{comb}, set of species to be excluded from the reference set
d, depth of the recursion
n, number of required EBRs

Output : R, a set of EBRs to estimate the enthalpy of formation for species s

Purpose : Recursively identify n EBRs for species s.

function `identifyEBRs` (r, S^{ref}, s, c^{r}, S^{comb}, d, n)

```
// empty set
R ← ∅;
// check progress of the identification for the current path
if isCompleted (d, n) then
    // return empty set
    return R;
end
// determine species combinations for EBR r
X^comb(r) ← getSpeciesCombinations (r);
// iterate through each species combination
foreach S ∈ X^comb(r) do
    // current species combination
    S̃^comb ← S^comb ∪ S;
    // check if species combination has been treated already
    if !isTreated (S̃^comb) then
        /* identify single EBR of class c^r, where reaction class c^r is either isogyric,
           isodesmic, etc                                                              */
        r ← identifyEBR (S^ref \ S̃^comb, s, c^r);
        // check whether we want to keep this reaction
        if validateEBR (r) then
            // add valid EBR to R
            R ← R ∪ {r};
            // check if termination criteria has been reached
            if !identificationComplete (d, n − 1) then
                // step deeper into the recursion
                R ← R ∪ identifyEBRs (S^ref, s, c^r, S̃^comb, d + 1, n − 1);
                // update n
                n ← n − |R|;
            end
        end
    end
end
// set of identified EBRs
return R;
```

5.2.2 Global Cross-Validation

Algorithm 3: Initialisation.

Input : S^{ref}, all reference species
n, number of required EBRs
C, an ordered list of EBR classes

Output : S, set of consistent reference species
S^{rej}, set of inconsistent reference species

Purpose : Analyse the reference species and classify species into a set of consistent and inconsistent reference species.

function `analyseReferenceData` (S^{ref}, n, C)

```
// initialisation of the reference set
S ← S^ref;
// iterator variable
i ← 0;
// iterate until convergence criterion is achieved
while !converged (S^ref, S, i) do
    // use previously identified consistent reference set
    S^ref ← S;
    // perform validation and determine error metrics (see Algorithm 4)
    (R, S^rej, ε) ← dataPreProcessing (S^ref, n, C);
    // perform reference data analysis (see Algorithm 5)
    S ←initialDataAnalysis (S^ref, S^rej, R, ε, n, C);
    // increment control variable
    i = i + 1;
end
// set of inconsistent species
S^rej ← S^ref \ S;
// return set of consistent and inconsistent species
return (S, S^rej);
```

Algorithm 4: Data pre-processing.

Input : S^{ref}, all reference data
n, number of required EBRs
C, an ordered list of EBR classes

Output : R, all EBRs identified during the validation process
S^{rej}, a list of recommended rejected species
ε, calculated error metrics

Purpose : Determine the error metrics for reactions and species. Recommend a list of rejected species.

function `dataPreProcessing` (S^{ref}, n, C)

```
// validate each species independently
```
$R \leftarrow$ `performCrossValidation` (S^{ref}, n, C);
```
/* calculate the error metrics for reactions and species as defined by Equations (4)-(12)
   */
```
$\varepsilon \leftarrow$ `calculateErrorMetrics` (R);
```
// determine a list of consistent species
```
$S \leftarrow$ `determineConsistentSpecies` $(S^{\text{ref}}, \varepsilon)$;
```
// determine a list of rejected species
```
$S^{\text{rej}} \leftarrow S^{\text{ref}} \setminus S$;
```
// sort by calculated species error contribution
```
$S^{\text{rej}} \leftarrow$ `descendingSort` $(S^{\text{rej}}, \varepsilon)$;
```
// return calculated data
```
return $(R, S^{\text{rej}}, \varepsilon)$;

Algorithm 5: Initial data analysis.

Input : S^{ref}, all reference data
S^{rej}, an ordered list of rejected species
R, all previously identified EBRs
ε, calculated error metric
n, number of required EBRs
C, an ordered list of EBR classes

Output : S, an updated list of reference species

Purpose : Verify whether better estimates of the enthalpy of formation can be achieved by excluding the rejected species from the reference data.

function `initialDataAnalysis` (S^{ref}, S, S^{rej}, R, ε, n, C)

```
// iterate through the set of rejected species
foreach s ∈ S^rej do
    // identify n EBRs for s (see Algorithm 1)
    R̂^new(s) ← identifyEBRs (S, s, n, C);
    // determine error metric for the n new EBRs
    if !hasLowerError (R(s), R̂^new(s)) then
        /* perform an extended data analysis to identify which species are essential to
           obtain improved EBRs (see Algorithm 6)                                      */
        S ← S ∪ extendedDataAnalysis (S^ref, S^rej, S, s, ε, R, n, C);
    end
end
// updated set of reference species
return S;
```

Algorithm 6: Extended data analysis.

Input : S^{ref}, all reference data
S^{rej}, an ordered list of rejected species
S, a set of species excluded from the list of rejected species
s, species under investigation
R, all previously identified EBRs
ε, calculated error metric
n, number of required EBRs
C, an ordered list of EBR classes

Output : S, an updated list of reference species

Purpose : Check whether improved estimates of the enthalpy of formation can be achieved by modifying the list of rejected species.

function `extendedDataAnalysis` (S^{ref}, S^{rej}, S, s, ε, R, n, C)

```
/* identification of rejected species based on the definitions of Equations (18) and (19)
   */
S̃^rej(s) ← getRejectedSpecies (S, R, s, ε);
// sort species list ascending their species error contribution
S̃^rej(s) ← ascendingSort (S̃^rej(s), ε);
// iteration through all rejected species co-appearing in reactions with species s
foreach s_i ∈ S̃^rej(s) do
    /* prepare modified reference set as defined by Equations (22) and (23)          */
    S ← prepareReferenceSet (S, S̃^rej(s), ε, s_i);
    // identify n EBRs for s (see Algorithm 1)
    R^new(s) ← identifyEBRs (S, s, n, C);
    // determine error metric for the n new EBRs
    if hasLowerError (R(s_i), R^new(s)) then
        // update and return reference set
        return updateReferenceSet (R^new (s), S);
    end
end
// update and return reference set
return updateReferenceSet (R (s), S);
```

References

[1] S. Abramowitz and M. Chase. Thermodynamic Properties of Gas Phase Species of Importance to Ozone Depletion. *Pure Appl. Chem.*, 63(10):1449–1454, 1991.

[2] Active Thermochemical Tables. Version 1.118, 2016. URL `http://atct.anl.gov/`. Retrieved July 03, 2016.

[3] M. D. Allendorf and C. F. Melius. Theoretical Study of Thermochemistry of Molecules in the Silicon-Carbon-Chlorine-Hydrogen System. *J. Phys. Chem.*, 97 (3):720–728, 1993. doi:10.1021/j100105a031.

[4] American Society of Mechanical Engineers. *Test Uncertainty: ASME PTC 19.1-2005*. American National Standard. The American Society of Mechanical Engineers, Three Park Avenue, New York, NY 10016-5990, 2005. ISBN 0791830101.

[5] R. Atkinson, D. L. Baulch, R. A. Cox, R. F. Hampson, J. A. Kerr, M. J. Rossi, and J. Troe. Evaluated Kinetic and Photochemical Data for Atmospheric Chemistry, Organic Species: Supplement VII. *J. Phys. Chem. Ref. Data*, 28(2):191–393, 1999. doi:10.1063/1.556048.

[6] S. W. Benson and J. H. Buss. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.*, 29(3):546–572, 1958. doi:10.1063/1.1744539.

[7] W. J. Bloss, S. L. Nickolaisen, R. J. Salawitch, R. R. Friedl, and S. P. Sander. Kinetics of the ClO Self-Reaction and 210 nm Absorption Cross Section of the ClO Dimer. *J. Phys. Chem. A*, 105(50):11226–11239, 2001. doi:10.1021/jp012429y.

[8] A. D. Boese, J. M. L. Martin, and N. C. Handy. The Role of the Basis Set: Assessing Density Functional Theory. *J. Chem. Phys.*, 119(6):3005–3014, 2003. doi:10.1063/1.1589004.

[9] P. Buerger, D. Nurkowski, J. Akroyd, S. Mosbach, and M. Kraft. First-Principles Thermochemistry for the Thermal Decomposition of Titanium Tetraisopropoxide. *J. Phys. Chem. A*, 119(30):8376–8387, 2015. doi:10.1021/acs.jpca.5b01721.

[10] P. Buerger, D. Nurkowski, J. Akroyd, and M. Kraft. A Kinetic Mechanism for the Thermal Decomposition of Titanium Tetraisopropoxide. *Proc. Combust. Inst., In Press*, 2016. doi:10.1016/j.proci.2016.08.062.

[11] P. Buerger, J. Akroyd, J. W. Martin, and M. Kraft. A Big Data Framework to Validate Thermodynamic Data for Chemical Species. *Combust. Flame*, 176:584–591, 2017. doi:10.1016/j.combustflame.2016.11.006.

[12] Cbc. Coin-Or Branch and Cut, 2016. URL https://projects.coin-or.org/Cbc/.

[13] CCDBDB. NIST Computational Chemistry Comparison and Benchmark DataBase, Standard Reference Database Number 101, 2015. URL http://cccbdb.nist.gov/. Retrieved May 13, 2016.

[14] M. W. J. Chase. *NIST-JANAF Thermochemical Tables, 4th Edition*. American Institute of Physics, New York, 1998.

[15] J. Cioslowski, M. Schimeczek, G. Liu, and V. Stoyanov. A Set of Standard Enthalpies of Formation for Benchmarking, Calibration, and Parametrization of Electronic Structure Methods. *J. Chem. Phys.*, 113(21):9377–9389, 2000.

[16] N. Cohen. Revised Group Additivity Values for Enthalpies of Formation (at 298 K) of Carbon-Hydrogen and Carbon-Hydrogen-Oxygen Compounds. *J. Phys. Chem. Ref. Data*, 25(6):1411–1481, 1996. doi:http://dx.doi.org/10.1063/1.555988.

[17] CPLEX Optimizer, 2016. URL www.ibm.com/software/commerce/optimization/cplex-optimizer/.

[18] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.*, 106(3):1063–1079, 1997. doi:http://dx.doi.org/10.1063/1.473182.

[19] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople. Assessment of Gaussian-3 and Density Functional Theories for a Larger Experimental Test Set. *J. Chem. Phys.*, 112(17):7374–7383, 2000. doi:http://dx.doi.org/10.1063/1.481336.

[20] L. A. Curtiss, P. C. Redfern, and K. Raghavachari. Assessment of Gaussian-3 and Density-Functional Theories on the G3/05 Test Set of Experimental Energies. *J. Chem. Phys.*, 123(12), 2005. doi:http://dx.doi.org/10.1063/1.2039080.

[21] G. B. Dantzig. Linear Programming. *Oper. Res.*, 50(1):42–47, 2002. doi:10.1287/opre.50.1.42.17798.

[22] W. B. DeMore, S. P. Sander, D. Golden, R. F. Hampson, M. J. Kurylo, C. Howard, A. Ravishankara, C. Kolb, and M. Molina. Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling. Evaluation No. 12. *J. Org. Chem.*, 1997.

[23] N. J. DeYonker, K. A. Peterson, G. Steyl, A. K. Wilson, and T. R. Cundari. Quantitative Computational Thermochemistry of Transition Metal Species. *J. Phys. Chem. A*, 111(44):11269–11277, 2007. doi:10.1021/jp0715023.

[24] N. J. DeYonker, T. G. Williams, A. E. Imel, T. R. Cundari, and A. K. Wilson. Accurate Thermochemistry for Transition Metal Complexes from First-Principles Calculations. *J. Chem. Phys.*, 131(2):024106, 2009. doi:10.1063/1.3160667.

[25] D. E. Edwards, D. Y. Zubarev, A. Packard, W. A. Lester, and M. Frenklach. Interval Prediction of Molecular Properties in Parametrized Quantum Chemistry. *Phys. Rev. Lett.*, 112:253003, Jun 2014. doi:10.1103/PhysRevLett.112.253003.

[26] R. Feeley, P. Seiler, A. Packard, and M. Frenklach. Consistency of a Reaction Dataset. *J. Phys. Chem. A*, 108(44):9573–9583, 2004. doi:10.1021/jp047524w.

[27] R. Feeley, M. Frenklach, M. Onsum, T. Russi, A. Arkin, and A. Packard. Model Discrimination Using Data Collaboration. *J. Phys. Chem. A*, 110(21):6803–6813, 2006. doi:10.1021/jp056309s.

[28] M. Frenklach, A. Packard, P. Seiler, and R. Feeley. Collaborative Data Processing in Developing Predictive Models of Complex Reaction Systems. *Int. J. Chem. Kinet.*, 36(1):57–66, 1 2004. doi:10.1002/kin.10172.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin,

K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision D.01, 2009.

[30] C. W. Gao, J. W. Allen, W. H. Green, and R. H. West. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.*, 203:212–225, 2016. doi:10.1016/j.cpc.2016.02.013.

[31] Y. Ge, D. DePrekel, K.-T. Lam, K. Ngo, and P. Vo. Assessing Density Functionals for the Prediction of Thermochemistry of $Ti-O-Cl$ Species. *J. Theor. Comput. Chem.*, 14(08):1550055, 2015. doi:10.1142/S0219633615500558.

[32] J. L. Gearhart, K. L. Adair, R. J. Detry, J. D. Durfee, K. A. Jones, and N. Martin. Comparison of Open-Source Linear Programming Solvers. Technical report, Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2013.

[33] P. George, M. Trachtman, C. W. Bock, and A. M. Brett. An Alternative Approach to the Problem of Assessing Stabilization Energies in Cyclic Conjugated Hydrocarbons. *Theor. Chim. Acta*, 38(2):121–129, 1975. doi:10.1007/BF00581469.

[34] P. George, M. Trachtman, C. W. Bock, and A. M. Brett. Homodesmotic Reactions for the Assessment of Stabilization Energies in Benzenoid and Other Conjugated Cyclic Hydrocarbons. *J. Chem. Soc., Perkin Trans. 2*, pages 1222–1227, 1976. doi:10.1039/P29760001222.

[35] P. George, M. Trachtman, A. M. Brett, and C. W. Bock. Comparison of Various Isodesmic and Homodesmotic Reaction Heats with Values Derived from Published *Ab Initio* Molecular Orbital Calculations. *J. Chem. Soc., Perkin Trans. 2*, pages 1036–1047, 1977. doi:10.1039/P29770001036.

[36] P. E. Gill, W. Murray, M. A. Saunders, J. A. Tomlin, and M. H. Wright. George B. Dantzig and Systems Optimization. *Discrete Optimization*, 5(2):151–158, 2008. doi:10.1016/j.disopt.2007.01.002.

[37] GLPK. GNU Linear Programming Kit, Version 4.58, 2016. URL `http://www.gnu.org/software/glpk/glpk.html`.

[38] W. H. Green and R. H. West. Reaction Mechanism Generator. Open-Source Software., 2016. URL `http://reactionmechanismgenerator.github.io`. Retrieved July 07, 2016.

[39] Gurobi Optimizer, 2016. URL http://www.gurobi.com/products/gurobi-optimizer/.

[40] M. W. D. Hanson-Heine, M. W. George, and N. A. Besley. Investigating the Calculation of Anharmonic Vibrational Frequencies Using Force Fields Derived from Density Functional Theory. *J. Phys. Chem. A*, 116(17):4417–4425, 2012. doi:10.1021/jp301670f.

[41] W. J. Hehre, R. Ditchfield, L. Radom, and J. A. Pople. Molecular Orbital Theory of the Electronic Structure of Organic Compounds. V. Molecular Theory of Bond Separation. *J. Am. Chem. Soc.*, 92(16):4796–4801, 1970. doi:10.1021/ja00719a006.

[42] D. L. Hildenbrand. Low-Lying Electronic Sates and Revised Thermochemistry of TiCl, $TiCl_2$, and $TiCl_3$. *J. Phys. Chem. A*, 113(8):1472–1474, 2009. doi:10.1021/jp807913c.

[43] P. Ho and C. F. Melius. Theoretical Study of the Thermochemistry of Fluorosilanes (SiF_n and SiH_nF_m) Compounds and Hexafluorodisilane. *J. Phys. Chem.*, 94(12): 5120–5127, 1990. doi:10.1021/j100375a066.

[44] B. Iglewicz and D. Hoaglin. *How to Detect and Handle Outliers*. ASQC Basic References in Quality Control. ASQC Quality Press, 1993. ISBN 9780873892476.

[45] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

[46] W. Jiang, N. J. DeYonker, J. J. Determan, and A. K. Wilson. Toward Accurate Theoretical Thermochemistry of First Row Transition Metal Complexes. *J. Phys. Chem. A*, 116(2):870–885, 2012. doi:10.1021/jp205710e.

[47] W. Jiang, M. L. Laury, M. Powell, and A. K. Wilson. Comparative Study of Single and Double Hybrid Density Functionals for the Prediction of 3d Transition Metal Thermochemistry. *J. Chem. Theory. Comput.*, 8(11):4102–4111, 2012. doi:10.1021/ct300455e.

[48] K. Joback and R. Reid. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun.*, 57(1-6):233–243, 1987. doi:10.1080/00986448708960487.

[49] A. Karton, P. R. Schreiner, and J. M. L. Martin. Heats of Formation of Platonic Hydrocarbon Cages by Means of High-Level Thermochemical Procedures. *J. Comput. Chem.*, 37(1):49–58, 2016. doi:10.1002/jcc.23963.

[50] M. L. Laury and A. K. Wilson. Performance of Density Functional Theory for Second Row (4d) Transition Metal Thermochemistry. *J. Chem. Theory. Comput.*, 9(9):3939–3946, 2013. doi:10.1021/ct400379z.

[51] T. J. Lee, C. M. Rohlfing, and J. E. Rice. An Extensive *Ab Initio* Study of the Structures, Vibrational Spectra, Quadratic Force Fields, and Relative Energetics of Three Isomers of Cl_2O_2. *J. Chem. Phys.*, 97(9):6593–6605, 1992. doi:10.1063/1.463663.

[52] W.-K. Li and C.-Y. Ng. Gaussian-2 *Ab Initio* Study of Isomeric Cl_2O_2 and $Cl_2O_2^+$ and Their Dissociation Reactions. *J. Phys. Chem. A*, 101(2):113–115, 1997. doi:10.1021/jp962253d.

[53] W.-K. Li, K.-C. Lau, C. Y. Ng, H. Baumgärtel, and K. M. Weitzel. Gaussian-2 and Gaussian-3 Study of the Energetics and Structures of Cl_2O_n and Cl_2O_n+, n = $1-7$. *J. Phys. Chem. A*, 104(14):3197–3203, 2000. doi:10.1021/jp993398y.

[54] P. J. Linstrom and W. G. Mallard, editors. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology (NIST), Gaithersburg MD, 20899, 2005. Retrieved May 13, 2016.

[55] lp_solve. Version 5.5.2.0, 2016. URL `http://lpsolve.sourceforge.net/`.

[56] T. Lu and C. K. Law. A Directed Relation Graph Method for Mechanism Reduction. *Proc. Combust. Inst.*, 30(1):1333–1341, 2005. doi:http://dx.doi.org/10.1016/j.proci.2004.08.145.

[57] T. Lu and C. K. Law. Linear Time Reduction of Large Kinetic Mechanisms with Directed Relation Graph: n-Heptane and iso-Octane. *Combust. Flame*, 144(1–2): 24–36, 2006. doi:http://dx.doi.org/10.1016/j.combustflame.2005.02.015.

[58] D. McQuarrie and J. Simon. *Molecular Thermodynamics*. University Science Books, Sausalito, CA, United States, 1999. ISBN 9781891389054.

[59] C. F. Melius and M. D. Allendorf. Bond Additivity Corrections for Quantum Chemistry Methods. *J. Phys. Chem. A*, 104(11):2168–2177, 2000. doi:10.1021/jp9914370.

[60] C. F. Melius and P. Ho. Theoretical Study of the Thermochemistry of Molecules in the Silicon-Nitrogen-Hydrogen-Fluorine System. *J. Phys. Chem.*, 95(3):1410–1419, 1991. doi:10.1021/j100156a070.

[61] J. P. Merrick, D. Moran, and L. Radom. An Evaluation of Harmonic Vibrational Frequency Scale Factors. *J. Phys. Chem. A*, 111(45):11683–11700, 2007. doi:10.1021/jp073974n.

[62] MolHub, 2016. URL http://como.cheng.cam.ac.uk/molhub/compchem/. Retrieved July 07, 2016.

[63] Nano, 2016. URL http://nano.nature.com/. Retrieved July 07, 2016.

[64] Q. N. Nguyen, D. L. Myers, N. S. Jacobson, and E. J. Opila. Experimental and Theoretical Study of Thermodynamics of the Reaction of Titania and Water at High Temperatures. *NASA Technical Memorandum*, 2014. NASA/TM–2014-218372.

[65] S. L. Nickolaisen, R. R. Friedl, and S. P. Sander. Kinetics and Mechanism of the Chlorine Oxide ClO + ClO Reaction: Pressure and Temperature Dependences of the Bimolecular and Termolecular Channels and Thermal Decomposition of Chlorine Peroxide. *J. Phys. Chem.*, 98(1):155–169, 1994. doi:10.1021/j100052a027.

[66] K. E. Niemeyer and C.-J. Sung. On the Importance of Graph Search Algorithms for DRGEP-Based Mechanism Reduction Methods. *Combust. Flame*, 158(8):1439–1443, 2011. doi:http://dx.doi.org/10.1016/j.combustflame.2010.12.010.

[67] D. Nurkowski, P. Buerger, J. Akroyd, and M. Kraft. A Detailed Kinetic Study of the Thermal Decomposition of Tetraethoxysilane. *Proc. Combust. Inst.*, 35(2):2291–2298, 2015. doi:10.1016/j.proci.2014.06.093.

[68] D. Nurkowski, S. J. Klippenstein, Y. Georgievskii, M. Verdicchio, A. W. Jasper, J. Akroyd, S. Mosbach, and M. Kraft. *Ab Initio* Variational Transition State Theory and Master Equation Study of the Reaction $(OH)_3SiOCH_2 + CH_3 \longleftrightarrow (OH)_3SiOC_2H_5$. *Z. Phys. Chem.*, 229(5):691–708, 2015. doi:10.1515/zpch-2014-0640.

[69] D. Nurkowski, P. Buerger, J. Akroyd, S. Mosbach, and M. Kraft. Skeletal Chemical Mechanism of High-Temperature TEOS Oxidation in Hydrogen-Oxygen Environment. *Combust. Flame*, 166:243–254, 2016. doi:10.1016/j.combustflame.2016.01.025.

[70] P. Pepiot-Desjardins and H. Pitsch. An Efficient Error-Propagation-Based Reduction Method for Large Chemical Kinetic Mechanisms. *Combust. Flame*, 154(1–2): 67–81, 2008. doi:http://dx.doi.org/10.1016/j.combustflame.2007.10.020.

[71] W. Phadungsukanan, S. Shekar, R. Shirley, M. Sander, R. H. West, and M. Kraft. First-Principles Thermochemistry for Silicon Species in the Decomposition of Tetraethoxysilane. *J. Phys. Chem. A*, 113(31):9041–9049, 2009. doi:10.1021/jp905494s.

[72] W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust. The Semantics of Chemical Markup Language (CML) for Computational Chemistry: CompChem. *J. Cheminform.*, 4(15):1–16, 2012. doi:10.1186/1758-2946-4-15.

[73] J. A. Pople, L. Radom, and W. J. Hehre. Molecular Orbital Theory of the Electronic Structure of Organic Compounds. VII. Systematic Study of Energies, Conformations, and Bond Interactions. *J. Am. Chem. Soc.*, 93(2):289–300, 1971. doi:10.1021/ja00731a001.

[74] J. A. Pople, M. J. Frisch, B. T. Luke, and J. S. Binkley. A Møller-Plesset Study of the Energies of AHn Molecules (A = Li to F). *Int. J. Quantum Chem.*, 24(S17): 307–320, 1983. doi:10.1002/qua.560240835.

[75] PrIMe: Process Informatics Model, 2016. URL `http://www.primekinetics.org/prime_data_warehouse/`. Retrieved June 29, 2016.

[76] R. O. Ramabhadran and K. Raghavachari. Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy. *J. Chem. Theory. Comput.*, 7(7):2094–2103, 2011. doi:10.1021/ct200279q.

[77] S. Rayne and K. Forest. Estimated Gas-Phase Standard State Enthalpies of Formation for Organic Compounds Using the Gaussian-4 (G4) and W1BD Theoretical Methods. *J. Chem. Eng. Data*, 55(11):5359–5364, 2010. doi:10.1021/je100768s.

[78] P. C. Redfern, P. Zapol, L. A. Curtiss, and K. Raghavachari. Assessment of Gaussian-3 and Density Functional Theories for Enthalpies of Formation of C_1-C_{16} Alkanes. *J. Phys. Chem. A*, 104(24):5850–5854, 2000. doi:10.1021/jp994429s.

[79] P. Refaeilzadeh, L. Tang, and H. Liu. *Encyclopedia of Database Systems*, chapter Cross-Validation, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9.

[80] ReSpecTh, 2016. URL http://respecth.hu/. Retrieved July 07, 2016.

[81] K. E. Riley and K. M. Merz Jr. Assessment of Density Functional Theory Methods for the Computation of Heats of Formation and Ionization Potentials of Systems Containing Third Row Transition Metals. *J. Phys. Chem. A*, 111(27):6044–6053, 2007. doi:10.1021/jp0705931.

[82] D. W. Rogers, A. A. Zavitsas, and N. Matsunaga. Determination of Enthalpies ('Heats') of Formation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3(1):21–36, 2013. doi:10.1002/wcms.1109.

[83] E. Rühl, U. Rockland, H. Baumgärtel, O. Lösking, M. Binnewies, and H. Willner. Photoionization Mass Spectrometry of Chlorine Oxides. *Int. J. Mass Spectrom.*, 185–187:545–558, 1999. doi:10.1016/S1387-3806(98)14137-4.

[84] T. Russi, A. Packard, R. Feeley, and M. Frenklach. Sensitivity Analysis of Uncertainty in Model Prediction. *J. Phys. Chem. A*, 112(12):2579–2588, 2008. doi:10.1021/jp076861c.

[85] T. Russi, A. Packard, and M. Frenklach. Uncertainty Quantification: Making Predictions of Complex Reaction Systems Reliable. *Chem. Phys. Lett.*, 499(1–3):1–8, 2010. doi:10.1016/j.cplett.2010.09.009.

[86] M. Saeys, M.-F. Reyniers, G. B. Marin, V. van Speybroeck, and M. Waroquier. *Ab Initio* Calculations for Hydrocarbons: Enthalpy of Formation, Transition State Geometry, and Activation Energy for Radical Reactions. *J. Phys. Chem. A*, 107 (43):9147–9159, 2003. doi:10.1021/jp021706d.

[87] R. Shirley, Y. Liu, T. S. Totton, R. H. West, and M. Kraft. First-Principles Thermochemistry for the Combustion of a $TiCl_4$ and $AlCl_3$ Mixture. *J. Phys. Chem. A*, 113(49):13790–13796, 2009. doi:10.1021/jp905244w.

[88] J. Sicre and C. Cobos. Thermochemistry of the Higher Chlorine Oxides ClO_x (x=3, 4) and Cl_2O_x (x=3-7). *J. Mol. Struc.-THEOCHEM*, 620(2–3):215–226, 2003. doi:10.1016/S0166-1280(02)00602-4.

[89] K. P. Somers and J. M. Simmie. Benchmarking Compound Methods (CBS-QB3, CBS-APNO, G3, G4, W1BD) Against the Active Thermochemical Tables: Formation Enthalpies of Radicals. *J. Phys. Chem. A*, 119(33):8922–8933, 2015. doi:10.1021/acs.jpca.5b05448.

[90] A. Tajti, P. G. Szalay, A. G. Császár, M. Kállay, J. Gauss, E. F. Valeev, B. A. Flowers, J. Vázquez, and J. F. Stanton. HEAT: High Accuracy Extrapolated *Ab Initio* Thermochemistry. *J. Chem. Phys.*, 121(23):11599–11613, 2004. doi:10.1063/1.1811608.

[91] S. M. Tekarli, M. L. Drummond, T. G. Williams, T. R. Cundari, and A. K. Wilson. Performance of Density Functional Theory for 3d Transition Metal-Containing Complexes: Utilization of the Correlation Consistent Basis Sets. *J. Phys. Chem. A*, 113(30):8607–8614, 2009. doi:10.1021/jp811503v.

[92] W. R. Thompson. On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation. *Ann. Math. Stat.*, 6 (4):214–219, 1935.

[93] T. S. Totton, R. Shirley, and M. Kraft. First-Principles Thermochemistry for the Combustion of $TiCl_4$ in a Methane Flame. *Proc. Combust. Inst.*, 33(1):493–500, 2011. doi:10.1016/j.proci.2010.05.011.

[94] V. van Speybroeck, R. Gani, and R. Meier. The Calculation of Thermodynamic Properties of Molecules. *Chem. Rev.*, 39:1764–1779, 2010.

[95] R. J. Vanderbei. *Linear Programming: Foundations and Extensions*. Department of Operations and Research and Financial Engineering, Princeton University, 2001. ISBN 978-1-4614-7630-6.

[96] T.-H. Wang, A. M. Navarrete-López, S. Li, D. A. Dixon, and J. L. Gole. Hydrolysis of $TiCl_4$: Initial Steps in the Production of TiO_2. *J. Phys. Chem. A*, 114(28):7561–7570, 2010. doi:10.1021/jp102020h.

[97] M. N. Weaver, J. Kenneth M. Merz, D. Ma, H. J. Kim, and L. Gagliardi. Calculation of Heats of Formation for Z_n Complexes: Comparison of Density Functional Theory, Second Order Perturbation Theory, Coupled-Cluster and Complete Active Space Methods. *J. Chem. Theory. Comput.*, 9(12):5277–5285, 2013. doi:10.1021/ct400856g.

[98] R. Weber and A. K. Wilson. Do Composite Methods Achieve Their Target Accuracy? *Comp. Theor. Chem.*, 1072:58–62, 2015. doi:10.1016/j.comptc.2015.08.015.

[99] R. H. West, G. J. O. Beran, W. H. Green, and M. Kraft. First-Principles Thermochemistry for the Production of TiO_2 from $TiCl_4$. *J. Phys. Chem. A*, 111(18): 3560–3565, 2007. doi:10.1021/jp0661950.

[100] R. H. West, M. S. Celnik, O. R. Inderwildi, M. Kraft, G. J. O. Beran, and W. H. Green. Toward a Comprehensive Model of the Synthesis of TiO_2 Particles from $TiCl_4$. *Ind. Eng. Chem. Res.*, 46(19):6147–6156, 2007. doi:10.1021/ie0706414.

[101] S. E. Wheeler, K. N. Houk, P. v. R. Schleyer, and W. D. Allen. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *J. Am. Chem. Soc.*, 131(7):2547–2560, 2009. doi:10.1021/ja805843n.

[102] M. D. Wodrich, C. Corminboeuf, and S. E. Wheeler. Accurate Thermochemistry of Hydrocarbon Radicals Via an Extended Generalized Bond Separation Reaction Scheme. *J. Phys. Chem. A*, 116(13):3436–3447, 2012. doi:10.1021/jp212209q.

[103] J. Yang and M. P. Waller. JACOB: A Dynamic Database for Computational Chemistry Benchmarking. *J. Chem. Inf. Model.*, 52(12):3255–3262, 2012. doi:10.1021/ci300374g.

[104] A. A. Zavitsas, N. Matsunaga, and D. W. Rogers. Enthalpies of Formation of Hydrocarbons by Hydrogen Atom Counting. Theoretical Implications. *J. Phys. Chem. A*, 112(25):5734–5741, 2008. doi:10.1021/jp801152t.

[105] Y. Zhou, J. Wu, and X. Xu. How Well Can B3LYP Heats of Formation be Improved by Dispersion Correction Models? *Theor. Chem. Acc.*, 135(2):1–15, 2016. doi:10.1007/s00214-015-1801-9.

Citation index