

MIT Open Access Articles

*An Extended Group Additivity Method for
Polycyclic Thermochemistry Estimation*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Han, Kehang et al. "An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation." *International Journal of Chemical Kinetics* 50, 4 (February 2018): 294–303 © 2018 Wiley Periodicals, Inc

As Published: <http://dx.doi.org/10.1002/kin.21158>

Publisher: Wiley Blackwell

Persistent URL: <http://hdl.handle.net/1721.1/114934>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation

Kehang Han^a, Adeel Jamal^a, Colin A. Grambow^a, Zachary J. Buras^a, William H. Green^{a,*}

^a*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States*

Abstract

Automatic kinetic mechanism generation, virtual high-throughput screening, and automatic transition state search are currently trending applications requiring exploration of a large molecule space. Large scale search requires fast and accurate estimation of molecules' properties of interest, such as thermochemistry. Existing approaches are not satisfactory for large polycyclic molecules: considering the number of molecules being screened, quantum chemistry (even cheap DFT methods) can be computationally expensive, and group additivity, though fast, is not sufficiently accurate. This paper provides a fast and moderately accurate alternative by proposing a polycyclic thermochemistry estimation method that extends the group additivity method with two additional algorithms: similarity match and bicyclic decomposition. It significantly reduces $H_f(298\text{ K})$ estimation error from over 60 kcal/mol (group additivity method) to around 5 kcal/mol, $C_p(298\text{ K})$ error from 9 cal/mol/K to 1 cal/mol/K, and $S(298\text{ K})$ error from 70 cal/mol/K to 7 cal/mol/K. This method also works well for heteroatomic polycyclics. A web application for estimating thermochemistry by this method is made available at http://rmg.mit.edu/molecule_search.

Keywords: polycyclics, thermochemistry, ring corrections

1. Introduction

Automatic kinetic mechanism generation [1], virtual high-throughput screening in drug discovery [2], and automatic transition state search [3] are a few examples of current trending applications where a large space of molecules is to be explored. For instance, RMG [1], an automatic kinetic mechanism generation package, typically scans 10,000 \sim 1,000,000 of species for every single run. Such large scale of screening for which even cheap DFT methods are not affordable, requires fast estimation of molecular properties, such as thermochemistry. Group additivity method has been widely used since they were invented by Benson [4] as described by Eq. 1 (taking enthalpy as an example). Applications like RMG use it as a primary estimation method, benefiting from its convenience and interpretability.

*Corresponding author
Email address: whgreen@mit.edu (William H. Green)

$$H_f(298K) = \sum_{i=1}^{N_{atom}} GAV_i \quad (1)$$

where GAV_i is group additivity value for i^{th} atom centered group.

However, due to its underlying assumption that each atom-based group is independent and their contributions are additive, group additivity methods have difficulty estimating the thermochemistry of cyclic molecules, since ring strain is a joint effect among many ring atoms that is beyond single-atom-based scope.

To handle this problem, Benson [4] and others [5, 6] proposed ring corrections on top of the normal atom-based group additivity scheme (Eq. 2).

$$H_f(298K) = \sum_{i=1}^{N_{atom}} GAV_i + \sum_{j=1}^{N_{ring\ cluster}} (ring\ correction)_j \quad (2)$$

where $(ring\ correction)_j$ is additional strain contributed by ring cluster j as a whole.

Note a ring cluster may consist of several individual rings that share at least one atom with at least one other individual ring in the cluster. To make accurate predictions, Eq. 2 requires correction data for every ring cluster in each molecule.

Since each ring cluster structure has its specific ring correction, and there are an extremely large number of possible fused ring clusters, this group additivity method only gives accurate predictions for molecules whose ring structures have been studied in the past. Estimation accuracy drops significantly when dealing with molecules with ring cluster structures not included in the database.

The root problem is that one cannot list the infinite number of possible ring clusters and prepare all the ring corrections. Due to the difficulty in acquiring data from ab initio calculations or experimental measurements, less data is available for molecules with larger ring clusters than for those with smaller ones. However, as cluster size increases more possible structural variations exist, which worsens the situation for estimating large polycyclics.

Therefore, we divide the problem of accurately estimating the thermochemistry of a polycyclic into two sub-problems based on the size of the ring cluster (number of smallest rings in the cluster, using Fan’s algorithm [7] of Smallest Set of Smallest Rings) in the molecule:

- small cyclics (\leq 2-ring molecules) and
- large cyclics (\geq 3-ring molecules)

For the former problem, we calculate and organize the available ring corrections into a functional group tree that can find similar matches for any new small cyclics. For the latter problem, we develop a bicyclic-decomposition model which estimates large polycyclic ring cluster corrections by

decomposing them into smaller ones and adding up the contributions from the fragments. Overall, we managed to bring down group additivity thermo prediction error from over 60 kcal/mol in some cases (original group additivity method in cases where the ring cluster structure of interest had not been studied previously) to 5 kcal/mol for both small cyclics and large cyclics as judged using the dataset of Ramakrishnan, et al. [8].

In this paper, we discuss our similarity match approach in Section 3 and our bicyclic-decomposition approach in Section 4. Additionally, to power these algorithms, we organize and precalculate ring corrections for a list of frequently seen ring cluster structures, with more details in Section 2.

2. Pre-calculation

The precalculated list of hydrocarbon molecules covers molecules with mostly small ring cluster structures (1-ring and 2-ring clusters, see Supporting Information). Some example molecules are shown in Figure 1.

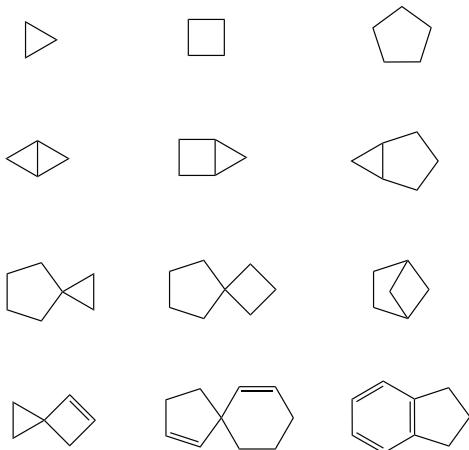


Figure 1: Example small cyclics in our database

To automate the data preparation process, a 3-step scheme was used as shown in Figure 2. Firstly, molecular identifiers are fed into RDKit Chem module [9] to generate initial XYZ coordinates. A GAUSSIAN job creator receives the molecular coordinates, composes GAUSSIAN 09 [10] job inputs and launches quantum chemistry jobs. To optimize the geometry and to compute vibrational frequencies at the optimized geometry XYZ_{opt} , we used the DFT method at M06-2X/cc-pVTZ level of theory [11]. Once the quantum chemistry calculation finishes, RMG’s Cantherm module [1] (additional cantherm information can be found at: <http://reactionmechanismgenerator.github.io/RMG-Py/users/>

cantherm/index.html) parses the output GAUSSIAN log file and calculates the thermochemical parameters such as $H_f(298\text{ K})$, C_p and $S(298\text{ K})$ using the Rigid Rotor Harmonic Oscillator (RRHO) approximation. At each step, molecular representations (SMILES, XYZ, and optimized XYZ coordinates) are converted into RMG species objects and RMG’s isomorphism check ensures that they still represent the same molecule.

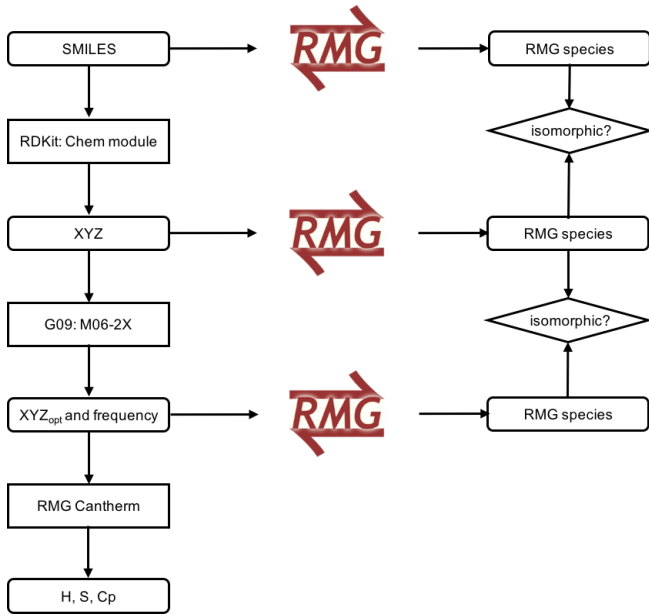


Figure 2: Quantum calculation scheme for small cyclic thermochemistry

Note that the single structure RRHO approach employed here only considers one conformer, and ignores anharmonicity, so it is expected to underestimate S and C_p for floppy rings.

3. Similarity Match

As discussed in the previous section, Eq. 2 needs exact matches of target ring clusters (otherwise no correction is applied at all), and thus requires extensive data to ensure high prediction accuracy.

Here we propose a similarity match algorithm that can find a similar ring with similar thermochemistry for situations where no exact matches are available.

3.1. Cyclic trees

The similarity algorithm greatly relies on data organization. The ring structures are organized into trees (we have two trees so far: monocyclic tree and polycyclic tree), see a sub-tree example in Figure 3. Nodes further down the tree have more specific structural details. The top layer defines the skeleton frame, for instance, s1_3_6 represents a bicyclic consisting of 3-member ring and 6-member ring with 1

atom shared, and the next layer defines categories such as alkane, alkene, diene or aromatics. Finally, the bottom layer lists the most specific ring structures.

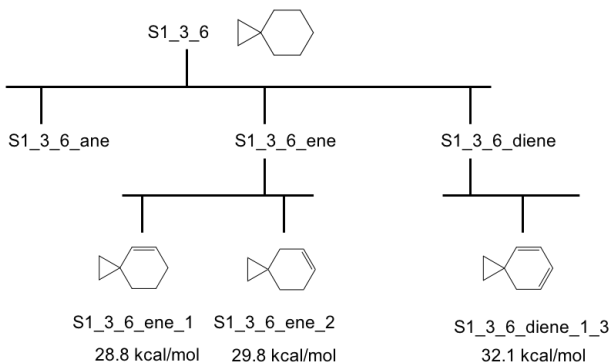


Figure 3: Example sub-tree that organizes polycyclic ring corrections with derived ring correction for enthalpy of formation

With this design, if a new molecule does not exactly match a known ring cluster, it can be classified as similar to some other nodes in the tree, and assigned the average of their values. For instance the molecule in Figure 4 most closely matches the second-layer node `s1_3_6_diene` in Figure 3. Since there is no exact match, the algorithm will use the average of the ring corrections of the children of `s1_3_6_diene` as its estimated correction.



Figure 4: Example molecule which does not exactly match any node in the tree in Figure 3. The tree gives enthalpy estimation of 32.1 kcal/mol, while its real enthalpy of formation from quantum calculation in this study is 27 kcal/mol, leading to around 5 kcal/mol error

3.2. Model test

To evaluate the performance of this similarity match algorithm, an external large quantum calculation dataset [8] is used. This dataset contains 134,000 molecules and has enabled several interesting big data studies [12, 13] that connect machine learning models to molecular property estimation. This paper selects cyclic molecules in that dataset as the test dataset and categorizes the cyclics into small cyclics (1-ring and 2-ring molecules, see example in Figure 5(a)), large linear cyclics (at least 3-ring molecules, and atoms are at most shared by two rings, see example in Figure 5(b)) and large fused cyclics (at least 3-ring molecules, rings are heavily fused (having atoms shared by at least 3 rings), see example in Figure 5(c)).

With the polycyclic tree (small cyclics calculated by the authors of this paper), the similarity match algorithm can successfully reduce the mean absolute error of $H_f(298\text{ K})$ from 32 kcal/mol

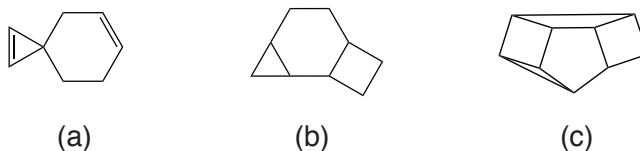


Figure 5: Example cyclics in each category: (a) small cyclics, (b) large linear cyclics, and (c) large fused cyclics

(original group additivity method in cases where the ring cluster structure of interest had not been studied previously) to 3 kcal/mol for small cyclics in the test dataset by Ramakrishnan, et al. [8] (see Table 1 and Figure 6), which is expected since the tree includes many pre-calculated small cyclic corrections.

Table 1: Mean absolute error (kcal/mol) of $H_f(298\text{ K})$ for each category in validation dataset

Method	small cyclics	large linear cyclics	large fused cyclics
Group Additivity Method	32	65	80
+ Similarity Match	3	29	40
+ Bicyclic Decomposition	3	4.9	9.8

Even though there are no large polycyclics in the tree, this similarity match approach also improves the predictions for large polycyclics by sub-molecule isomorphism (see an example in the Supporting Information for how a 3-ring molecule matches a 2-ring node if no 3-ring node available in tree), cutting the mean absolute error by about a factor of two (Table 1).

To further improve the accuracy of large cyclics thermochemistry prediction, obvious approaches would require pre-calculated data on large cyclics. However, the number of possible large polycyclics increases rapidly with the number of rings. To avoid this poor scaling, we built a model that estimates ring corrections of polycyclics from known corrections for bicyclics, as discussed in Section 4.

4. Bicyclic Decomposition

4.1. Method development

Having available ring correction data mostly for small cyclics, a model that estimates the thermochemistry of large cyclics from small cyclic building blocks is needed. We tried three methods, as shown in Figure 7.

Method (a) simply sums up ring correction contributions from single rings that make up the targeted large cyclics. One main drawback of this method is that it only counts the ring strain

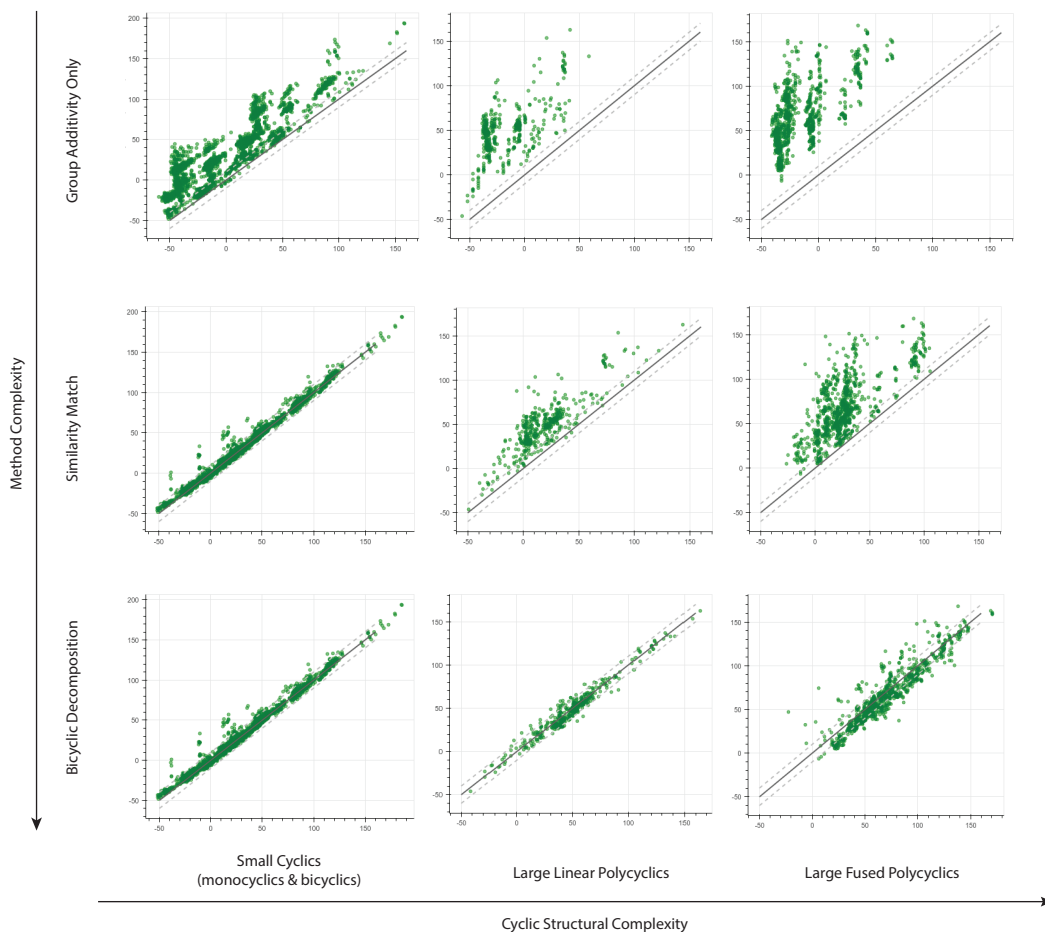


Figure 6: Predictions of enthalpy of formation (x axis) vs. quantum mechanics values (y axis, calculated by Ramakrishnan [8] using DFT method B3LYP/6-31G(2df,p)) with various models and cyclic types, unit: kcal/mol

contributions from individual rings and overlooks the extra strain from the fused part of the bicyclic ring AB in Figure 8. That is the reason method (a) underestimates the ring corrections by over 60 kcal/mol in some cases (Figure 7). Significant discrepancy is observed between actual ring corrections needed to accurately estimate bicyclics and the sums of individual ring strain corrections (Figure 9).

Method (b) in Figure 7 divides a large cyclic into bicyclic components (called bicyclic decomposition), which automatically captures thermo contributions from fused parts. For instance, it decomposes a tricyclic into two bicyclics and estimates its ring corrections by taking the sum of bicyclic corrections.

Method (b) reduces the error in predicted enthalpy of formation of tricyclo-octane to 27 kcal/mol by adding ring strain contributions from two fused parts in the tricyclic (Figure 7). However, it always

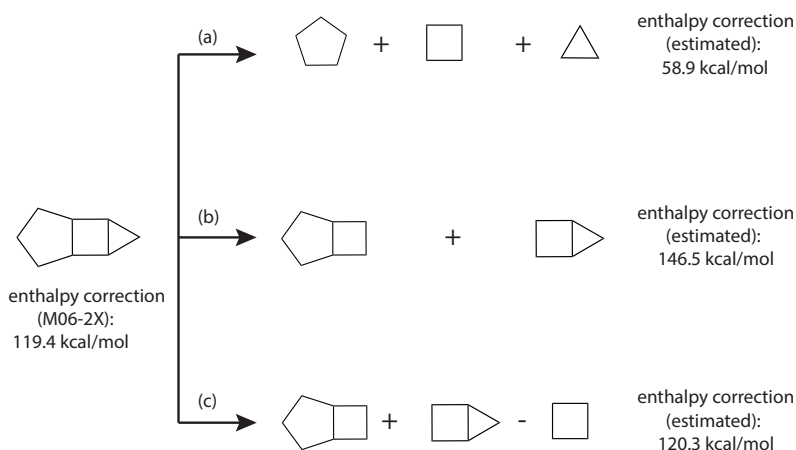


Figure 7: Large cyclic corrections estimation method evolution: (a) sum of individual single ring corrections, (b) sum of bicyclic corrections, and (c) sum of bicyclic corrections with overlapping ring correction subtraction

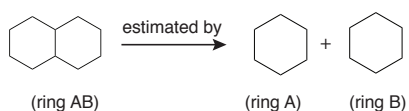


Figure 8: bicyclic AB correction estimated as sum of single ring A and B's corrections

over-predicts the enthalpy corrections, due to the fact that method (b) double counts the contribution of the middle 4-member ring. By eliminating the overlapped ring correction, method (c) shown in Figure 7 calculates a ring strain that agrees well with "true" ring correction (here we use our M06-2X calculations as "true" values). In this particular example, the prediction error is remarkably reduced to 0.9 kcal/mol by adopting the bicyclic decomposition approach. In Subsection 4.3, we conducted a more thorough test of the performance of method (c).

4.2. Bicyclic Correction Estimation

We attempted to calculate commonly seen bicyclics and store them in the database. But polycyclics may have bicyclic components that are not registered in our database. Often these are highly strained (e.g., consecutive double bonds in a ring) such as the examples in Figure 10.

To maintain high accuracy, adding relevant bicyclic clusters into the database would be a long term solution. In this study we developed an additional layer (Figure 11) in the original bicyclic decomposition method for cases where requested bicyclic components are not available or matched nodes are not similar enough to the target bicyclics.

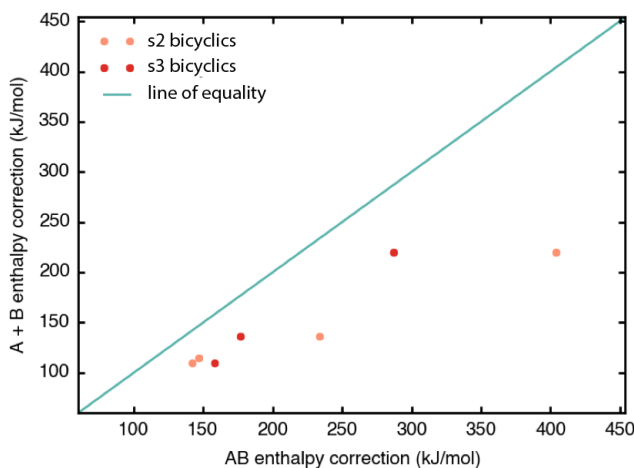


Figure 9: Ring corrections from bicyclics are very different from sum of corrections of individual single rings that make up the bicyclics. Note s2 bicyclics are those with 2-atom bridges, and s3 bicyclics are those with 3-atom bridges.

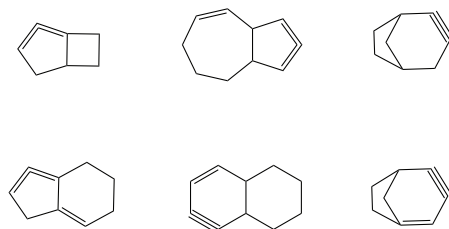


Figure 10: Bicyclic structures like these with high ring strain are usually not recorded in databases, but they can be formed as intermediates during reaction network exploration

4.3. Model test

Using the same polycyclic dataset, the bicyclic decomposition method significantly reduces prediction error for both small and large cyclics (Figure 6).

For small cyclics (≤ 2 -ring molecules), bicyclics decomposition automatically falls back to the similarity match method, which guarantees the prediction accuracy achieved by similarity match, see Table 1. For large polycyclics, bicyclic decomposition outperforms similarity match, bringing the error down to 4.9 kcal/mol and 9.8 kcal/mol for large linear cyclics and large fused cyclics, respectively.

Our algorithm was also tested against the data set by Osmont, et al. [14] who used B3LYP/6-31g(d,p) to calculate enthalpy of formation for propellanes. Our bicyclic decomposition algorithm, without running any further quantum chemistry calculations, was able to get enthalpies of dispiro[2.0.2.1]heptane, trispiro[2.0.2.0.2.0]nonane, trispiro[2.0.0.2.1.1]nonane and tetrspiro[2.0.0.0.2.1.1.1]undecane with DFT accuracy, as shown in Table 2.

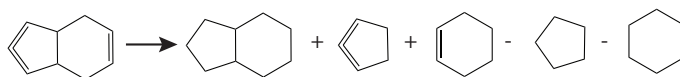


Figure 11: Bicyclic correction estimation scheme for bicyclics missing from the database

Table 2: Experimental and calculated enthalpy of formation at 298K (kcal/mol) for spirocyclic compounds

Structure	Name	Experimental value	Osmont [14] DFT data	This work
	spirocyclic pentane	44.3±0.2	39.0	44.3
	dispiro[2.0.2.1]heptane	72.4±0.8	67.7	69.0
	trispiro[2.0.0.2.1.1]nonane	102.7	96.5	100.1
	trispiro[2.0.2.0.2.0]nonane	101.2±1.2	96.5	97.1
	tetraspiro[2.0.0.0.2.1.1.1]undecane	130.0±1.6	125.3	131.3

For users that are interested in using this method to estimate polycyclic thermochemistry, a web application (http://rmg.mit.edu/molecule_search) is made available to allow users to input molecules (with elements of C, H, O) using species identifiers such as SMILES, InChI, CAS number or species name and to compute that molecule’s thermochemistry. A screenshot illustrating the output from this web tool is shown in Figure 12.

5. Discussion

The new thermochemistry estimator based on group additivity, which combines similarity match and the bicyclic decomposition method, predicts polycyclic thermochemistry more accurately. This extension makes the group additivity method generalizable for polycyclics without requiring much pre-calculated data.

5.1. Large fused cyclics

For bicyclics and linear polycyclics, the typical error of 3 ~ 5 kcal/mol is generally acceptable as a first approximation, especially when dealing with molecules with more than 10 carbons. The underlying reason that such a simple model performs well is that the decomposed bicyclic components act relatively independently; the thermodynamic contribution from the inter-bicyclic interaction is small compared with the ring strain contributions from the bicyclic itself.

by this paper to estimate heteroatom polycyclic thermochemistry.

Predictions made in this way for oxygen-embedded polycyclics agree well with the quantum mechanically calculated values for over 18,000 oxygen-embedded polycyclics [8] (Table 3). Predictions for hetero-atom polycyclics achieve similar accuracy to hydrocarbon polycyclics.

Table 3: Mean absolute error (kcal/mol) of $H_f(298\text{ K})$ for each category of oxygen-embedded polycyclics

Method	small cyclics	large linear cyclics	large fused cyclics
Group Additivity Method	44	78	84
+ Similarity Match	5	34	40
+ Bicyclic Decomposition	5	6.6	10.6

5.3. Heat capacity and standard entropy predictions

Table 4: Mean absolute error (cal/mol/K) of $C_p(298\text{ K})$ for each category of polycyclics

Method	small cyclics	large linear cyclics	large fused cyclics
Group Additivity Method	6.1	10.0	10.5
+ Similarity Match	1.1	2.0	2.7
+ Bicyclic Decomposition	1.1	0.7	1.7

Besides $H_f(298\text{ K})$, the methods proposed by this paper also improve heat capacity prediction accuracy by a factor of $6 \sim 10$ for all three categories of polycyclics (Table 4). This can be crucial for chemical systems operated at temperatures other than 298 K . For standard entropy $S(298\text{ K})$ predictions (Table 5), we also observed a similar accuracy boost using the bicyclic decomposition method. The good agreement seems to suggest the contributions to entropy and heat capacity from ring strains are also additive (although we note the entropy prediction for large fused cyclics has large uncertainties).

Table 5: Mean absolute error (cal/mol/K) of $S(298\text{ K})$ for each category of polycyclics

Method	small cyclics	large linear cyclics	large fused cyclics
Group Additivity Method	44.5	93.3	103.5
+ Similarity Match	3.6	36.7	47.6
+ Bicyclic Decomposition	3.6	4.9	11.9

6. Conclusion

The similarity match algorithm combined with the bicyclic decomposition model can estimate unknown ring corrections and can be applied to various kinds of polycyclics. By assuming bicyclic ring strain contributions are independent and additive, the proposed method is both interpretable and effective (mean absolute error of $H_f(298\text{ K})$: $3 \sim 5$ kcal/mol) for bicyclics and linear polycyclics. It also achieves moderate accuracy (mean absolute error of $H_f(298\text{ K})$: 10 kcal/mol) for large heavily fused cyclics using the test dataset [8]. The method also shows good performance in heat capacity and entropy predictions. To further improve prediction accuracy, one might consider adding more terms that describe interactions between decomposed bicyclics. In addition, this method also accurately estimates the thermochemistry of some heteroatomic polycyclics (tested on oxygen-embedded polycyclics).

This proposed method provides a quick and moderately accurate way of estimating thermochemistry of large unknown polycyclics where quantum mechanical calculation may be significantly more expensive. We recommend using it in applications where a large molecular space has to be scanned/explored with limited time such as during automatic mechanism generation, drug discovery, automatic transition state search, etc. We have made software implementing this new method freely available at http://rmg.mit.edu/molecule_search.

Acknowledgments

We gratefully acknowledge financial support from the DOE Gas Phase Chemical Physics program, grant No. DE-SC0014901. This research used resources of the National Energy Research Scientific (NERSC) Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Supporting Information Available

An example of querying a big cyclic molecule using similarity match algorithm and a collection of 190 computed thermochemistry of bicyclics are provided.

References

- [1] C. W. Gao, J. W. Allen, W. H. Green, R. H. West, [Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms](#), *Computer Physics Communications* 203 (2016) 212–225. doi:10.1016/j.cpc.2016.02.013.
URL <http://www.sciencedirect.com/science/article/pii/S0010465516300285>

- [2] E. Gawehn, J. A. Hiss, G. Schneider, [Deep Learning in Drug Discovery](#), *Molecular Informatics* 35 (1) (2016) 3–14. doi:10.1002/minf.201501008.
URL <http://onlinelibrary.wiley.com/doi/10.1002/minf.201501008/abstract>
- [3] Y. V. Suleimanov, W. H. Green, [Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods](#), *Journal of Chemical Theory and Computation* 11 (9) (2015) 4248–4259. doi:10.1021/acs.jctc.5b00407.
URL <http://dx.doi.org/10.1021/acs.jctc.5b00407>
- [4] S. W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Haugen, H. E. O'Neal, A. S. Rodgers, R. Shaw, R. Walsh, [Additivity Rules for the Estimation of Thermochemical Properties](#), *Chem. Rev.* 69 (1969) 279–324.
URL <http://kinetics.nist.gov/kinetics/Detail?id=1969BEN/CRU279-324:0>
- [5] E. R. Ritter, J. W. Bozzelli, [THERM: Thermodynamic property estimation for gas phase radicals and molecules](#), *International Journal of Chemical Kinetics* 23 (9) (1991) 767–778. doi:10.1002/kin.550230903.
URL <http://onlinelibrary.wiley.com/doi/10.1002/kin.550230903/abstract>
- [6] T. H. Lay, T. Yamada, P.-L. Tsai, J. W. Bozzelli, [Thermodynamic Parameters and Group Additivity Ring Corrections for Three- to Six-Membered Oxygen Heterocyclic Hydrocarbons](#), *The Journal of Physical Chemistry A* 101 (13) (1997) 2471–2477. doi:10.1021/jp9629497.
URL <http://dx.doi.org/10.1021/jp9629497>
- [7] B. T. Fan, A. Panaye, J. P. Doucet, A. Barbu, [Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table](#), *Journal of Chemical Information and Computer Sciences* 33 (5) (1993) 657–662. doi:10.1021/ci00015a002.
URL <http://dx.doi.org/10.1021/ci00015a002>
- [8] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. v. Lilienfeld, [Quantum chemistry structures and properties of 134 kilo molecules](#), *Scientific Data* 1 (2014) 140022. doi:10.1038/sdata.2014.22.
URL <http://www.nature.com/articles/sdata201422>
- [9] G. Landrum, et al., [Rdkit: Open-source cheminformatics](#).
URL <http://www.rdkit.org>
- [10] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Rob, J. R. Cheeseman, et al., *Gaussian 09 Revision C.01*, Gaussian Inc. Wallingford CT 2016.

- [11] Y. Zhao, D. G. Truhlar, [The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals](#), *Theoretical Chemistry Accounts* 120 (1) (2008) 215–241. doi:10.1007/s00214-007-0310-x.
URL <https://link.springer.com/article/10.1007/s00214-007-0310-x>
- [12] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, [Fast and accurate modeling of molecular atomization energies with machine learning](#), *Physical Review Letters* 108 (5) (2012) 058301. doi:10.1103/PhysRevLett.108.058301.
URL <http://link.aps.org/doi/10.1103/PhysRevLett.108.058301>
- [13] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, [Big data meets quantum chemistry approximations: The machine learning approach](#), *Journal of Chemical Theory and Computation* 11 (5) (2015) 2087–2096. doi:10.1021/acs.jctc.5b00099.
URL <http://dx.doi.org/10.1021/acs.jctc.5b00099>
- [14] A. Osmont, L. Catoire, I. Gkalp, [Physicochemical Properties and Thermochemistry of Propellanes](#), *Energy & Fuels* 22 (4) (2008) 2241–2257. doi:10.1021/ef8000423.
URL <http://dx.doi.org/10.1021/ef8000423>

Supporting Information for An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation

1. Computed Thermochemistry of Bicyclics

The 190 pre-calculated molecules are mostly bicyclics. A full collection can be found in two of the attached files: `precalculated_data.py`, which is already in RMG thermo library format, and `precalculated_data.pdf` in PDF format with molecular structures. The level theory employed is M06-2X/cc-pVTZ using RRHO partition functions.

2. Similarity Match Example

Here is an example where a big cyclic molecule (see Figure 2) that does not have exact match in tree (see Figure 1). The similarity match algorithm will still match an existing node in the tree for the molecule.

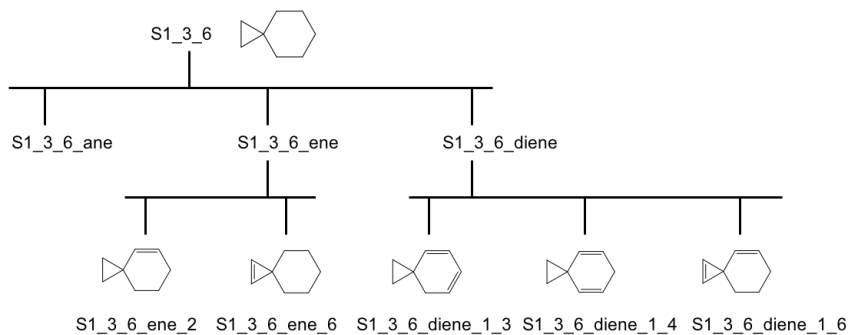


Figure 1: Example sub-tree that organizes polycyclic ring corrections

It uses sub-graph isomorphism check to find which nodes are contained by the big molecule and selects the correction of the first matched node. In this case, both `s1_3_6_ene.2` and `s1_3_6_diene.1_4` are contained in the example big molecule, but the correction of first match `s1_3_6_ene.2` will be applied. This explains necessity of bicyclic decomposition algorithm.

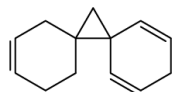



















































Figure 2: Example tricyclic that does not have exact match in the tree































Precalculated Molecule Structures






























This table below records 190 pre-calculated molecules (mostly bicyclics). The level theory employed is M06-2X/cc-pVTZ using RRHO partition functions. Note the values listed here are thermochemistry for molecules.




























Ring corrections derived from these row values and Benson group values used for the derivation are recorded in RMG-database, which is hosted on Github. Specifically, ring corrections are stored at <https://github.com/ReactionMechanismGenerator/RMG-database/blob/master/input/thermo/groups/polycyclic.py> and Benson group values are at <https://github.com/ReactionMechanismGenerator/RMG-database/blob/master/input/thermo/groups/group.py>.































Structure	Label	$\Delta_f H(298\text{K})$	$S(298\text{K})$	$C_p(300\text{K})$	$C_p(1000\text{K})$	$C_p(1500\text{K})$
	s1_3_3_ane	40.87	67.23	21.09	53.04	61.46
	s1_3_3_ene	95.61	67.64	20.20	45.94	52.32
	s1_3_4_ane	31.71	75.52	25.26	65.68	76.42
	s1_3_4_ene	64.25	73.35	23.59	58.31	67.12
	s1_3_5_ane	8.12	79.72	29.10	78.09	91.19
	s1_3_5_diene_1_3	55.75	74.45	25.61	63.37	72.54
	s1_3_5_ene_1	33.86	78.67	27.63	70.79	81.91
	s1_3_5_ene_2	33.86	78.68	27.63	70.79	81.91
	s1_3_6_ane	-3.92	82.94	33.38	90.63	106.03
	s1_3_6_diene_1_3	49.92	81.16	30.36	75.97	87.44
	s1_3_6_diene_1_4	48.56	79.91	30.29	75.88	87.40
	s1_3_6_ene_1	23.18	82.32	31.90	83.26	96.72
	s1_3_6_ene_2	23.18	82.32	31.90	83.26	96.72
	s1_4_4_ane	26.49	78.15	29.14	78.23	91.32
	s1_4_4_diene_1_5	91.71	75.80	26.19	63.58	72.80
	s1_4_4_ene_1	58.98	78.38	27.69	70.94	82.08
	s1_4_5_ane	3.81	85.56	33.58	90.71	106.14
	s1_4_5_diene_1_3	55.69	80.94	29.97	76.09	87.55

	s1_4_5_diene_1_6	62.88	81.79	30.22	76.05	87.58
	s1_4_5_diene_2_6	63.38	81.12	30.12	76.10	87.62
	s1_4_5_ene_1	30.23	82.76	31.71	83.42	96.87
	s1_4_5_ene_2	30.53	82.26	31.52	83.42	96.88
	s1_4_5_ene_6	36.82	84.58	32.08	83.39	96.85
	s1_4_6_ane	-7.63	87.69	37.76	103.35	121.02
	s1_4_6_diene_1_3	47.30	85.37	34.63	88.63	102.25
	s1_4_6_diene_1_4	46.94	86.59	34.75	88.62	102.29
	s1_4_6_diene_1_7	53.01	85.77	34.70	88.62	102.32
	s1_4_6_diene_2_7	52.80	85.32	34.63	88.63	102.32
	s1_4_6_ene_1	20.39	86.92	36.24	95.95	111.69
	s1_4_6_ene_2	20.38	86.97	36.23	95.94	111.69
	s1_4_6_ene_7	25.39	86.30	36.09	95.90	111.69
	s1_5_5_ane	-17.99	88.86	37.57	103.18	120.97
	s1_5_5_diene_1_3	33.47	86.09	34.21	88.50	102.32
	s1_5_5_diene_1_6	35.22	86.50	34.33	88.56	102.38
	s1_5_5_diene_1_7	35.17	85.55	34.13	88.56	102.40
	s1_5_5_diene_2_7	206.39	88.46	37.12	87.89	100.95
	s1_5_5_ene_1	8.54	89.98	35.82	95.88	111.66
	s1_5_5_ene_2	8.10	87.37	35.88	95.91	111.69
	s1_5_6_ane	-28.08	92.09	41.62	115.67	135.76
	s1_5_6_diene_1_3	25.34	90.33	38.84	101.06	117.06
	s1_5_6_diene_1_4	26.21	92.85	38.81	101.01	117.07
	s1_5_6_diene_1_7	25.76	90.07	38.77	101.06	117.09
	s1_5_6_diene_1_8	25.68	89.64	38.71	101.13	117.12
	s1_5_6_diene_2_7	24.78	89.46	38.69	101.09	117.10
	s1_5_6_diene_2_8	25.41	89.34	38.56	101.09	117.11
	s1_5_6_diene_7_9	23.54	88.66	38.29	100.99	117.12
	s1_5_6_ene_1	-0.61	93.01	40.37	108.41	126.42
	s1_5_6_ene_2	-1.37	91.03	40.28	108.34	126.38
	s1_5_6_ene_7	-2.01	90.65	40.26	108.43	126.47

	s1_5_6_ene_8	-2.01	90.64	40.26	108.43	126.47
	s1_6_6_ane	-37.61	94.71	46.05	128.31	150.63
	s1_6_6_diene_1_3	15.68	93.57	43.27	113.71	131.96
	s1_6_6_diene_1_4	16.45	94.96	43.15	113.52	131.88
	s1_6_6_diene_1_7	17.41	93.09	43.21	113.55	131.90
	s1_6_6_diene_1_8	17.05	94.02	43.21	113.61	131.92
	s1_6_6_diene_2_8	15.41	91.61	43.12	113.65	131.96
	s1_6_6_ene_1	-9.99	94.70	44.61	120.89	141.21
	s1_6_6_ene_2	-10.87	93.89	44.50	120.90	141.24
	s2_3_3_ane	49.72	61.87	15.50	40.24	46.60
	s2_3_3_ene	113.84	62.54	14.33	32.88	37.40
	s2_3_4_ane	37.61	66.26	19.20	52.73	61.40
	s2_3_4_ene_1	79.26	66.23	17.88	45.58	52.21
	s2_3_5_ane	13.73	73.13	23.58	65.33	76.29
	s2_3_5_ene_1	38.45	70.48	21.72	57.99	66.97
	s2_3_6_ane	8.53	79.47	28.42	77.88	91.12
	s2_3_6_ben	88.52	73.33	22.72	55.40	63.14
	s2_3_6_diene_1_3	50.99	74.00	24.63	63.08	72.45
	s2_3_6_ene_1	29.65	75.58	26.35	70.42	81.75
	s2_3_6_ene_2	29.87	75.52	26.60	70.46	81.82
	s2_4_4_ane	34.40	72.46	23.74	65.51	76.44
	s2_4_4_ene_1	30.00	70.98	22.48	57.70	66.89
	s2_4_5_ane	9.44	76.07	27.63	77.91	91.20
	s2_4_5_diene_0_3	73.79	75.63	25.29	63.34	72.69
	s2_4_5_diene_4_6	119.13	74.98	24.40	61.45	70.74
	s2_4_5_ene_1	34.05	75.66	25.88	70.53	81.89
	s2_4_6_ane	0.56	81.37	32.09	90.41	106.03
	s2_4_6_ben	50.11	77.61	26.83	68.03	78.02
	s2_4_6_diene_1_3	51.90	79.35	28.92	75.70	87.41
	s2_4_6_diene_1_6	58.62	79.41	29.05	75.73	87.46

	s2_4_6_diene_2_6	59.96	80.39	29.41	75.83	87.53
	s2_4_6_diene_5_7	97.66	81.98	30.89	76.06	87.51
	s2_4_6_ene_1	25.59	80.44	30.50	83.03	96.71
	s2_4_6_ene_2	29.14	82.16	30.60	83.04	96.74
	s2_4_6_ene_6	47.27	79.70	30.39	83.00	96.72
	s2_5_5_ane	-14.47	79.90	31.54	90.31	105.96
	s2_5_5_diene_0_2	40.26	79.67	28.81	75.72	87.45
	s2_5_5_diene_0_3	36.35	79.52	29.22	75.74	87.45
	s2_5_5_diene_0_4	37.29	81.85	29.47	75.53	87.37
	s2_5_5_diene_0_5	42.58	79.04	28.72	75.64	87.43
	s2_5_5_diene_0_6	39.38	78.91	28.62	75.62	87.42
	s2_5_5_diene_1_5	37.56	79.21	28.38	75.64	87.41
	s2_5_5_diene_1_6	38.78	79.53	28.49	75.70	87.45
	s2_5_5_diene_m_2	37.66	80.30	29.23	75.59	87.41
	s2_5_5_ene_0	14.24	81.61	30.65	82.99	96.73
	s2_5_5_ene_1	11.24	81.29	30.15	83.04	96.74
	s2_5_5_ene_m	12.38	81.49	31.07	82.79	96.67
	s2_5_5_tetraene_0_2_4_6	96.36	76.11	25.38	60.81	68.72
	s2_5_6_ane	-23.45	87.64	36.72	102.94	120.89
	s2_5_6_ben	18.42	81.95	30.95	80.50	92.82
	s2_5_6_diene_0_2	27.17	84.14	33.47	88.20	102.27
	s2_5_6_diene_0_3	29.41	85.93	33.67	88.16	102.25
	s2_5_6_diene_0_4	26.86	84.97	33.95	88.17	102.25
	s2_5_6_diene_0_5	23.69	85.82	33.83	88.07	102.20
	s2_5_6_diene_0_6	29.83	84.03	33.39	88.17	102.27
	s2_5_6_diene_0_7	26.08	84.00	33.33	88.15	102.25
	s2_5_6_diene_1_3	29.95	86.71	33.21	88.15	102.21
	s2_5_6_diene_1_5	32.37	85.27	33.77	88.25	102.30
	s2_5_6_diene_1_6	30.28	83.85	33.17	88.21	102.27

	s2_5_6_diene_1_7	31.10	83.86	33.11	88.20	102.26
	s2_5_6_diene_2_5	29.36	86.20	33.76	88.18	102.28
	s2_5_6_diene_2_6	32.24	83.19	33.10	88.24	102.31
	s2_5_6_diene_5_7	25.30	83.49	33.19	88.22	102.26
	s2_5_6_diene_5_8	24.67	83.45	33.84	88.28	102.31
	s2_5_6_diene_m_1	26.13	85.93	34.06	88.15	102.21
	s2_5_6_diene_m_2	25.06	86.21	33.94	88.03	102.18
	s2_5_6_diene_m_7	23.24	84.68	33.83	88.24	102.31
	s2_5_6_ene_0	1.75	86.31	35.29	95.54	111.57
	s2_5_6_ene_1	4.36	85.92	34.83	95.49	111.52
	s2_5_6_ene_2	6.22	85.10	34.74	95.54	111.53
	s2_5_6_ene_5	6.01	86.68	35.45	95.54	111.55
	s2_5_6_ene_6	6.61	85.77	34.95	95.59	111.58
	s2_5_6_ene_m	-1.05	87.16	35.58	95.38	111.56
	s2_6_6_ane	-35.51	89.21	41.18	115.60	135.82
	s2_6_6_ben	10.52	87.14	35.56	93.04	107.67
	s2_6_6_ben_ene_1	34.45	85.01	33.78	85.66	98.34
	s2_6_6_diene_0_2	21.86	89.50	38.24	100.73	117.00
	s2_6_6_diene_0_3	17.87	89.18	37.97	100.70	117.10
	s2_6_6_diene_0_4	18.90	88.14	38.38	100.72	117.02
	s2_6_6_diene_0_5	14.26	87.86	38.14	100.62	116.97
	s2_6_6_diene_0_6	72.90	88.59	38.55	101.21	117.21
	s2_6_6_diene_0_7	20.57	89.64	38.16	100.69	117.02
	s2_6_6_diene_0_8	17.87	88.86	37.95	100.67	117.07
	s2_6_6_diene_1_3	19.90	86.51	37.77	100.83	117.16
	s2_6_6_diene_1_6	20.92	87.02	37.80	100.73	117.09
	s2_6_6_diene_1_7	129.14	86.94	38.44	101.49	117.44

	s2_6_6_diene_1_8	21.64	88.10	37.74	100.76	117.11
	s2_6_6_diene_2_7	20.67	87.98	37.79	100.79	117.13
	s2_6_6_diene_m_1	16.93	90.17	38.45	100.65	117.03
	s2_6_6_diene_m_2	16.49	89.51	38.44	100.59	117.05
	s2_6_6_ene_0	-8.08	90.03	39.58	108.04	126.38
	s2_6_6_ene_1	-0.62	89.55	39.37	108.12	126.39
	s2_6_6_ene_2	-7.55	88.77	39.71	108.20	126.48
	s2_6_6_ene_m	-9.82	88.32	39.87	107.91	126.32
	s3_4_4_ane	49.94	62.34	17.96	52.71	61.48
	s3_4_4_diene_0_2	152.09	63.37	16.06	38.12	42.95
	s3_4_4_ene_0	112.10	66.61	18.17	45.60	52.28
	s3_4_5_ane	18.81	68.86	22.23	65.27	76.30
	s3_4_5_diene_0_2	118.33	67.48	20.06	50.66	57.77
	s3_4_5_diene_0_3	127.78	73.35	22.67	51.04	57.85
	s3_4_5_diene_1_3	125.49	69.87	20.62	50.94	57.91
	s3_4_5_diene_3_4	145.71	69.18	21.10	50.90	57.89
	s3_4_5_ene_0	130.15	70.23	22.32	58.53	67.19
	s3_4_5_ene_1	60.44	67.42	20.81	58.08	67.05
	s3_4_5_ene_3	99.29	73.46	23.12	58.44	67.27
	s3_4_6_ane	9.37	75.59	26.99	77.75	91.13
	s3_4_6_diene_0_2	115.97	72.72	24.41	63.48	72.74
	s3_4_6_diene_0_3	82.15	73.20	24.32	63.26	72.47
	s3_4_6_diene_0_4	93.02	77.39	26.87	63.54	72.57
	s3_4_6_diene_1_4	110.24	72.61	24.23	63.42	72.72
	s3_4_6_diene_1_5	116.53	73.06	24.42	63.41	72.74
	s3_4_6_ene_0	80.33	73.56	25.91	70.88	81.86
	s3_4_6_ene_1	35.62	73.59	25.21	70.43	81.83
	s3_4_6_ene_4	80.56	74.66	26.02	70.67	81.96
	s3_5_5_ane	-6.59	73.65	26.40	77.69	91.07
	s3_5_5_diene_1_4	64.50	70.04	23.44	63.30	72.57
	s3_5_5_ene_1	26.24	72.98	24.81	70.45	81.79

























	s3_5_6_ane	-17.27	79.25	31.03	90.24	105.94
	s3_5_6_diene_1_5	42.66	76.43	28.15	75.81	87.27
	s3_5_6_ene_1	10.51	78.43	29.42	82.84	96.61
	s3_5_6_ene_5	14.24	77.61	29.35	82.89	96.62
	s3_6_6_ane	-22.57	82.55	35.34	102.51	120.64
	s3_6_6_diene_0_2	49.78	81.32	32.60	88.17	102.22
	s3_6_6_diene_0_3	65.54	80.99	32.94	88.37	102.27
	s3_6_6_diene_0_4	60.78	80.07	33.13	88.35	102.26
	s3_6_6_diene_0_5	43.85	81.27	32.55	88.19	102.22
	s3_6_6_diene_0_6	110.77	80.51	32.82	88.32	102.20
	s3_6_6_diene_0_m	125.35	81.56	33.56	88.78	102.52
	s3_6_6_diene_1_5	26.80	80.01	32.33	87.99	102.15
	s3_6_6_diene_1_6	30.44	83.28	32.60	87.98	102.13
	s3_6_6_diene_1_8	77.43	81.93	33.06	88.36	102.32
	s3_6_6_diene_1_m	78.57	81.61	33.00	88.42	102.36
	s3_6_6_ene_0	54.35	82.41	34.43	95.71	111.60
	s3_6_6_ene_1	0.10	82.73	34.04	95.44	111.51
	s3_6_6_ene_4	95.55	86.73	37.12	96.46	111.86
	s3_6_7_ane	-23.91	88.98	40.63	115.40	135.68
	s3_6_7_diene_6_9-0	118.04	86.85	38.44	101.29	117.32
	s3_6_7_ene_6	28.52	86.49	38.49	107.94	126.24
	s4_6_8_ane	26.56	91.69	42.02	115.68	135.62
	s4_6_8_diene_7_9	41.54	86.07	37.66	100.86	117.12
	s4_6_8_ene_7	15.63	88.48	39.52	108.21	126.43

Table 1: Pre-calculated thermo-properties for bicyclic hydrocarbon compounds