State-of-the-Art Computational Approaches for Characterizing Reaction Properties and Predicting Unknown Reactivity

Brett M. Savoie
Davidson Associate Professor of Chemical Engineering,
Purdue University

Students: Qiyuan Zhao, Tyler Pasut

P2SAC Fall Conference, Purdue University, 12/14/22

“Known Unknowns” and “Unknown Unknowns”

A → B

- To safely plan a known reaction, we need access to solid thermodynamic data (e.g., ΔH_f, S^o, C_v) to understand and classify risks.

- This is a “known unknown” in that we know the reaction, A → B, but we need values for a few unknown variables.

A → ? → B ; A → B + ? ; A → ?

- **A → ? → B**, means that we know the net reaction, but there may be a consequential (e.g., potentially reactive) intermediate. Even if we have accurate thermodynamic data on A/B, neglecting the intermediate could be disastrous.

- The **A → B + ?** (unknown side-reaction) and **A → ?** (unknown main product), problems have similar “unknown unknown” characteristics.

“Known Unknowns” and “Unknown Unknowns”

A → B

- To assess a known reaction, we need access to solid thermodynamic data (e.g., ΔH_f, S°, C_v) to understand and classify risks.
- This is a “known unknown” in that we know the reaction, A → B, but we need values for a few unknown variables.

TAFFI Component Increment Theory (TCIT)

A → ? → B ; A → B + ? ; A → ?

- **A → ? → B**, means that we know the net reaction, but there may be a consequential (e.g., potentially reactive) intermediate. Even if we have accurate thermodynamic data on A/B, neglecting the intermediate could be disastrous.
- The **A → B + ?** (unknown side-reaction) and **A → ?** (unknown main product), problems have similar “unknown unknown” characteristics.

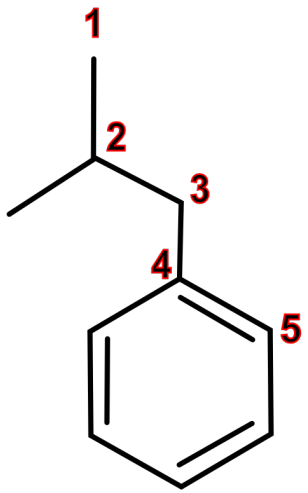
Yet Another Reaction Program (YARP)

Challenges of Contemporary Group Theories

Benson Group Theory:

- The idea is to decompose molecular properties (ΔH_f, S^o, C_v) as the sum of "group" contributions.
- Group contributions are calculated based on trusted experimental or computational data, and transferability is assumed.

From Anslyn and Dougherty's Textbook



1) C -(C)(H)$_3$ . . . . . 2(-10.20)
2) C -(C)$_3$(H) . . . . . -1.90
3) C -(C$_B$)(C)(H)$_2$ . . . . . -4.86
4) C$_B$ -(C) 5.51
5) C$_B$ -(H) . . . . . 5(3.30)

-5.15 kcal/mole
(-21.6 kJ/mole)

Experimental ΔH_f: -5.15 +/- 0.34 kcal/mol

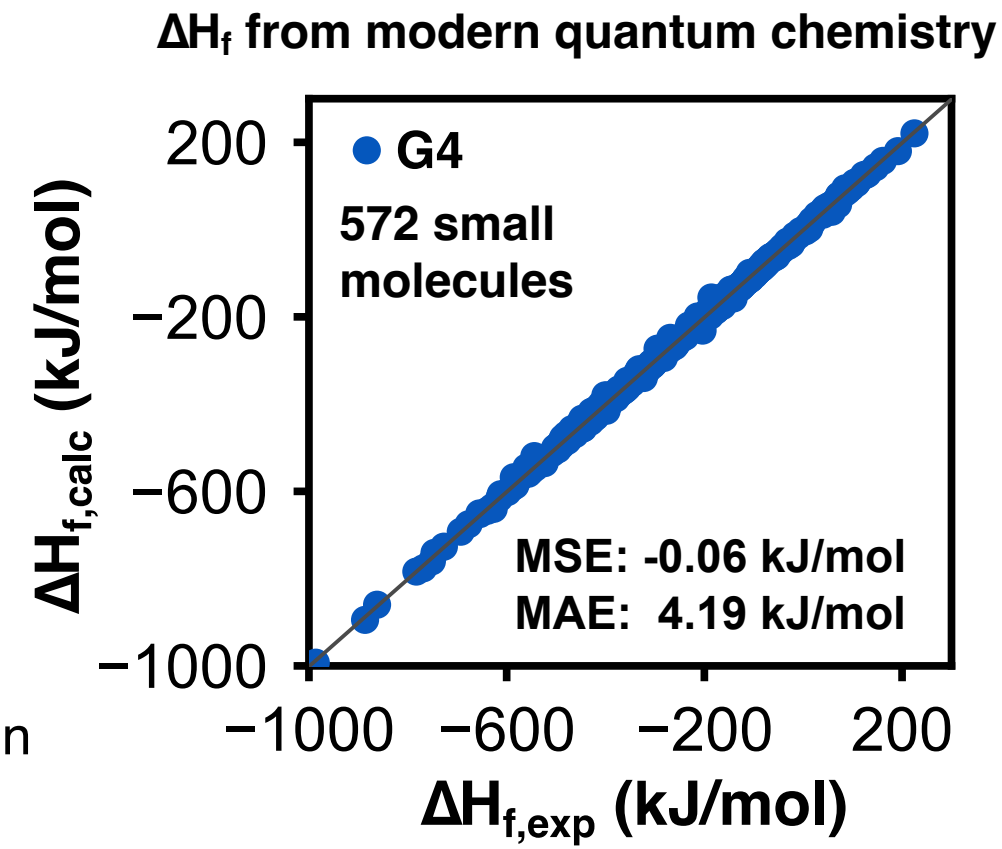
Problems we want to address:

- **Specificity:** the definition of a "group" has never been formalized and inconsistent granularity is applied.
- **Provenance:** inconsistent thermodynamic data is available/used to determine group contributions.
- **Extensibility:** because of the provenance and specificity problems, it isn't possible to develop new groups in a consistent way.

Challenges of Contemporary Group Theories

Benson Group Theory:

- The idea is to decompose molecular properties (ΔH_f, S^o, C_v) as the sum of "group" contributions.
- Group contributions are calculated based on trusted experimental or computational data, and transferability is assumed.



Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Problems we want to address:

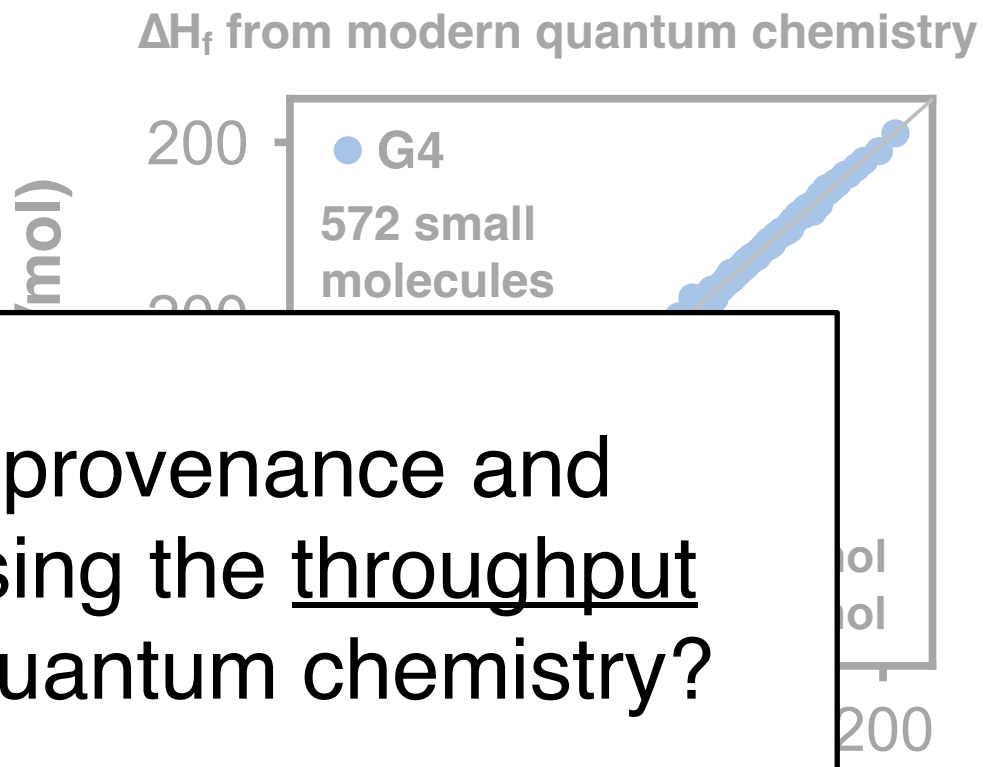
- **Specificity:** the definition of a "group" has never been formalized and inconsistent granularity is applied.
- **Provenance:** inconsistent thermodynamic data is available/used to determine group contributions.
- **Extensibility:** because of the provenance and specificity problems, it isn't possible to develop new groups in a consistent way.

Challenges of Contemporary Group Theories

Benson Group Theory:

- The idea is to decompose molecular properties (ΔH_f, S^o, C_v) as the sum of "group" contributions.

- Grou
on trus
data, a



Can we circumvent the provenance and extensibility challenges using the throughput and accuracy of modern quantum chemistry?

Prob

- **Spec**
formal

- **Provenance:** inconsistent thermodynamic data is available/used to determine group contributions.

Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

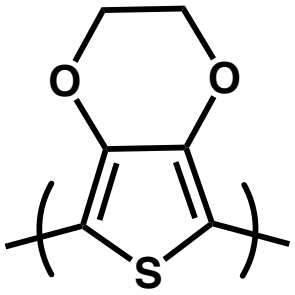
- **Extensibility:** because of the provenance and specificity problems, it isn't possible to develop new groups in a consistent way.

TAFFI Component Increment Theory (TCIT)

The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.

TCIT is a <u>component</u> theory (2-bond specific)



Topology **A**utomated **F**orce **F**ield **I**nteractions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

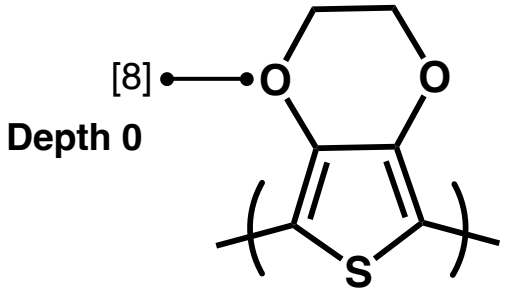
P2SAC Publications

TAFFI Component Increment Theory (TCIT)

The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.

TCIT is a <u>component</u> theory (2-bond specific)



Depth 0

Topology **A**utomated **F**orce **F**ield **I**nteractions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

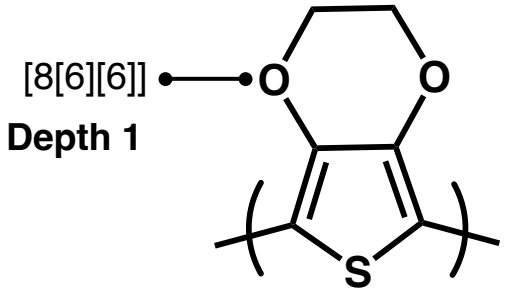
P2SAC Publications

TAFFI Component Increment Theory (TCIT)

The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.

TCIT is a <u>component</u> theory (2-bond specific)



Topology **A**utomated **F**orce **F**ield **I**nteractions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

P2SAC Publications

TAFFI Component Increment Theory (TCIT)

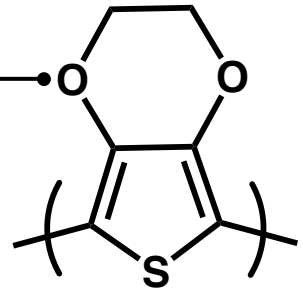
The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.

TCIT is a <u>component</u> theory (2-bond specific)

[8[6[6][1][1]][6[6][6]]]

Depth 2



Topology Automated Force Field Interactions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

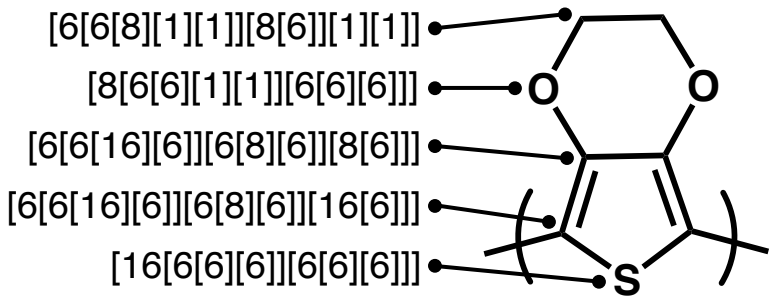
P2SAC Publications

TAFFI Component Increment Theory (TCIT)

The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.

TCIT is a <u>component</u> theory (2-bond specific)



Topology Automated Force Field Interactions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

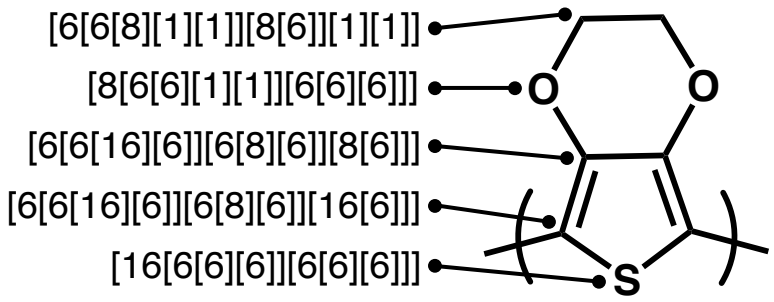
P2SAC Publications

TAFFI Component Increment Theory (TCIT)

The fundamental idea

- Systematize component-definitions and model compound selection with rigorous graph-based typing.
- Two-bond specificity should improve both the accuracy and transferability of the resulting components.
- Parameterizing a component model **would not be feasible with only experimental data**.

TCIT is a <u>component</u> theory (2-bond specific)



Topology **A**utomated **F**orce **F**ield **I**nteractions

graph/structure equivalence

$$
\begin{array}{c}
S \\ C \\ C \\ C \\ C \\ O \\ C \\ C \\ H \\ H \\ O \\ H \\ H
\end{array}
\begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Adjacency matrix for PEDOT monomer

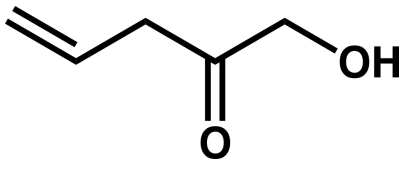
Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Zhao, Q.; Iovanac, N.; Savoie, B. M.; "Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds" *J. Chem. Info. Model.* **2021**, 61, 5013-5027

Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

P2SAC Publications

Graphical Decomposition of Model Compounds

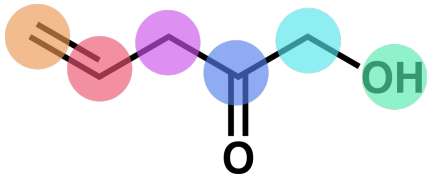


1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

Graphical Decomposition of Model Compounds

identify components*

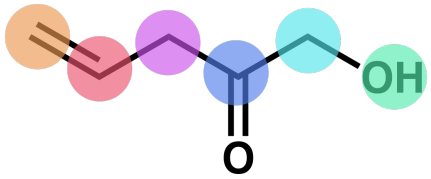


1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

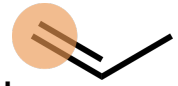
Graphical Decomposition of Model Compounds

identify components*



1-hydroxy-pent-2-ene-2-one

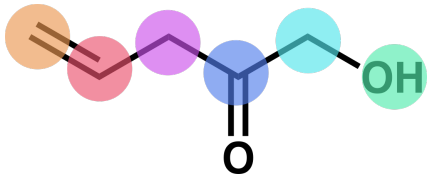
Recursively generate smallest acyclic model compounds



How will we select molecules for parameterizing TCIT components?

Graphical Decomposition of Model Compounds

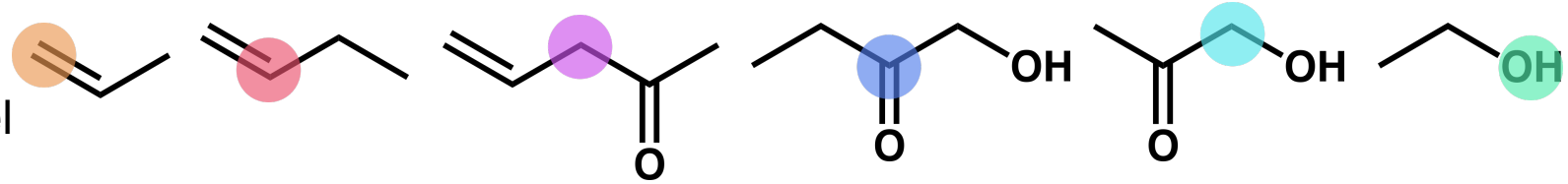
identify components*



1-hydroxy-pent-2-ene-2-one

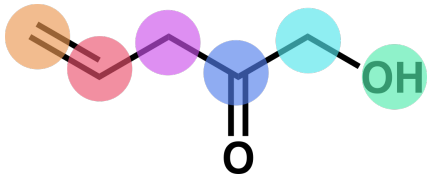
How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds



Graphical Decomposition of Model Compounds

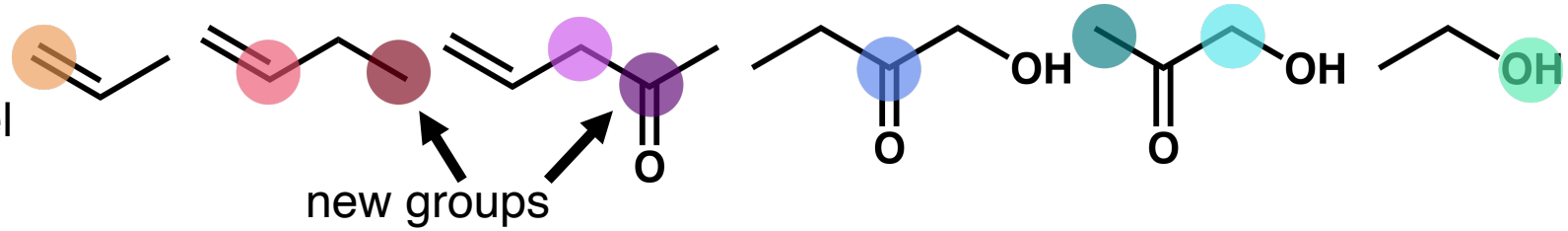
identify components*



1-hydroxy-pent-2-ene-2-one

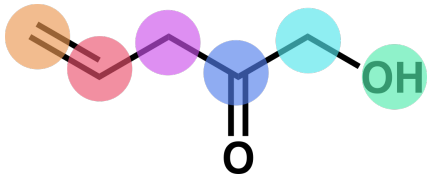
How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds



Graphical Decomposition of Model Compounds

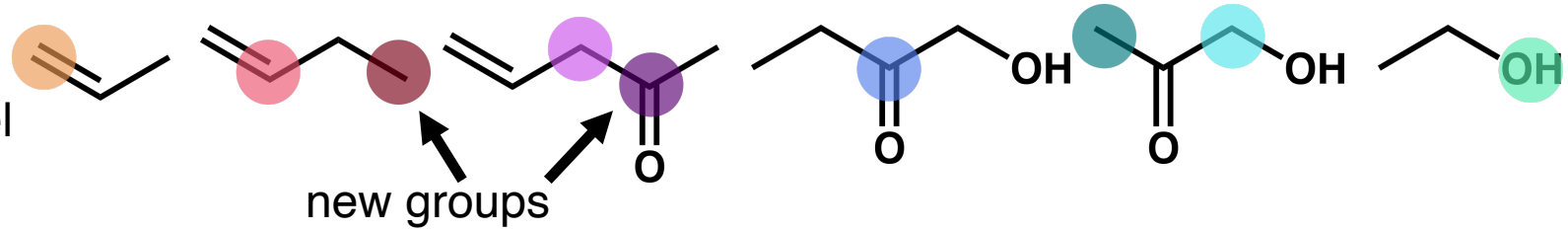
identify components*



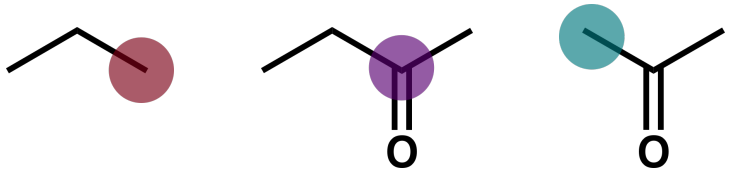
1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds

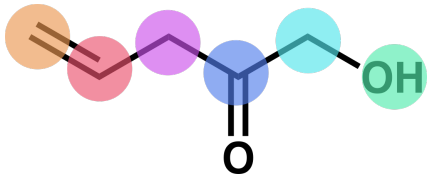


Resolve dependencies



Graphical Decomposition of Model Compounds

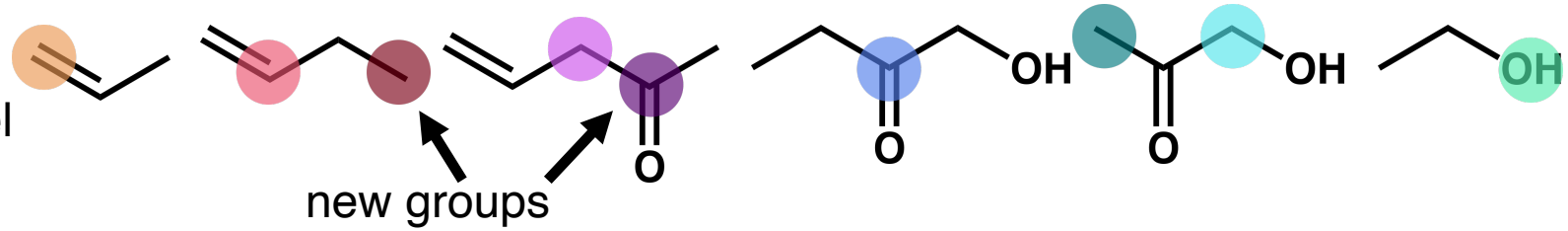
identify components*



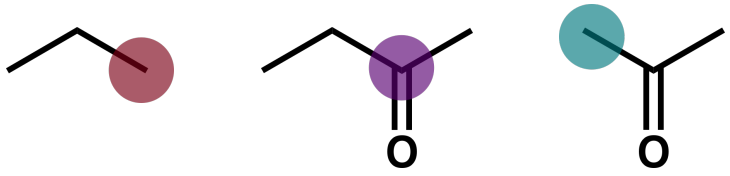
1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

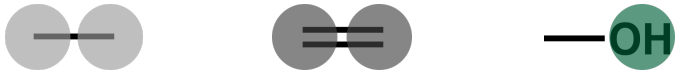
Recursively generate smallest acyclic model compounds



Resolve dependencies

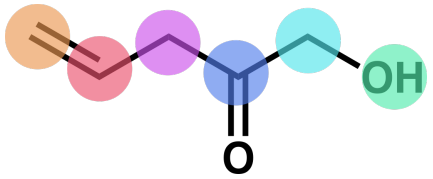


Resolve rank deficiency with elementary constraints



Graphical Decomposition of Model Compounds

Prediction target:



1-hydroxy-pent-2-ene-2-one

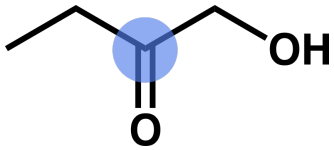
$\Delta H_{f,G4}$ = -259.9 kJ/mol
$\Delta H_{f,TCIT}$ = -259.3 kJ/mol

no experimental data

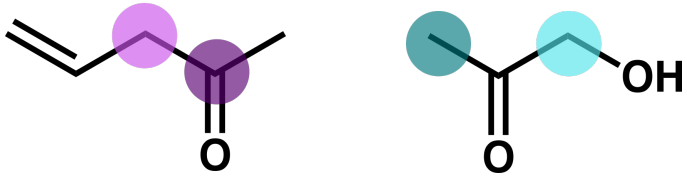
Topologically sort dependency graph

(Automatically handled by TCIT software)

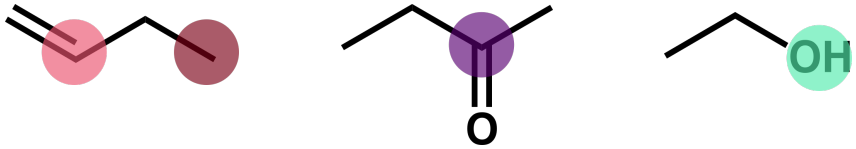
Gen 4:



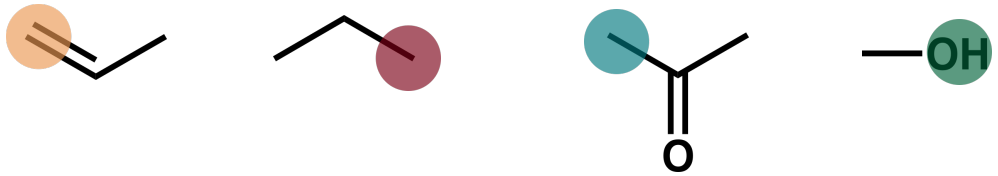
Gen 3:



Gen 2:



Gen 1:



Gen 0:

Model compounds are small enough to perform the highest quality quantum chemistry calculations (G4 throughout)

Graphical Decomposition of Model Compounds

Have we solved the specificity problem?

All components are unique out to a graph depth of two, no exceptions.

Have we solved the provenance problem?

All ΔH_f data is calculated at the G4 composite level, no exceptions.

Have we solved the extensibility problem?

Model compounds exist for all conceivable components, no exceptions.

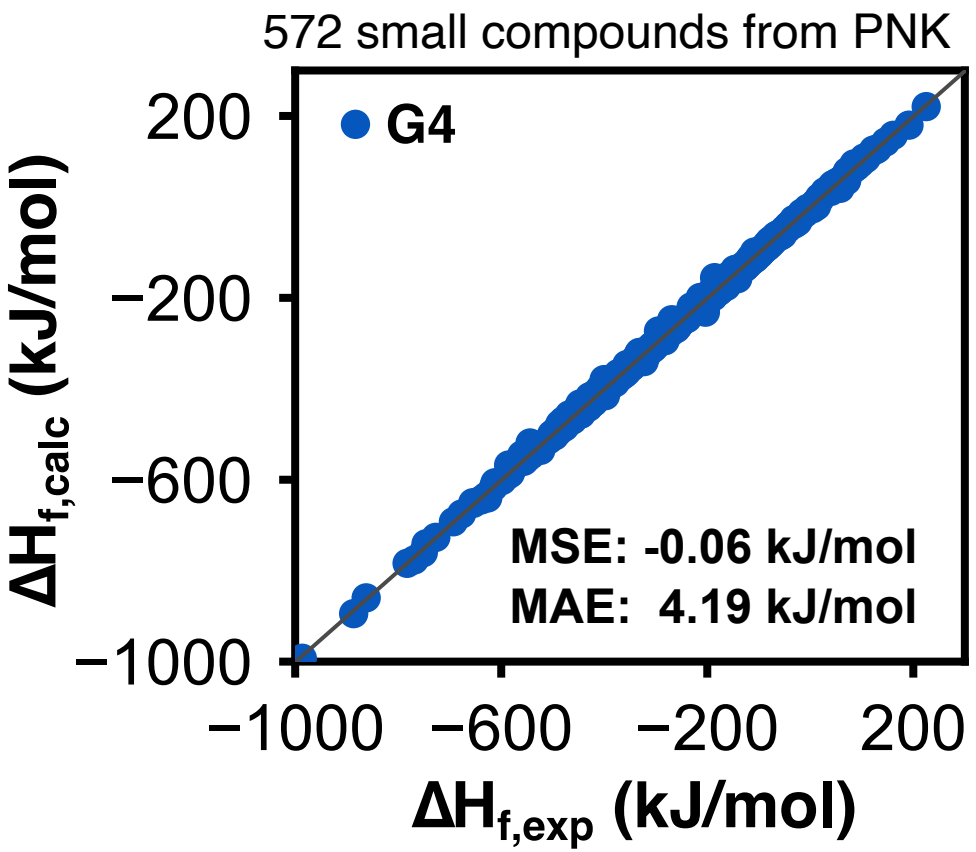
Benchmarking $\Delta H_{f,gas}$ Predictions Against the PNK Dataset

- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

- PNK is a core dataset for fitting Benson groups

- ~600 PNK compounds are small enough for G4 calculations and comparison with experiment.



Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a Fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Benchmarking $\Delta H_{f,gas}$ Predictions Against the PNK Dataset

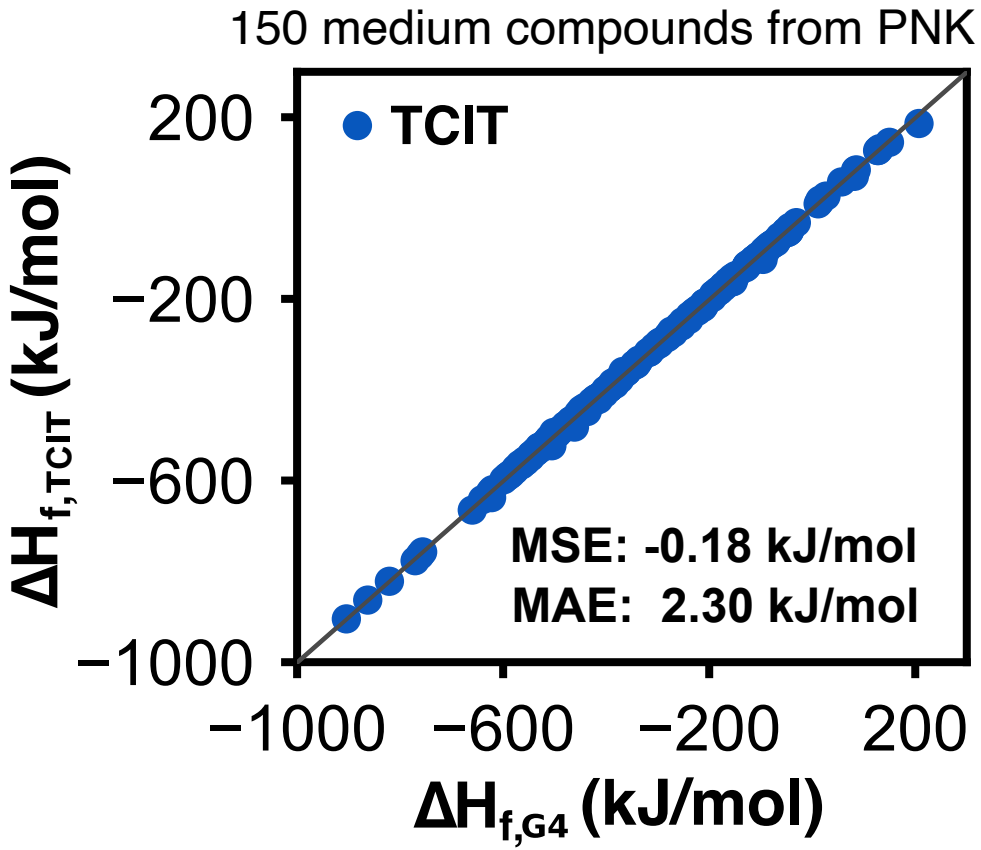
- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

- PNK is a core dataset for fitting Benson groups

- ~600 PNK compounds are small enough for G4 calculations and comparison with experiment.

- ~150 PNK compounds are large enough for direct G4 calculation and comparison with TCIT.



Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a Fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Benchmarking $\Delta H_{f,gas}$ Predictions Against the PNK Dataset

- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

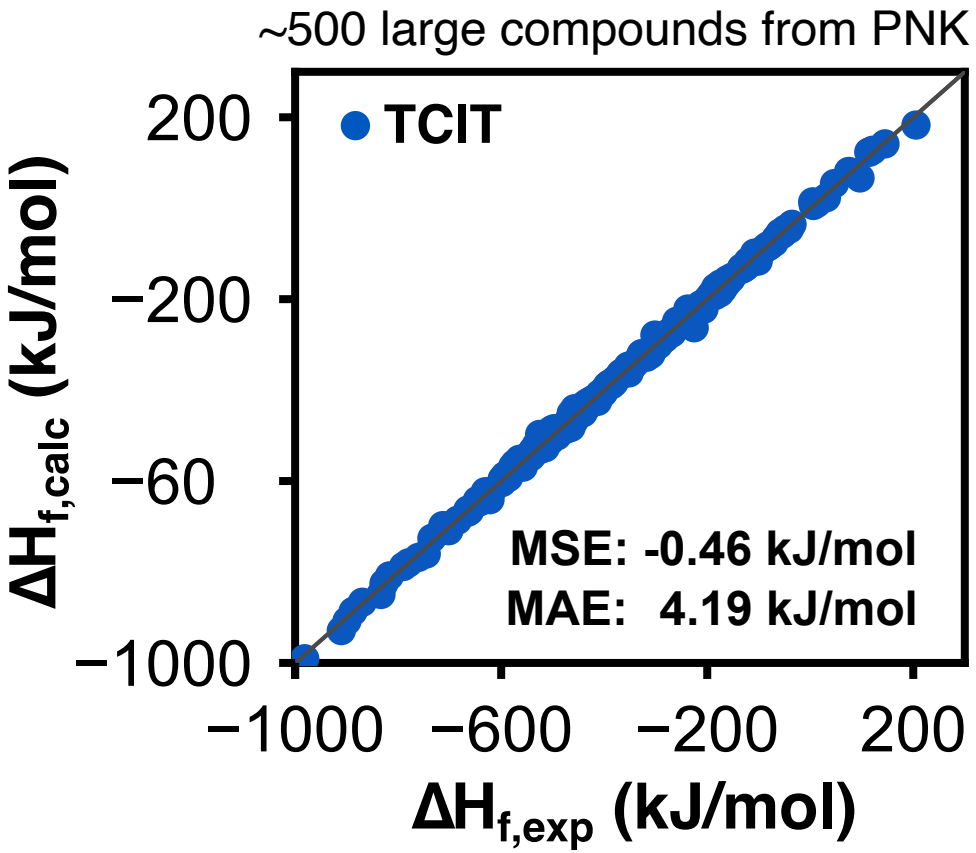
(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

- PNK is a core dataset for fitting Benson groups

- ~600 PNK compounds are small enough for G4 calculations and comparison with experiment.

- ~150 PNK compounds are large enough for direct G4 calculation and comparison with TCIT.

- ~500 PNK compounds are large enough to evaluate the predictive accuracy of the increment theories.



Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a Fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

Benchmarking $\Delta H_{f,gas}$ Predictions Against the PNK Dataset

- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

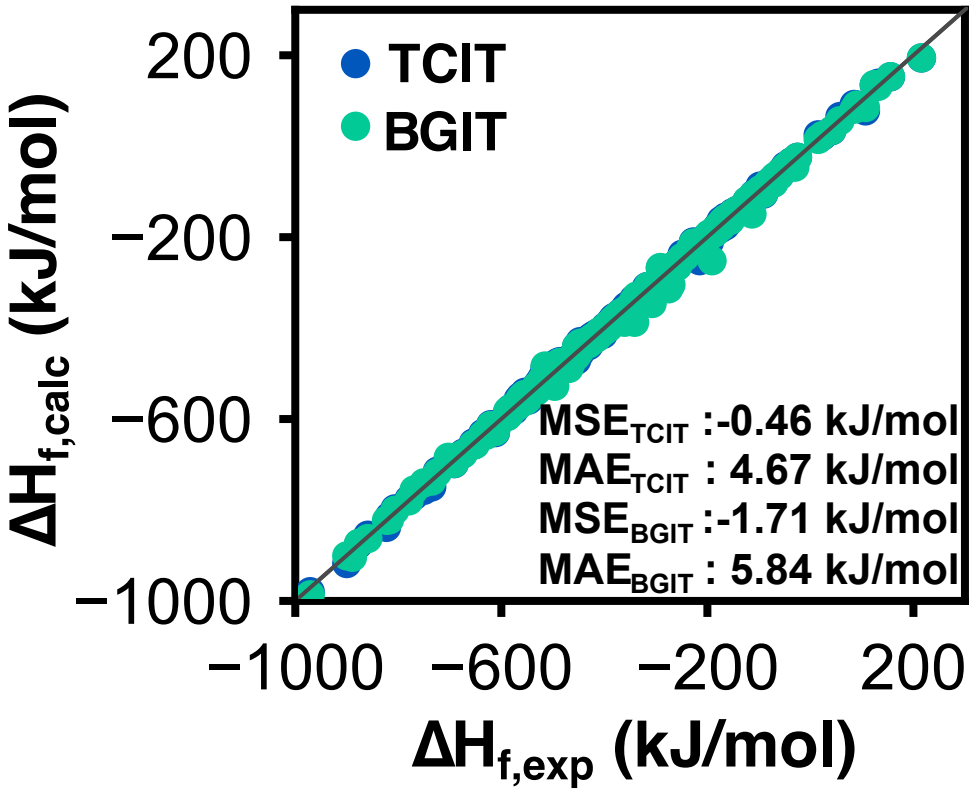
- PNK is a core dataset for fitting Benson groups

- ~600 PNK compounds are small enough for G4 calculations and comparison with experiment.

- ~150 PNK compounds are large enough for direct G4 calculation and comparison with TCIT.

- ~500 PNK compounds are large enough to evaluate the predictive accuracy of the increment theories.

~500 large compounds from PNK



Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a Fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

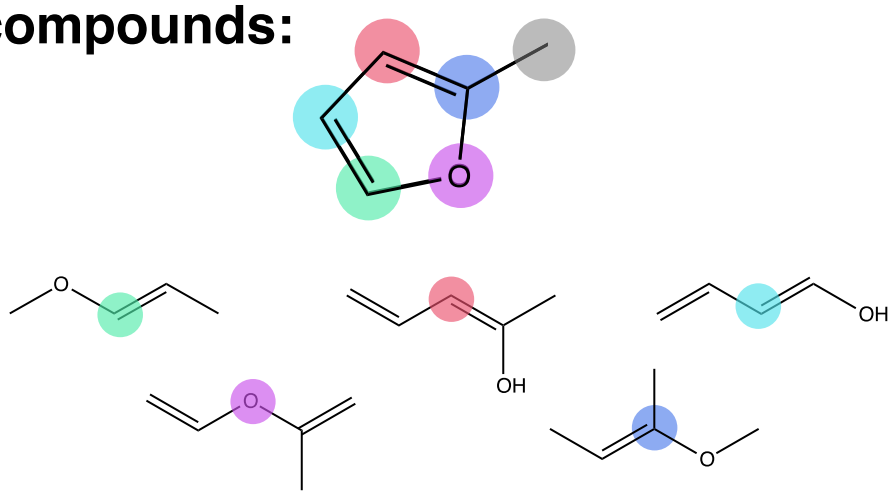
TCIT shows comparable performance to BGIT/CHETAH but is derived exclusively from extensible G4 data.

Extension to Ring-Containing Molecules

- **Ring-containing molecules have additional strain and/or conjugation corrections that exacerbate the extensibility issues of Benson Theory.**

- **In TCIT we are addressing this through chemically specific ring corrections that account for differences in substitution pattern and topology:**

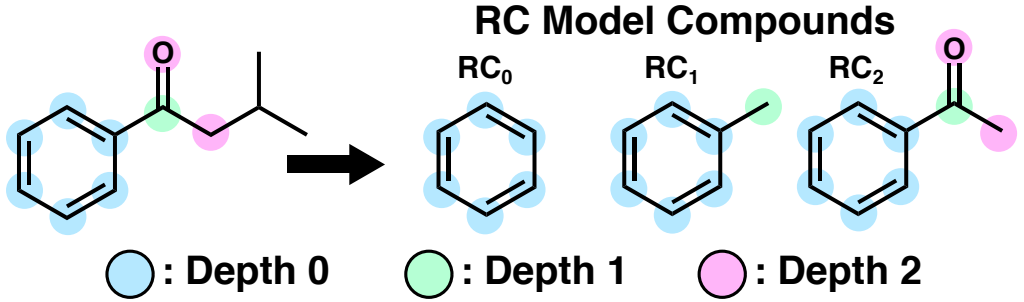
1. Decompose ring into acyclic model compounds:



2. Add ring correction (RC) to final prediction:

$$\mathrm{RC} = H_f(ring) - H_f(\bullet) - H_f(\bullet) - H_f(\bullet) - H_f(\bullet) - H_f(\bullet) - H_f(\bullet)$$

Technical Developments



Method 1: Use RC_1 based model parameterized to G4 data.

Method 2: Use graph-NN to predict RC_0-RC_2

Benchmarking Ring-Correction Performance

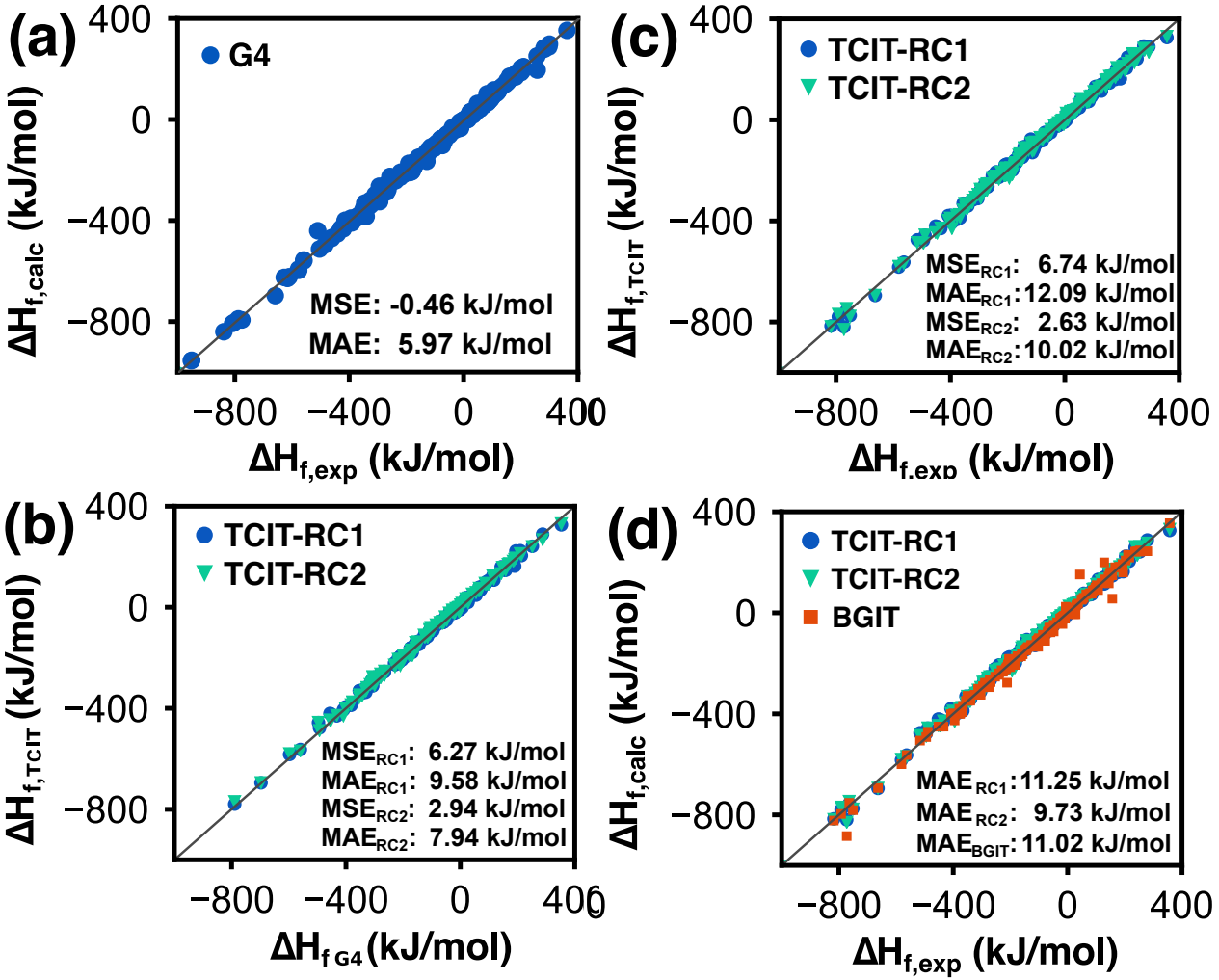
(a) G4 errors are marginally larger for ring-containing compounds but still very accurate

(b) The neural-network based ring-correction exhibits excellent reproduction of the G4 predictions (MSE: ~3kJ/mol; MAE: ~8 kJ/mol).

(c) TCIT is completely transferable to new testing compounds that are experimentally characterized. Errors are consistent with G4 comparison

(d) The TCIT-R2 model outperforms BGIT on the large molecule benchmark while being extensible. Significantly, <u>these compounds are within BGIT's training data</u>.

~120 ring-containing compounds from PNK (excluding training)



BGIT cannot make predictions for ~2% of PNK compounds

Benchmarking Ring-Correction Performance

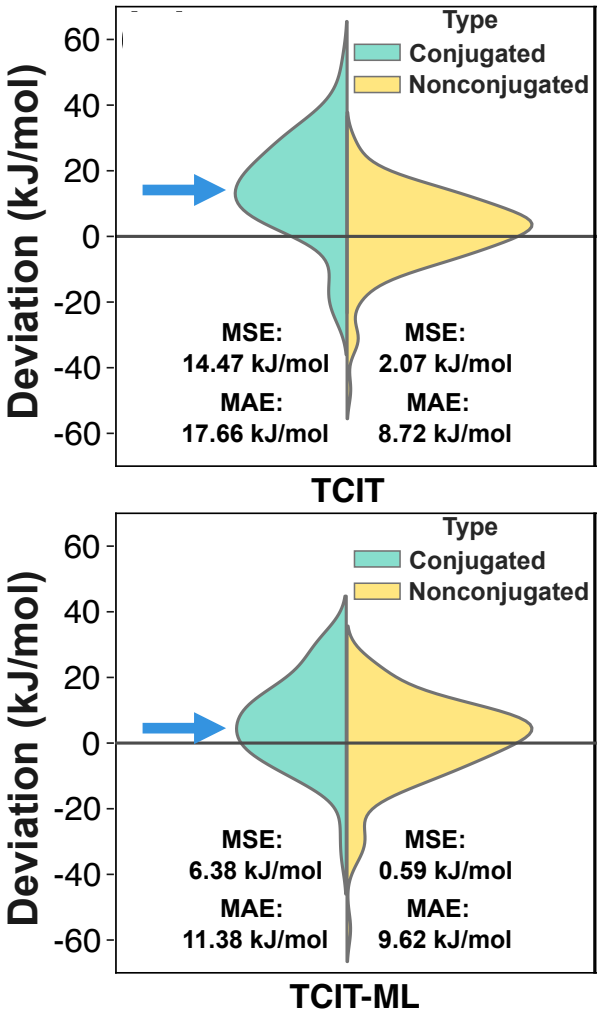
Breaking down distinct subsets:

- Conjugated systems are challenging to accurately predict with a local ring correction.

- BGIT has excellent performance on benzene rings due to the prevalence of experimental data, but poor performance on novel rings.

- The ML ring-correction shows the strongest overall performance. This strategy could also be used to generically correct for long-range conjugation effects.

~120 ring-containing compounds from PNK (excluding training)



BGIT cannot make predictions for ~2% of PNK compounds

TCIT Extension to Other Properties and Phases

Condensed Phases: The condensed-phase and gas-phase standard enthalpies of formation differ by the heats of sublimation and vaporization[1]:

$$\Delta_{\mathrm{f}} H^{\circ}_{(\mathrm{s})} = \Delta_{\mathrm{f}} H^{\circ}_{(\mathrm{g})} - \Delta_{\mathrm{sub}} H^{\circ}$$
$$\Delta_{\mathrm{f}} H^{\circ}_{(\ell)} = \Delta_{\mathrm{f}} H^{\circ}_{(\mathrm{g})} - \Delta_{\mathrm{vap}} H^{\circ}$$

We have implemented group contribution models for heat of vaporization[2] and sublimation[3], respectively. The group assignments and group values associated with these models have been automated within the context of TCIT.

Standard Molar Entropy (S°) and heat capacity (C_v): The molar entropies and constant volume heat capacities are accessible from quantum chemistry using the harmonic oscillator approximation for the molecular partition function and corrections based on the number of rotatable bonds (N_{rot}) and molecular symmetry:

$$S^{\circ} = \langle S^{\circ}_{\mathrm{harm}} \rangle + R N_{\mathrm{rot}} + R \log \sigma \qquad C_{\mathrm{v}} = \langle C_{\mathrm{v,harm}} \rangle + \alpha N_{\mathrm{rot}} + \beta$$

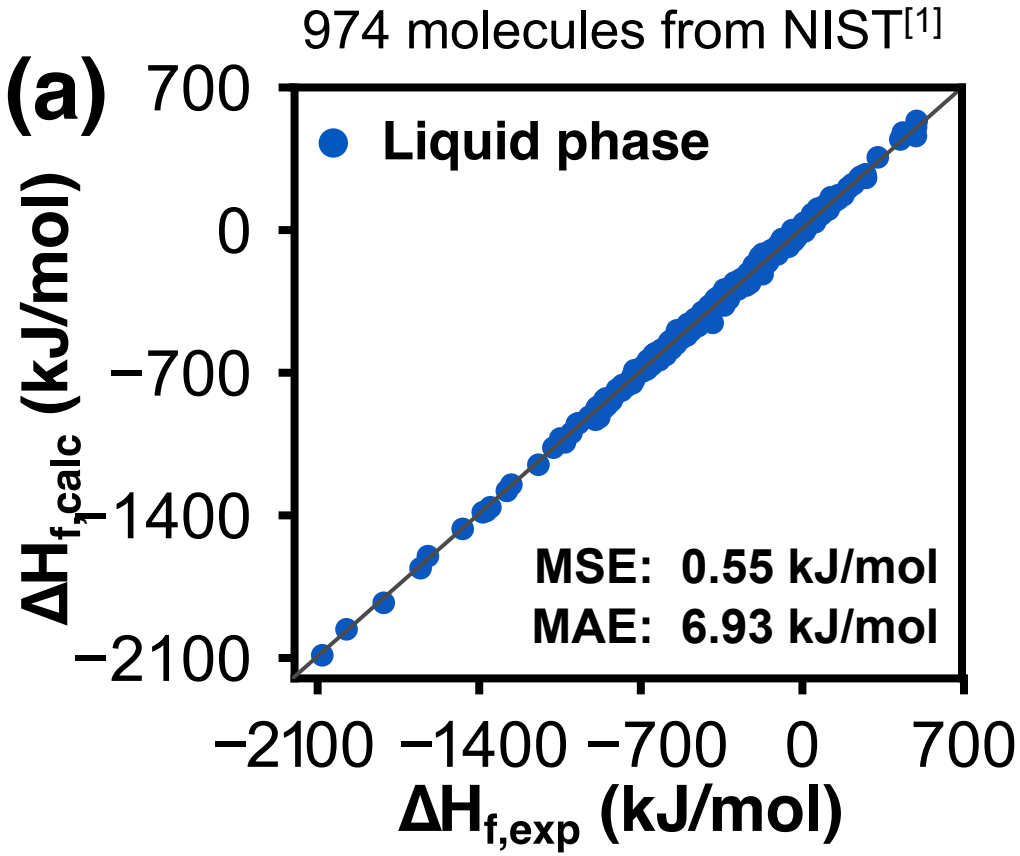
[1] Murray, J.S., Brinck, T. and Politzer, P., **1996**. *Chemical physics*, 204, 289-299.

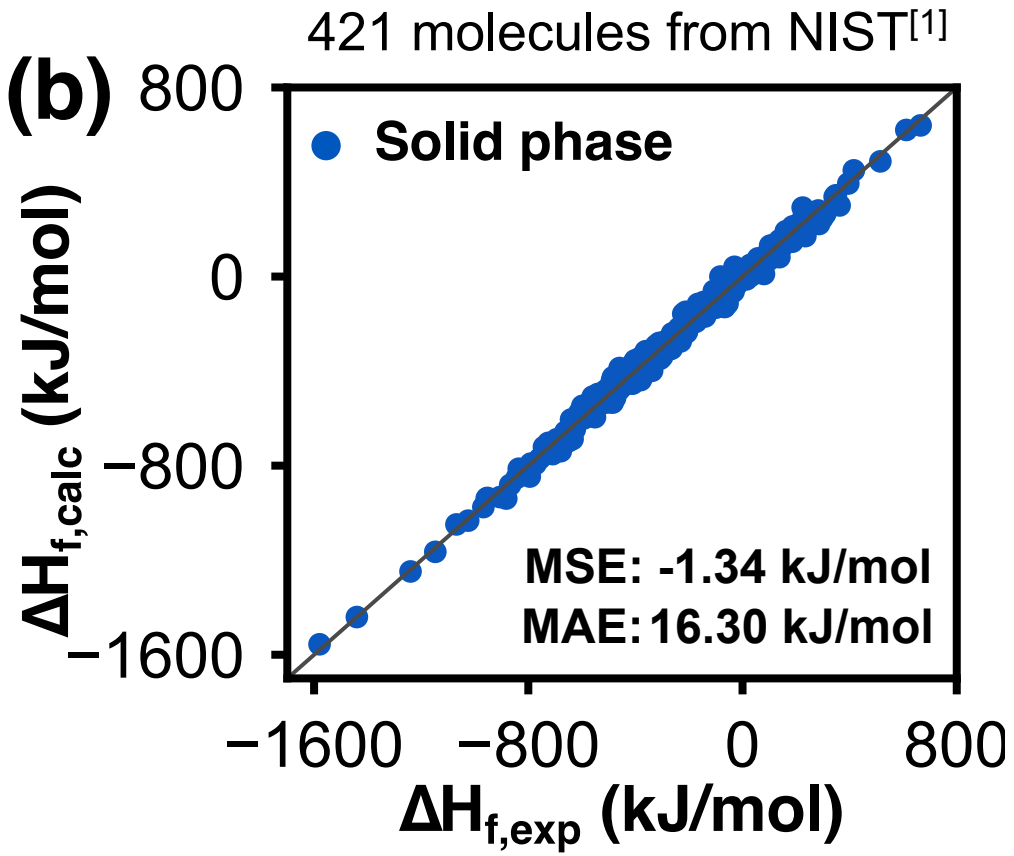
[2] Pankow, J.F. and Asher, W.E., **2008**. *Atmospheric Chemistry and Physics.*

[3] Bagheri, M.; Gandomi, A. H.; Golbraikh, A. **2012**, *Thermochim. Acta*, 543, 96–106

<.> Indicates conformational averaging
R: ideal gas constant
σ: symmetry number
α, β: regressed constants

Benchmarking Condensed Phase ΔH_f Predictions





- Testing set includes both linear and cyclic compounds with number of heavy atoms varying from 1 to 30.

- Low MSE indicates no systematic bias, larger absolute errors result from the quality of the ΔH_{vap} and ΔH_{sub} models.

[1] Linstrom, P.J. and Mallard, W.G., 2001. *Journal of Chemical & Engineering Data*, 46(5), pp.1059-1063.

Benchmarking TCIT S° and C_v Predictions

(a) G4/TCIT S° comparison for 314 medium sized molecules.

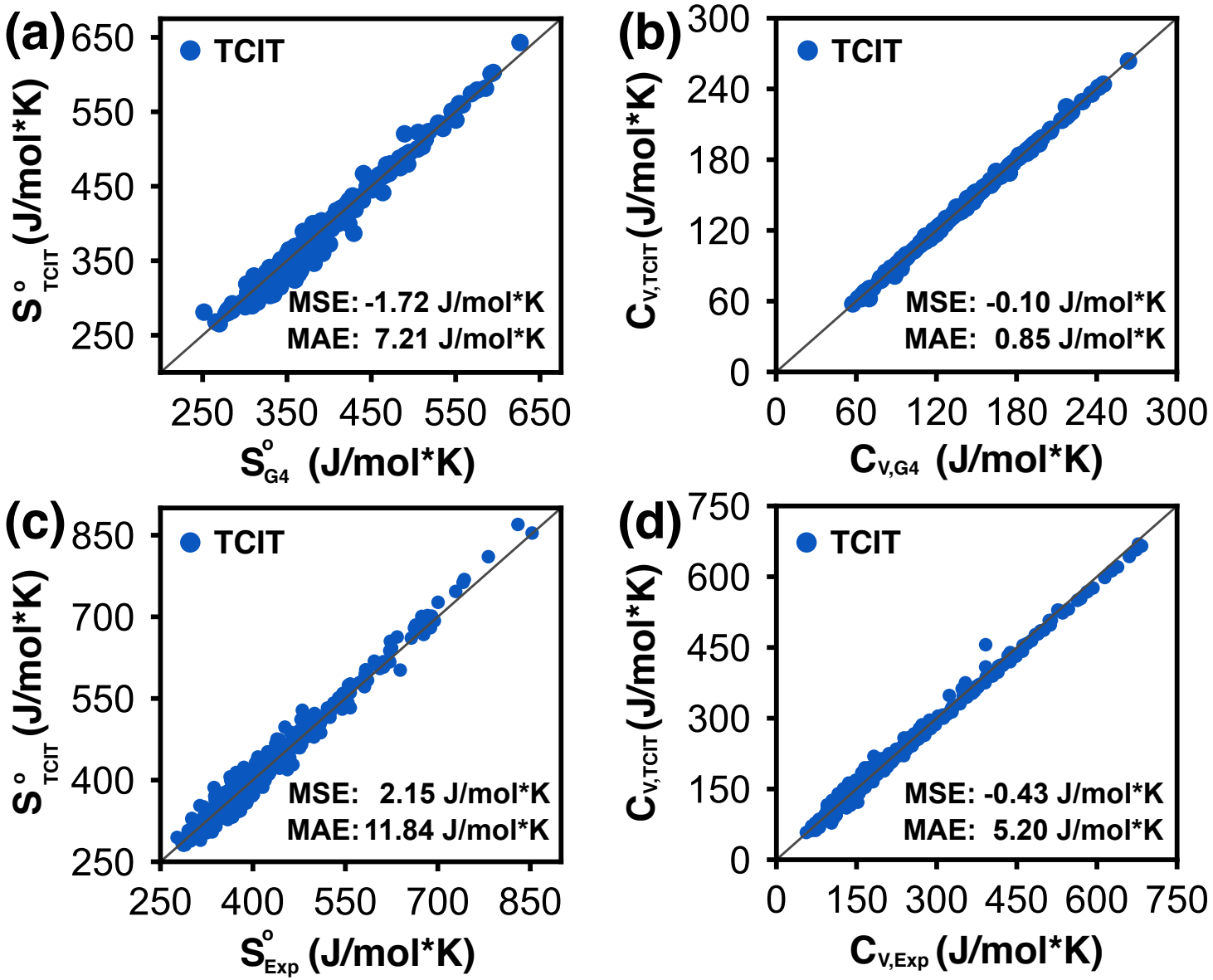
(b) G4/TCIT S° comparison for 314 medium sized molecules.

(c) TCIT S° comparison for 439 large molecules from NIST[1]

(d) TCIT heat capacity comparison for 904 large molecules from NIST[1]

- The TCIT errors are consistent with error propagation of G4:exp and TCIT:G4 errors.

TCIT now supports S° and C_v predictions with accuracies comparable to G4 model chemistry.



[1] Linstrom, P.J. and Mallard, W.G., 2001. *Journal of Chemical & Engineering Data*, 46(5), pp.1059-1063.

Benchmarking TCIT S° and C_v Predictions

(a) G4/TCIT S° comparison for 314 medium sized molecules.

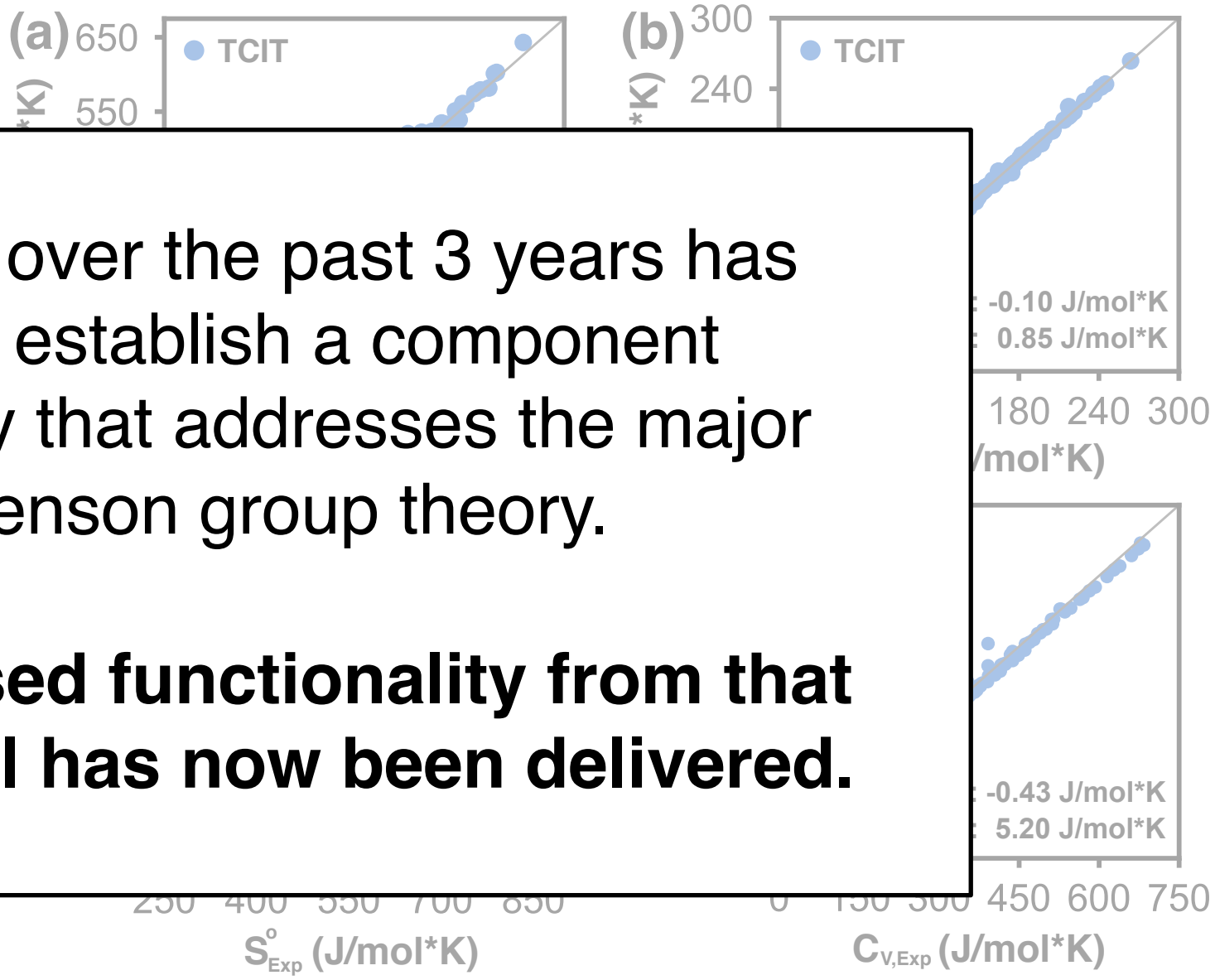
(b) G4/
medium

(c) TCIT
molecul

(d) TCIT
904 larg

• The T
error p
TCIT:G4

TCI
predictions with accuracies
comparable to G4 model chemistry.

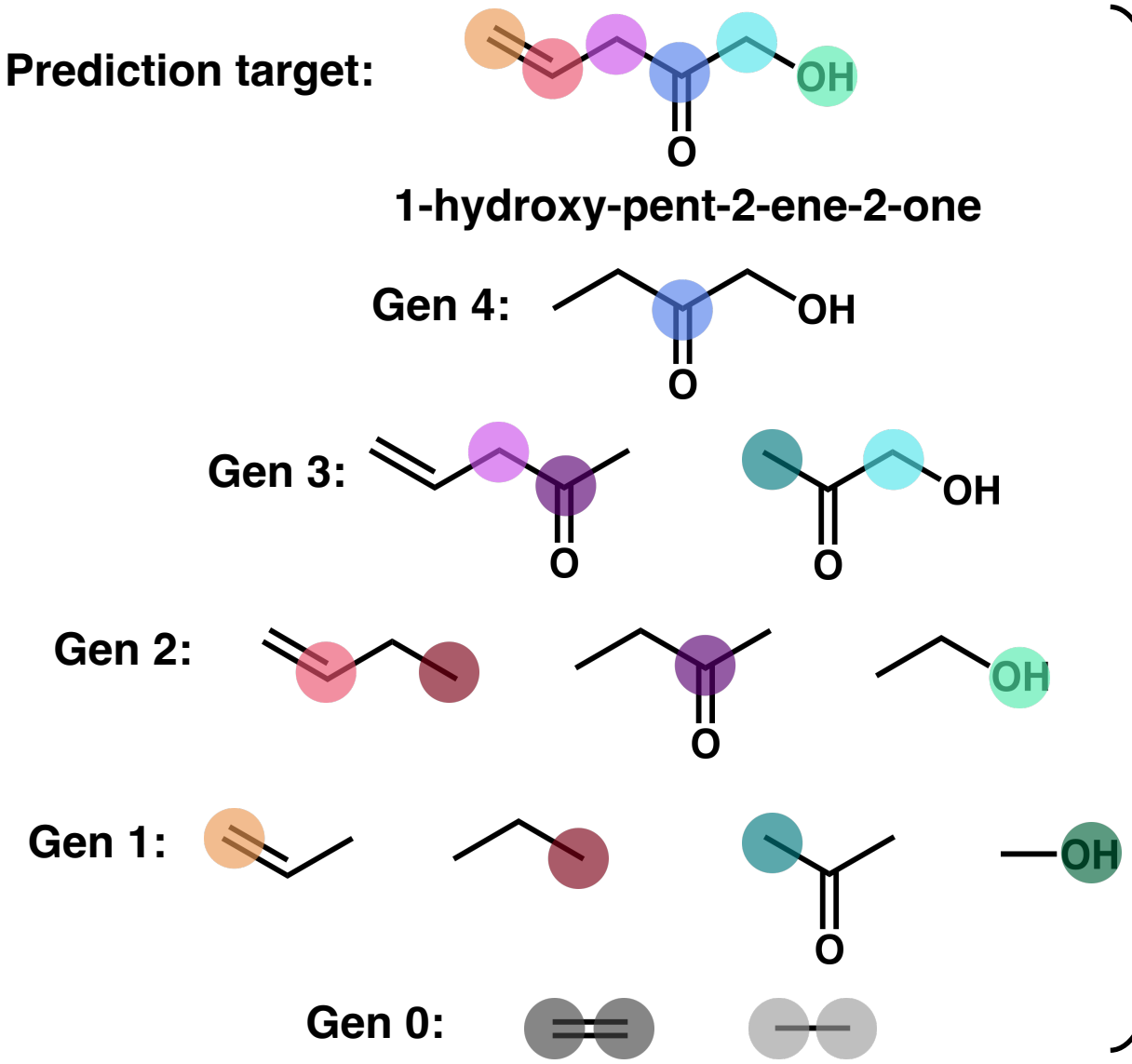


P2SAC funding over the past 3 years has allowed us to establish a component increment theory that addresses the major gaps in Benson group theory.

All of the promised functionality from that original proposal has now been delivered.

[1] Linstrom, P.J. and Mallard, W.G., 2001. *Journal of Chemical & Engineering Data*, 46(5), pp.1059-1063.

How Many Components are Possible?



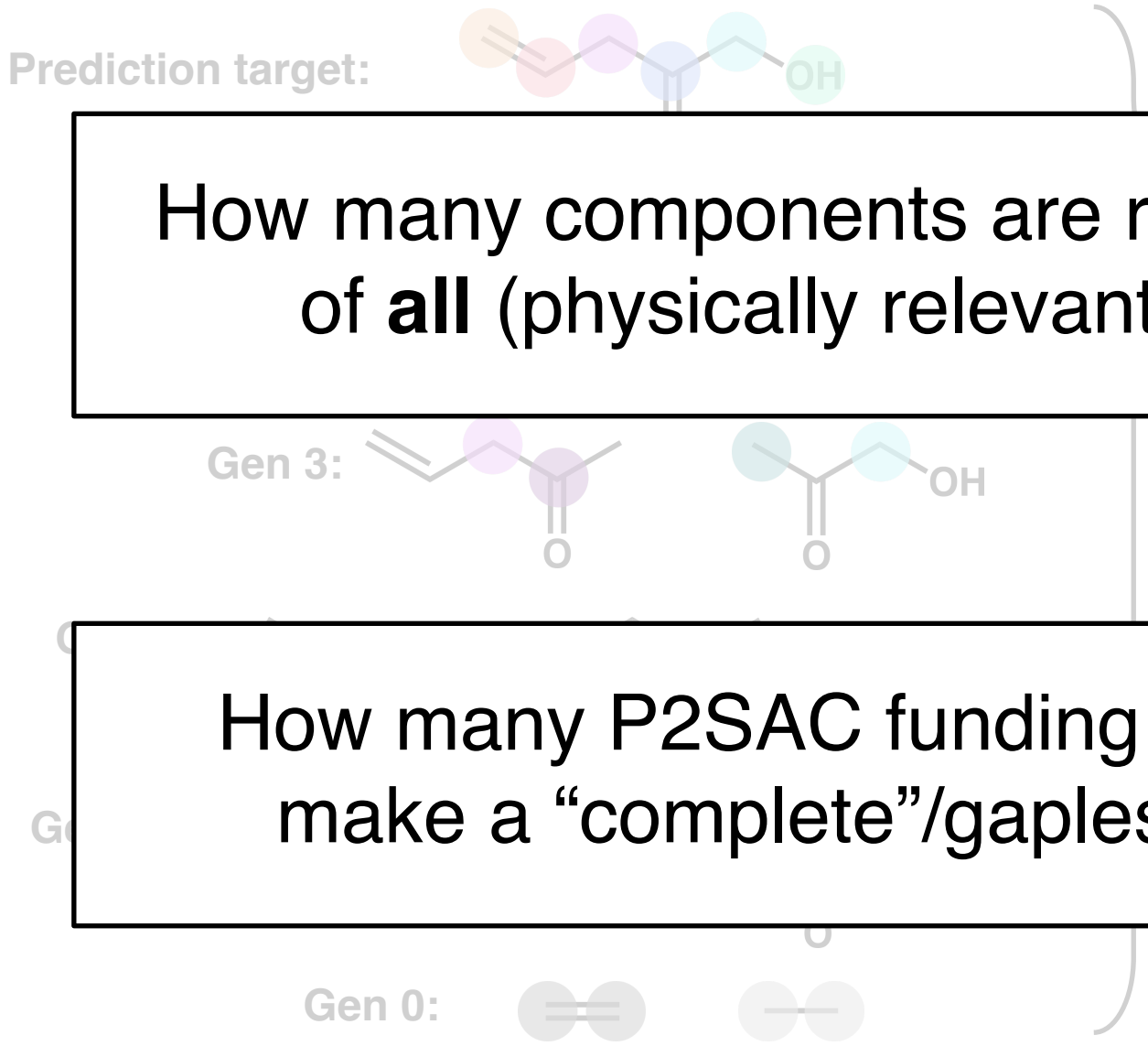
We database all model compounds and components for reuse.

Over the past two years, we have parameterized new components in response to distinct project needs **(<u>many from P2SAC Pharma Members</u>)**

Current Database:

- ~35k distinct components for ΔH_f relevant to organic chemistry
- ~35k distinct G4 calculations on organic molecules.
- ~450 distinct ring corrections

How Many Components are Possible?



We database all model compounds

(many from P2SAC Pharma Members)

• ~450 distinct ring corrections

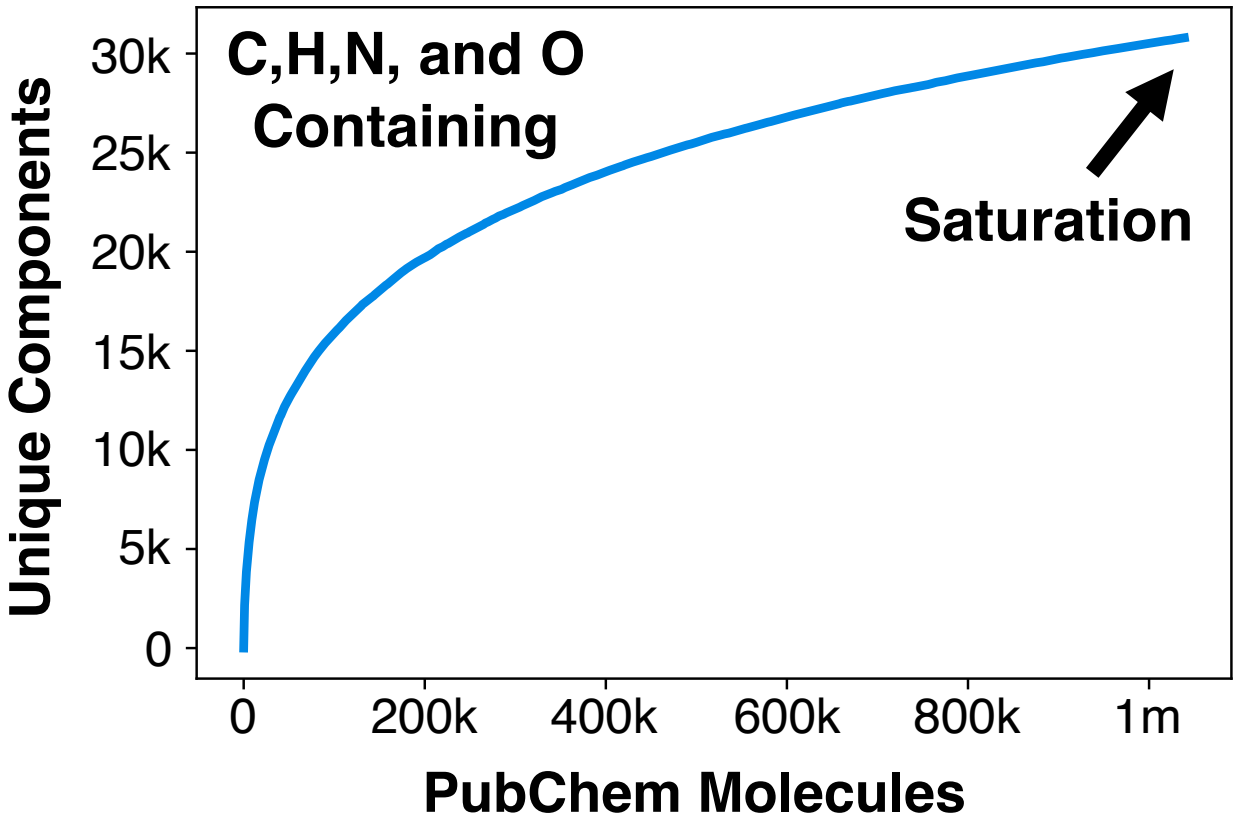
How many components are required to predict the ΔH_f of **all** (physically relevant) organic molecules?

How many P2SAC funding periods would it take to make a “complete”/gapless component theory?

Treating PubChem as a Model of Organic Chemical Space

PubChem is a repository of chemical properties that contains many millions of organic species ranging from small molecules to oligonucleotides.

We recently started mining PubChem's H,C,N, and O containing molecules for distinct components and the model compounds necessary to predict ΔH_f

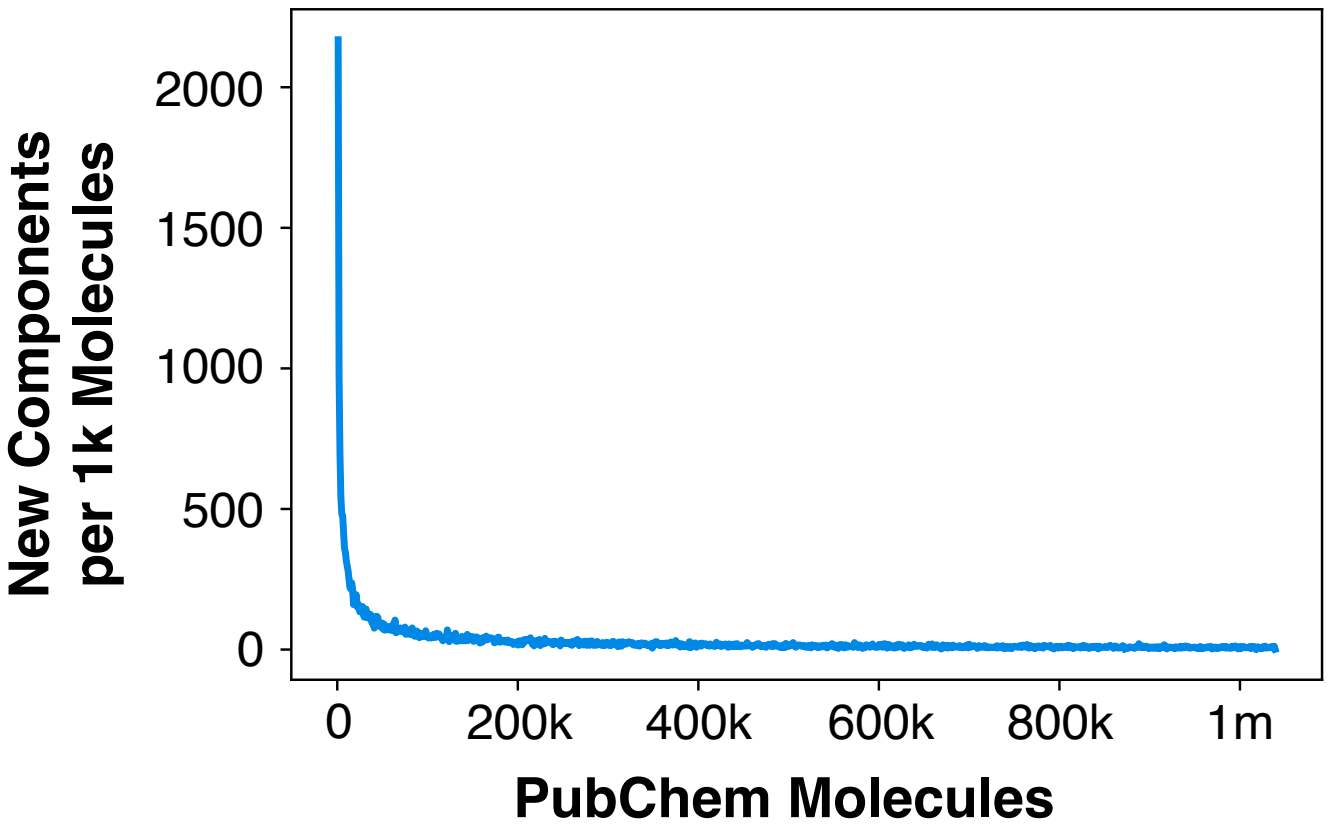


Treating PubChem as a Model of Organic Chemical Space

PubChem is a repository of chemical properties that contains many millions of organic species ranging from small molecules to oligonucleotides.

We recently started mining PubChem's H,C,N, and O containing molecules for distinct components and the model compounds necessary to predict ΔH_f

The derivative plot shows that TCIT initially generates ~2 new components per molecule, but by the end of the sampling ~100 molecules need to be sampled to find a new component.

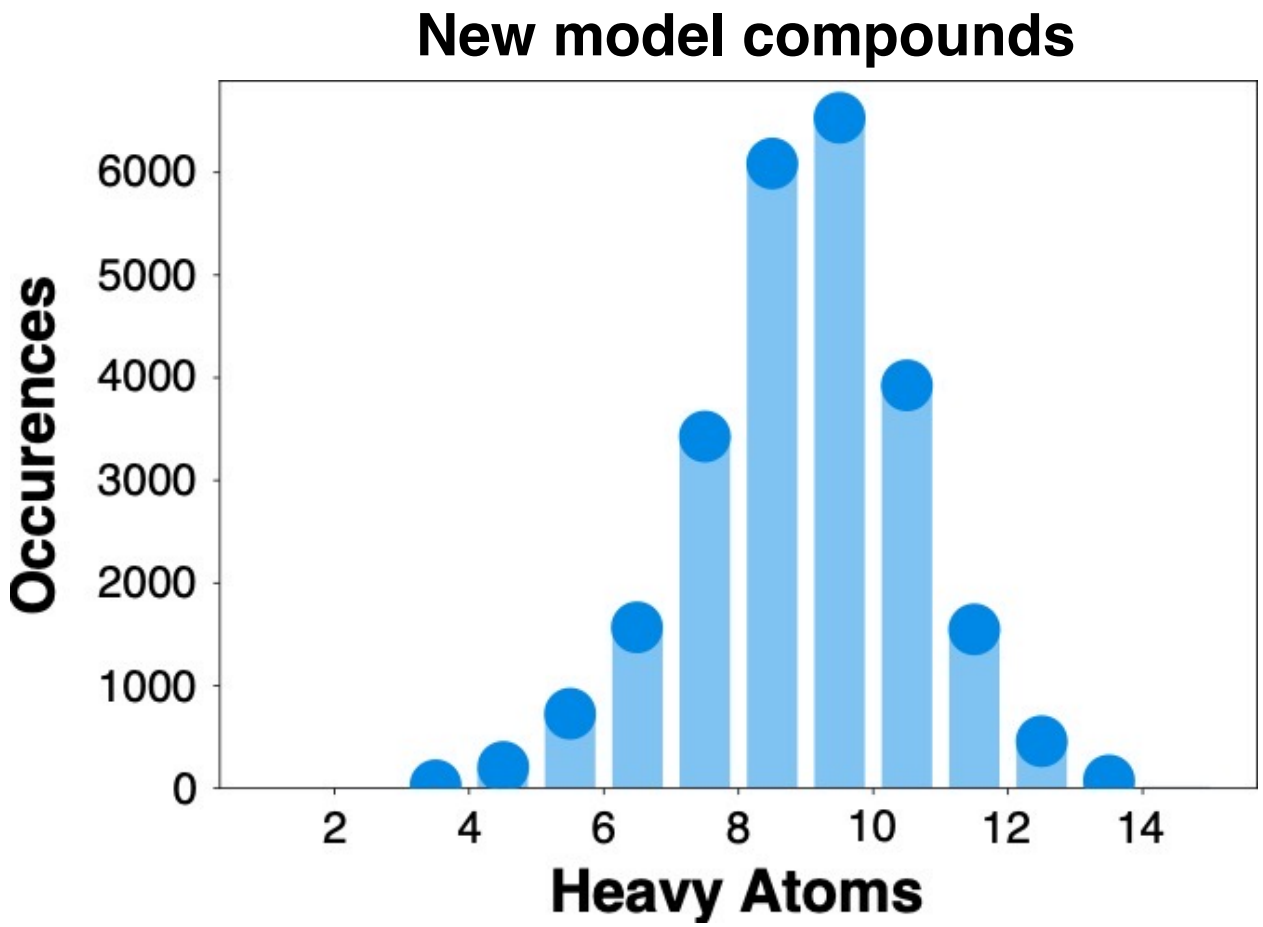


Treating PubChem as a Model of Organic Chemical Space

PubChem is a repository of chemical properties that contains many millions of organic species ranging from small molecules to oligonucleotides.

We recently started mining PubChem's H,C,N, and O containing molecules for distinct components and the model compounds necessary to predict ΔH_f

The derivative plot shows that TCIT initially generates ~2 new components per molecule, but by the end of the sampling ~100 molecules need to be sampled to find a new component.



Treating PubChem as a Model of Organic Chemical Space

In the past six months we have generated all training data necessary to make predictions on all N, H, O, and C-containing molecules in pubchem. **This is the largest repository of G4 calculations on large molecules in the world.**

It is foreseeable that we could complete all B, F, Cl, S, and P containing structures over the next few years.

“Known Unknowns” and “Unknown Unknowns”

A → B

- To safely plan a known reaction, we need access to solid thermodynamic data (e.g., ΔH_f, S^o, C_v) to understand and classify risks.

- This is a “known unknown” in that we know the reaction, A → B, but we need values for a few unknown variables.

A → ? → B ; A → B + ? ; A → ?

- **A → ? → B**, means that we know the net reaction, but there may be a consequential (e.g., potentially reactive) intermediate. Even if we have accurate thermodynamic data on A/B, neglecting the intermediate could be disastrous.

- The **A → B + ?** (unknown side-reaction) and **A → ?** (unknown main product), problems have similar “unknown unknown” characteristics.

The Reaction Prediction Problem

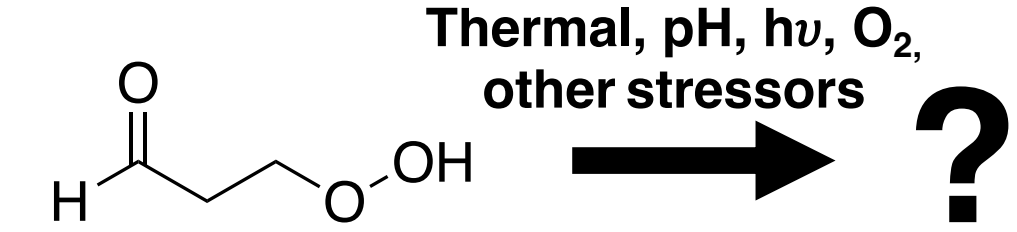
A → B : When we know the reactants and products, mature quantum chemistry tools exist to characterize transition states and establish pathways

A → ? : For degradation reactions, plausible reactions are often unknown.

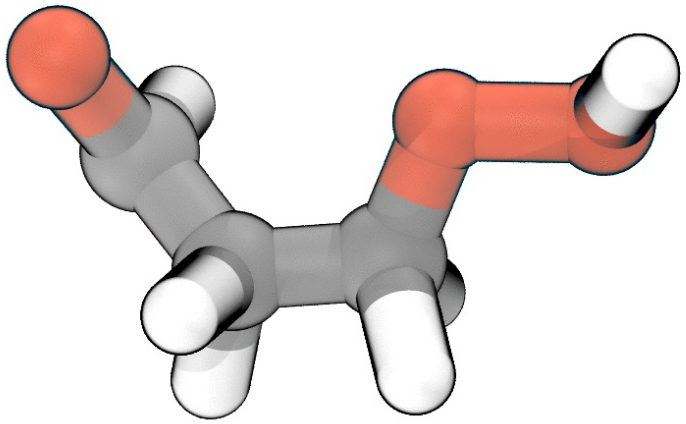
The Reaction Prediction Problem

A → B : When we know the reactants and products, mature quantum chemistry tools exist to characterize transition states and establish pathways

A → ? : For degradation reactions, plausible reactions are often unknown.



3-hydroperoxypropanal

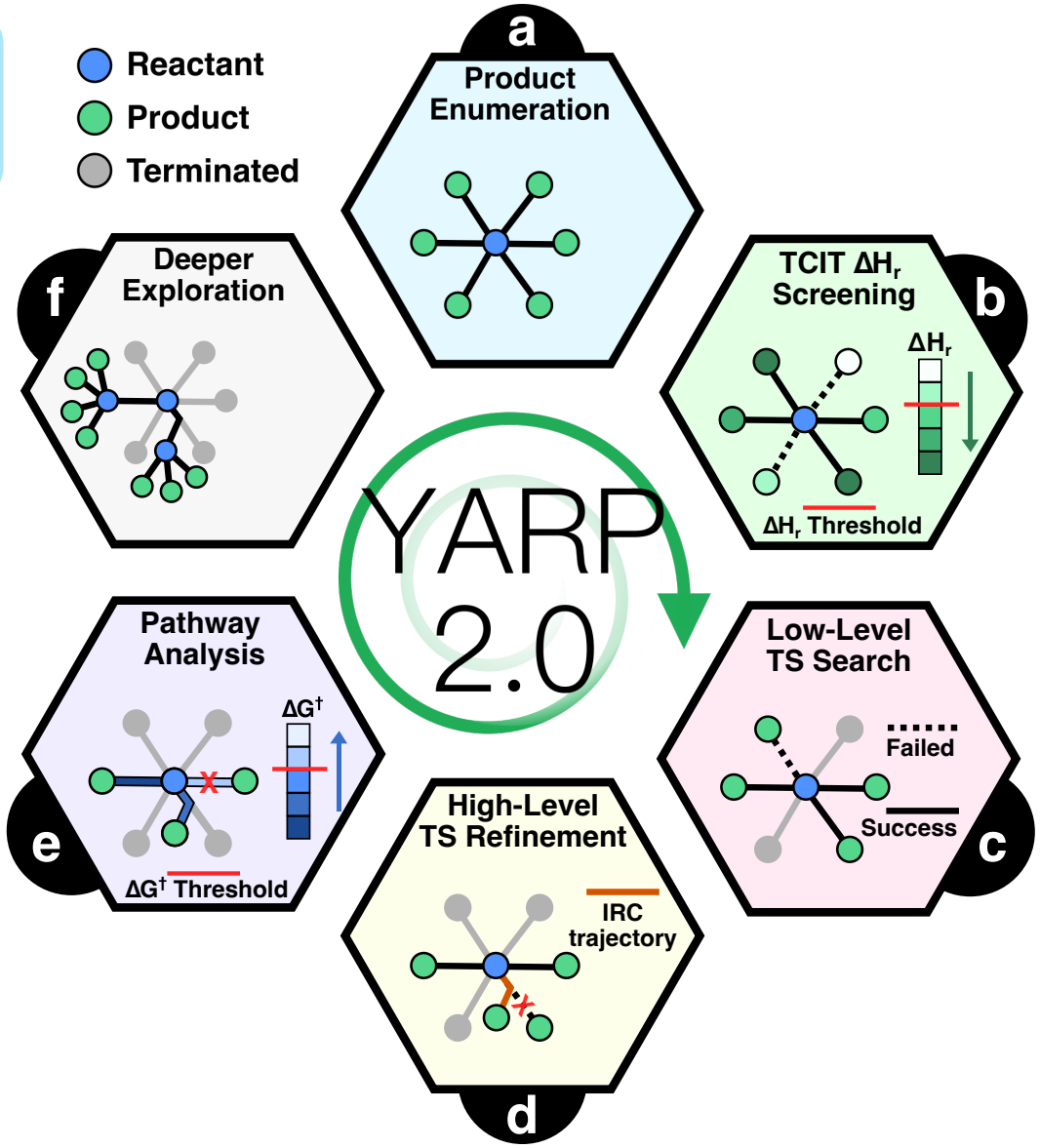


Yet Another Reaction Program (YARP)

Idea: Turn the **A→?** problem into tractable (and parallelizable) **A→B** problems.

Observations:

- Product enumeration is easier than transition state enumeration.
- Transition state algorithms for A→B problems are mature. Let the TS algorithm identify physical reactions.
- Recent developments in semi-empirical models and ML create opportunities.
- If you are fast enough, you can brute force the **A→?** problem.



Zhao, Q.; Savoie, B. M. *Nature Comput. Sci.* 2021 ; Zhao, Q.; Savoie, B. M. *Angew. Chemie* 2022.

Converting Reactions into a Machine-Readable Grammar

Bond-Electron Matrix Formalism: matrix representation of molecules with bond order indicated in off-diagonal elements and lone electrons along the diagonal.

Ugi, I. et al. "New Applications of Computers in Chemistry." *Angew. Chem.* **1979**, 18 (2), 111–123.

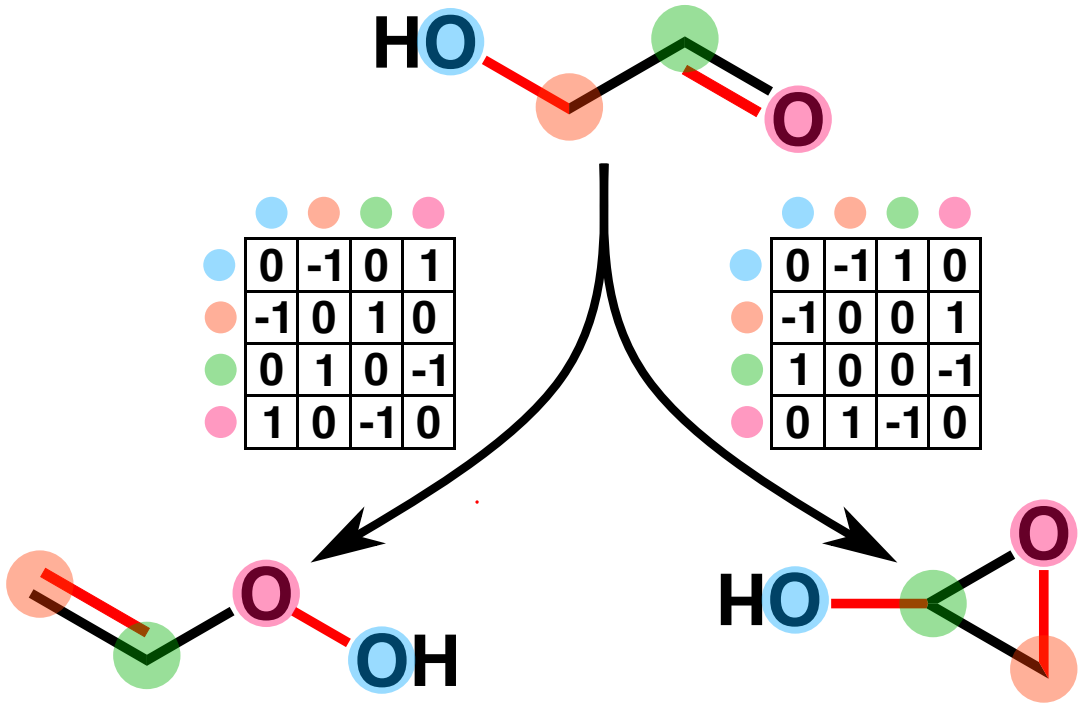
R = B-A

$$\begin{array}{c} \mathbf{O} \\ \mathbf{C} \\ \mathbf{C} \\ \mathbf{H} \\ \mathbf{H} \\ \mathbf{O} \\ \mathbf{H} \\ \mathbf{H} \end{array} \begin{bmatrix} 4 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{array}{c} \mathbf{O} \\ \mathbf{C} \\ \mathbf{C} \\ \mathbf{H} \\ \mathbf{H} \\ \mathbf{O} \\ \mathbf{H} \\ \mathbf{H} \end{array} \begin{bmatrix} 4 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{array}{c} \mathbf{O} \\ \mathbf{C} \\ \mathbf{C} \\ \mathbf{H} \\ \mathbf{H} \\ \mathbf{O} \\ \mathbf{H} \\ \mathbf{H} \end{array} \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

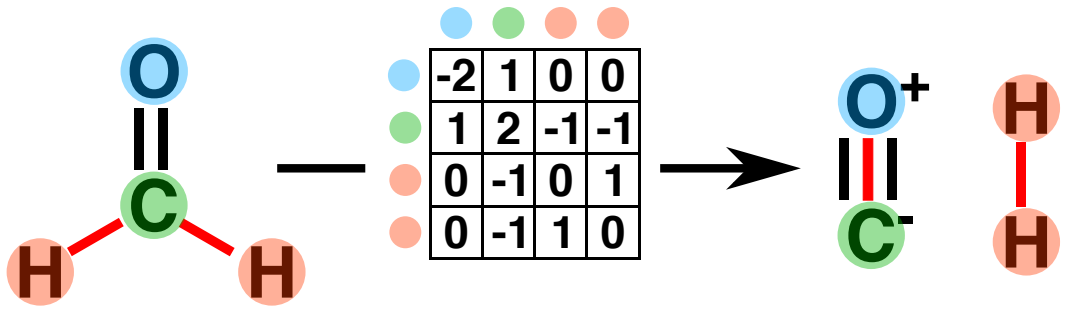
This is essentially a way of converting arrow pushing into something we can automate and interpret by a program

YARP: Elementary Reaction Step(s)

For full-octet neutral organic molecules, “break 2 bonds form 2 bonds” (**b2f2**) is the simplest ERS that yields non-trivial closed-shell neutral products.



Example with concomitant e-transfer

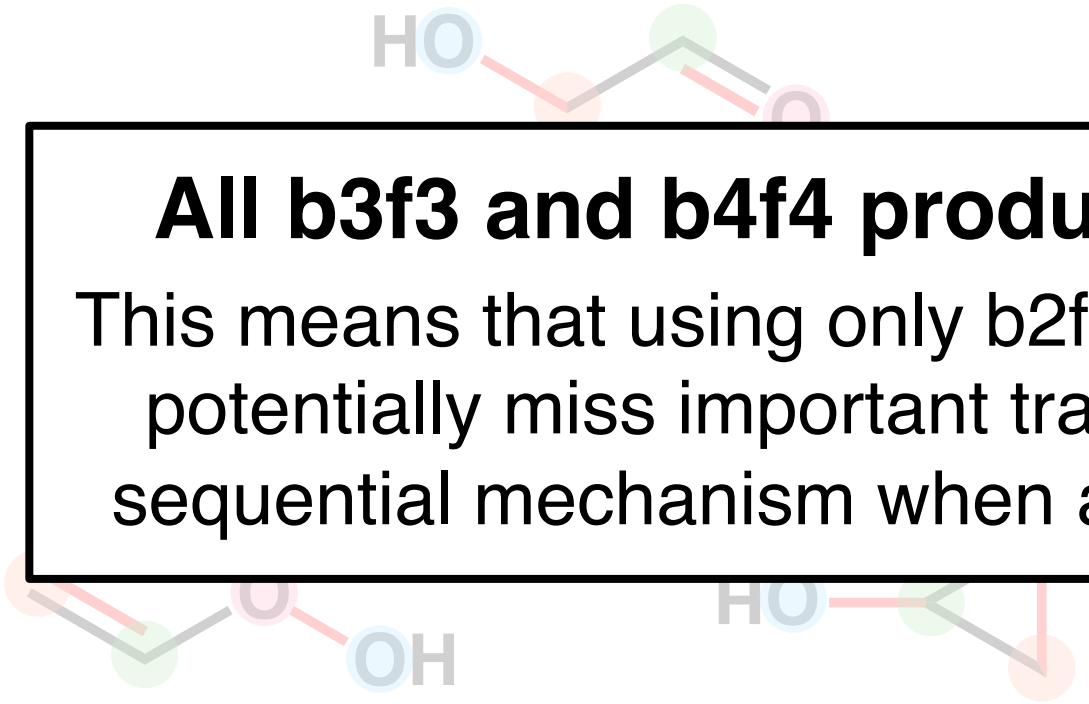


Note: for each pair of broken bonds two distinct reactions are possible

YARP supports a more general suite of ERS(s), but **b2f2** strikes a useful balance that scales as N^2 with reactant size and generates **b1** and **b2f1** type products for free.

YARP: Elementary Reaction Step(s)

For full-octet neutral organic molecules, “break 2 bonds form 2 bonds” (**b2f2**) is the simplest ERS that yields non-trivial closed-shell neutral products.



Example with concomitant e-transfer

Note: for each pair of broken bonds two distinct reactions are possible

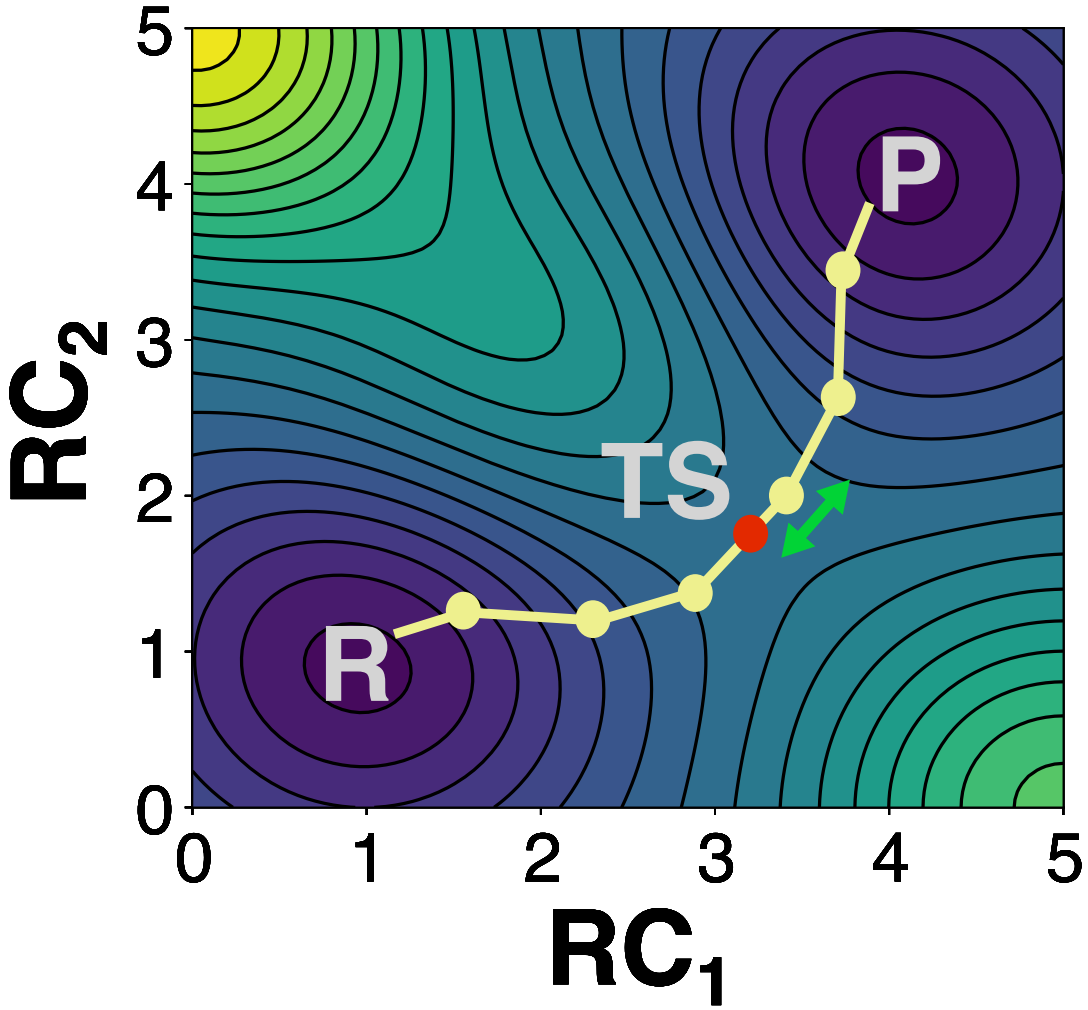
All b3f3 and b4f4 products are b2f2 decomposable

This means that using only b2f2 won’t miss any products, but it will potentially miss important transition states (i.e., by predicting a sequential mechanism when a concerted mechanism is favored)

YARP supports a more general suite of ERS(s), but **b2f2** strikes a useful balance that scales as N^2 with reactant size and generates **b1** and **b2f1** type products for free.

YARP: Pseudo-1D Transition State Searches

Searching for saddle points in 3N-6 space is expensive.



Double-Ended Searches: Using the product and reactant geometries as a constraint, the search can be recast as an effective 1-D search about the connecting coordinate.

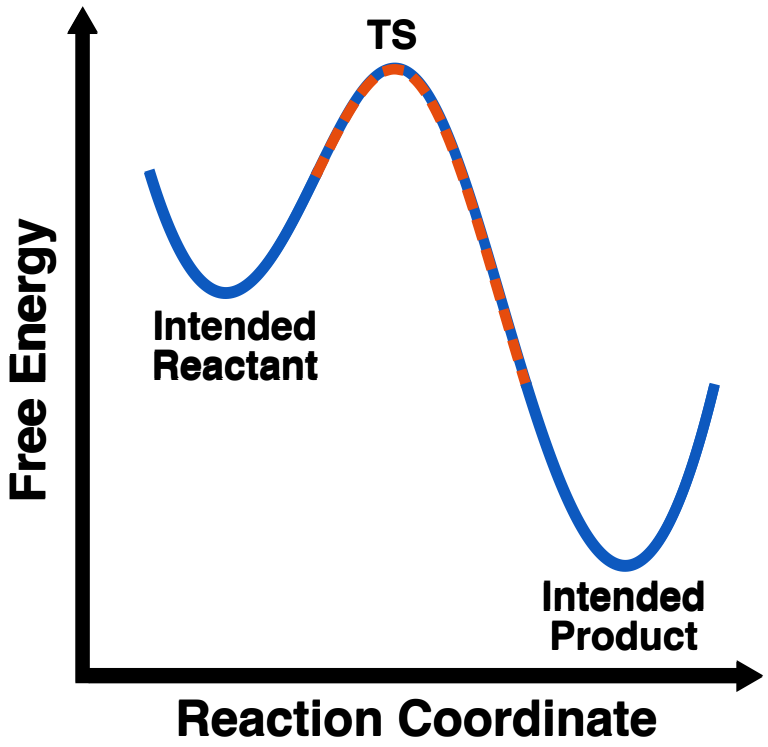
Many Flavors:

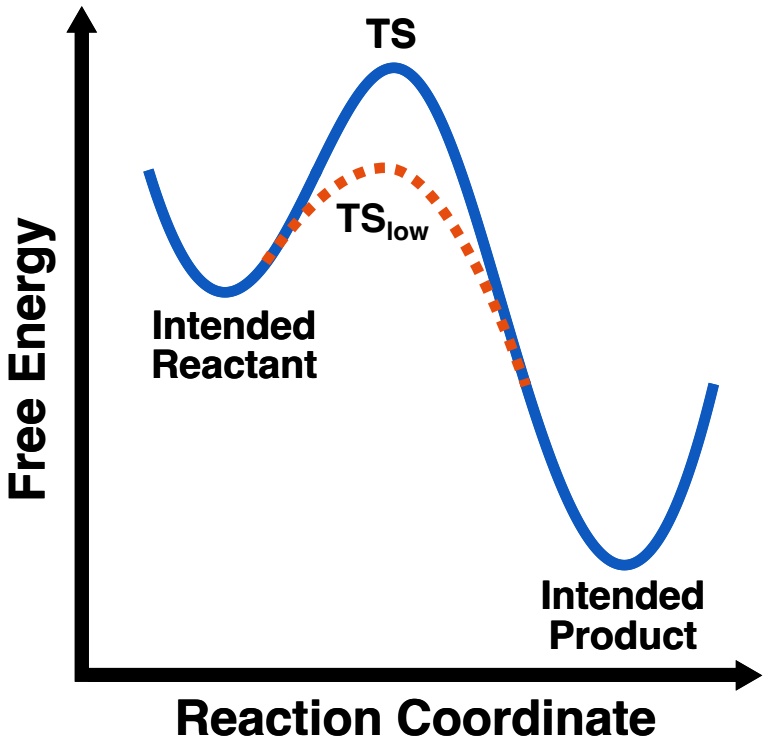
- (Growing, Freezing) String methods
- Nudged-Elastic band methods

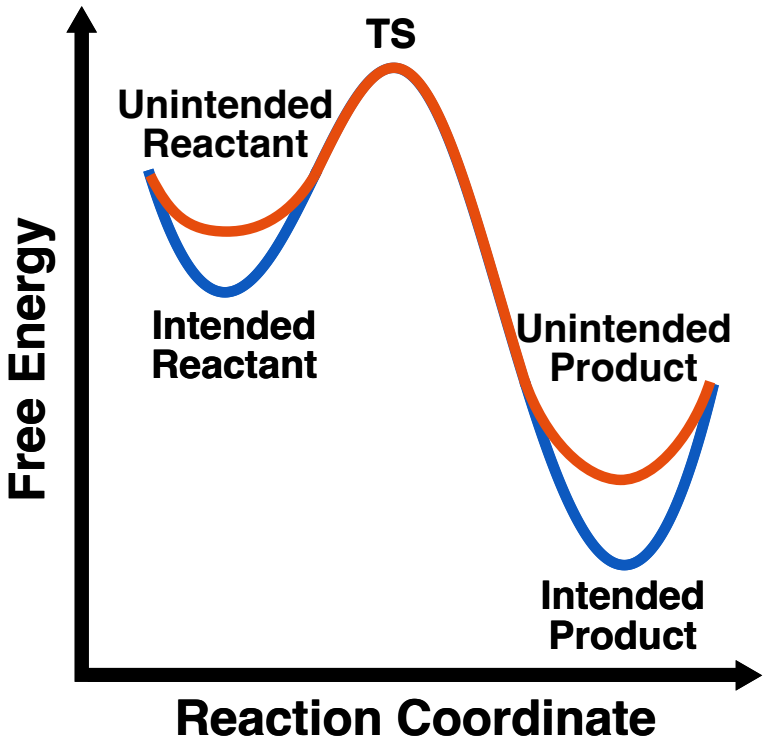
Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, 120 (17), 7877–7886.

We like strings because they avoid the bad initial pathway problems of bands.

Three Sources of Error in TS Searches







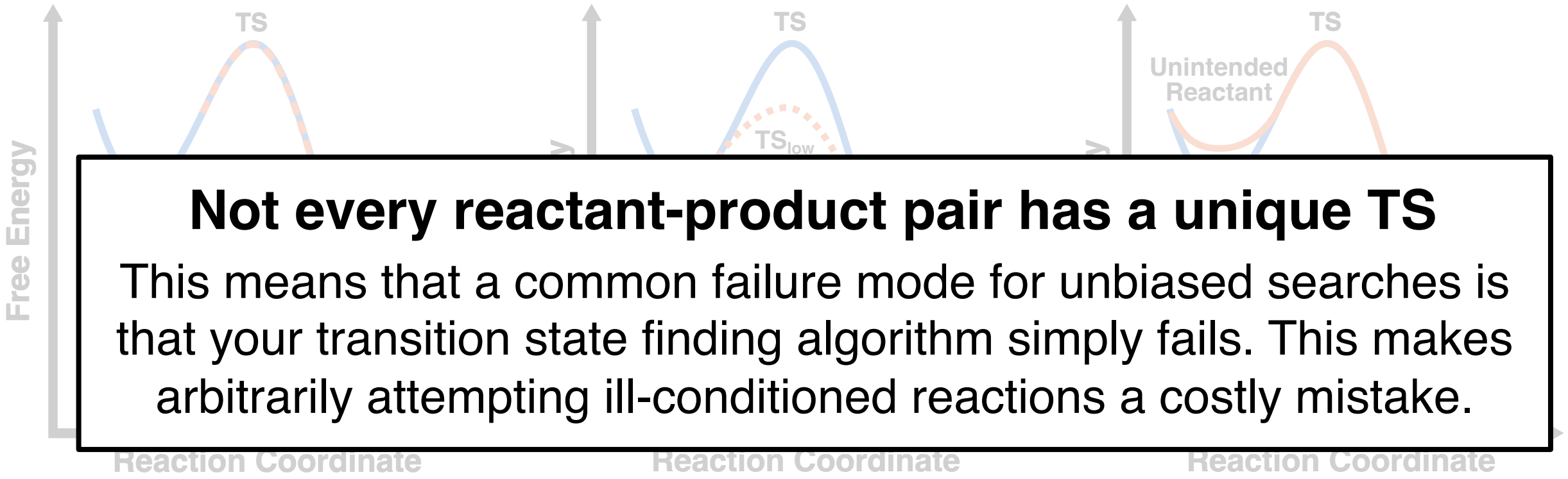
Failure to localize a transition state for a given A→B reaction

Failure to localize the kinetically relevant TS for a given reaction

Failure to localize an intended TS for a given reaction

Zhao, Q.; Savoie, B. M. *Nature Comput. Sci.* 2021 ; Zhao, Q.; Savoie, B. M. *Angew. Chemie* 2022.

Three Sources of Error in TS Searches



Not every reactant-product pair has a unique TS

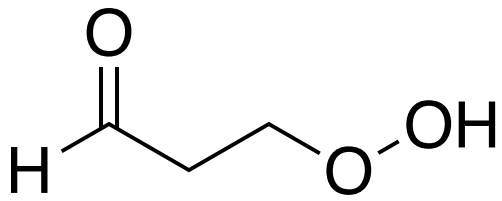
This means that a common failure mode for unbiased searches is that your transition state finding algorithm simply fails. This makes arbitrarily attempting ill-conditioned reactions a costly mistake.

Failure to localize a transition state for a given A→B reaction

Failure to localize the kinetically relevant TS for a given reaction

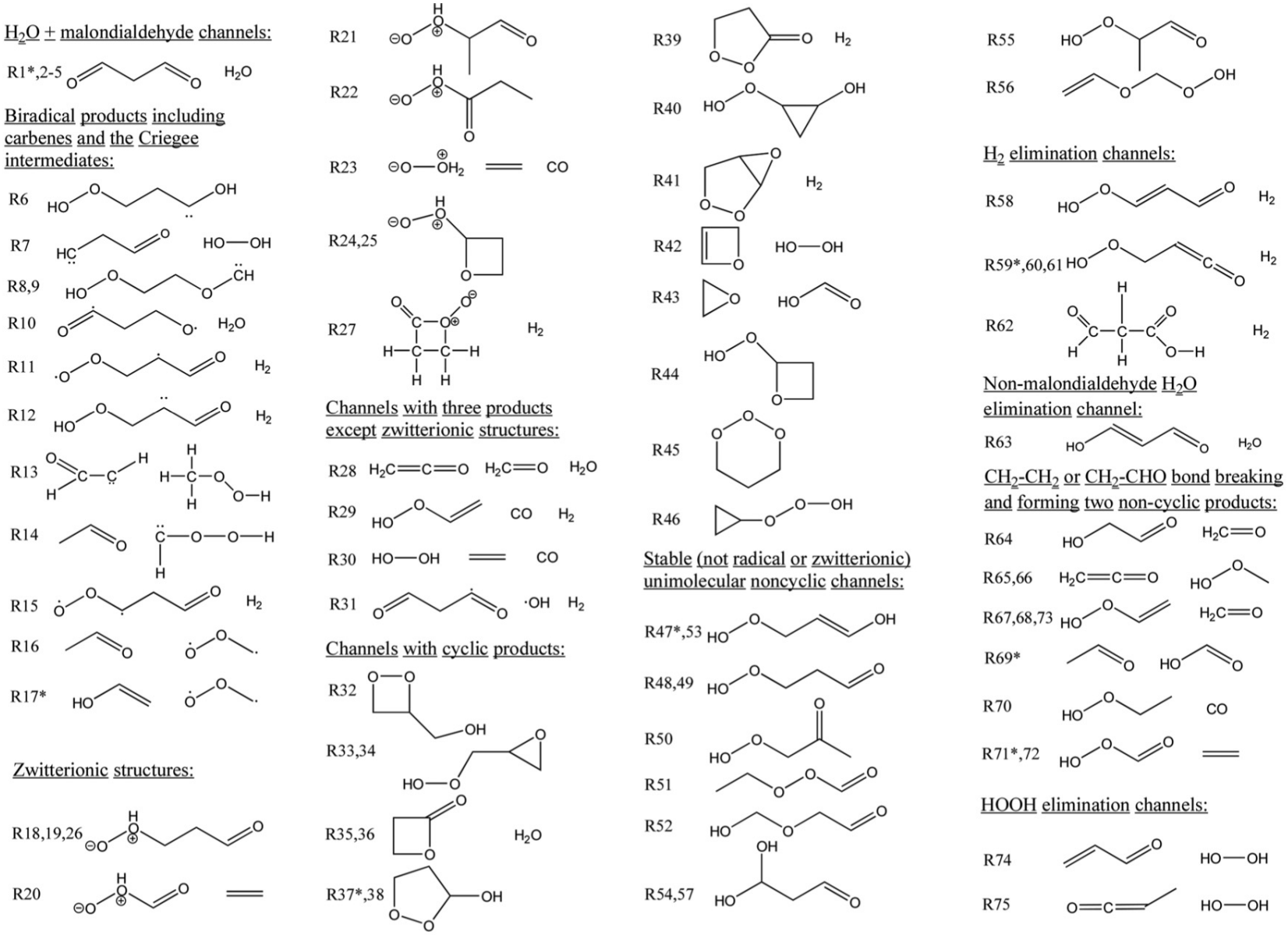
Failure to localize an intended TS for a given reaction

Testing YARP on a Unimolecular Decomposition Problem

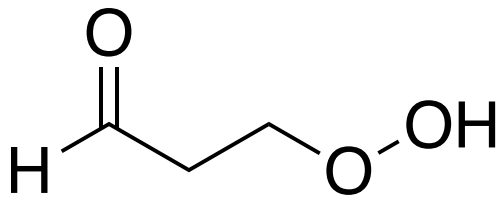


The 3-hydroperoxypropanal reaction network out to b4f4 was recently published as a benchmark for 5 reaction discovery methods.

Grambow, C. A, Suleimanov, Y. V. et al. *J. Am. Chem. Soc.* **2018**, 140 (3), 1035–1048.

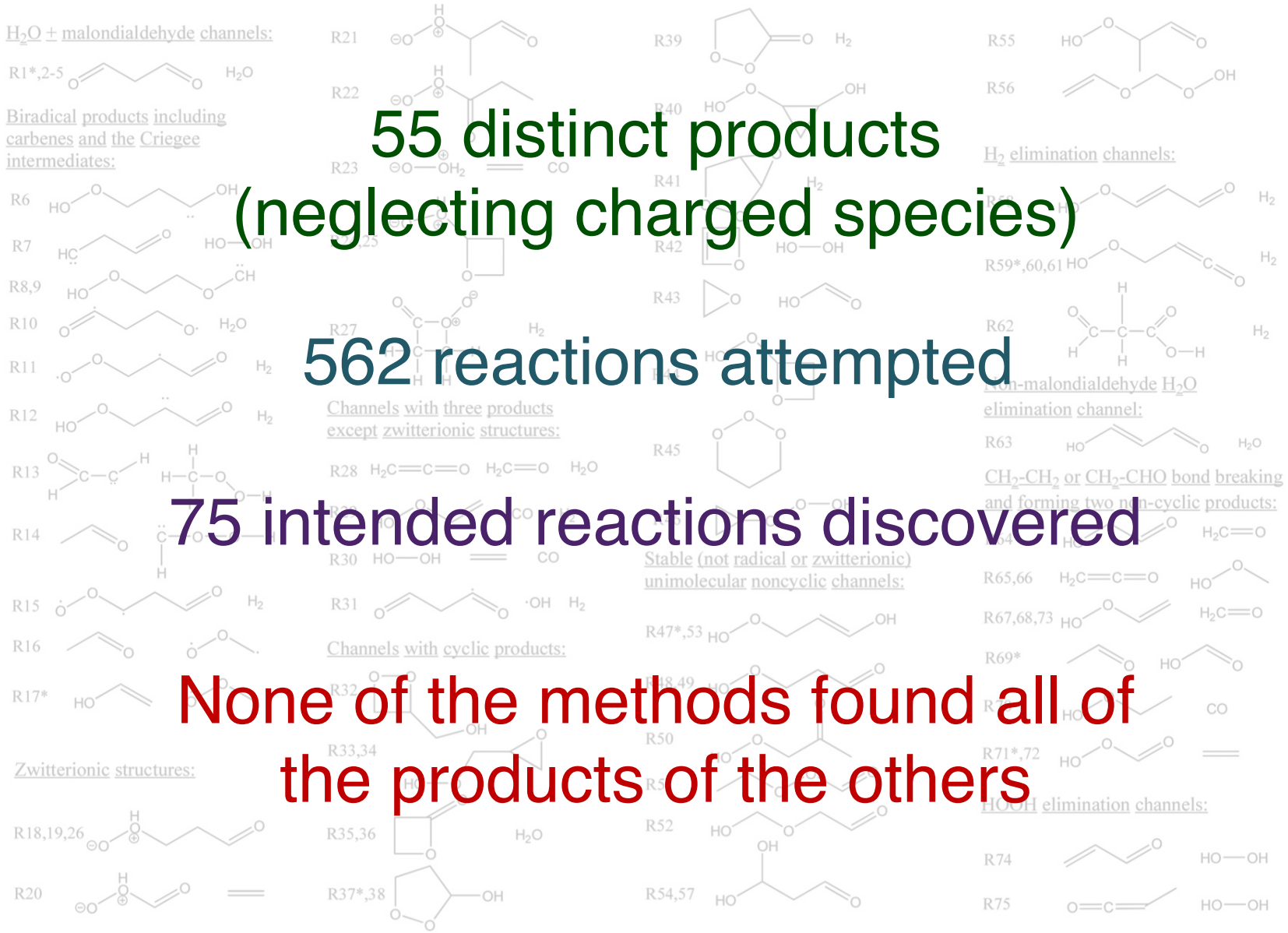


Testing YARP on a Unimolecular Decomposition Problem



The 3-hydroperoxypropanal reaction network out to b4f4 was recently published as a benchmark for 5 reaction discovery methods.

Grambow, C. A, Suleimanov, Y. V. et al. *J. Am. Chem. Soc.* **2018**, 140 (3), 1035–1048.



55 distinct products (neglecting charged species)

562 reactions attempted

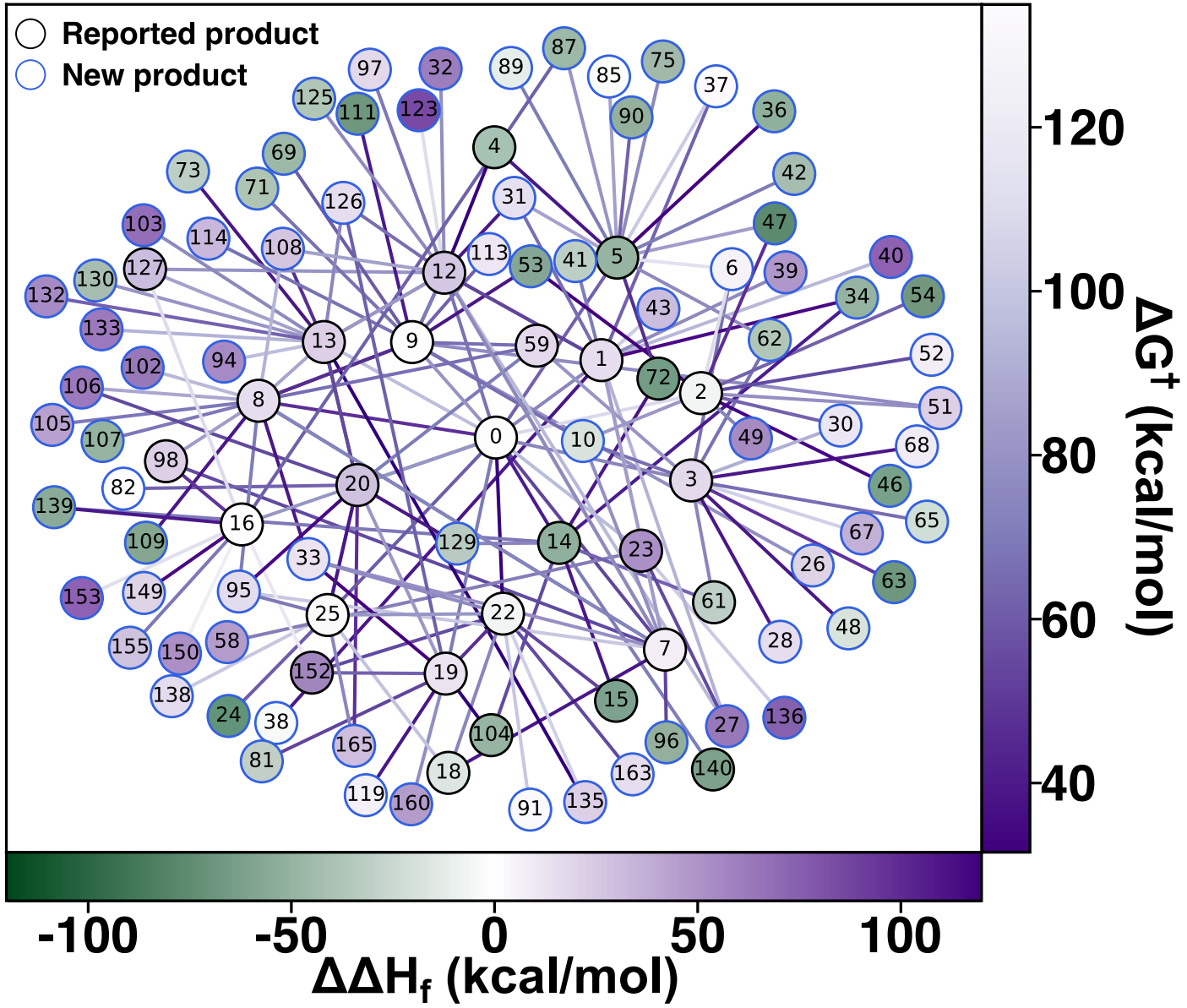
75 intended reactions discovered

None of the methods found all of the products of the others

3-Hydroperoxypropanal - Reaction Network

We used YARP to recursively elucidate the 3-hydroperoxypropanal unimolecular thermal degradation network for comparison with Grambow et al.

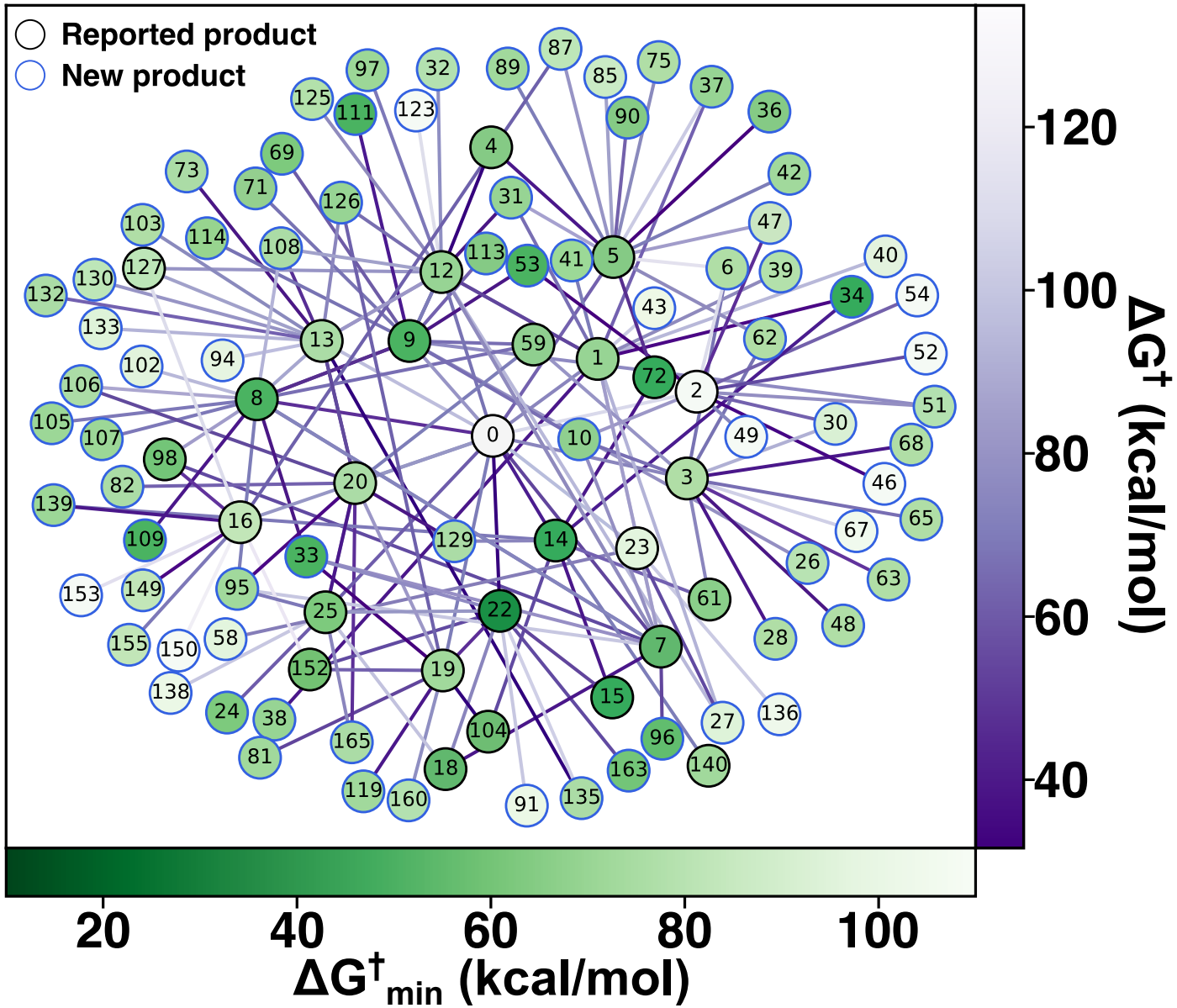
YARP finds **all known products** of this thermal decomposition network, as well as new products (77), and new reactions (157).



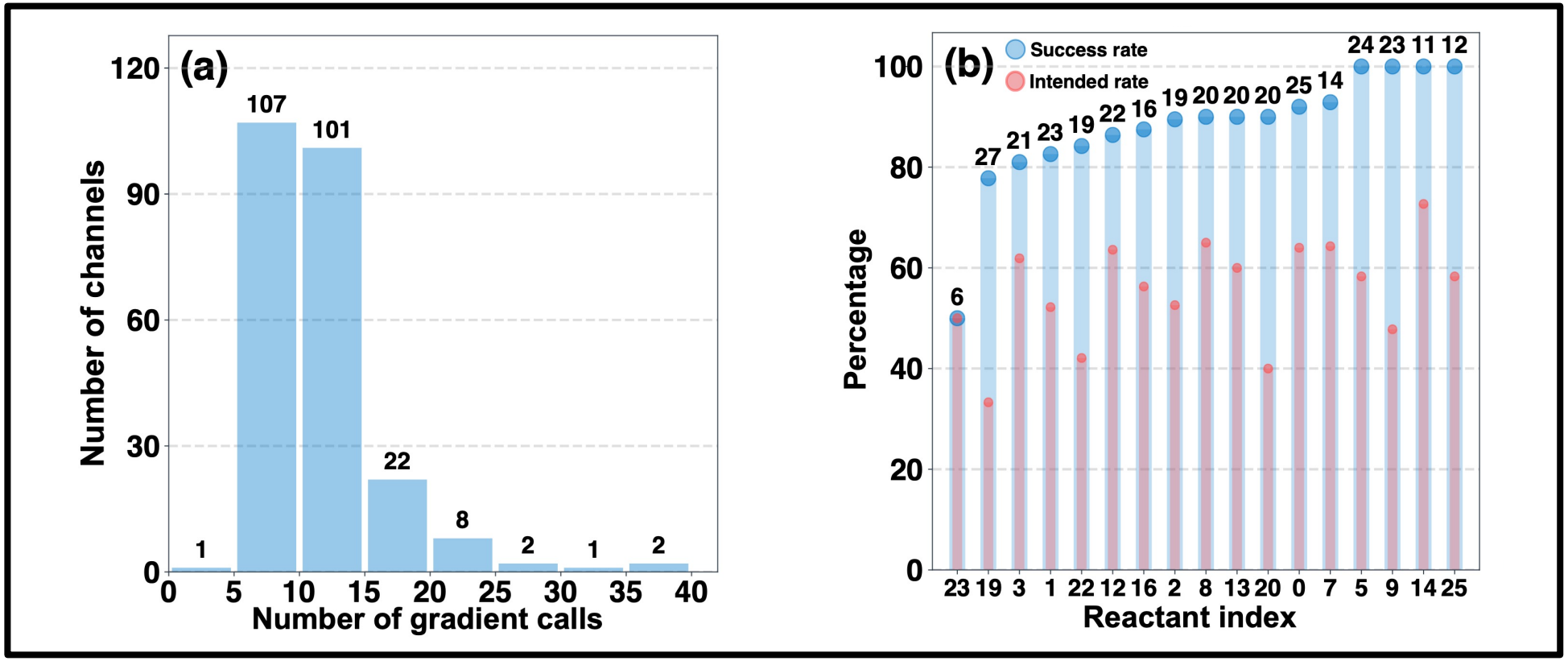
3-Hydroperoxypropanal - Reaction Network

We used YARP to recursively elucidate the 3-hydroperoxypropanal unimolecular thermal degradation network for comparison with Grambow et al.

YARP finds **all known products** of this thermal decomposition network, as well as new products (77), and new reactions (157).



Predicting More (Reactions) with Less (Cost)



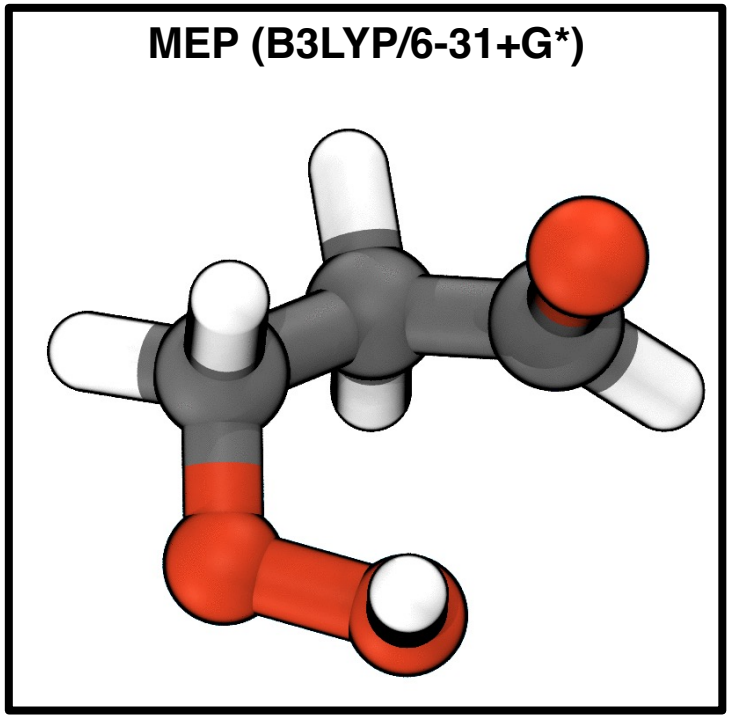
Constructing the whole network required **8364** DFT gradient calls for YARP compared with **756,227** for the earlier benchmark (**100-fold reduction**)

Average success and intended rates for YARP are **81.4%** and **41.1%**, respectively, compared with **38%** and **4%**, in the earlier benchmark.

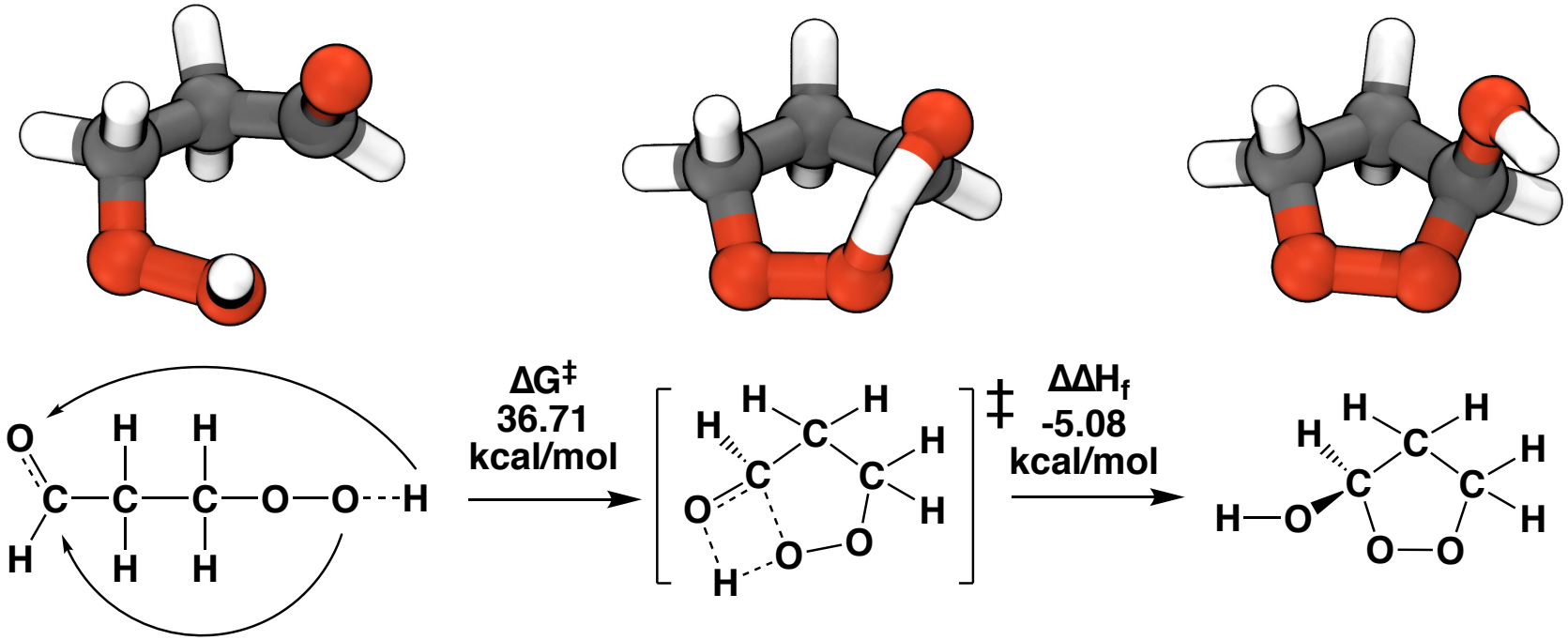
What Happens First?

Jensen, R. K.; Korcek, S.; Mahoney, L. R.; Zinbo, M. *JACS* **1979**, 101, 7574

The Korcek Mechanism



According to YARP, this is the lowest barrier degradation product.



Validated 30 years later by Green and Truhlar:

Jalan, A.; Alecu, I. M.; Meana-Pañeda, R.; Aguilera-Iparraguirre, J.; Yang, K. R.; Merchant, S. S.; Truhlar, D. G.; Green, W. H. *JACS* **2013**, *135* (30), 11100–11114.

Recent A → ? and A → ? → B Case Studies

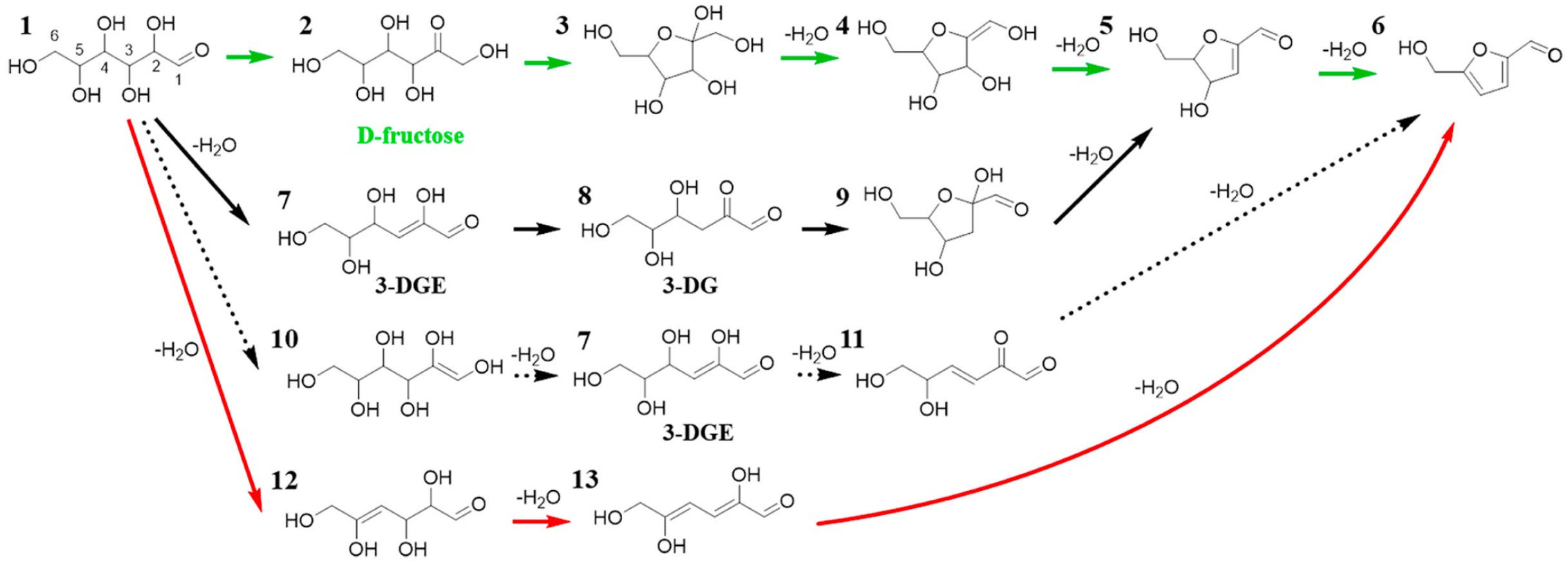
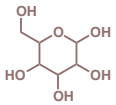
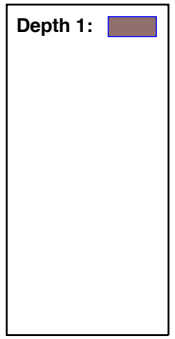


Figure 1. Proposed pathways in literature from glucose to HMF, namely the fructose path (green), 3-DG paths (black and black dotted), and direct path (red). The molecules are indicated by numbers and some key molecules are named as follows: **1**. D-glucose; **2**. D-fructose; **3**. D-fructofuranose; **6**. 5-hydroxymethylfurfural (5-HMF); **7**. 3-deoxyglucos-2-ene (3-DGE); **8**. 3-deoxyglucosone (3-DG); and **10**. hex-1-ene-1,2,3,4,5,6-hexaol (enol form of glucose).

Glucose Pyrolysis Network from YARP Exploration





To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

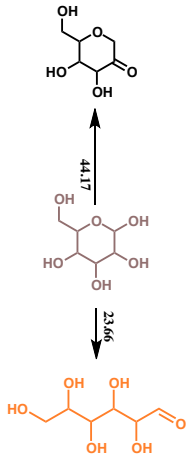
At each iteration:

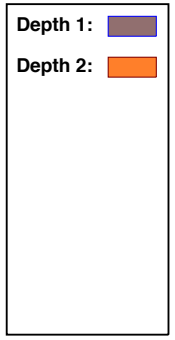
(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration





To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

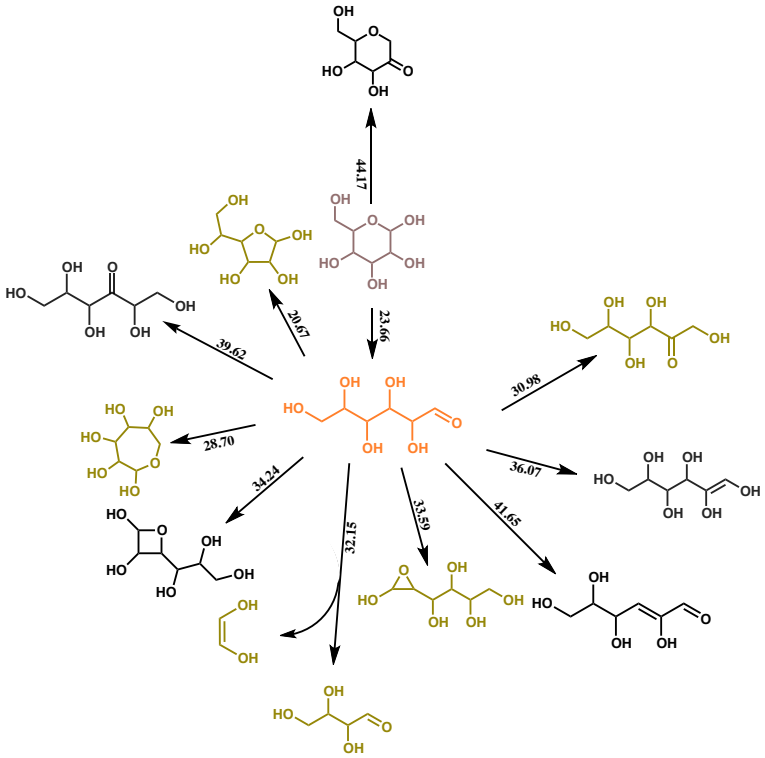
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



Depth 1:
Depth 2:
Depth 3:

To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

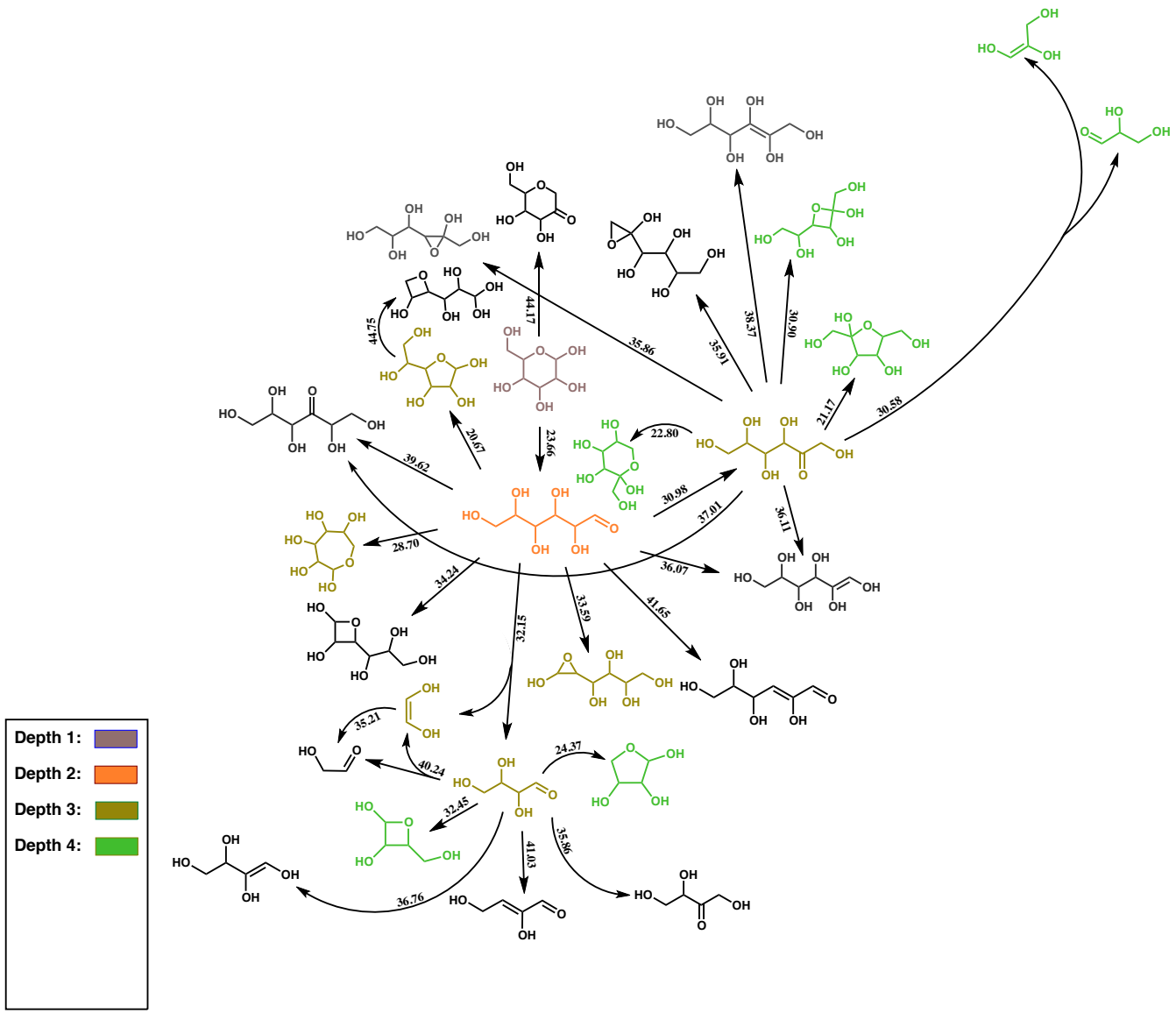
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

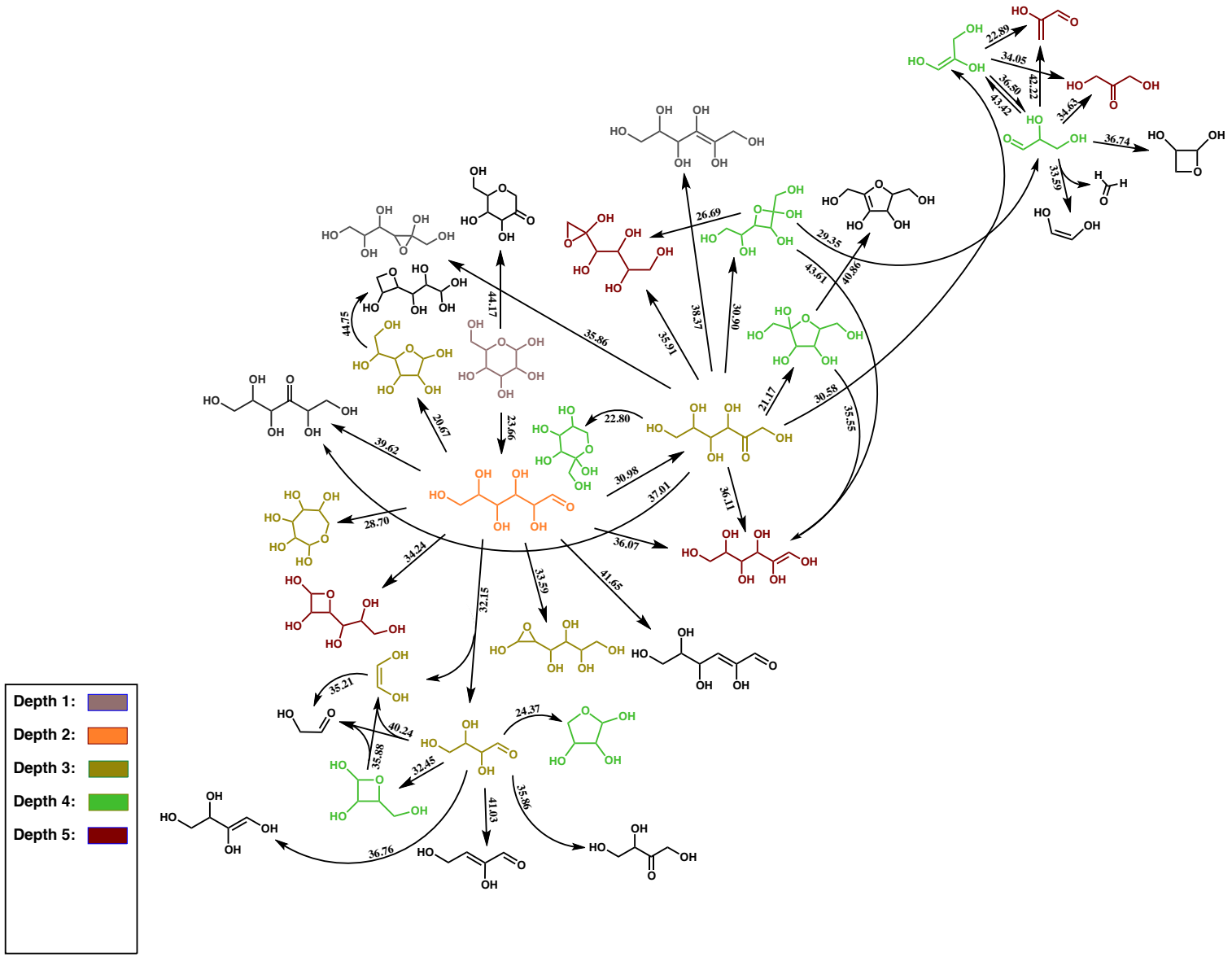
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

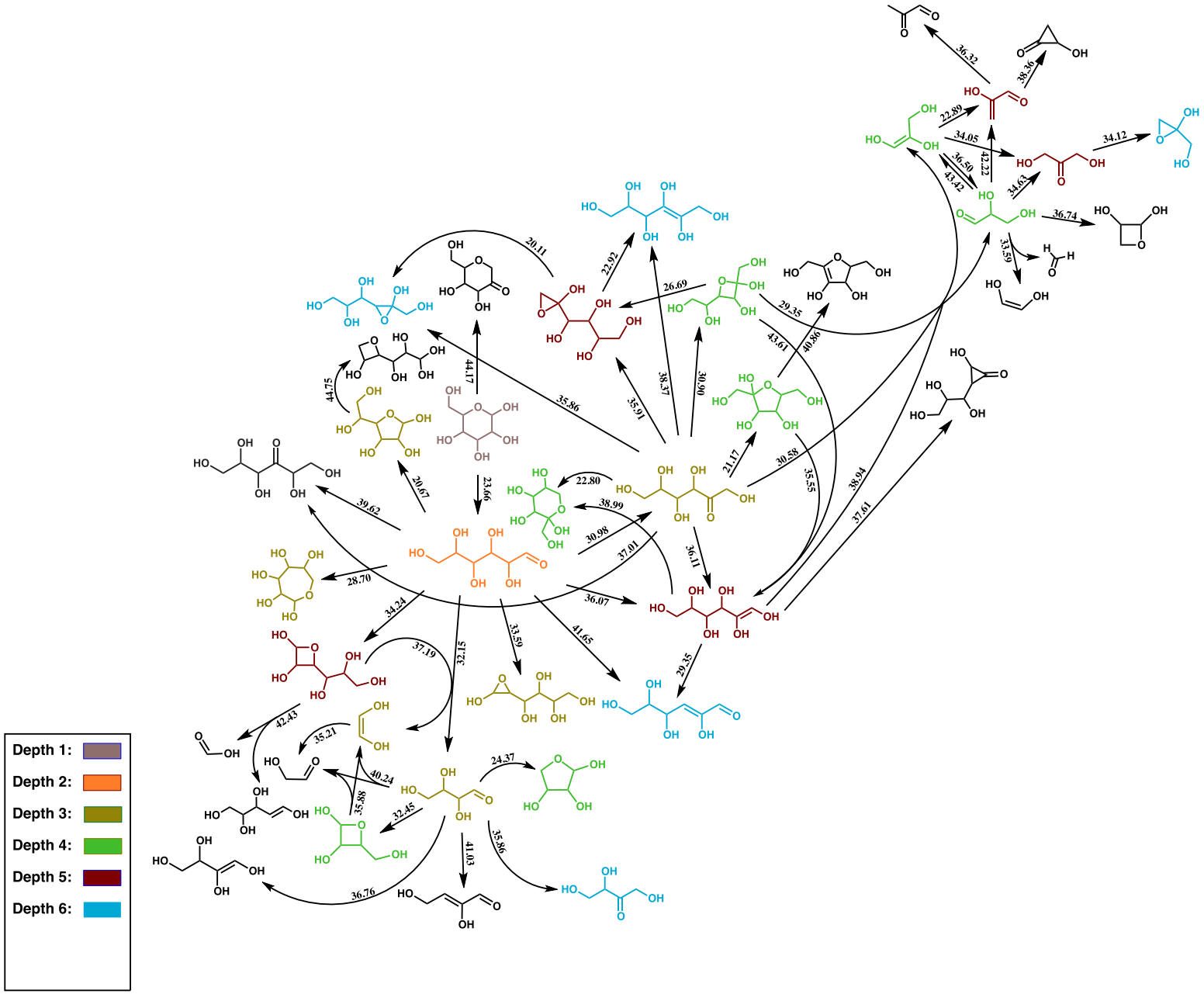
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

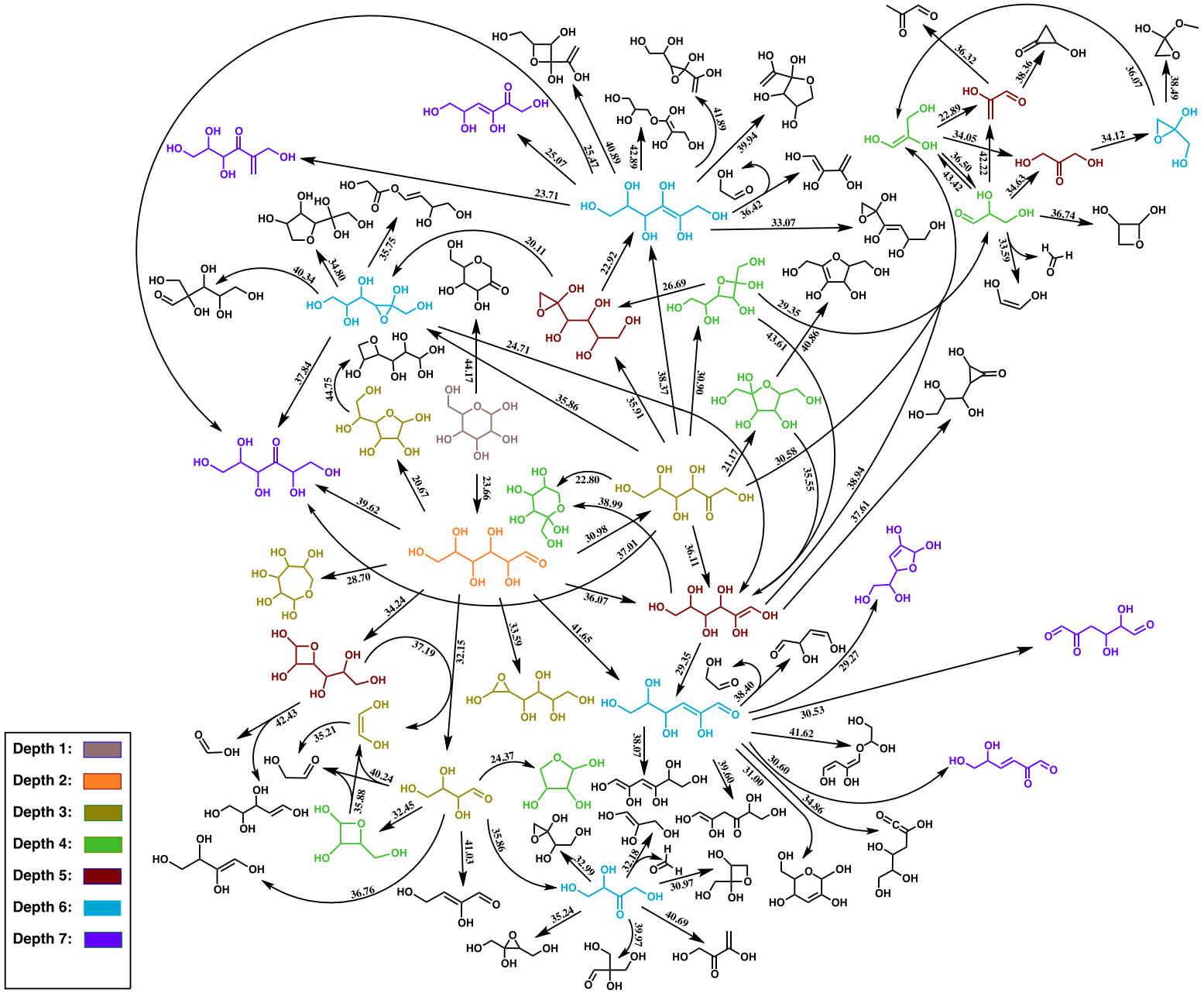
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

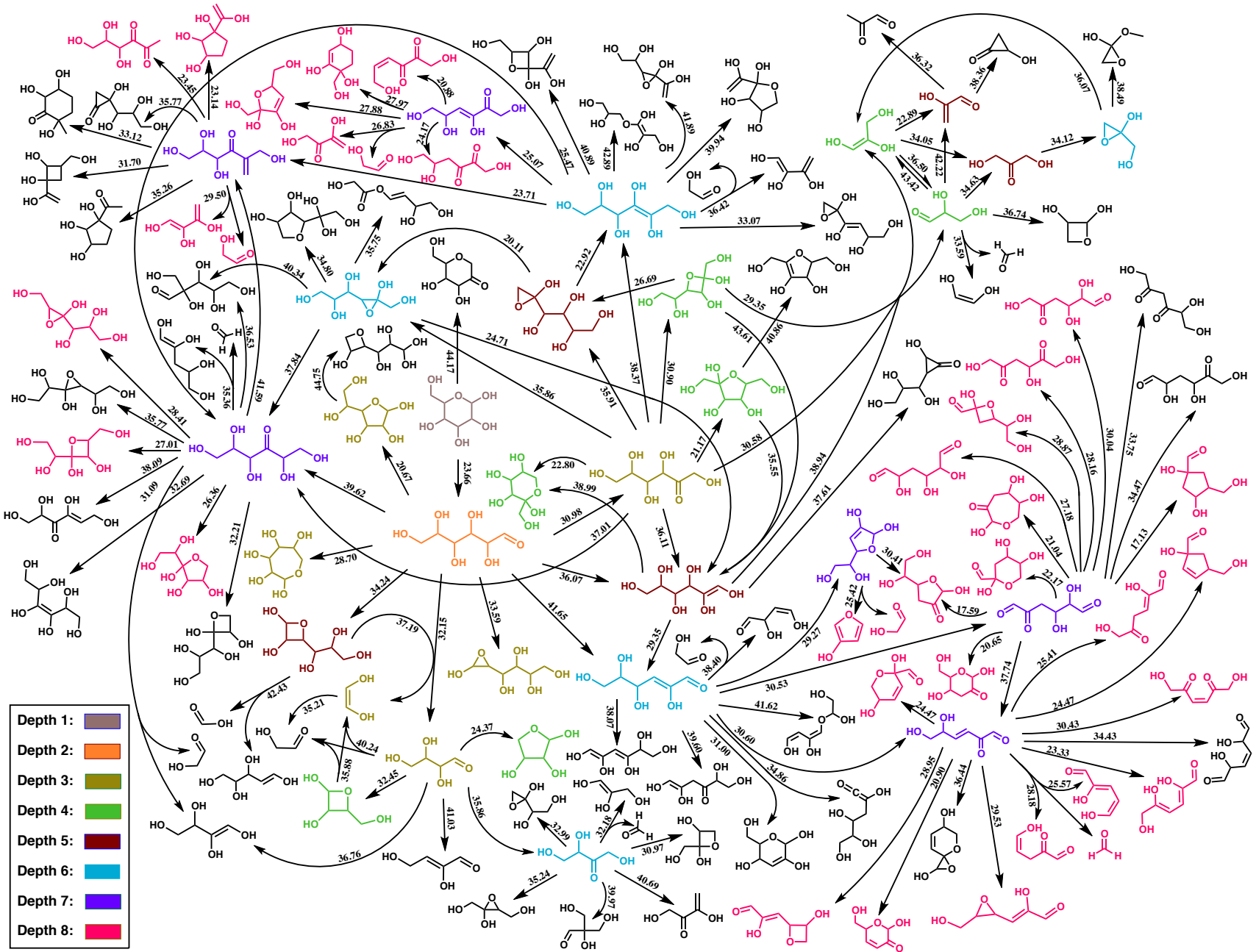
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's algorithm

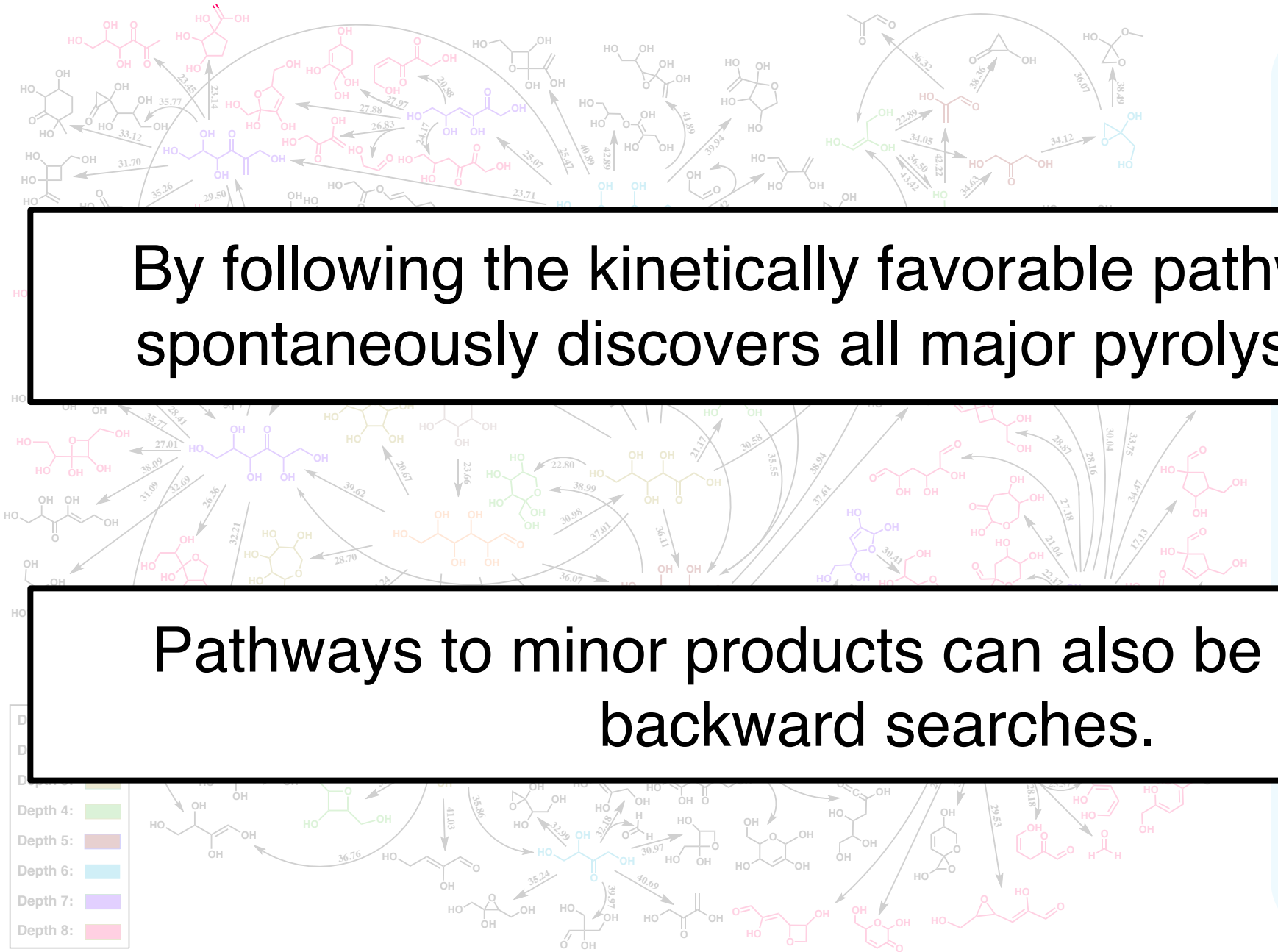
At each iteration:

(1) all b2f2 reactions are explored for active nodes.

(2) Active nodes are determined by the minimum barrier to a given product (with a window)

(3) Water catalyzed reactions are considered for all H-transfers

Glucose Pyrolysis Network from YARP Exploration



To perform a deep network exploration, we've implemented a modified version of Dijkstra's

By following the kinetically favorable pathways, YARP spontaneously discovers all major pyrolysis products.

(1) all b2f2 reactions are explored for active nodes.

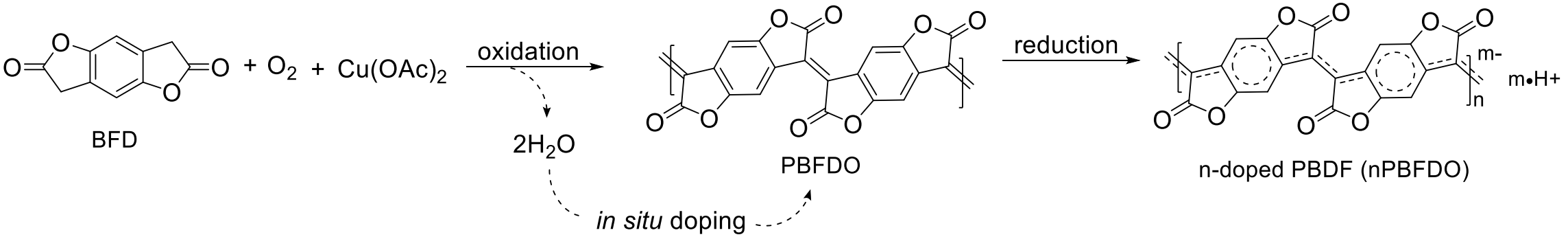
(2) Active nodes are

Pathways to minor products can also be found using backward searches.

(3) Water catalyzed reactions are considered for all H-transfers

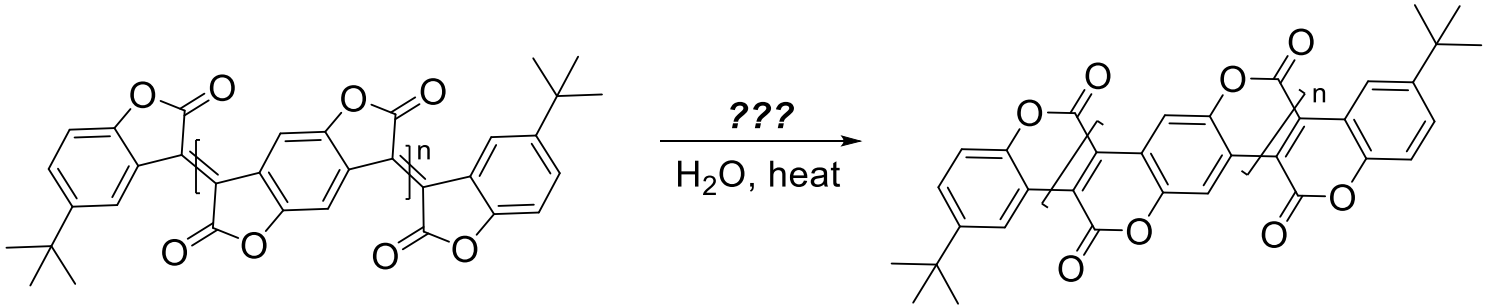
A Recent A → ? → B Case Study

PBDF is a new record breaking n-type semi-conducting plastic

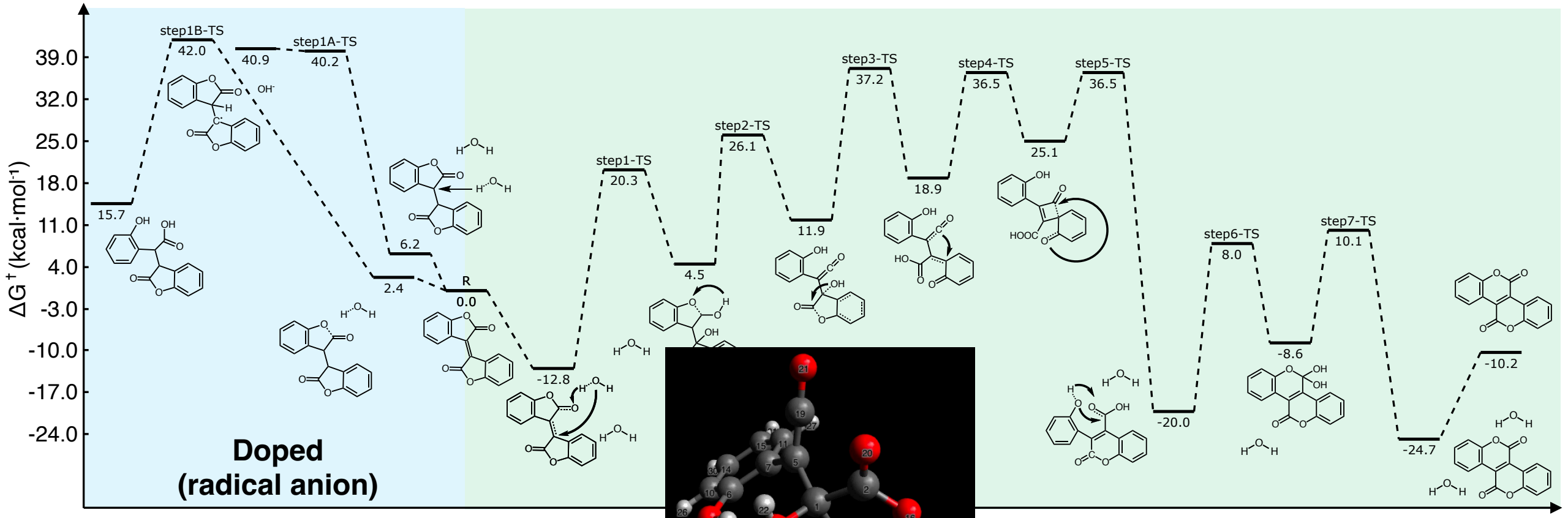


Small molecule studies show a significant synthetic side product

To what extent is this occurring in the polymer? or why don't we see it in the polymer?

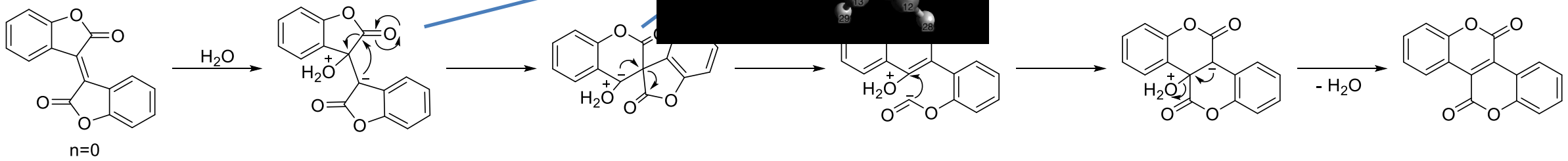


Pathways identified by YARP



✓ Proposed reaction pathways

k after DFT geometry optimization)



Outlook and Acknowledgements

Students: Qiyuan Zhao, Tyler Pasut

State-of-the-art:

- The accurate calculation of thermodynamic properties has become routine in many scenarios. Major opportunities lie in automation, systemization, and low-cost models.

- Practical solutions to the A→?→B, A→B+?, and A→? problems are now available. We envision black-box tools for non-experts in the near future that will assist in hypothesis generation and potentially reactivity screening.

- P2SAC and ONR for funding.
- Ray Mentzer (Purdue)
- Katherine Young (Purdue UG)

P2SAC
Purdue Process Safety & Assurance Center