UNIVERSITY OF LEEDS

CYPHER SUMMER SCHOOL

Analysis, uncertainty quantification, validation, optimization and reduction of combustion kinetics mechanisms

Professor Alison Tomlin
University of Leeds
School of Chemical
and Process Engineering

1-3
CONSTRUCTION OF DETAILED REACTION MECHANISMS

Constructing Chemical Mechanisms - Manual

- Historically mechanisms result from careful development work by experts.
- Begins with the selection of important species:
  - **reactants and products**
  - important **intermediates** necessary to predict production rates of key products or key quantities such as ignition delays, flame speeds or dynamic features such as extinction and oscillations.
- **Types of reactions** that can occur between these coupled groups of species then need to be specified along with appropriate descriptions of rate coefficients, and thermodyamic data.
- Growing expertise led to **protocols** for different types of application.
  - Indicate ***reaction classes*** for each category of important species.
- Typically, certain reaction classes ignored if
  - rates **very slow** compared to overall time-scales of interest,
  - they are **too endothermic or too complex** (e.g. too many bonds are broken or products produced (Yoneda, 1979; Németh et al., 2002)),
  - pathways to **minor products** also often ignored (Saunders et al., 2003a).

The concept of reaction classes

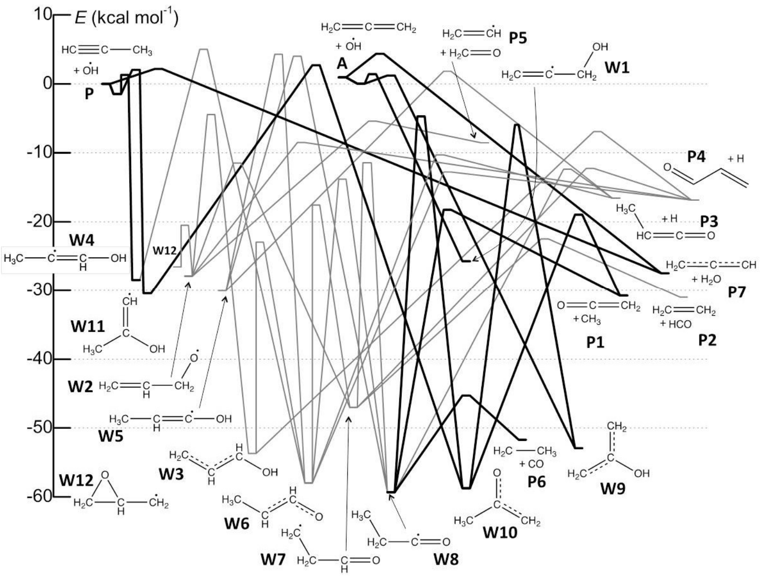
(Blurock, Battin-LeClerc, 2013)

- Developing detailed combustion mechanisms for oxidation of fuels/atmospheric species with a large number of C atoms presents challenges in mechanism production philosophy.
  - Not possible to source parameters exclusively from experiments/detailed theory.
  - Estimates of reaction rate constants/thermo must come from physical/chemical principles based on fundamental kinetic studies for a smaller number of fuels.
- One way of encompassing general principles into specific reactive properties is to define ***reaction classes***.
- ***Reaction classes*** - kinetic generalisations that systematically embody analogies and physical principles a modeller uses to estimate rate constants where no specific evidence exists.
- Based on a local set of functional features around the reactive centre of molecules that are significant when determining numeric value of rate constant.
- Used in both ***automatic*** and ***manual generation*** of reaction mechanisms e.g. n-hexadecane mechanism of Westbrook et al. (2009).

How to define reaction classes

- A reaction class has three sets of information:
  1. A **pattern or rule** to recognise within the chemical reactants (can be more than one) when the reaction class should be applied.
  2. A **transformation** of how the specific reactants are converted to products.
  3. The **rate coefficients** associated with the transformation.
- Generally built from years of chemical experience and intuition.
- May also be suggested by automatic computer codes designed to explore chemical pathways automatically for reactions that are relevant in gas phase chemical problems e.g. KinBot (Zador & Van De Vijver https://www.osti.gov/biblio/1464498-kinbot).

- KinBot uses a chemical network approach coupled with knowledge of the **potential energy surface** determined for the particular system.



Examples of high temperature reaction classes

(Sarathy et al., 2011, Curran et al., 1998)

1. Unimolecular fuel decomposition

Fuel molecule breaks apart

2. H-atom abstraction from the fuel

Something pulls off a H atom leaving alkyl radical

3. Alkyl radical decomposition
4. Alkyl radical isomerization

Alkyl radical either breaks apart or internally reorganises to new structure etc. etc.

5. H-atom abstraction reactions from alkenes
6. Addition of radical species O and OH to alkenes
7. Reactions of alkenyl radicals with HO_2, CH_3O_2, and $C_2H_5O_2$
8. Alkenyl radical decomposition
9. Alkene decomposition
10. Retroene decomposition reactions

Examples of Low Temperature Reaction Classes

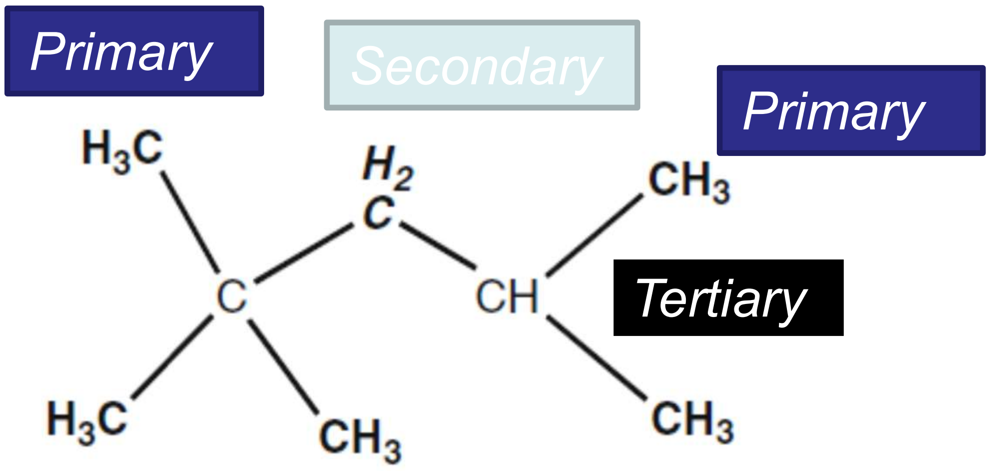
(e.g. RH any general alkane, Sarathy et al. 2011)

11. **Addition of O_2 to alkyl radicals ($R + O_2 = ROO$)**
12. $R + ROO = RO + RO$
13. $R + HO_2 = RO + OH$
14. $R + CH_3OO = RO + CH_3O$
15. **Alkyl peroxy radical isomerization ($ROO = QOOH$)**
16. Concerted eliminations (ROO = alkene + HO_2)
17. $ROO + HO_2 = ROOH + O_2$
18. $ROO + H_2O_2 = ROOH + HO_2$
19. $ROO + CH_3O_2 = RO + CH_3O + O_2$
20. $ROO + ROO = RO + RO + O_2$
21. $ROOH = RO + OH$
22. RO decomposition.
23. QOOH = cyclic ether + OH (cyclic ether formation)
24. QOOH = alkene + HO_2 (radical site beta to OOH group)
25. QOOH = alkene + carbonyl + OH (radical site gamma to OOH group)
26. **Addition of O_2 to QOOH ($QOOH + O_2 = OOQOOH$)**
27. **Isomerization of OOQOOH and formation of ketohydroperoxide and OH**
28. **Decomposition of ketohydroperoxide to form oxygenated radical species and OH**
29. Cyclic ether reactions with OH and HO_2
30. Decomposition of large carbonyl species and carbonyl radicals

Typical low temperature chain branching route for alkanes

Rate constants and functional groups

- Every chemical environment, meaning an atom and its bonding, has an effect on the neighbouring atoms and bonds.
- For example, a radical on a carbon atom is more energetically stable on a **tertiary** carbon atom than on a **primary** carbon atom which has the consequence that a **tertiary** hydrogen atom is more easily extracted from the carbon atom.



- Mechanisms for larger fuels can be built using this **concept of reaction classes** and populated by data based partly on experimental measurements or detailed theory calculations and partly on **extrapolations** of this data to larger and larger molecules using the concept of **functional groups**.

Use of functional groups: example of hydrogen atom abstraction from the fuel

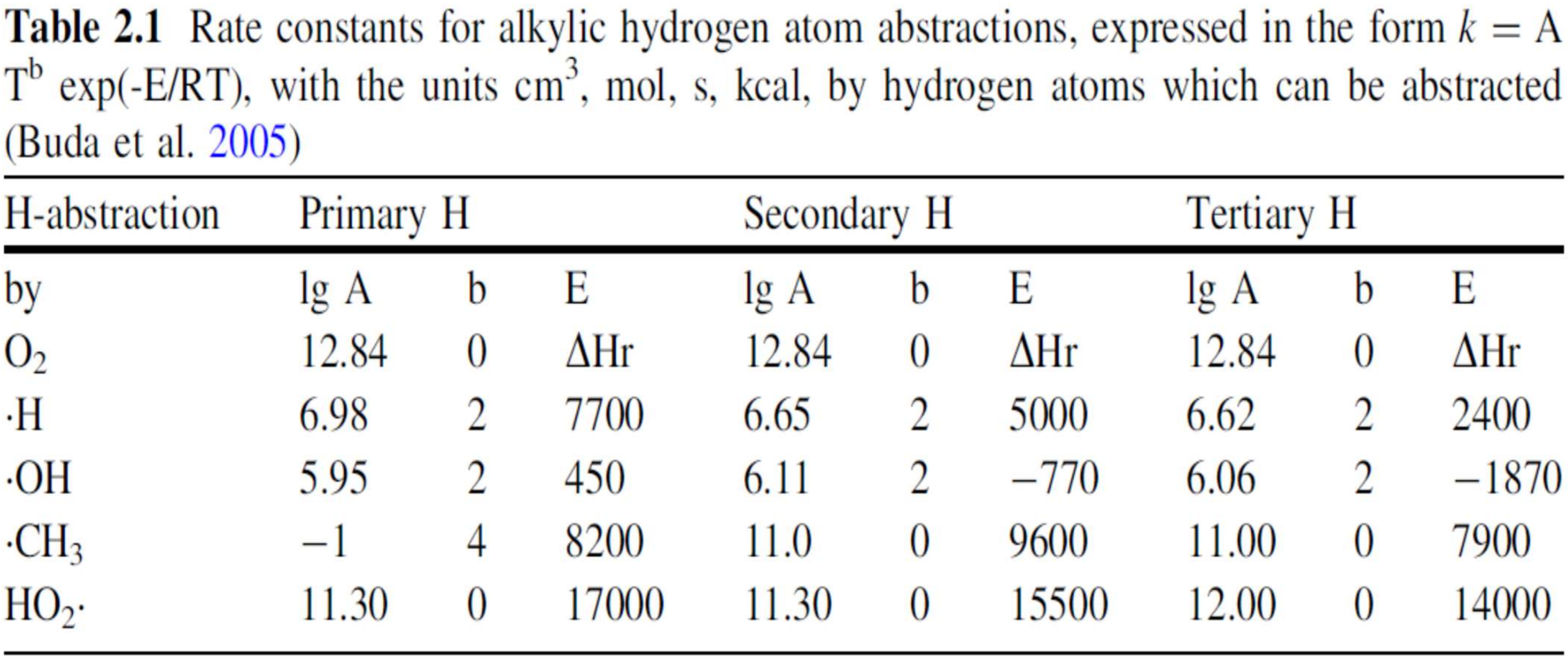


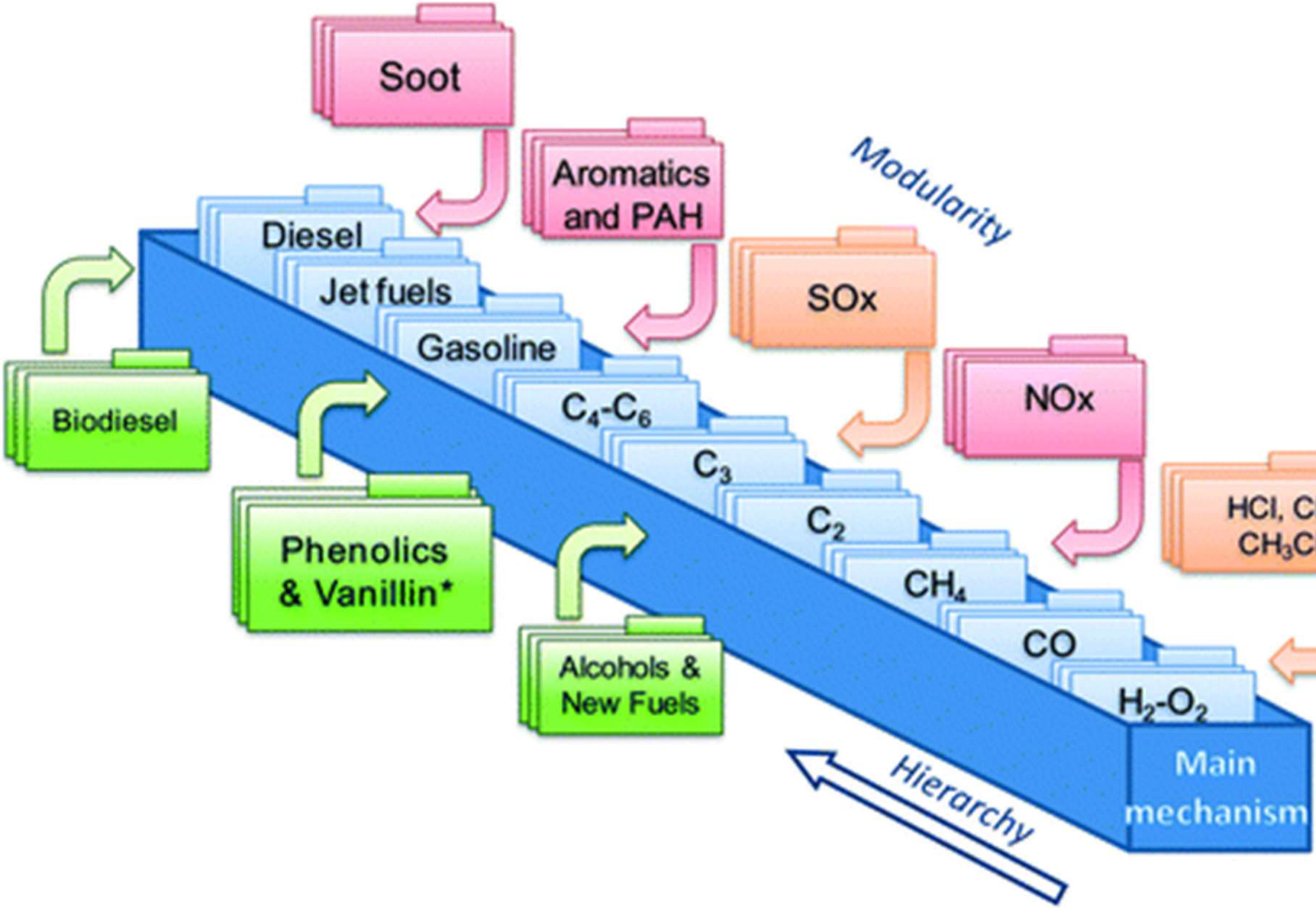
Table 2.1 Rate constants for alkylic hydrogen atom abstractions, expressed in the form k = A T^b exp(-E/RT), with the units cm^3, mol, s, kcal, by hydrogen atoms which can be abstracted (Buda et al. 2005)

H-abstraction	Primary H			Secondary H			Tertiary H		
by	lg A	b	E	lg A	b	E	lg A	b	E
O_2	12.84	0	ΔHr	12.84	0	ΔHr	12.84	0	ΔHr
·H	6.98	2	7700	6.65	2	5000	6.62	2	2400
·OH	5.95	2	450	6.11	2	−770	6.06	2	−1870
$\cdot CH_3$	−1	4	8200	11.0	0	9600	11.00	0	7900
$HO_2\cdot$	11.30	0	17000	11.30	0	15500	12.00	0	14000

The structure of reaction mechanisms

- Reaction classes can vary with temperature and hence size of required mechanism can be reduced by, e.g., restricting to low *T* classes (e.g. ignition problems), or high *T* mechanisms (e.g. flame propagation).
- Additional classification of sub-mechanisms can be based on:
- ***Hierarchical sub-mechanisms*** based on size of reactants: within a given sub-mechanism, only species of a given size are consumed. Smaller products (produced but not consumed within this sub-mechanism) are consumed by sub-mechanisms 'lower' in the hierarchy.
- **Primary**, ***secondary***, and ***base*** mechanisms: a special case of the hierarchical structure.
  - The **primary mechanism** - reactions of initial reactants and directly derived radicals.
  - The **secondary mechanism** – consumes products of primary mechanism. It would be possible to define iteratively tertiary and even n-ary mechanisms, but in practice in most combustion models, secondary mechanisms are designed to lead to intermediate species, which are finally consumed in a **base mechanism**.
- **Pathways**: A chain of reactions or reaction classes. The remaining species at the end of this chain should be consumed by other sub-mechanisms.

Hierarchical development of mechanisms



Pelucchi , 2019

The base mechanism

- Usually, a well-validated detailed mechanism of smaller species (e.g. up to C2-C4), which includes reactions taken from databases.
- Has usually been validated under the conditions being considered.
- **Estimated rate constants are not usually used** within base mechanisms, rather data is obtained from ***measurements***, ***theory calcs***, ***evaluations*** or even from ***optimised mechanisms***.
- Likely to be known with **lower uncertainty** than the reaction pathways for the larger hydrocarbons.
- Needs to be updated frequently but often in larger mechanisms "legacy" mechanisms may still be present.
- Care needs to be taken when updating base mechanisms within larger schemes since other reaction steps may have been "tuned" based on the existing base scheme.
- Example:
  - Aramco mechanism (2.0 http://www.nuigalway.ie/c3/aramco2/frontmatter.html)

AramcoMech2.0

- A C1-C4 mechanism that has been developed in a hierarchical way ’from the bottom up’
  - starting with a H_2/O_2 sub-mechanism,
  - followed by a C1 sub-mechanism
  - grown to include larger carbon species such as ethane, ethylene, acetylene, allene, propyne, propene, n-butane, isobutane, isobutene, 1-butene and 2-butene, and oxygenated species including formaldehyde, acetaldehyde, methanol, ethanol, and dimethyl ether.
- Has been **validated against a large array of experimental** measurements including data from shock tubes, rapid compression machines, flames, jet-stirred and plug-flow reactors.

Primary and secondary mechanisms

- **Primary mechanism** represents reactions of the primary fuels and their derived radicals.
  - Usually kept in detail.
- **Secondary mechanism** consumes the products of the primary mechanism forming smaller species.
- In secondary mechanisms often simplifications are made even at the generation stage to keep the number of reactions as low as possible:
  - **Vertical reaction lumping** is applied so that reactants go directly to smaller products via one reaction step without passing through intermediates (*see later for methodology*).
  - **Species lumping** where parallel pathways of similar isomers are grouped (*see later*).
  - Reaction classes of **low importance** can be removed.

Automatic Reaction Generation Methods

- Several reasons why this is important for mechanisms describing the oxidation of larger and more complex fuels:
  - simply too large a task for a single human
  - humans make mistakes
  - the production of larger mechanisms has to be careful and systematic to generate what could be mechanisms with thousands of species and reactions
  - data for individual reactions is unlikely to be obtained from experiment/evaluation. Estimations based on Reaction Class rules will be required.

- **Why not use the help of a computer informed by decades of human knowledge?**

Principles of Automatic Generators

- Expert systems using a database of chemical principles to systematically and efficiently produce large detailed mechanisms (Blurock et al., 2013, Cleaner Combustion, p59-92).
- The developer or modeller determines which sub-mechanisms and reaction classes should be generated.
- Therefore **expert system** based on **similar rules and reaction classes** discussed earlier but these are now encoded rather than applied by hand.
- Reduces errors and apply rules in a systematic way.
- If rate constants are changed for a whole class then should be easier to regenerate the mechanism.
  - EXGAS – Developed at CNRS Nancy (Côme et al., 1996).
  - RMG – Developed at MIT (Green et al., 2001; Van Geem et al., 2006).
  - REACTION – Developed by Ned Blurock (Blurock,1995; Moreac et al., 2006).
  - MAMOX++ – Developed by Milan (Ranzi et al., 1995).

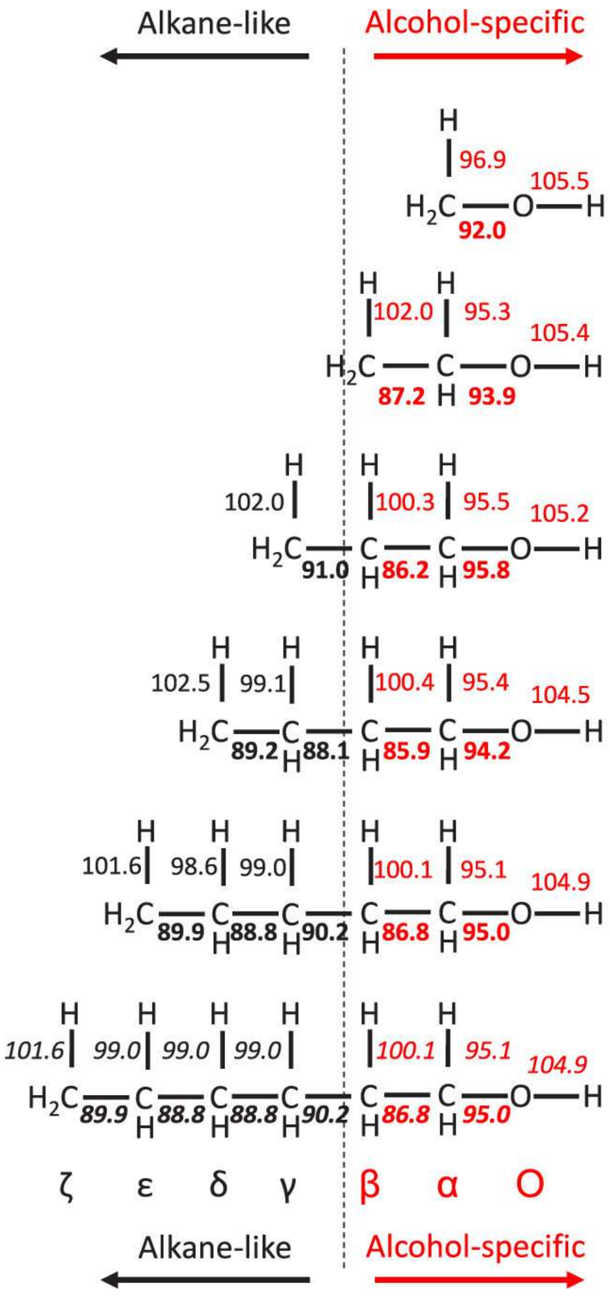
Different AMG codes and specificities

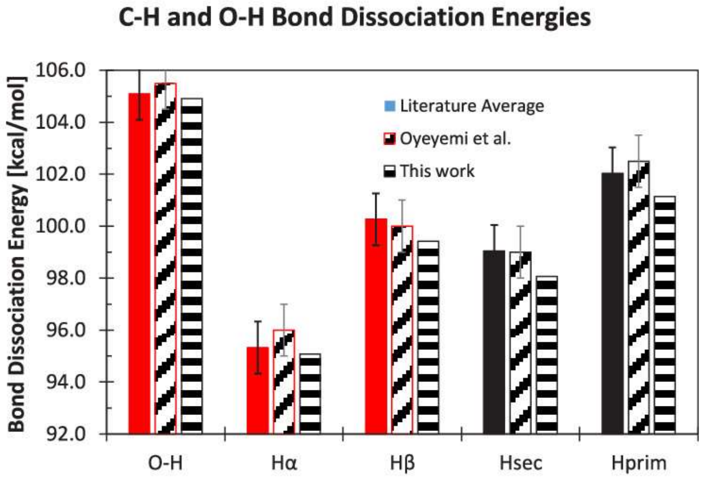
- MAMOX ++
  - Produces hierarchy of (highly) **lumped mechan**isms derived numerically from automatically generated detailed mechanisms.
- EXGAS
  - Has comprehensive reaction class database and large choice given to user for **mechanism tailoring**: e.g. low *T* vs high *T*, degree of lumping used etc.
- RMG
  - Uses a unique ''**generate and test**'' algorithm which generates a fundamental mechanistic step, estimates rate constants and then uses an ''on-the-fly'' reduction processes to determine whether the reaction should be included in the final mechanism.
  - Publicly distributed automatic generator of pressure-dependent reaction networks.
- REACTION
  - Uses concept of Reaction Pathways rather than exhaustive list of Reaction Classes.
  - Fundamental chemical information solely based on external databases so that it can be updated without modifying or recompiling the software.

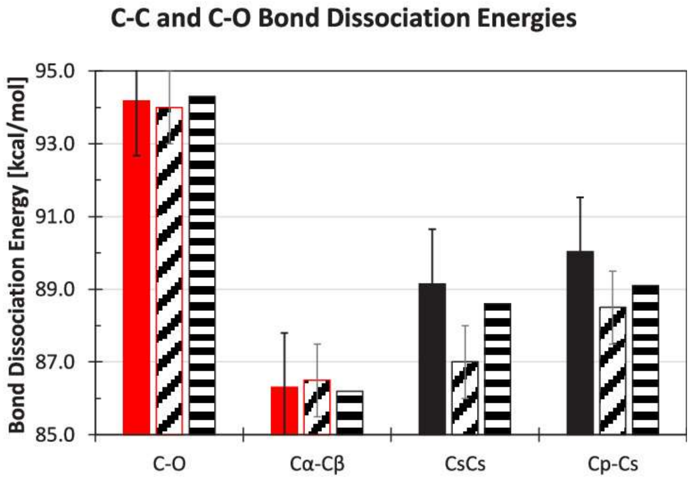
Particular Challenges Posed by Biofuels

- **AMG codes initially based on alkanes.**
- Wide range of biofuels now being used for applications in vehicles e.g. as additives or in blends with gasoline and diesel.
- Most common examples include:
  - Alcohols e.g. ethanol, butanol isomers, methanol
  - Methyl Esters e.g. in biodiesel, furans, etc.
- **Molecules contain oxygen** and have **different functional groups and bond energies** compared to e.g. alkanes.
- Modifications need to be made in terms of
  - Reaction classes
  - Relevant rate data for existing classes compared to alkanes, alkenes
  - Species present, groups included for group additivity calculations.
- The existence of measured data for the reactions of such compounds is pretty SCARCE!

Bond energies for alcohols (Pelucchi, 2020)



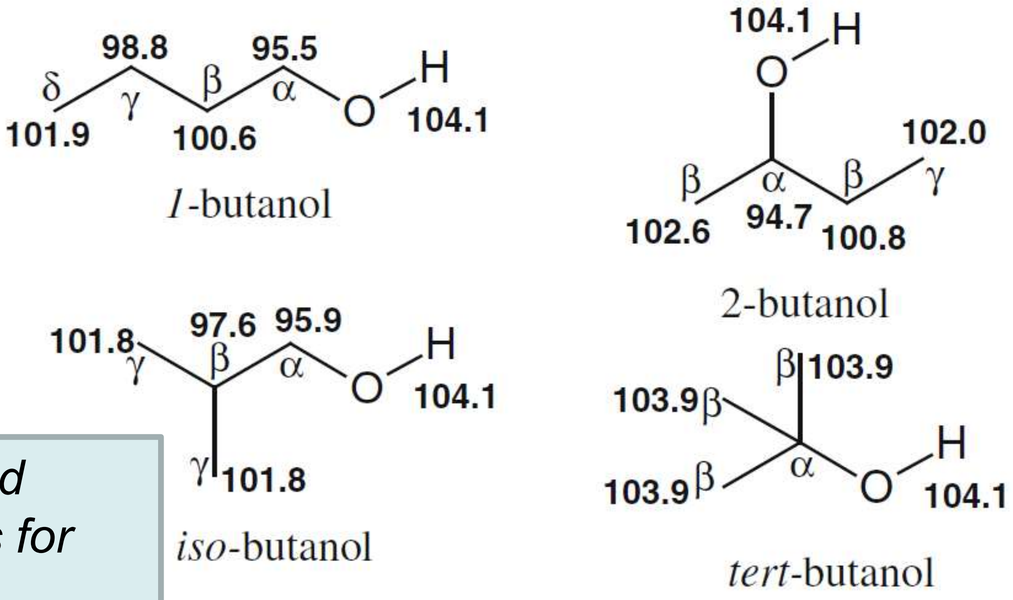




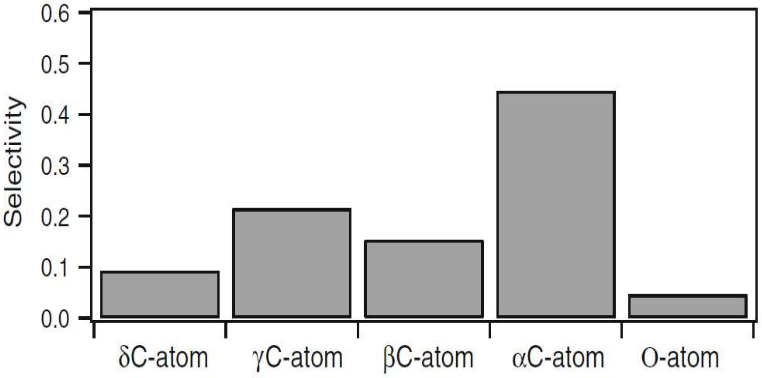
- Derived from theory and ATChTs.
- Will affect H abstraction rates at different *T*s, and therefore low *T* pathways.

Example of H abstraction Reactions

- We saw for alkanes that H abstraction rates were determined based on whether the H was attached to a primary, secondary or tertiary carbon atom.
- For oxygenated species there are more types of H atom.



C-H bond energies for butanol isomers

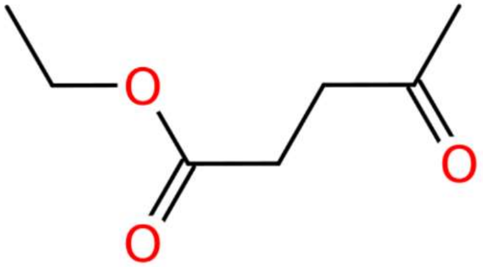


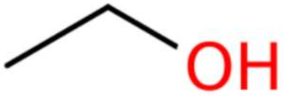
Selectivities for abstraction by OH
(Frassoldati et al., 2012)

RMG Case Study: Advanced Oxygenated Biofuels

Christian Michelbach

- We want to predict the combustion behaviour of advanced biofuel blends composed of ethyl levulinate, ethanol, and diethyl ether.
  - There is an interest in such fuels for use in SI and CI engines as they can be made from lignocellulosic biomass (2nd gen).
  - This requires predictions that cover a wide range of temperature, pressure, stoichiometry, and fuel blending conditions.

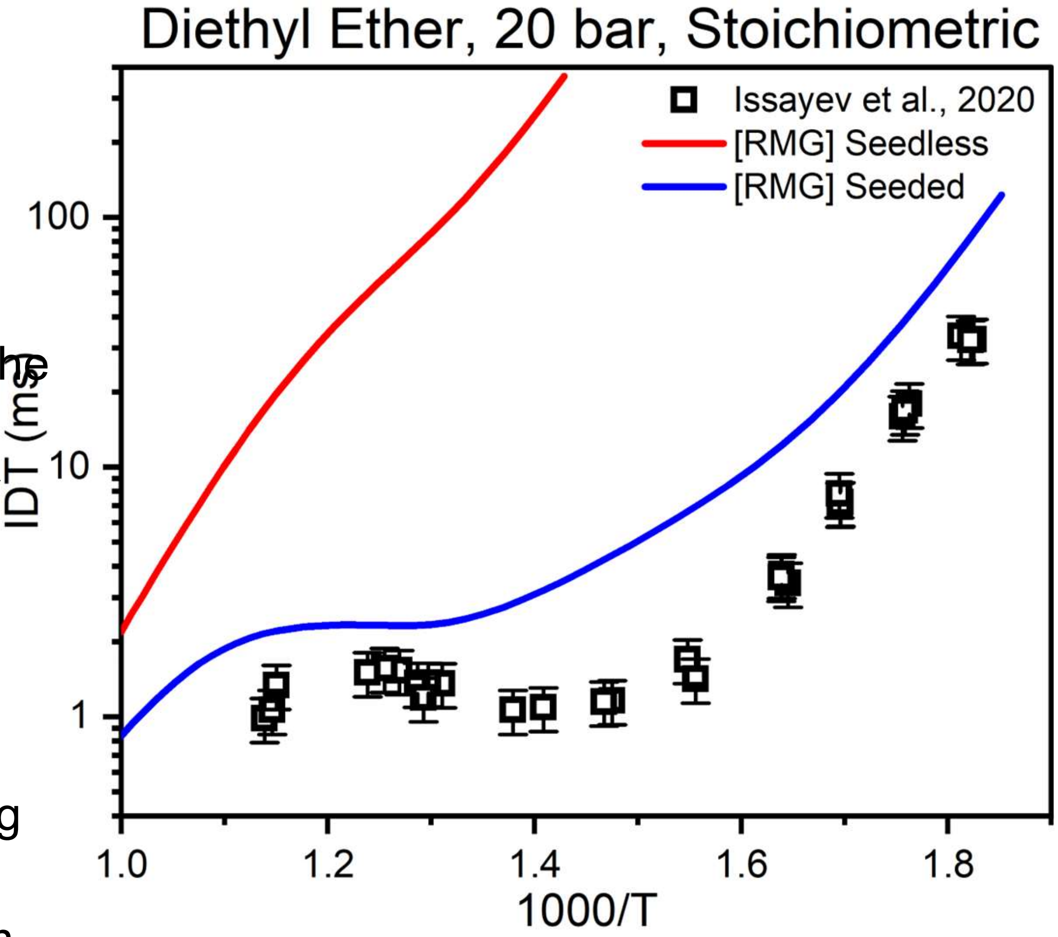




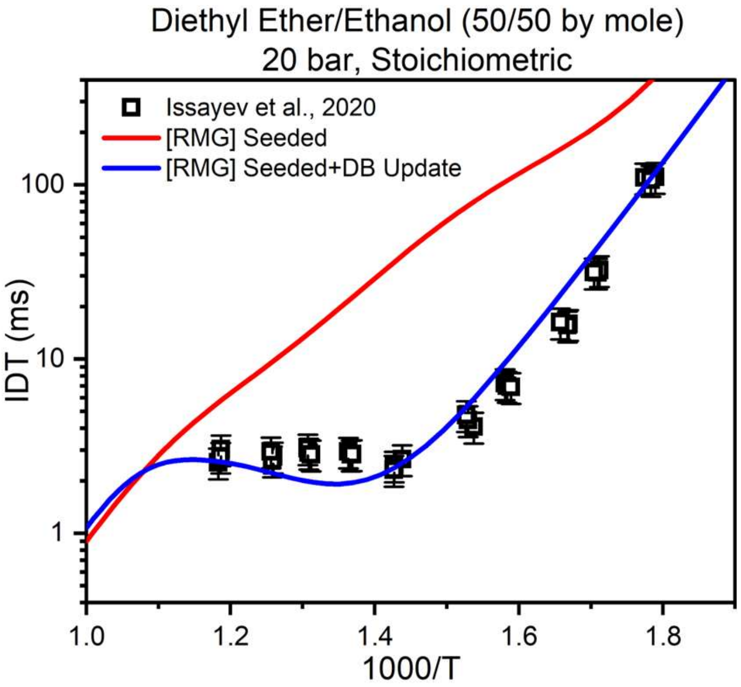
- Using RMG, can we produce a detailed kinetic mechanism that suits our needs?
- Can we then extend the methodology to a butanolic version where less data is available.

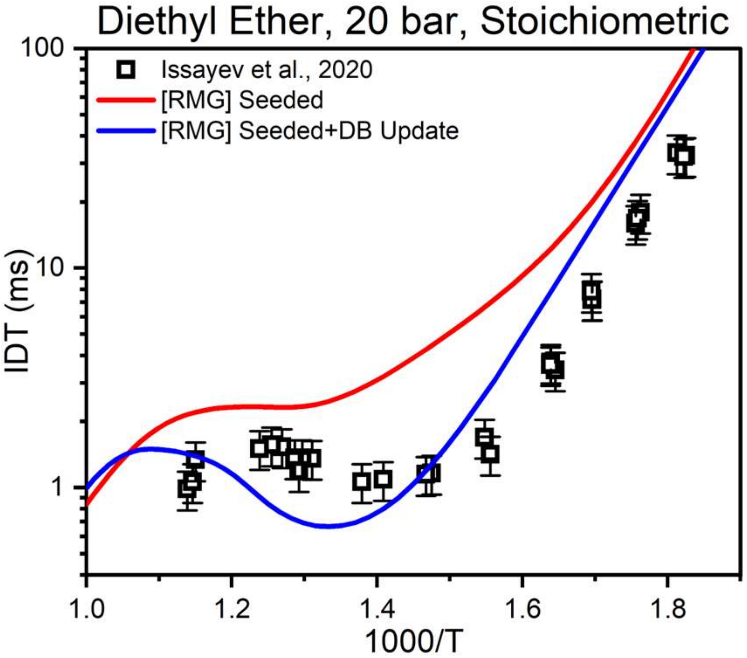
RMG Case Study: Seeding

- Building RMG models is an iterative process. Requires user to analyse produced model, making incremental improvements with each step.
- An initial model generated, using only RMG database reaction families and training reactions.
  - Model clearly insufficient, as shown by the IDTs of diethyl ether.
  - Holes in the current kinetic database for oxygenated species.
- Introducing seed mechanism provides key reaction steps (e.g. initiation and chain branching).
  - RMG typically good at filling in remaining propagation and termination steps.
  - Including a diethyl ether sub-mechanism (Tran et al., 2019) greatly improves IDT predictions.
  - Still room for improvement.



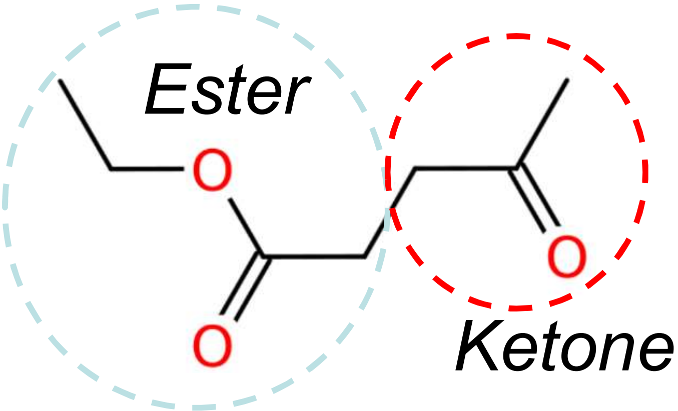
RMG Case Study: Kinetic Database



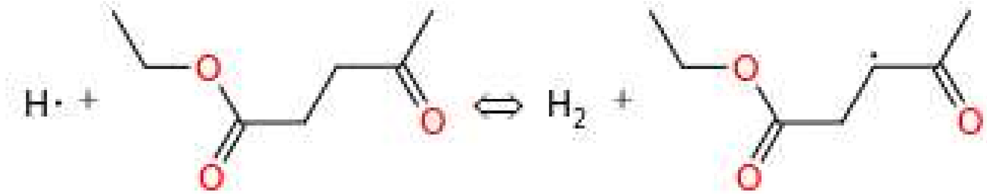


- Prediction of component IDTs has improved, but what about blends?
- Even with kinetic and thermodynamic seeding for diethyl ether and ethanol, the prediction of blended IDTs is poor.
- RMG database is largely lacking training reactions and groups specific to oxygenated species.
  - Alcohols are reasonably covered.
  - Ethers, esters, and ketones need database updates for many reaction families (i.e. H abstraction, intra H migration, cyclic ether formation, radical recombination).
  - Rate constants can be sourced from literature – be sure to consider uncertainty when adding data.
- After making database updates for oxygenated species, predictions are improved significantly.

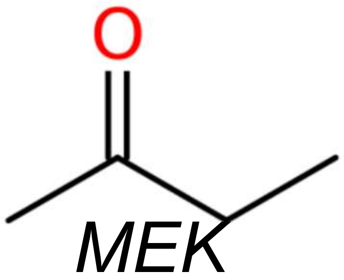
RMG Case Study: Database Updates

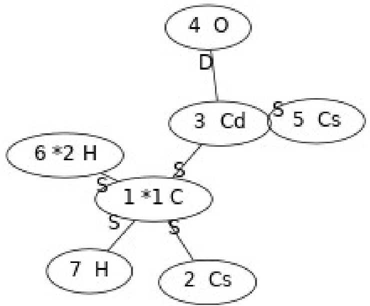


- For database training, ethyl levulinate can be split into a ketone and ester functional group.
- We can use literature to find appropriate training data for these groups.
- As an example reaction, consider hydrogen abstraction by a H radical, to form H_2 and EL3J.

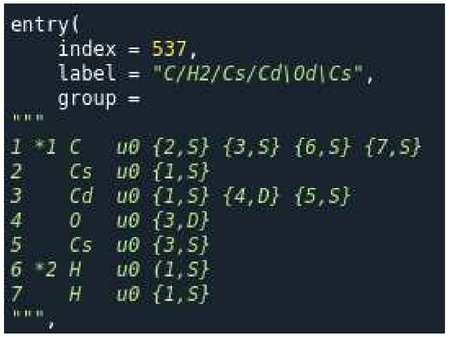


- Methylethyl ketone (MEK) can represent this ketone group.
  - Thion (2017) calculated H abstraction rates for MEK.
  - Calculations performed at the G3//MP2/aug-cc-pVDZ level of theory.
  - In RMG, create a library of new training reactions. Then use the ‘kinetics_library_to_training.ipynb’ tool.
- May need to add new group structures to RMG.
  - Tedious and must be done ‘by-hand’, using the adjacency list format.
  - Prevents ambiguous group definition and incorrect training reaction selection during reaction generation.
  - High potential for human error. Needs care.





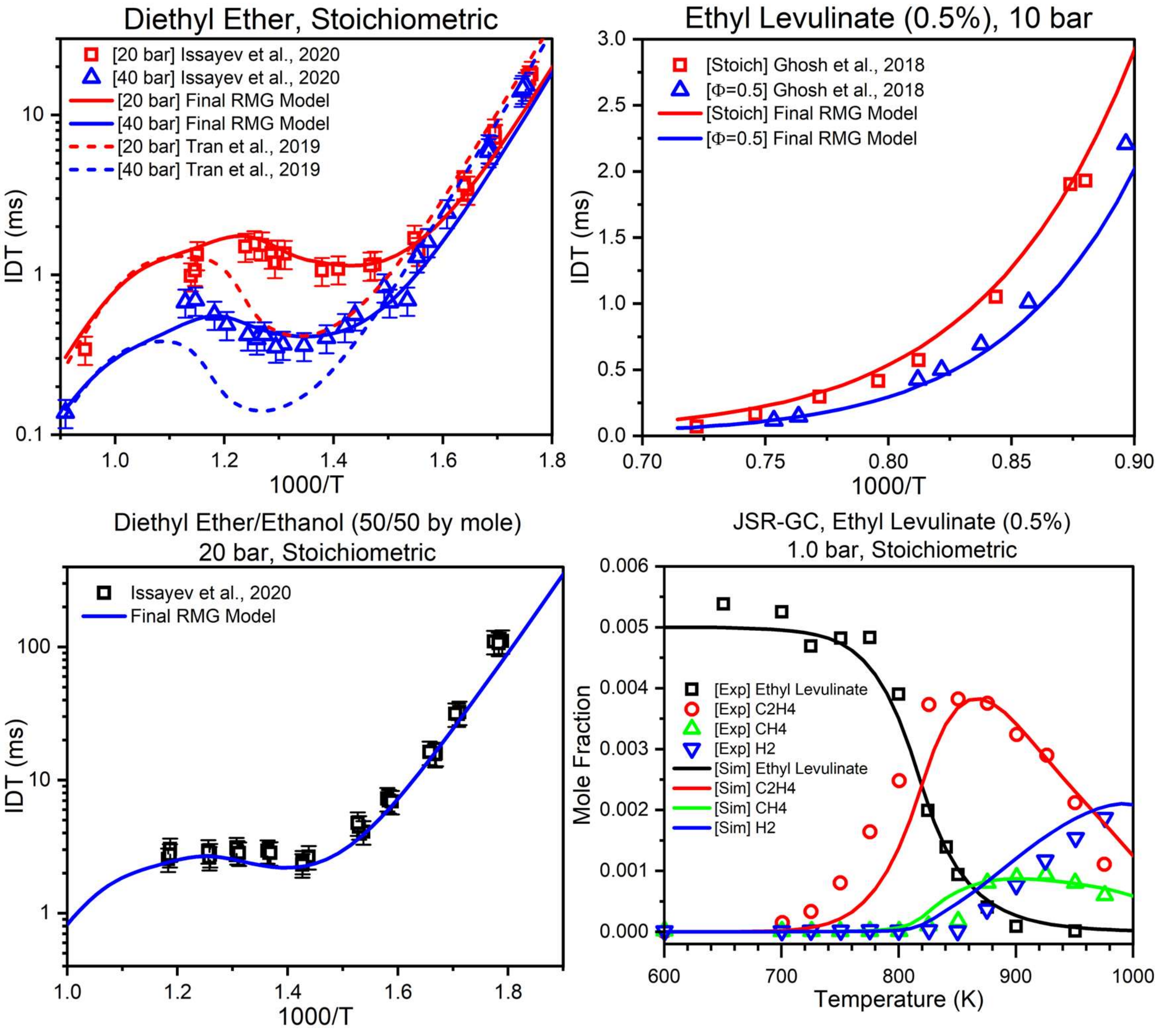
Group Drawing (RMG)



```
entry(
    index = 537,
    label = "C/H2/Cs/Cd\Od\Cs",
    group =
"""
1 *1 C   u0 {2,S} {3,S} {6,S} {7,S}
2    Cs  u0 {1,S}
3    Cd  u0 {1,S} {4,D} {5,S}
4    O   u0 {3,D}
5    Cs  u0 {3,S}
6 *2 H   u0 {1,S}
7    H   u0 {1,S}
""",
```

Group Definition

RMG Case Study: Final Model



- Using the outlined process, we can produce a final model that performs extremely well for complex fuels.
- Model can be extrapolated beyond regime of original seed mechanisms, outperforming them.

What we learned…

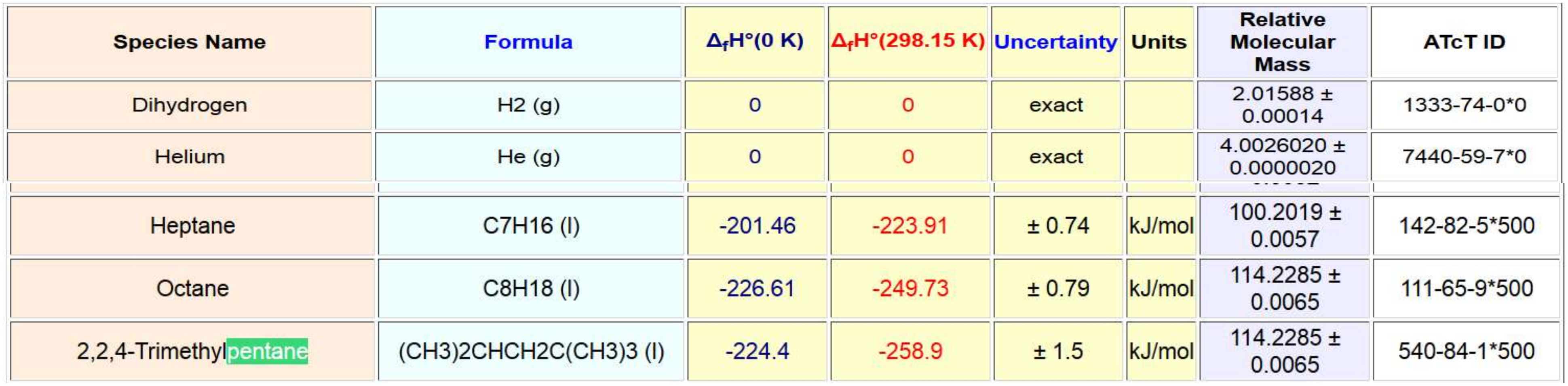
- Robust and accurate mechanism generation can contain some automisation but also requires careful human interaction.
- A reasonable **seed mechanism** was required.
  - Composed using human expertise – particularly when new reaction classes are needed.
- **New groups/training** data were required – we need an automatic way to pull this in. LLMs?
- Physical reality checks were needed outside the developed region of the seed mechanisms.
  - Many violations of **collision limits** found.
- **Sensitivity analysis** and rate constant updates (based on **high level theory**) were needed to get final good accuracy.
- **Validation data** is sparser than we would like – particularly for mixtures.

Importance of thermochemistry

- Many target outputs from combustion systems depend on accurate thermochemistry e.g:
  - Calculation of reverse rates.
  - Low temperature oxidation routes for hydrocarbon fuels involving RO_2 and QOOH species.
  - The prediction of heat release rates.
  - Prediction of adiabatic flame temperatures.
- Large molecules are challenging.
- Goldsmith et al. (2012) presented a method and data for 200 molecular species of interest in combustion chemistry.
- **A bond additivity correction (BAC)** was developed to account for shortcomings in the treatment of multiple bonds and to remove systematic errors that appeared for different bond types compared to Active Tables (see later): C—H, C—C, C═C, C≡C, O—H, C—O, C═O, and O—O.
  - **2σ uncertainties of 0.58 kcal/mol**
- A high level of theory can produce <0.2 kcal/mol uncertainty but at large CPU cost.

Active Tables

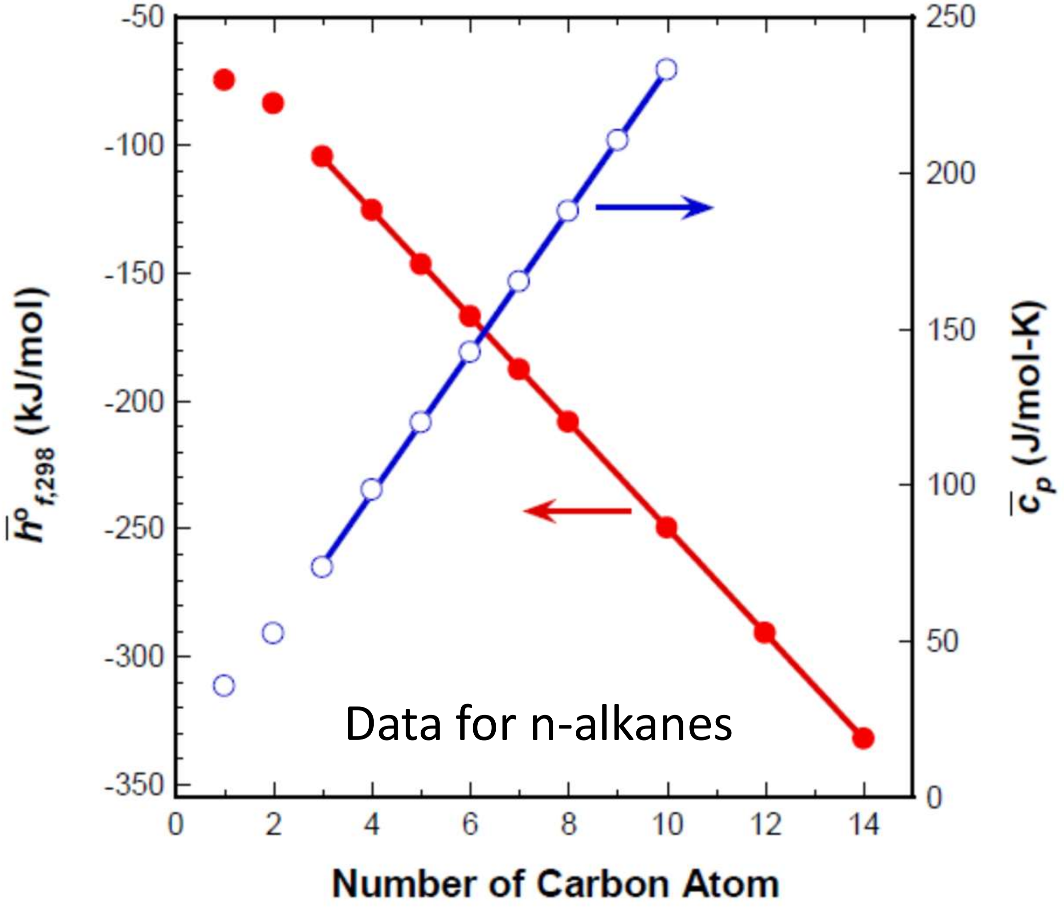
- New paradigm to develop accurate, reliable, and internally consistent thermochemical values for stable, reactive, and transient chemical species by utilizing to the fullest all available experimental measurements as well as state-of-the art theoretical data.
- **ATcT** is based on constructing, analysing, and solving the underlying **Thermochemical Network (TN).**
- Brings together both experimental and theoretical studies (**see earlier**) to reduce uncertainties in data (Burcat & Ruscic, 2005).
- Network of Computed Reaction Enthalpies to Atom-Based thermochemistry **(NEAT)** *Csaszar and Furtenbacher (2010)*
- **Results in highly correlated parameters – be careful of the effects of neglecting such correlations!**



Species Name	Formula	$\Delta_f H°(0\ K)$	$\Delta_f H°(298.15\ K)$	Uncertainty	Units	Relative Molecular Mass	ATcT ID
Dihydrogen	H2 (g)	0	0	exact		2.01588 ± 0.00014	1333-74-0*0
Helium	He (g)	0	0	exact		4.0026020 ± 0.0000020	7440-59-7*0
Heptane	C7H16 (l)	-201.46	-223.91	± 0.74	kJ/mol	100.2019 ± 0.0057	142-82-5*500
Octane	C8H18 (l)	-226.61	-249.73	± 0.79	kJ/mol	114.2285 ± 0.0065	111-65-9*500
2,2,4-Trimethylpentane	(CH3)2CHCH2C(CH3)3 (l)	-224.4	-258.9	± 1.5	kJ/mol	114.2285 ± 0.0065	540-84-1*500

Group Additivity

- Experimentally, for n-alkanes it is observed that *H*, *S*, and C_p all vary linearly with the number of Carbons.
- One can assign a value to the increments caused by inserting one more CH_2 group into the alkane chain.
- This approach works for many different chemical functional groups: adding the group to the molecule adds a set amount to *H, S, C_p* called a GAV.
- For *S*, need to add a symmetry correction to the sum of the GAV.



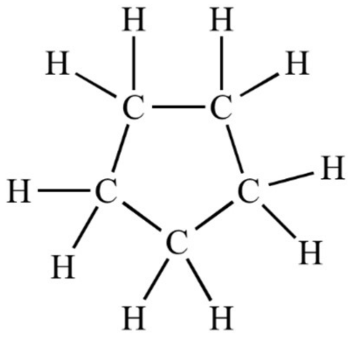
S.W. Benson constructed tables of these Group Additivity Values (GAV). Several researchers, especially Bozzelli and Green, have added to these tables using quantum chemistry to fill in gaps in experimental data.

Treatment of Radicals

- Typically the **hydrogen bond increment (HBI)** approach used, as implemented in THERM.
- Radical thermochemistry based on thermochemistry of **corresponding parent molecule** by adding a so-called **bond dissociation (BD) group**, that accounts for difference in thermochemistry between radical and its parent due to broken hydrogen bond.
- $\Delta h_{\mathrm{f},j,}^{radical}(298\text{ K})$

  $= \Delta h_{\mathrm{f},j,}^{parent}(298\text{ K}) + \Delta h_{\mathrm{f},j,}^{BD}(298\text{ K}) - \Delta h_{\mathrm{f},j,}^{H}(298\text{ K})$
- $\Delta h_{\mathrm{f},j,}^{H}(298\text{ K}) = 217.998\text{ kJmol}^{-1}$, enthalpy of formation of the abstracted hydrogen atom.

Problems with Group Additivity



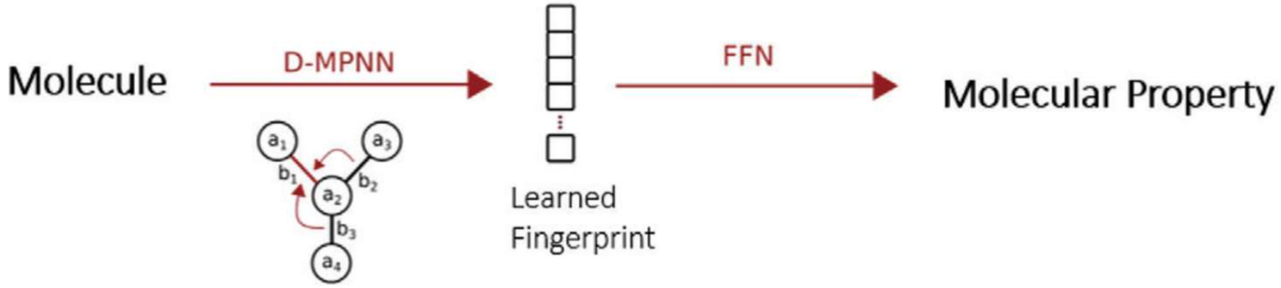
- While the group additivity method is intuitively simple, it has its drawbacks stemming from the need to consider **higher-order correction terms** for a large number of molecules.
- Take e.g. cyclopentane, the addition of group contributions yields H^o = –103 kJ/mol, yet the experimental value is –76 kJ/mol. Difference is caused by **ring strain**, not accounted for in the group value of C–(C2,H2) obtained from unstrained, straight-chain alkane molecules.
- **Cyclics are biggest problem** for group additivity, but some other species also do not work well, e.g. halogenated compounds, and some highly branched compounds.
- Very small molecules are often unique (e.g. CO, OH), so group additivity does not help with those.
- Species with different resonance forms can also cause problems, e.g. propargyl CH2CCH can be written with a triple bond or two double bonds, not clear which should be used when determining groups.
- ***Many methods have been developed using straight chain alkanes but future fuels might not look like this….***

Use of Machine Learning in automated mechanism generation

- For almost all complex fuels (beyond C4 say) the rate constants (and in some cases equilibrium constants) used for elementary reactions describing their oxidation are **based on estimates**.
- Detailed data for smaller molecules can be used to estimate rates for larger molecules with similar chemical structures e.g. in RMG.
- However, data is often sparse and potentially a **mixture of experimental and ab initio/theory computations.**
- Machine-learned models trained on large datasets can improve the accuracy of estimates, allowing a better integration of quantum chemistry and experimental data (Green, 2024).
- There are still challenges – in particular for newer fuels (potentially oxygenated) that have more complex structures and for which **available training data is sparse**.
- Care needed in application of ML – we can't just throw all available data to a neural network (NN) and expect the best fits. **Improved estimates gained from utilizing information on chemical structure and reaction classes that are used in traditional estimation methods.**

Use of Deep Learning Methods for rate coefficient prediction across reaction classes

- Li et al. (2024) A machine learning method to predict rate constants for various reactions in combustion kinetic models.
- Use a **generalised deep learning method** which operates **across reaction classes** – using natural language processing (NLP) methods to infer reaction classes using only the text-based information of reactions (simplified molecular-input line-entry system, SMILES).
- Training data mostly based on high-level quantum chemistry calculations for 9 reaction classes across 8 fuels – 242 reactions.
- Reaction fingerprints serve as inputs of deep neural network (DNN) models to predict modified Arrhenius parameters ln A, n, and E_a.



- Of course the inputs from theory contain uncertainties – but in this work they are assumed to be zero, although the impact of this assumption is tested.

Reaction Classes

- Reaction classes are typical of those found in complex fuel oxidation mechanisms.
- See earlier Low T reaction classes.

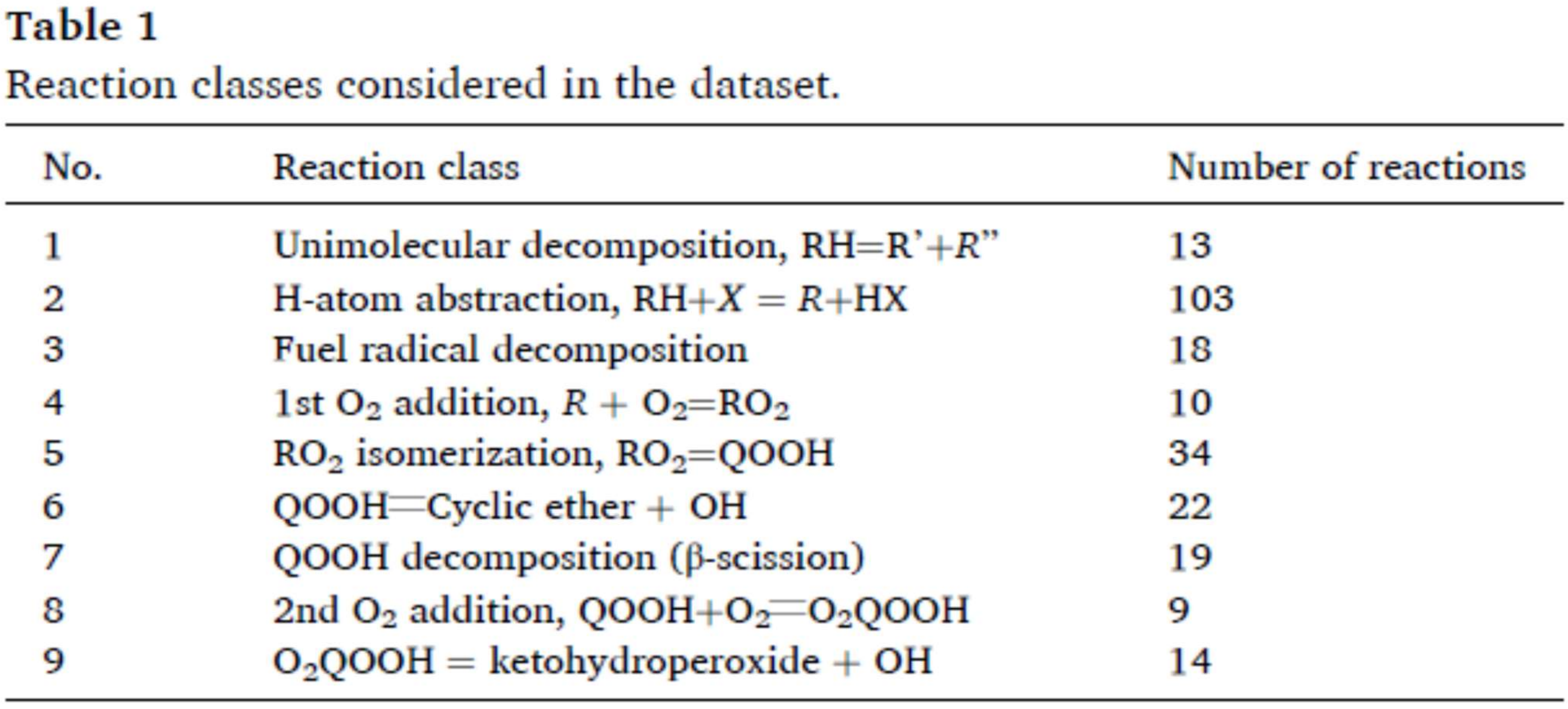


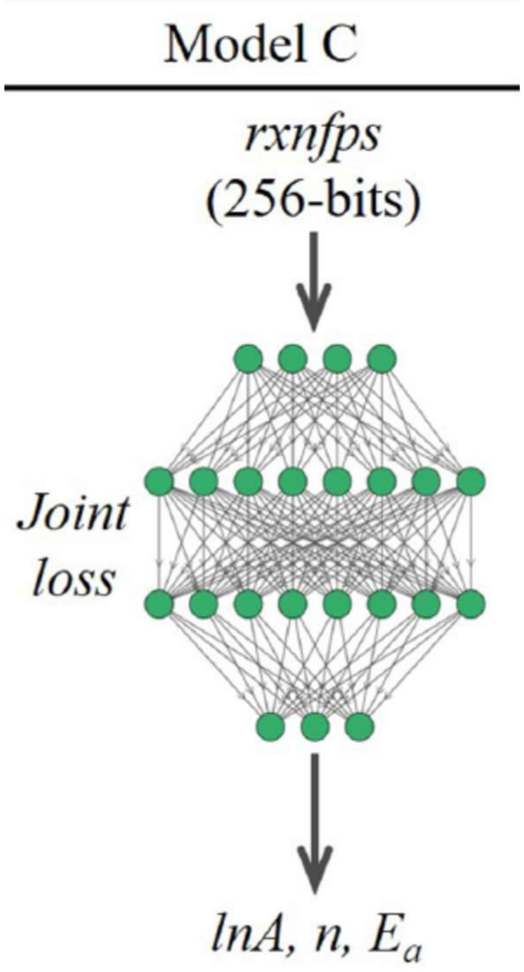
Table 1

Reaction classes considered in the dataset.

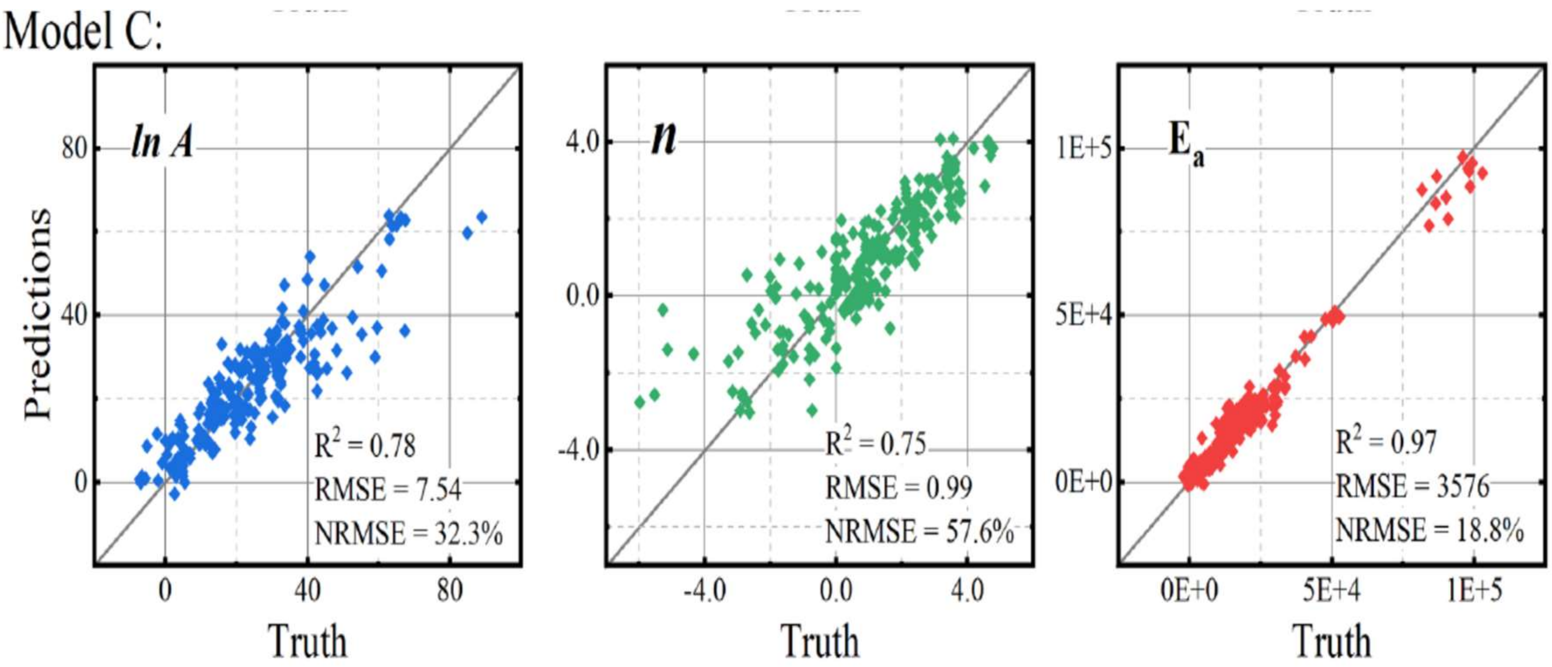
No.	Reaction class	Number of reactions
1	Unimolecular decomposition, RH=R'+R"	13
2	H-atom abstraction, RH+X = R+HX	103
3	Fuel radical decomposition	18
4	1st O_2 addition, $R + O_2 = RO_2$	10
5	RO_2 isomerization, RO_2=QOOH	34
6	QOOH=Cyclic ether + OH	22
7	QOOH decomposition (β-scission)	19
8	2nd O_2 addition, QOOH+O_2=O_2QOOH	9
9	O_2QOOH = ketohydroperoxide + OH	14

Approach

- Using this general approach and reaction fingerprints carries the risk of **overfitting** because the dimensionality of the input features is comparable to the size of the dataset.
- Hence mitigating overfitting during the model training phase is an important part of the methodology using a drop-out method.
- An automatic method (OPtuna) is used to optimize the hyper-parameters of the NN:
  - the number of hidden layers,
  - the number of neurons per layer,
  - learning rate, batch size,
  - drop out ratio,
  - weight decay in the optimizer.
- Reaction classes were almost always successfully identified from fingerprint data.



Outcomes



- Note that this is *ln A* and there are some fairly large errors in predictions.
- Activation energy is more successfully estimated than the pre-exponential factor.

Uncertainties in predictions

- The uncertainty factor for most reactions (about 70%) is less than 4, roughly comparable with uncertainties of high-level calculations.
- Uncertainty factors for some reactions (about 18 %) are 4 – 10. Small number of predicted rate constants with uncertainty factors above one order of magnitude, mostly from the complex low-temperature reaction classes – likely due to sparsity of training data.

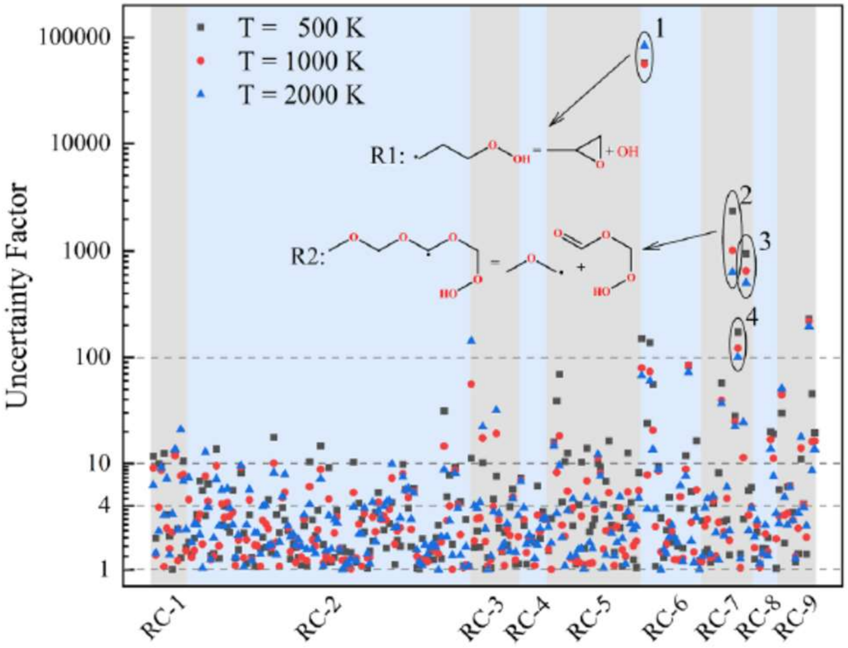


Fig. 10. Uncertainty factor of rate constants predicted by model C. Horizontal coordinates are reactions sorted by reaction class.

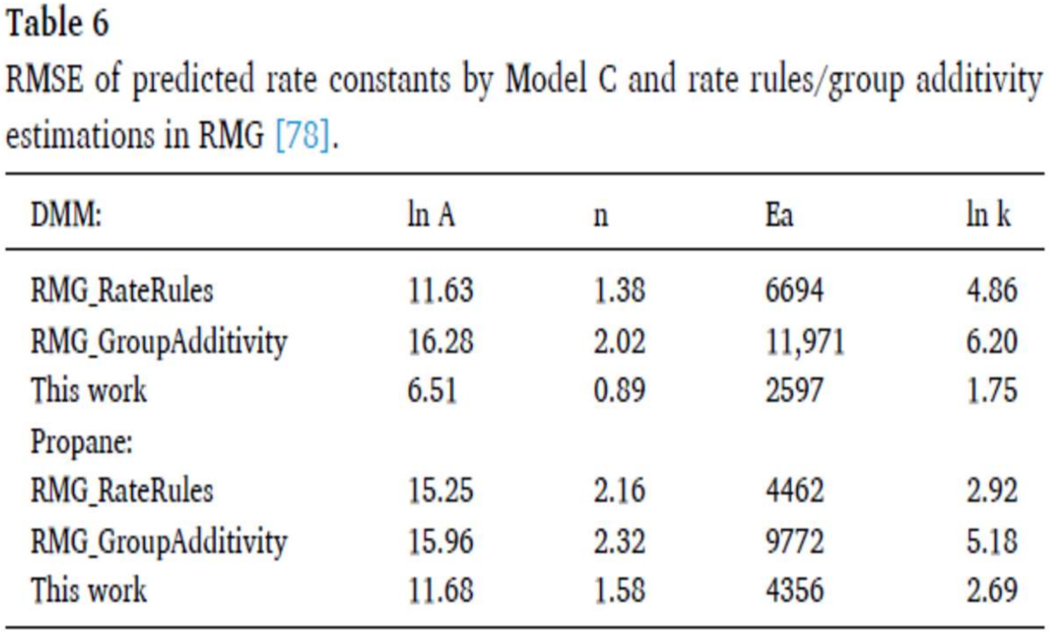


Table 6
RMSE of predicted rate constants by Model C and rate rules/group additivity estimations in RMG [78].

DMM:	ln A	n	Ea	ln k
RMG_RateRules	11.63	1.38	6694	4.86
RMG_GroupAdditivity	16.28	2.02	11,971	6.20
This work	6.51	0.89	2597	1.75
Propane:				
RMG_RateRules	15.25	2.16	4462	2.92
RMG_GroupAdditivity	15.96	2.32	9772	5.18
This work	11.68	1.58	4356	2.69

Better than traditional rate rules approach

What is this approach lacking?

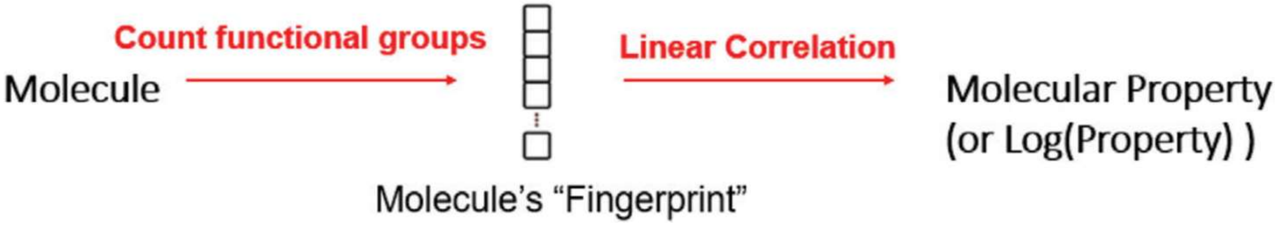
- It is dependent on the reaction classes being successfully identified – but the list of included classes is predetermined.
- It is not actually building a mechanism from scratch and so is not an automated reaction generator.
- The uncertainties remain high for systems with sparse training data sets.
- **How can machine learning (ML) be used to generate reaction mechanisms and populate their required data even for systems where data is sparse?**

What do we need to build a mechanism for a particular process? (Johnson and Green, 2024)

- A systematic method both to propose candidate reactions and species, and to decide which are actually important.
  - Build new reaction classes by systematically generating all possible reactions from important species, then numerically test and prune reactions calculated to be too slow to be important based on **selected error tolerance**.
  - Requires calculations (or estimations) for large numbers of reaction rates. Number of reactions can build quickly. Also, some slow reactions may lead to important intermediates and be missed with the wrong tolerances.
  - Sometimes there are so many possible isomers that it is difficult for a human to correctly enumerate them all. KinBot attempts to overcome this, but still challenging to solve numerical eqs resulting from complex systems with thousands of species.
  - Currently not possible to compute accurate rates for all possible reactions in a system using higher level quantum methods. So even in order to build a mechanism we rely on accurate estimates.

Problems with sparse data

- As we have said, traditional approaches are based on identifying functional groups within molecule, or bond changes, and then using simple (e.g. least squares regression) fits to data from smaller molecules.



- For example, one entry in the fingerprint could be the number of a certain functional group in the molecule.
- Methods such as Li et al. and Chemprop use machine learning to improve estimates over simple regression approaches.
- However, we saw problems where data sets are sparse and the newer the system of interest, the sparser the data sets will be.
- Often in kinetics, the data available is very "clumpy", with many data for certain types of molecule or reactions, and zero data on some other types of molecules or reactions (e.g. new systems or fuels).

Can this be fully automated?

- Will require:
  - – Building an appropriate reaction set from reactants, through intermediates, through to products.
  - – Pruning away unimportant reactions at appropriate reaction conditions to avoid combinatorial explosion.
  - – Estimating appropriate thermo for all species in the mechanism.
  - – Estimating appropriate *T* and or *P* dependent reaction rate coefficients for each reaction based on available training data.
  - – Without human intervention!

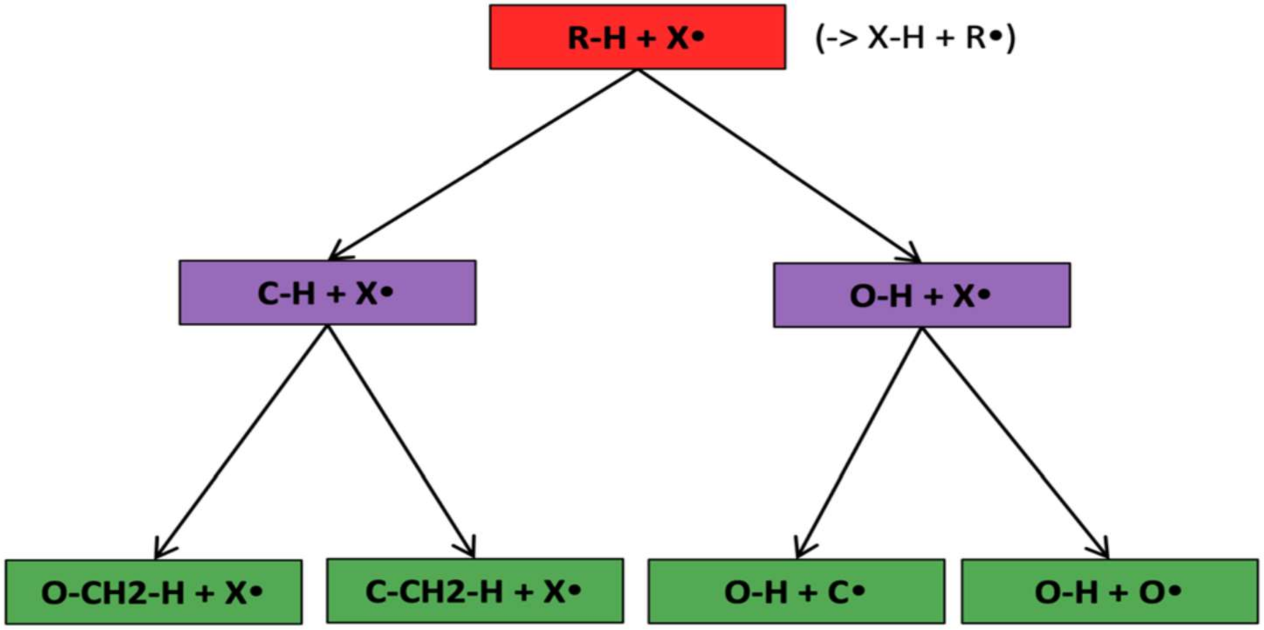
What would a good ML algorithm for building a mechanism and predicting rate coefficients look like:

- Fully automated.
- **Good optimization using fairly sparse data** for some reaction classes.
- Human readable and able to formulated rates in a format readable by commonly used simulation codes.
- Relatively easy to maintain and extend and scalable.
- Best use of all available data; combining experimental and theory calculations.
- Able to incorporate qualitative information from experts where appropriate.
- Able to provide **uncertainty estimates** for predicted rates over a range of appropriate *T,P*. Extremely important because we need to know where there are cases where the uncertainty is so large for the estimate not to be useful.

(Johnson and Green, 2024)

Using tree structures

- Typical successful applications of NNs for predicting rate constants can be based on datasets containing more than 10,000 reactions.
- Typical reaction families in AMG codes such as RMG have to be estimated based on data from fewer than 20 reactions.
- According to Johnson and Green (2024) the best way to use such sparse training data is to incorporate ML within a similar tree structure to that used for developing RRs e.g. as used in RMG.
- Have developed an approach based on the use of decision trees.



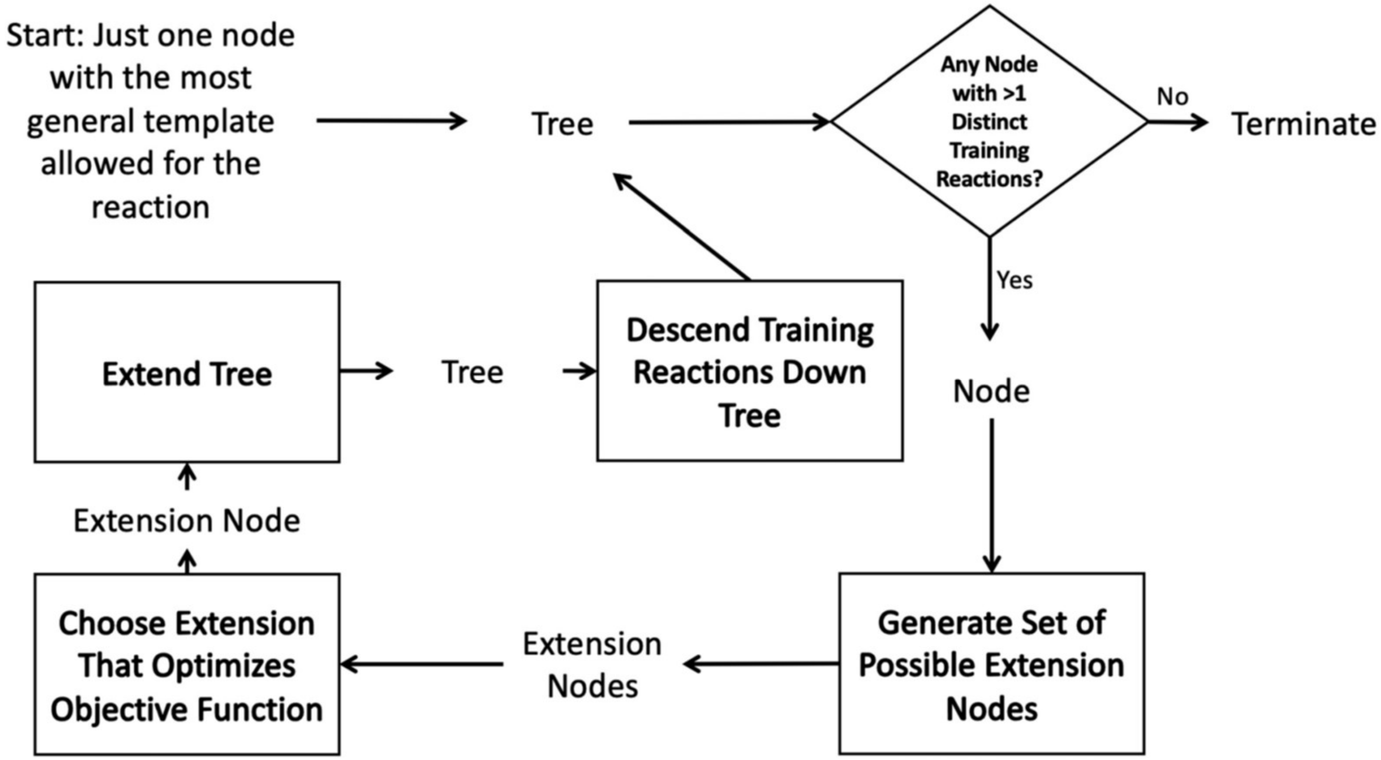
Example of a subgraph isomorphic reaction template decision tree.

What are decision trees?

- Decision trees are classifiers that start with an item at a single root node with no parents.
- At each node the item is checked against each descendant “child” node and moves to the first child node that it matches and so on until it reaches a node where it doesn't match any children.
- That final node becomes its classification.
- In this context the items are **reactions, i.e. molecular graphs of reactant(s) and product(s) with atom mapping** with matching done by subgraph isomorphism checks.
- **Automation:**
  - iterate starting from a single root node adding the new node that best optimizes the tree each iteration
  - end tree generation process when some termination criterion related to the number of tree nodes or tree depth is satisfied
  - new nodes are added that best divide those reactions into groups with similar rate estimates.

The mechanism building process

- Extension chosen that minimizes $\Pi = N_1\sigma_1 + N_2\sigma_2$ where N_i is the number of reactions in partition *i* and σ_i is the standard deviation in log(k(1000 K)) within partition *i*.
- I.e. choose the extension that clusters reactions with similar rates into the same partitions at chosen *T* of 1000 K where the training data is expected to be the most accurate.


Start: Just one node with the most general template allowed for the reaction
Tree
Any Node with >1 Distinct Training Reactions?
No
Terminate
Yes
Node
Extend Tree
Tree
Descend Training Reactions Down Tree
Extension Node
Choose Extension That Optimizes Objective Function
Extension Nodes
Generate Set of Possible Extension Nodes

…

- A rule is then defined at each node. All training reactions matching a given node (top node has all reactions) are fitted to a to estimate Arrhenius parameters essentially using a least squares method.
- Uncertainties need to be estimated for the input reactions (e.g. from kinetics experiments, theory, from optimization studies incorporating bulk experimental data such as flame speeds), the interpolant, and for *k(T)* estimates for a new reactions where no data is available. Challenging!

- Normal distribution for $\Delta\log(k_i) \sim N(\log(k_{\text{best fit},i}) - \log(k_{\text{true},i}), \sigma^2_{\text{rxn,i}} + \sigma^2_{\text{model,i}})$

- i.e. uncertainties are due to limitations in the model’s ability to represent the chemical space and the errors in the training data.
- At lower nodes there will be fewer training reactions, but a much smaller chemical space hence fitting errors will be small and the uncertainty in the reaction rate for the rule would be expected to be smaller than that for the individual reactions.

…

- At higher nodes there will be more training reactions and a larger chemical space. Hence, $\log(k_{\text{best fit,i}}) - \log(k_{\text{true,i}})$ will be large and errors in $k_{\mathit{model},\text{i}}$ will dominate.
- Balance between selecting parent nodes with more training reactions but a larger fitting error vs. child modes with a smaller chemical space but fewer training reactions.
- **Estimation of uncertainties allows the appropriate node to be selected for rate constant estimation.**
- **At nodes near the top** of the tree there are many reactions making it possible to accurately calculate the reaction rate parameters, however, the chemical space spanned can be quite large making it difficult for a single model fit to represent all of the involved reactions.
- **At nodes near the bottom** of the tree the chemical space spanned is much smaller and the model can better represent the space, but there are fewer reactions, making the fit more sensitive to errors in the training data.

Testing against RRs

- Method has been tested for several commonly used reaction families where Rate Rules have previously been applied and outperforms RR methods.

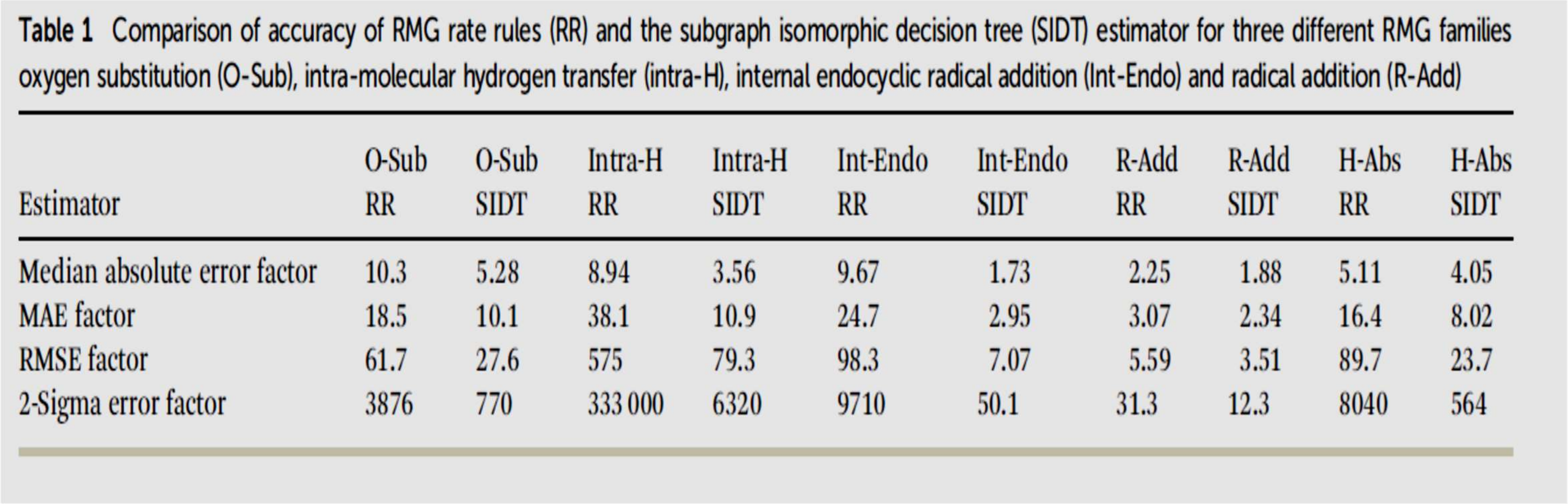


Table 1 Comparison of accuracy of RMG rate rules (RR) and the subgraph isomorphic decision tree (SIDT) estimator for three different RMG families oxygen substitution (O-Sub), intra-molecular hydrogen transfer (intra-H), internal endocyclic radical addition (Int-Endo) and radical addition (R-Add)

	O-Sub	O-Sub	Intra-H	Intra-H	Int-Endo	Int-Endo	R-Add	R-Add	H-Abs	H-Abs
Estimator	RR	SIDT	RR	SIDT	RR	SIDT	RR	SIDT	RR	SIDT
Median absolute error factor	10.3	5.28	8.94	3.56	9.67	1.73	2.25	1.88	5.11	4.05
MAE factor	18.5	10.1	38.1	10.9	24.7	2.95	3.07	2.34	16.4	8.02
RMSE factor	61.7	27.6	575	79.3	98.3	7.07	5.59	3.51	89.7	23.7
2-Sigma error factor	3876	770	333 000	6320	9710	50.1	31.3	12.3	8040	564

Final remarks

- There is definitely significant scope for ML methods to assist in building mechanisms for new fuels and processes.
- BUT!
- The success of such methods depends on
  - The available training data, its quality and how well it covers the space of interest.
  - Training data from both quantum and experimental methods is needed.
  - Effective data sharing and data curation will be important. Not an easy and an often thankless task.
  - The use of ML methods which maximise the value in this training data, including in cases where it can be sparse.
- This means using approaches that effectively partition the data to give optimal error minimisation.
- Finally (**at the risk of sounding like a broken record**) we need to track uncertainties, and consistently!