POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Fostering automation in chemical kinetics: a protocol for bond energy computation and the implementation of a hierarchical approach for thermochemistry calculations

TESI MAGISTRALE IN CHEMICAL ENGINEERING – INGEGNERIA CHIMICA

AUTHOR: MARCELLO FERRARO

ADVISOR: CARLO ALESSANDRO CAVALLOTTI

CO-ADVISOR: ANDREA DELLA LIBERA

ACADEMIC YEAR: 2022-2023

1. Introduction

The automation of chemical kinetics has reached, in the past years, a stage where predictive kinetics can override postdictive kinetics, as suggested by Green [1]. Automatic exploration of Potential Energy Surfaces (PES) and accurate estimation of thermochemical parameters are two fundamental challenges to fully automate calculation procedures in predictive chemical kinetics.

The study of PESs focuses on the determination of the main reaction channels and the identification of transition state molecular structures. Not all the possible reaction pathways are equally important; for example, the energy required for breaking a bond can be quite different for distinct bonds.

The estimation of thermochemical parameters implemented by modern software like Auto-Mech [2], RMG [3], KinBot [4], Genesys [5] and Arkane [6] relies either on group contribution (GC) methods, such as those proposed by Joback [7], Costantinou and Gani [8], and Benson [9], or computational chemistry calculations, usually exploiting the application of atomization schemes.

In the present work three pieces of software, which work in synergy with our in-house software EStokTP [10], were implemented.

The first software automatizes the creation of input data and files for EStokTP calculations. In particular the automatic generation of a Z-matrix from InChI identifiers suitable to describe even complex geometries is necessary in order to overcome the tedious and error generating (i.e., bad first guess structure and ordering of atoms) procedure which was previously required a direct intervention the user. The Python code, named InChI2data, has two versions: for single **./data** subdirectory generation, starting from a single InChI identifier, and for multiple **./data** subdirectories generation, starting from a list of InChI identifiers.

The second software implements the calculation of bond energies of a molecule by successive fragmentation of each covalent bond (avoiding ring breakage at this stage of implementation) using the Python code FragsGen (Fragments Generator).

The third protocol implements the estimation of thermochemical parameters $C_P^0(T)$, $S^0(T)$, $H^0(T)$ and $G^0(T)$ in the form of NASA polynomials in CHEMKIN format, using the Python code CHEMTP (Connectivity Hierarchy Estimation Model for Thermochemical Parameters).

2. Algorithms description

2.1 InChI2data

Starting from an InChI identifier, using the open-source toolkit RDKit, the Mol file of the molecule of interest is generated. It contains Cartesian information of the 3D molecular geometry. The Mol file of ethane is reported in Scheme (1) as an example:

```
10  9  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0
    1.0186    0.0000   -0.3968 H   0  0  0  0  0  0
   -0.5097   -0.8818   -0.3968 H   0  0  0  0  0  0
   -0.5088    0.8825   -0.3966 H   0  0  0  0  0  0
    0.0000    0.0000    1.5261 C   0  0  0  0  0  0
   -1.0186    0.0015    1.9228 H   0  0  0  0  0  0
    0.5101    0.8818    1.9227 H   0  0  0  0  0  0
    0.5084   -0.8825    1.9232 H   0  0  0  0  0  0
```

Scheme 1: Mol file for the ethane molecule

After extraction of the Cartesian information (highlighted in bold in Scheme (1)), an iterative generation of Z-matrix is performed by successive rearrangement of the atomic definition order in the Cartesian matrix. The Z-matrix generation is performed using x2z, a Python code that makes heuristic analysis of molecular bonding based on Cartesian information and produces electronic configurations, internal rotation structure and Z-matrix, developed by Klippenstein et al. Once the definition of the Z-matrix is satisfactory (i.e., priority definition of non-hydrogen atoms with respect to hydrogen atoms for the molecule backbone description and the correct identification of rotors), InChI2data copies the other input files containing the levels of theory, the EStokTP modules requested for the calculation, the input for the Master Equation Solver and all the other required input data. The procedure is repeated for every InChI identifier passed to the code. The flowchart of InChI2data algorithm is reported in Figure 1.

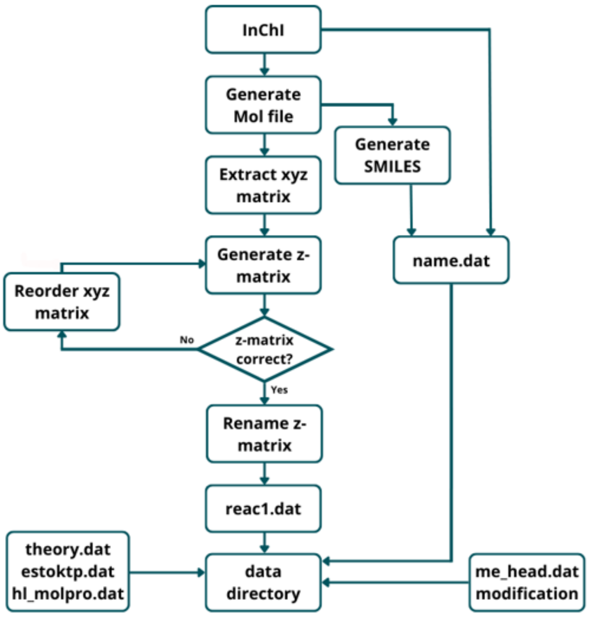

Figure 2: Flowchart of InChI2data algorithm

2.2 FragsGen

FragsGen uses the toolkit RDKit for the identification of the bond to break and the fragments generated by the rupture of the ith bond. Starting from the InChI identifier, the code checks if a bond belongs to a ring structure; if not, it generates the Mol files of the fragments generated by the rupture of such bond. FragsGen also checks if an equivalent fragmentation has already been analysed, to avoid unnecessary EStokTP calculations. If a new fragmentation pathway is found, FragsGen calls the code InChI2data to generate input data for EStokTP calculations for each fragment. Bond energy is calculated as difference between electronic energy and Zero Point Energy (ZPE) of fragmentation products and the original molecule as:

$$E_{bond} = \left[\sum_{i=1}^{2}(E_{el} + ZPE)_i\right] - (E_{el} + ZPE)_{orig.mol.} \quad (1)$$

The flowchart of the FragsGen algorithm is reported in Figure 2.

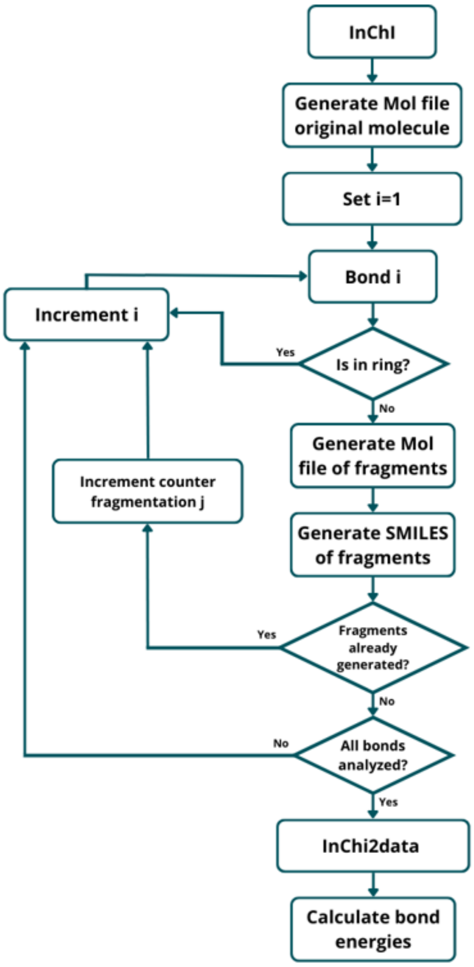
InChI
Generate Mol file original molecule
Set i=1
Bond i
Increment i
Is in ring?
Yes
No
Generate Mol file of fragments
Increment counter fragmentation j
Generate SMILES of fragments
Fragments already generated?
Yes
No
All bonds analyzed?
No
Yes
InChi2data
Calculate bond energies

Figure 2: Flowchart of FragsGen algorithm

2.3 CHEMTP

The Connectivity Based Hierarchy method, developed by Ramabhadran and Raghavachari [11], is implemented for the estimation of $\Delta H^0(0\,K)$. The atomization schemes used in the CBH method are the isogyric scheme (CBH-0), the isodesmic scheme (CBH-1) and the homodesmotic scheme (CBH-2). CBH is a convenient method for estimating $\Delta H^0(0\,K)$, based on electronic calculations on the molecule of interest and a database of reference species, containing electronic energy+ZPE and $\Delta H^0(0\,K)$ information. At each rung, the reaction corresponding to the appropriate atomization scheme is generated using SMILES notation. The reaction energy is calculated both as difference between $\Delta H^0(0\,K)$ of reactants and products, and as difference between the sum of electronic energy and ZPE or reactants and products, as reported in Equation (2).

$$\Delta H_R^o(0\,K) = \sum_{prod} \Delta H_i^o(0\,K) - \sum_{react} \Delta H_i^o(0\,K) = \\ = \sum_{prod} (E_{El} + ZPE)_i - \sum_{react} (E_{El} + ZPE)_i \tag{2}$$

The only unknown is the $\Delta H^0(0\,K)$ of the molecule of interest. The reference species $E_{El} + ZPE$ and $\Delta H^0(0\,K)$ are reported in an in-house built database.

The CBH-0, CBH-1, CBH-2 reactions of 2-hydroperoxybutyl are reported in Scheme (2) as an example.

```
[CH2]C(CC)OO + 5 [H][H] --> 3 C + [CH3] + 2 O
[CH2]C(CC)OO + 3 C + O --> 2CC + [CH2]C +CO +OO
[CH2]C(CC)OO + CC + CO --> CCC + COO + [CH2]C(C)O
```

Scheme 2: CBH-0, CBH-1 and CBH-2 rungs' working reactions for 2-hydroperoxybutyl.

CBH-0 products are constructed by saturating every non-hydrogen atom with H, keeping stable and radical species distinct; a proper amount of hydrogen molecules is added to the reactant side to balance the reaction.

CBH-1 initial reactants are CBH-0 products, corrected to account for the presence of terminal moieties and ramifications (as observed in Scheme (2), CBH-1 reactants and CBH-0 products of 2-hydroperoxybutyl differ for the presence of a [CH3] and a O). CBH-1 products are generated preserving every atom-bond-atom structure in terms of atomic stability and single/double/triple bond type.

CBH-2 products are determined by analyzing each non-hydrogen bond vicinity and by generating the proper SMILES structure. CBH-2 reactants (which are obtained from CBH-1 products) are corrected for terminal moieties (i.e., non-hydrogen atom bonded to a single non-hydrogen atom); ramifications do not need corrections since CBH-2 is an atomic-centred scheme, which considers the surrounding space of each non-hydrogen atom, as opposed to the CBH-1 bond-centred scheme, which does not take automatically into account ramifications.

Once $\Delta H^0(0\,K)$ is estimated, it needs to be corrected to estimate $\Delta H^0(298.15\,K)$; this is necessary because molecular partition functions can not be computed at 0 K. The extrapolation scheme used, reported by Ochterski [12], uses experimental thermal corrections to enthalpy of atomic species by Pople et al [13].

$C_P^0(T)$ and $S^0(T)$ are estimated using molecular partition functions contribution by Ochterski [12] and explicit 1D hindered rotor treatment of structures such as methyl groups.

The obtained data are summarized in the NASA polynomial format, defined as:

$$\frac{C_P^0(T)}{R} = a_0 + a_1T + a_2T^2 + a_3T^3 + a_4T^4 \quad (3.a)$$

$$\frac{H^0(T)}{RT} = a_0 + \frac{1}{2}a_1T + \frac{1}{3}a_2T^2 + \frac{1}{4}a_3T^3 + \frac{1}{5}a_4T^4 + \frac{a_5}{T} \quad (3.b)$$

$$\frac{S^0(T)}{R} = a_0\ln(T) + a_1T + \frac{1}{2}a_2T^2 + \frac{1}{3}a_3T^3 + \frac{1}{4}a_4T^4 + a_6 \quad (3.c)$$

The estimated values of $C_P^0(T)$ over a sufficiently large temperature interval are used for a double non-linear regression of Equation (3.a) to estimate the first 5 coefficients for low-T and high-T intervals. Coefficient a_6 is calculated with two values of $S^0(T)$, one at low T and one at high T, making it explicit from Expression (3.c). Coefficient a_5 at low temperature is calculated making it explicit from Expression (3.b) using $\Delta H^0(298.15\ K)$. H^0 at high T (2500 K) is estimated exploiting the relation between $C_P^0(T)$ and $H^0(T)$:

$$H^0(2500\ K) = \Delta H^0(298.15\ K) + \int_{298.15}^{T_{split}} C^0_{P,lowT}(T)\,dT + \int_{T_{split}}^{2500} C^0_{P,highT}(T)\,dT \quad (4)$$

T$_{split}$ in Equation (4) is the temperature value that separate low and high T ranges.

Once all coefficients are determined, they are formatted in the CHEMKIN format and saved in a text file named *nasa_polyn.out*.

The flowchart of CHEMTP algorithm is reported in Figure 3.

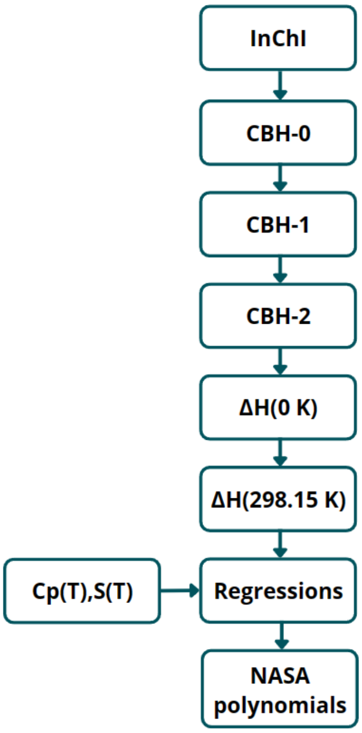

Figure 3: Flowchart of CHEMTP algorithm

3. Results and conclusion

3.1. The performances of the developed software have been tested through the simulation of experimental and theoretical literature data.

3.2. 1,3-butadiene-2-ol fragmentation

The code FragsGen was tested on 1,3-butadiene-2-ol, obtaining a total of 8 different fragmentations, reported in Figure (4).

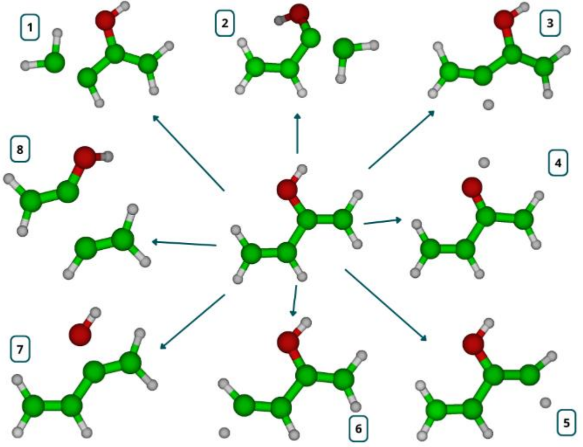

Figure 4: 1,3-butadiene-2-ol fragmentation products

Bond energies were calculated at the ωB97X-D/jun-cc-pVTZ (level 1) and CCSD(T) (HL) levels of theory with extrapolation to the complete basis set limit and correction for core electrons correlation (high level). They are reported in Table (1).

Table 1: 1,3-butadiene-2-ol bond energies

N. Fragments	Level 1 [kcal mol^{-1}]	High level [kcal mol^{-1}]
1	144.28	144.77
2	155.64	154.91
3	107.43	109.86
4	81.54	85.40
5	112.43	113.94
6	109.44	111.32
7	105.29	107.84
8	110.17	113.98

The protocol implemented by FragsGen allowed the correct determination of all possible fragmentations of 1,3-butadiene-2-ol, with good agreement between the bond energy values estimated by FragsGen and literature estimations. The bond energy of the O-H bond was estimated to be 85.40 [kcal mol^{-1}] and identified as the weakest bond. This is in good agreement with the literature value of 84.50 [kcal mol-1] reported by So et al. [14]. Also, the energy of a C-H bond in ethylene molecule, about 110.7 [kcal mol-1], is in line with the bond energies of 109.86, 113.94 and 111.32 [kcal mol-1] calculated by FragsGen for fragmentations 3, 5 and 6.

Based on the results obtained from 1,3-Butadiene-2-ol fragmentation, FragsGen algorithm test was considered to be successful. It can thus be used for determination of the most important reaction channels involving the rupture of a bond in a molecule and it is an important step forward towards the implementation of an automatic approach for PES investigation for unimolecular decomposition.

3.3. CHEMTP estimations

The CHEMTP algorithm was tested on 142 species from the thermochemical database of Klippenstein et. al [15], which includes alkanes, alkyls, alkylhydroperoxides, alkylperoxides and hydroperoxy-alkyls, which was used as a reference for the comparison of the values of the estimated $\Delta H^0(0\ K)$ values. The standard deviations of the absolute error and the relative percentage error were 0.54 [kcal mol-1] and 5.09, respectively; the estimated absolute and relative percentage error, with a confidence of 95%, were 0.39 ± 0.54 [kcal mol-1] and 3.39 ± 5.09 %, which is below the chemical precision of 1 [kcal mol-1] for most of the species.

The relative percentage error between the estimated $\Delta H^0(0\ K)$ of the entire set of 142 species using the code CHEMTP and the $\Delta H^0(0\ K)$ reported by Klippenstein et at. [15] is reported in Figure (5).

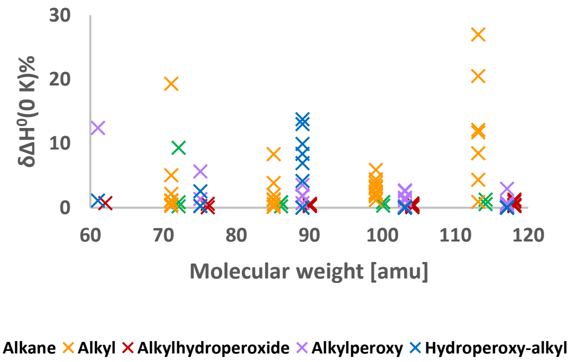

Figure 5: Relative percentage error between CHEMTP and Klippenstein et al. [15]

The estimation of NASA polynomials of the isoprene molecule was performed and compared with the calculation with RMG by Green et al. [3], which implement Benson's GC method.

$C_P^0(T)$ and $S^0(T)$ were calculated within the full rigid rotor harmonic oscillator approximation (RRHO) and with explicit treatment of 1D hindered rotors (RRHO-1DHR).

$C_P^0(T)$, $S^0(T)$, $H^0(T)$ and $G^0(T) = H^0(T) - TS^0(T)$ in the range of 200-3500 [K] are reported in Figures (6-9).

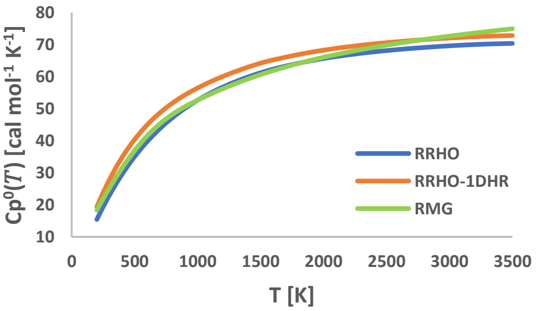

Figure 6: Isoprene $C_P^0(T)$

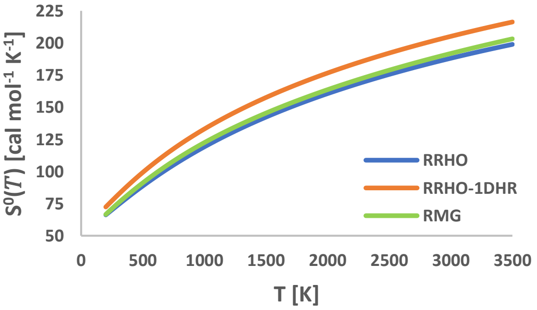

Figure 7: Isoprene $S^0(T)$

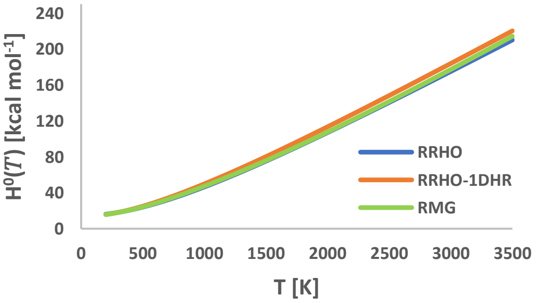

Figure 8: Isoprene $H^0(T)$

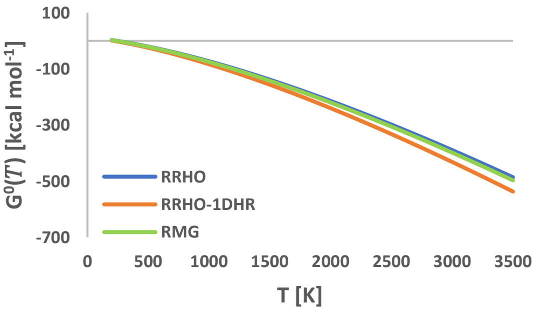

Figure 9: Isoprene $G^0(T)$

In the low temperature regime, CHEMTP shows a good agreement with RMG estimates, both in the full RRHO approximation and with explicit treatment of 1D hindered rotors. The agreement with the data from RMG decreases progressively by moving to higher temperatures; this is due to the incorrect treatment of low frequency internal motions in the RRHO, which is likely to give poor results at high temperatures, while the rotors explicitly treated as hindered rotors in the RRHO-1HDR model assume more and more the characteristics of free rotors, requiring a different theoretical approach.

Bibliography

[1] Green, W.H., *Moving from postdictive to predictive kinetics in reaction engineering*, AIChE J. 66 (2020).

[2] Elliott, S.N., Moore, K.B., Copan, A.V., Keçeli, M., Cavallotti, C., Georgievskii, Y., Schaefer, H.F. Klippenstein, S.J., *Automated theoretical chemical kinetics: Predicting the kinetics for the initial stages of pyrolysis*, Proceedings of the Combustion Institute, Volume 38, Issue 1, Pages 375-384 (2021).

[3] Gao, C.W, Allen, J.W, Green, W.H., West, R.H., *Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms*, Computer Physics Communications, Volume 203, Pages 212-225, (2016).

[4] Van de Vijver, R., Zádor, J., *KinBot: Automated stationary point search on potential energy surfaces*, Computer Physics Communications, Volume 248 (2020).

[5] Vandewiele, N.M., Van Geem, K.M., Reyniers, M.-F., Marin, G.B., *Genesys: Kinetic model construction using chemo-informatics*, Chemical Engineering J., Volumes 207–208, Pages 526-538 (2012).

[6] Dana, A.G., Johnson, M.S., Allen, J.W., et al., *Automated reaction kinetics and network exploration (Arkane): A statistical mechanics, thermodynamics, transition state theory, and master equation software*. Int. J. Chem. Kinet., Volume 55, Pages 300–323, (2023).

[7] Joback, K.G., *A unified approach to physical property estimation using multivariate statistical techniques*, Diss. Massachusetts Institute of Technology (1984).

[8] Constantinou, L., Gani, R., *New group contribution method for estimating properties of pure compounds*, AIChE J., Volume 40, Pages 1697-1710, (1994).

[9] Benson, S. W., Cruickshank, F. R., Golden, D. M., Haugen, G. R., O'neal, H. E., Rodgers, A. S., Walsh, R., et al. *Additivity rules for the estimation of thermochemical properties.* Chemical Reviews, Volume 69 (3), Pages 279-324, (1969).

[10] Cavallotti, C., Pelucchi, M., Georgievskii, Y. & Klippenstein, S. J., *EStokTP: Electronic Structure to Temperature- and Pressure-Dependent Rate Constants—A Code for Automatically Predicting the Thermal Kinetics of Reactions*. J. Chem. Theory Comput. 15, Pages 1122–1145 (2019).

[11] Ramabhadran, R.O., Raghavachari, K., *Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy*, J. of Chemical Theory and Computation, Volume 7 (7), Pages 2094-2103, (2011).

[12] Ochterski, J.W., *Thermochemistry in Gaussian*, Gaussian Inc, (2000).

[13] Curtiss, L.A., Raghavachari, K., Redfern, P.C., Pople, J.A., *Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation*. J. Chem. Phys., Volume 106 (3), Pages 1063–1079, (1997).

[14] So, S., Wille, U., da Silva, G., *Photoisomerization of Methyl Vinyl Ketone and Methacrolein in the Troposphere: A Theoretical Investigation of Ground-State Reaction Pathways*, ACS Earth and Space Chemistry, Volume 2 (8), Pages 753-763, (2018).

[15] Elliott, S.N., Keçeli, M., Ghosh, M.K., Somers, K.P., Curran, H.J., Klippenstein, S.J., *High-Accuracy Heats of Formation for Alkane Oxidation: From Small to Large via the Automated CBH-ANL Method*, The J. of Physical Chemistry A, Volume 127-6, Pages 1512-1532, (2023).