A Big Data Framework to Validate Thermodynamic Data for Chemical Species

Philipp Buerger[a], Jethro Akroyd[a], Jacob W. Martin[a], Markus Kraft[a,b,*]

[a] *Department of Chemical Engineering and Biotechnology, University of Cambridge, New Museums Site, Pembroke Street, Cambridge CB2 3RA, United Kingdom*
[b] *School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore, 637459*

Abstract

The advent of large sets of chemical and thermodynamic data has enabled the rapid investigation of increasingly complex systems. The challenge, however, is how to validate such large databases. We propose an automated framework to solve this problem by identifying which data are consistent and recommending what future experiments or calculations are required. The framework is applied to validate data for the standard enthalpy of formation for 920 gas-phase hydrocarbon species retrieved from the NIST Chemistry WebBook. The concept of error-cancelling balanced reactions is used to calculate a distribution of possible values for the standard enthalpy of formation of each species. The method automates the identification and exclusion of inconsistent data. We find that this enables the rapid convergence of the calculations towards chemical accuracy. The method can exploit knowledge of the structural similarities between species and the consistency of the data to identify which species introduce the most error and recommend what future experiments and calculations should be considered.

Keywords: enthalpy of formation, heat of formation, error-cancelling balanced reactions, big data, validation

*Corresponding author
Email address: mk306@cam.ac.uk (Markus Kraft)

1. Introduction

The availability of large sets of chemical and thermodynamic data has enabled the investigation of increasingly complex reaction systems. For example, the Reaction Mechanism Generator [1, 2] automates the generation of chemical mechanisms for gas-phase systems containing carbon, hydrogen, oxygen, sulfur, and nitrogen. Other approaches have been used to develop models for the gas-phase chemistry of common precursors for the formation of various nanoparticles [3, 4, 5, 6].

The data sets associated with such tools are widely available via the internet. Some are based on data collated from the literature, for example, Nano [7] and the NIST WebBook [8]. Others are designed to enable benchmarking and comparison of computational methods, for example, the NIST Computational Chemistry Comparison and Benchmark Database [9]. Others exist as repositories for computational and experimental data, such as the Active Thermochemical Tables [10], PrIME database [11], ReSpecTh information system [12], and MolHub [13, 14].

The availability of so much data presents opportunities and challenges. For example, how do we check which data are consistent? Previous validation has typically been performed at a single-point level, where increasingly accurate methods are applied one-species-at-a-time. This is expensive and becomes intractable for large systems. Methods that exploit the data at a global level, on the other hand, leverage existing knowledge to provide cheaper and potentially more accurate estimates.

This paper considers a global method for the calculation and validation of the standard enthalpies of formation for a large set of gas-phase species. The standard enthalpy of formation is chosen to illustrate the method because it is used in many thermodynamic calculations, and is a key parameter in the development of kinetic mechanisms and understanding the chemistry of novel systems. Any problems in the data could lead to significant errors in the resulting mechanisms.

The estimation of enthalpies of formation using high level electronic structure calculations is computationally demanding. To compound the problem, care must be taken to choose the right level of theory [15, 16, 17], the errors in the method scale with the size of the species [18, 19], and various correction terms are needed to achieve accurate estimates [20]. The calculations become intractable for large molecules [17, 21] and such methods are not suitable for large scale analysis.

Fortunately the errors in electronic structure calculations are systematic. Methods such as bond additivity correction (BAC) [22, 23, 24, 25] and atom additivity correction (AAC) [26] seek to exploit this to cancel the systematic component of the error. Both rely on pre-determined parameters associated with the level of theory used for the electronic structure calculation.

Similarly, error-cancelling balanced reactions (EBRs) seek to exploit structural and electronic similarities between species to cancel the systematic error introduced when using electronic structure calculations to estimate the enthalpy of a species. The method does not introduce any additional parameters and has been applied to a wide variety of systems [see for example 27, 28, 29, 30, 31, 5].

The use of EBRs requires the calculated total electronic energies for all species in the reaction and the enthalpies of formation to be known (either experimentally or otherwise) for all except one species in the reaction, for which the unknown enthalpy of formation is to be estimated. The method requires the identification of suitable balanced reaction(s), given the set of species with known enthalpies of formation. A number of different types of EBRs have been proposed [32, 33, 34, 35, 36, 37, 38, 30, 39] and shown to be able to estimate the enthalpy of formation on the back of affordable electronic structure calculations.

The purpose of this paper is to present a framework for the systematic calculation and validation of enthalpies of formation using EBRs. The framework is applied to a set of 920 hydrocarbons, including species with oxygen. Reference data for the enthalpy of formation were taken from the NIST Chemistry WebBook [8]. The framework was able to assess the consistency of the reference data. The automatic exclusion of problematic data was shown to reduce signifi-

cantly the statistical uncertainty in the estimates of the enthalpies of formation, in many cases a difference smaller than 1.0 kcal mol^{-1} was observed compared to the reference data. We demonstrate how the information generated within the framework may be used to suggest what future experiments or computations might be considered to improve the quality of the data, and which methods may be most suitable.

2. Methodology

The framework is outlined in Figure 1. A species set is provided to an automated validation module. The module uses a cheap method, in this case error-cancelling balanced reactions (EBRs), to estimate a thermodynamic property of interest. It differs from traditional validation methods because it uses multiple overlapping subsets of the data, in this case multiple EBRs, to calculate a distribution of values for each species and performs a global cross-validation of the data.

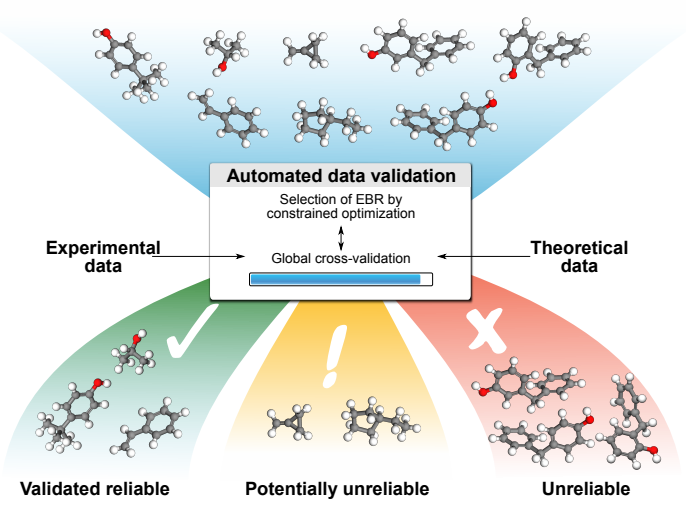

Figure 1: An automated data validation procedure to assess the consistency of experimental, computational and theoretical species data.

In order to demonstrate the framework, we validate data for the standard enthalpy of formation $\Delta_{\mathrm{f}}H^{\circ}_{298.15\ K}$ of 920 gas-phase hydrocarbons consisting

of carbon, oxygen, and hydrogen, taken from the NIST Chemistry WebBook [8].Open- and closed-shell species were considered. The largest species is composed of 32 carbon and 66 hydrogen atoms. We used the 3D geometries provided by NIST as an initial guess and recalculated the ground state geometry, *i.e.* the lowest energy conformer of the species, scaled frequencies [40] and total electronic energy of each species using density functional theory (DFT) at the B97-1/6-311+G(d,p) level of theory [41] using Gaussian09 [42]. A simple rigid-rotor harmonic-oscillator approximation was assumed [43] and defines the lower bound of accuracy for calculating the total energy. This gives an idea about the predictive power of the method because it presents the worst case scenario with respect to the accuracy of the total energy calculation.

The EBRs were identified by constrained optimisation, implemented using the GNU Linear Programming Kit [44] software library [45, 46, 47], which has been shown to perform well compared to other open-source solvers [48], to find combinations of reactant and product species that conserve structural and electronic similarities across each reaction, for example, the number of each type of bond. The set of species available to the constrained optimization was recursively adjusted to exclude species that had been used in other reactions to ensure the identification of a set of unique EBRs. .

Each EBR is used to calculate the standard enthalpy of formation of a species based on the application of Hess's Law to the reaction. This is analogous to methods for calculating the enthalpy of formation from experimental measurements of reaction enthalpy. The principle of the method is that systematic components of the error in the electronic structure calculations cancel out across the EBR. The extent to which the errors cancel depends on what properties are being conserved by the choice of EBR [see for example 38]. For ease of presentation, only isodesmic reactions [32, 33] are considered in this paper. These conserve the number of each type of bond on each side of the reaction. However, the method is general in the sense that it can be used with any type of EBR, for example, isogyric, isodesmic, hypohomodesmotic, homodesmotic, hyperhomodesmotic, and others [see for example 34, 35, 36, 37, 38, 39, 30].

The global cross-validation adapts techniques developed for data mining and statistical analysis [49, 50] to validate the data set. It uses the difference between the calculated data (in this case the standard enthalpies of formation) from multiple EBRs and the corresponding quantity in the reference data to isolate the error contribution from each species in the reference data set. The species with the largest error contributions are iteratively excluded from the calculation, and the cross-validation is repeated to analyse the impact of the excluded species. The algorithm converges rapidly and the exclusion of inconsistent data has a strong beneficial impact on the accuracy of the calculated data.

The framework is able to quantify the consistency of data and classify them according to whether or not they are consistent, or whether there is some ambiguity that merits closer examination. The framework is not limited to a single database or data set, and could, in principle, be applied to validate data spread over multiple data sets and multiple locations. The information generated during the validation may be used to suggest what future experiments or computations might be considered to improve the quality of the data, and which methods may be most suitable.

3. Results

3.1. Selection of error-cancelling balanced reactions by constrained optimization

Figure 2 shows example results for the standard enthalpy of formation of butanoic acid calculated using EBRs. The use of constrained optimization to automate the identification of multiple EBRs enables the construction of a histogram. It is clear that methods that rely on a single EBR can be problematic.

An example reaction from the set of 143 identified reactions for butanoic acid ($C_4H_8O_2$) is,

$$2\,C_4H_8O_2 \longleftrightarrow C_3H_8O_2 + C_5H_8O_2, \tag{1}$$

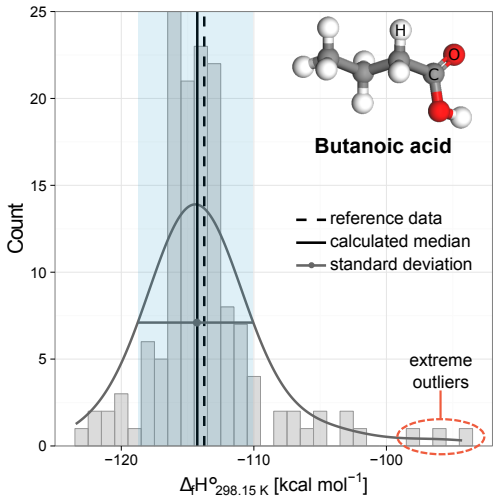

Figure 2: Histogram of the estimated values of the standard enthalpy of formation for butanoic acid ($C_4H_8O_2$). 143 distinct isodesmic reactions were identified. The reference value of -113.74 ± 0.96 kcal mol^{-1} (dashed line) for butanoic acid was taken from the NIST Chemistry WebBook [8]. Outliers giving particular poor estimates are highlighted.

where propylene glycol ($C_3H_8O_2$) and acetylacetone ($C_5H_8O_2$) are used. Although not the EBR that gives the best estimate of the enthalpy of formation for butanoic acid, it results in a deviation of just 0.73 kcal mol^{-1} from the reference value. On the other hand, the reaction,

$$C_4H_8O_2 + C_4H_6 \longleftrightarrow C_4H_8O + C_4H_6O, \tag{2}$$

where 1,3-butadiene (C_4H_6), butanal (C_4H_8O), and (Z)-1,3-butadienol (C_4H_6O) are used to estimate the standard enthalpy of formation for butanoic acid, results in a larger deviation of 6.81 kcal mol^{-1}. In this case the inconsistency is likely to be the result of an observed inconsistency originating from (Z)-1,3-butadienol. Estimating the standard enthalpy of formation for (Z)-1,3-butadienol shows an absolute difference of 7.81 kcal mol^{-1} from the reference value.

The distribution of values in Figure 2 enables the calculation of a central measure to provide a more accurate estimate of the standard enthalpy of formation. The width of the distribution provides some information about the

statistical uncertainty in the estimate. A further improvement can be achieved using techniques such as the modified Thompson-Tau [51] or modified z-score method [52] to identify and exclude outliers. This is valuable because the species in the reactions that contribute to the outliers are also potentially sources of inconsistent reference data. The identification and exclusion of species that contribute to the outliers is automated and exploited by the cross-validation.

3.2. Global cross-validation

Figure 3 (top panel) shows the decrease in the mean absolute error (defined over the full set of reference data) that is achieved by iteratively identifying and excluding inconsistent species. The bottom panel shows the number of excluded species at each iteration. The cross-validation requires the specification of a rejection threshold parameter, x_{rej}. This is the magnitude of the maximum acceptable error for each species. The error is the difference between the calculated data (in this case enthalpy of formation) and the corresponding quantity in the reference data.

The mean absolute error is observed to converge rapidly to an asymptotic value, and the results are shown to be repeatable between independent runs. The asymptotic values of the mean absolute error are significantly less than the rejection threshold for the cases where $x_{\mathrm{rej}} \geq 2.0$ kcal mol^{-1}. However, there are diminishing returns as the rejection threshold is decreased. This is because we reduce the number of possible EBRs as we reject more species, such that we start to lose the statistical benefits of using multiple EBRs. A mean absolute error of ≈ 1.2 kcal mol^{-1} is achieved for a rejection threshold of $x_{\mathrm{rej}} = 1.0$ kcal mol^{-1}.

There is an analogous trade-off when choosing the class of EBR, for example the isogyric, isodesmic, hypohomodesmotic, homodesmotic, and hyperhomodesmotic reaction class. A more restrictive class should give more accurate results for each individual EBR [38, 53]. However, the more restrictive the class, the fewer EBRs are available and the less we benefit from the statistics of using multiple EBRs.

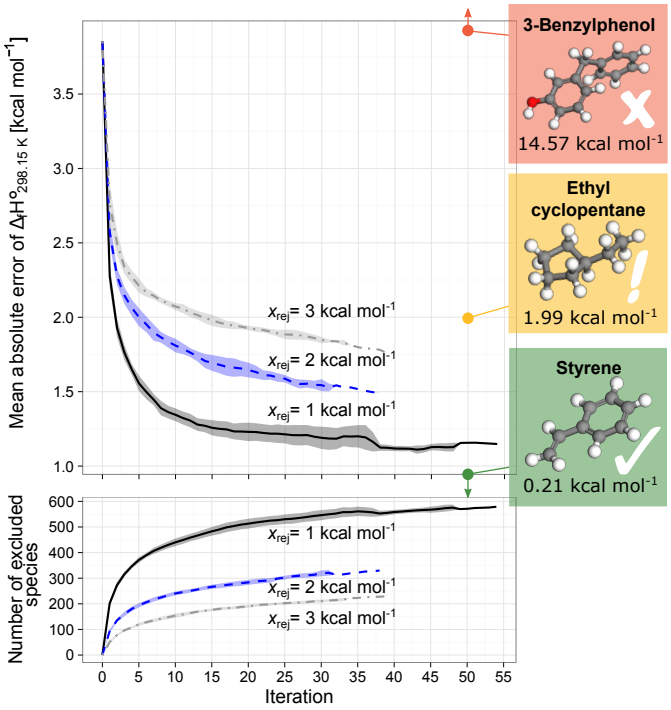

Figure 3: The mean absolute error (top panel) and the number of excluded species (bottom panel) for rejection thresholds of $x_{\mathrm{rej}} = 3.0$, 2.0, and 1.0 kcal mol^{-1}. The results were averaged over ten independent runs. The lines show the mean values. The shaded areas show the standard deviation. An example of a consistent species (styrene), a potentially inconsistent species (ethyl cyclopentane), and an inconsistent species (3-benzylphenol) identified by the cross-validation are shown.

The panels at the side of Figure 3 shows example output from the cross-validation. Styrene (C_8H_8) is classified as consistent because there is only 0.21 kcal mol^{-1} error between the calculated and reference values of the standard enthalpy of formation. Ethyl cyclopentane (C_7H_{14}) is classified as potentially inconsistent because the 1.99 kcal mol^{-1} error is similar to the error that might be expected from the B97-1/6-311+G(d,p) level of theory used in the calcula-

tion[1]. It cannot be determined automatically whether the discrepancy is due to the level of theory, the geometry, the reaction class or an issue with the reference data. Manual analysis of the EBRs used to calaculate the standard enthalpy of formation of ethyl cyclopentane showed that most of the reactions that led to more accurate results conserved the five-member ring on either side of the reaction. An example reaction which results in a better estimate for ethyl cyclopentane is defined by,

$$2\,C_7H_{14} \longleftrightarrow C_{14}H_{28}, \tag{3}$$

where nonyl cyclopentane ($C_{14}H_{28}$) is required. This reaction only leads to a deviation of just 0.43 kcal mol^{-1} from the reference value of ethyl cyclopentane. This shows that in this case the isodesmic reaction class is most likely insufficient and a higher order reaction class should be considered. 3-benzylphenol ($C_{13}H_{12}O$) is classified as inconsistent because the 14.57 kcal mol^{-1} error exceeds what might be expected from the level of theory. It is likely that there is an issue with the reference data, and in fact, Verevkin [54] and Miranda et al. [55] showed some evidence of discrepancies for related phenols from the same original experiment [56].

Detailed results for styrene, ethyl cyclopentane, and 3-benzylphenol are shown in Figure 4. In each case, the standard deviation of the estimated standard enthalpy of formation is observed to decrease significantly as we exclude outliers and inconsistent species. The histograms show tight distributions, and good agreement with the reference data for styrene. The reference data for ethyl cyclopentane is just within one standard deviation of the median when using the full data set, but falls outside this criteria as the outliers and inconsistent

[1]Different works investigated the accuracy of DFT methods to predict standard enthalpies of formation using EBRs. In the work of Wheeler et al. [38] the B3LYP/6-31G(d) level of theory yields for the reaction enthalpies a mean absolute error for selected hydrocarbons of 7.06 kcal mol^{-1} for conjugated and 2.67 kcal mol^{-1} for nonconjugated hydrocarbons using isodesmic reactions.

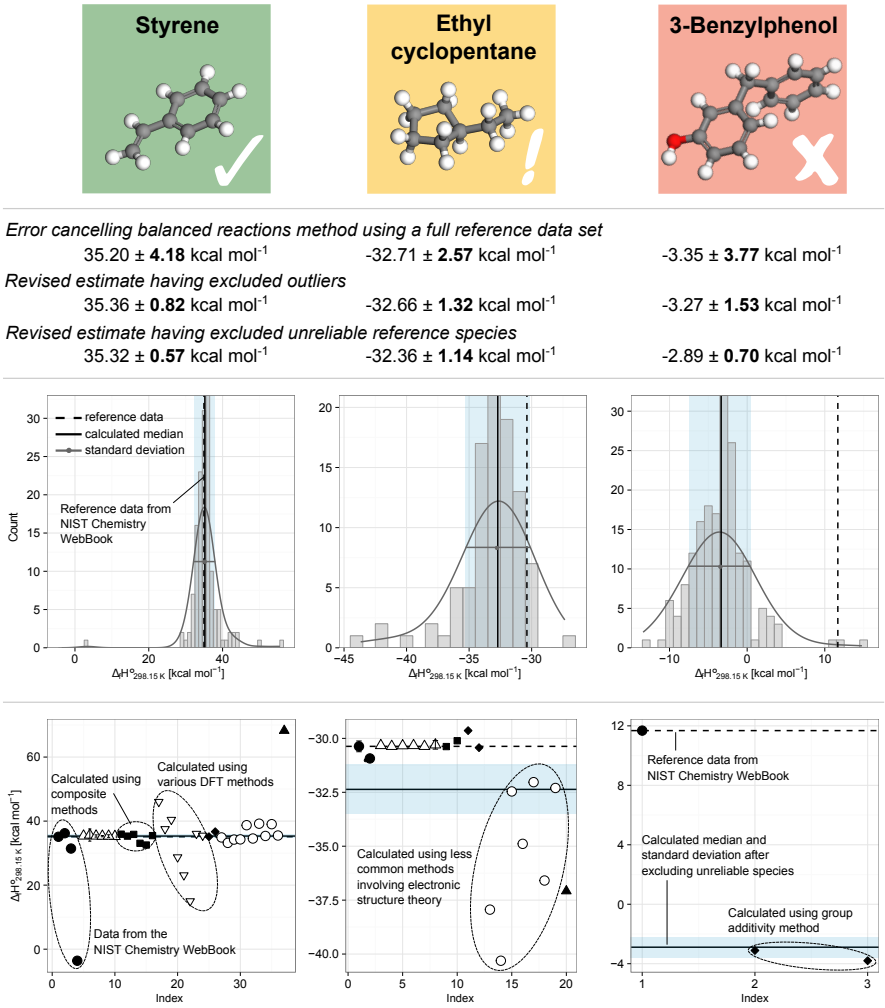

Figure 4: Detailed results for styrene, ethyl cyclopentane, and 3-benzylphenol. The text gives the median and standard deviation of the estimated standard enthalpies of formation using the full reference data set, after excluding outliers and after excluding inconsistent species. The histograms show the distributions of the estimated values of the standard enthalpy of formation using the full reference data set. The scatter plots compare the results (after excluding inconsistent species) with literature data [8, 57, 58, 59, 60, 61, 62, 26, 21, 19, 63, 64, 26, 65, 66, 67, 68, 69, 70, 71, 72, 73]. The clustering of the results calculated using other methods is highlighted. Full details of the reference data are provided as Supplementary Material.

species are removed, again suggesting that ethyl cyclopentane is worthy of further scrutiny. The reference data for 3-benzylphenol is nowhere close to the calculated values, and is considered to be inconsistent as discussed above.

The scatter plots in Figure 4 (bottom panel) show that there is often significant scatter in the literature data. This is of course expected. The availability of the distribution of estimated values enables an assessment of the reference data. The framework automates this process and is able to select the reference data that it deems most likely to be accurate, and allows the identification of data that may be less consistent and that merit further consideration. The analysis can be taken further and used to identify which methods may be appropriate for a particular species. For example, the clustering of the data imply that composite methods should be considered over DFT methods for styrene, and that group additivity methods[2] may be a suitable choice for 3-benzylphenol.

3.3. Discussion

A critical element of the framework is the method used to calculate the quantity of interest. In this paper, we use EBRs to exploit the similarities between a set of species to calculate the standard enthalpy of formation of one species from the set. The fact that we use multiple EBRs and that each EBR uses a set of species enables the framework to isolate and quantify the error due to each species. In principle, the framework is general and is not limited to EBRs. It could be used with any calculation that shares these properties. In abstract terms, it could be used with any calculation that permits the ability to use multiple overlapping subsets of the reference data to calculate a given quantity of interest.

It is proposed to make the framework available as a web application. The

[2]The group additivity calculations were performed *post hoc*. The data did not form part of the original reference data. The calculations were performed using the methods proposed by Benson and Buss [65] and Joback and Reid [66] as currently implemented by the NIST Chemsitry WebBook [8] and Cheméo [67].

concept is illustrated in Figure 5. The application would allow users to upload and validate their data, and would be linked to databases, for example MolHub [13], that allow the easy storage and retrieval of computational chemistry data.

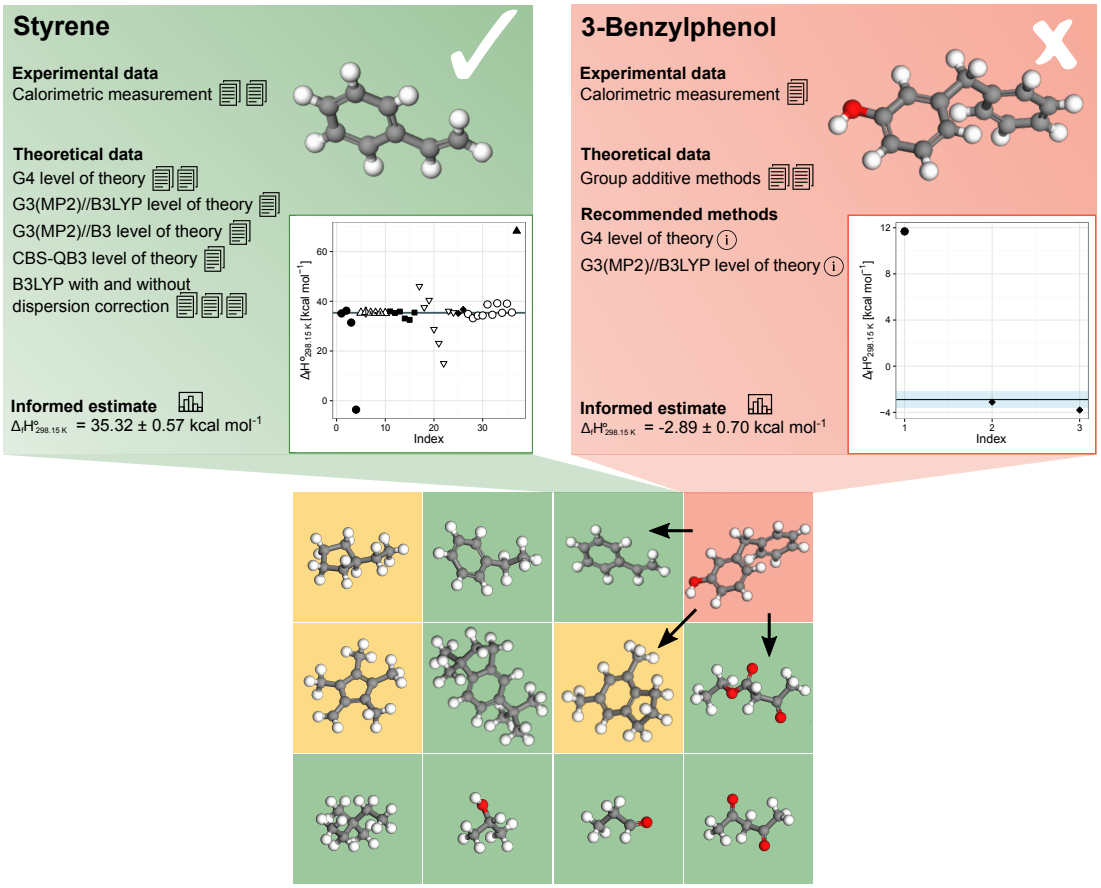

Figure 5: Conceptual illustration of the cross-validation web application. Each tile shows a species. The tiles are grouped according to the degree of similarity with neighbouring species and are coloured to indicate the results of the validation. Pop-up windows provide detailed information about the validation and reference data, and recommended methods to improve the quality of the data.

A colour code is used to indicate the results of the validation. Species are organised based on structural similarity, with species showing the highest degree

of similarity appearing closest to each other. For each species, users are able to access the results of the cross-validation as per Figure 2 and see an overview of the available experimental and theoretical data, together with links to view the data in more detail.

In the case of inconsistent species, for example 3-benzylphenol, it is envisaged that the ability to identify which methods worked well for consistent species (as per the scatter plots in Figure 4) and knowledge of the structural similarities between species could be combined to recommend what methods may be most suitable for similar inconsistent species. This could be to suggest what future experiments or computations might be considered to improve the quality of the data.

4. Conclusions

This paper demonstrates the application of a new framework to validate large sets of thermochemical data for chemical species. The framework implements a global cross-validation method that compares calculated values to corresponding quantities from a reference data set. The cross-validation enables the framework to assess the consistency of the reference data.

The framework may be used with any calculation that uses multiple overlapping subsets of the reference data to calculate the quantity of interest. The demonstration in this paper uses error-cancelling balanced reactions (EBRs) to validate data for the standard enthalpy of formation of 920 gas-phase hydrocarbons, including species containing oxygen, from the NIST Chemistry WebBook [8].

The EBRs were systematically identified using constrained optimization and the electronic structures of all species calculated using DFT at the B97-1/6-311+G(d,p) level of theory. There is a trade-off between the rejection threshold required by the global cross-validation and the accuracy of the calculated standard enthalpies of formation. The accuracy of calculations improves asymptotically at the expense of excluding more data as being inconsistent, as the

rejection threshold is tightened.

The framework offers many important advantages. Firstly, it calculates a distribution of estimates for each species. The width of the distribution provides a measure of the statistical uncertainty in the estimate, whilst the median provides a better estimate than would be obtained from a single sample from the distribution. Secondly, it identifies outliers and inconsistent reference data, and significantly improves the estimate by excluding these data. Thirdly, it is able to quantify the consistency of the species and recommend which ones should be investigated to most improve the data set.

It is proposed to make the cross-validation framework available as a web application. The application would allow users to upload and validate their data, and should be linked to databases that enable the easy storage and retrieval of computational chemistry data. This paper shows how the web application would recommend which *species* should be investigated to most improve the data set, and also how the structural similarities between species might be exploited to suggest which *methods* should be considered for these investigations. Full details of the methods used by the framework will be published in future work.

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.

References

[1] C. W. Gao, J. W. Allen, W. H. Green, R. H. West, Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms, Comput. Phys. Commun. 203 (2016) 212–25, doi:10.1016/j.cpc.2016.02.013.

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

[3] D. Nurkowski, P. Buerger, J. Akroyd, M. Kraft, A Detailed Kinetic Study of the Thermal Decomposition of Tetraethoxysilane, Proc. Combust. Inst. 35 (2) (2015) 2291–8, doi:10.1016/j.proci.2014.06.093.

[4] D. Nurkowski, S. J. Klippenstein, Y. Georgievskii, M. Verdicchio, A. W. Jasper, J. Akroyd, S. Mosbach, 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) (2015) 691–708, doi:10.1515/zpch-2014-0640.

[5] P. Buerger, D. Nurkowski, J. Akroyd, S. Mosbach, M. Kraft, First-Principles Thermochemistry for the Thermal Decomposition of Titanium Tetraisopropoxide, J. Phys. Chem. A 119 (30) (2015) 8376–87, doi:10.1021/acs.jpca.5b01721.

[6] P. Buerger, D. Nurkowski, J. Akroyd, M. Kraft, A Kinetic Mechanism for the Thermal Decomposition of Titanium Tetraisopropoxide, Accepted for publication in Proc. Combust. Inst. .

[7] Nano, URL http://nano.nature.com/, retrieved July 07, 2016, 2016.

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

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

[10] Active Thermochemical Tables, version 1.118, URL http://atct.anl.gov/, retrieved July 03, 2016, 2016.

[11] PrIMe: Process Informatics Model, URL http://www.primekinetics.org/prime_data_warehouse/, retrieved June 29, 2016, 2016.

[12] ReSpecTh, URL http://respecth.hu/, retrieved July 07, 2016, 2016.

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

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

[15] M. N. Weaver, J. Kenneth M. Merz, D. Ma, H. J. Kim, 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) (2013) 5277–85, doi:10.1021/ct400856g.

[16] K. P. Somers, 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) (2015) 8922–33, doi:10.1021/acs.jpca.5b05448.

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

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

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

[20] A. Tajti, P. G. Szalay, A. G. Császár, M. Kállay, J. Gauss, E. F. Valeev, B. A. Flowers, J. Vázquez, J. F. Stanton, HEAT: High Accuracy Extrapolated Ab Initio Thermochemistry, J. Chem. Phys. 121 (23) (2004) 11599–613, doi:10.1063/1.1811608.

[21] S. Rayne, 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) (2010) 5359–64, doi:10.1021/je100768s.

[22] P. Ho, C. F. Melius, Theoretical Study of the Thermochemistry of Fluorosilanes (SiF_n and SiH_nF_m) Compounds and Hexafluorodisilane, J. Phys. Chem. 94 (12) (1990) 5120–7, doi:10.1021/j100375a066.

[23] C. F. Melius, P. Ho, Theoretical Study of the Thermochemistry of Molecules in the Silicon-Nitrogen-Hydrogen-Fluorine System, J. Phys. Chem. 95 (3) (1991) 1410–9, doi:10.1021/j100156a070.

[24] M. D. Allendorf, C. F. Melius, Theoretical Study of Thermochemistry of Molecules in the Silicon-Carbon-Chlorine-Hydrogen System, J. Phys. Chem. 97 (3) (1993) 720–8, doi:10.1021/j100105a031.

[25] C. F. Melius, M. D. Allendorf, Bond Additivity Corrections for Quantum Chemistry Methods, J. Phys. Chem. A 104 (11) (2000) 2168–77, doi:10.1021/jp9914370.

[26] M. Saeys, M.-F. Reyniers, G. B. Marin, V. V. Speybroeck, 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) (2003) 9147–59, doi:10.1021/jp021706d.

[27] R. H. West, G. J. O. Beran, W. H. Green, M. Kraft, First-Principles Thermochemistry for the Production of TiO_2 from $TiCl_4$, J. Phys. Chem. A 111 (18) (2007) 3560–5, doi:10.1021/jp0661950.

[28] R. Shirley, Y. Liu, T. S. Totton, R. H. West, M. Kraft, First-Principles Thermochemistry for the Combustion of a $TiCl_4$ and $AlCl_3$ Mixture, J. Phys. Chem. A 113 (49) (2009) 13790–6, doi:10.1021/jp905244w.

[29] W. Phadungsukanan, S. Shekar, R. Shirley, M. Sander, R. H. West, M. Kraft, First-Principles Thermochemistry for Silicon Species in the Decomposition of Tetraethoxysilane, J. Phys. Chem. A 113 (31) (2009) 9041–9, doi:10.1021/jp905494s.

[30] R. O. Ramabhadran, K. Raghavachari, Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy, J. Chem. Theory. Comput. 7 (7) (2011) 2094–103, doi:10.1021/ct200279q.

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

[32] W. J. Hehre, R. Ditchfield, L. Radom, 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) (1970) 4796–801, doi:10.1021/ja00719a006.

[33] J. A. Pople, L. Radom, 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) (1971) 289–300, doi:10.1021/ja00731a001.

[34] J. A. Pople, M. J. Frisch, B. T. Luke, J. S. Binkley, A Møller-Plesset Study of the Energies of AHn Molecules (A = Li to F), Int. J. Quantum Chem. 24 (S17) (1983) 307–20, doi:10.1002/qua.560240835.

[35] P. George, M. Trachtman, C. W. Bock, A. M. Brett, An Alternative Approach to the Problem of Assessing Stabilization Energies in Cyclic

Conjugated Hydrocarbons, Theor. Chim. Acta 38 (2) (1975) 121–9, doi: 10.1007/BF00581469.

[36] P. George, M. Trachtman, C. W. Bock, A. M. Brett, Homodesmotic Reactions for the Assessment of Stabilization Energies in Benzenoid and other Conjugated Cyclic Hydrocarbons, J. Chem. Soc., Perkin Trans. 2 (1976) 1222–7doi:10.1039/P29760001222.

[37] P. George, M. Trachtman, A. M. Brett, 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 (1977) 1036–47doi:10.1039/P29770001036.

[38] S. E. Wheeler, K. N. Houk, P. v. R. Schleyer, W. D. Allen, A Hierarchy of Homodesmotic Reactions for Thermochemistry, J. Am. Chem. Soc. 131 (7) (2009) 2547–60, doi:10.1021/ja805843n.

[39] M. D. Wodrich, C. Corminboeuf, S. E. Wheeler, Accurate Thermochemistry of Hydrocarbon Radicals Via an Extended Generalized Bond Separation Reaction Scheme, J. Phys. Chem. A 116 (13) (2012) 3436–47, doi: 10.1021/jp212209q.

[40] J. P. Merrick, D. Moran, L. Radom, An Evaluation of Harmonic Vibrational Frequency Scale Factors, J. Phys. Chem. A 111 (45) (2007) 11683–700, doi: 10.1021/jp073974n.

[41] A. D. Boese, J. M. L. Martin, N. C. Handy, The Role of the Basis Set: Assessing Density Functional Theory, J. Chem. Phys. 119 (6) (2003) 3005–14, doi:10.1063/1.1589004.

[42] 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, D. J. Fox, Gaussian 09 Revision D.01, 2009.

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

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

[45] G. B. Dantzig, Linear Programming, Oper. Res. 50 (1) (2002) 42–7, doi:10.1287/opre.50.1.42.17798.

[46] P. E. Gill, W. Murray, M. A. Saunders, J. A. Tomlin, M. H. Wright, George B. Dantzig and Systems Optimization, Discrete Optimization 5 (2) (2008) 151–8, doi:10.1016/j.disopt.2007.01.002.

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

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

[49] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Sta-

tistical Learning: With Applications in R, Springer Publishing Company, Incorporated, ISBN 1461471370, 9781461471370, 2014.

[50] P. Refaeilzadeh, L. Tang, H. Liu, Encyclopedia of Database Systems, chap. Cross-Validation, Springer US, Boston, MA, ISBN 978-0-387-39940-9, 532–8, 2009.

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

[52] B. Iglewicz, D. Hoaglin, How to Detect and Handle Outliers, ASQC basic references in quality control, ASQC Quality Press, ISBN 9780873892476, 1993.

[53] S. E. Wheeler, A. Moran, S. N. Pieniazek, K. N. Houk, Accurate Reaction Enthalpies and Sources of Error in DFT Thermochemistry for Aldol, Mannich, and α-Aminoxylation Reactions, J. Phys. Chem. A 113 (38) (2009) 10376–84, doi:10.1021/jp9058565.

[54] S. P. Verevkin, Thermochemistry of Phenols: Buttress Effects in Sterically Hindered Phenols, J. Chem. Thermodyn. 31 (11) (1999) 1397–416, doi:10.1006/jcht.1999.0466.

[55] M. S. Miranda, J. C. E. da Silva, J. F. Liebman, Gas-Phase Thermochemical Properties of some Tri-Substituted Phenols: A Density Functional Theory Study, J. Chem. Thermodyn. 80 (2015) 65–72, doi:10.1016/j.jct.2014.08.025.

[56] G. Bertholon, M. Giray, R. Perrin, M. Vincent-Falquet-Berny, No. 532.-Etude Physicochimique des Phenols. IX.-Enthalpies de Combustion et Energies de Resonance des Alcoyletaryphenols, Bull. Soc. Chim. France 12 (1971) 3180–7.

[57] D. Lide, CRC Handbook of Chemistry and Physics, 84th Edition, CRC Handbook of Chemistry and Physics, Taylor & Francis, ISBN 9780849304842, 2003.

[58] A. I. Vitvitskii, Calculation of Heats of Formation for Carbocyclic Compounds, Theor. Exp. Chem. 3 (1) (1969) 44–7, doi:10.1007/BF00523731.

[59] A. Burcat, B. Ruscic, Third Millenium Ideal Gas and Condensed Phase Thermochemical Database for Combustion with Updates from Active Thermochemical Tables, Anl [series], Faculty of Aerospace Engineering, Technion - Israel Institute of Technology and Chemistry Division, Argonne National Laboratory, 2005.

[60] J. Pedley, Thermochemical Data of Organic Compounds, Springer Netherlands, ISBN 9780412271007, 1986.

[61] S. Lias, United States. National Bureau of Standards, Gas-Phase Ion and Neutral Thermochemistry, J. Phys. Chem. Ref. Data: Supplement, Published by the American Chemical Society and the American Institute of Physics for the National Bureau of Standards, ISBN 9780883185629, 1988.

[62] J. Cioslowski, M. Schimeczek, G. Liu, V. Stoyanov, A Set of Standard Enthalpies of Formation for Benchmarking, Calibration, and Parametrization of Electronic Structure Methods, J. Chem. Phys. 113 (21) (2000) 9377–89.

[63] E. Taskinen, Enthalpies of Formation and Isomerization of Aromatic Hydrocarbons and Ethers by G3(MP2)//B3LYP Calculations, J. Phys. Org. Chem. 22 (6) (2009) 632–42, doi:10.1002/poc.1494.

[64] G. Blanquart, H. Pitsch, Thermochemical Properties of Polycyclic Aromatic Hydrocarbons (PAH) from G3MP2B3 Calculations, J. Phys. Chem. A 111 (28) (2007) 6510–20, doi:10.1021/jp068579w.

[65] S. W. Benson, J. H. Buss, Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties, J. Chem. Phys. 29 (3) (1958) 546–72, doi:10.1063/1.1744539.

[66] K. Joback, R. Reid, Estimation of Pure-Component Properties from Group-Contributions, Chem. Eng. Commun. 57 (1-6) (1987) 233–43, doi:10.1080/00986448708960487.

[67] Cheméo, URL `https://www.chemeo.com/`, retrieved June 29, 2016, 2016.

[68] A. A. Voityuk, Accurate Treatment of Energetics and Geometry of Carbon and Hydrocarbon Compounds Within Tight-Binding Model, J. Chem. Theory Comput. 2 (4) (2006) 1038–44, doi:10.1021/ct600064m.

[69] A. A. Voityuk, Thermochemistry of Hydrocarbons. Back to Extended Hückel Theory, J. Chem. Theory Comput. 4 (11) (2008) 1877–85, doi:10.1021/ct8003222.

[70] P. Duchowicz, E. Castro, Hydrocarbon Enthalpies of Formation from Ab Initio Calculations Improved Through Bond Parameters, J. Korean Chem. Soc. 43 (6) (1999) 621–7.

[71] M. P. Repasky, J. Chandrasekhar, W. L. Jorgensen, Improved Semiempirical Heats of Formation Through the Use of Bond and Group Equivalents, J. Comput. Chem. 23 (4) (2002) 498–510, doi:10.1002/jcc.10023.

[72] D. Bond, Computational Methods in Organic Thermochemistry. 1. Hydrocarbon Enthalpies and Free Energies of Formation, J. Org. Chem. 72 (15) (2007) 5555–66, doi:10.1021/jo070383k.

[73] D. F. DeTar, Experimental Formal Steric Enthalpy. 1. Alkanes and Cycloalkanes, J. Org. Chem. 56 (4) (1991) 1463–70, doi:10.1021/jo00004a023.