Predicting ΔH_f and Degradation Pathways from First Principles

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

P2SAC May Meeting, 5/10/22

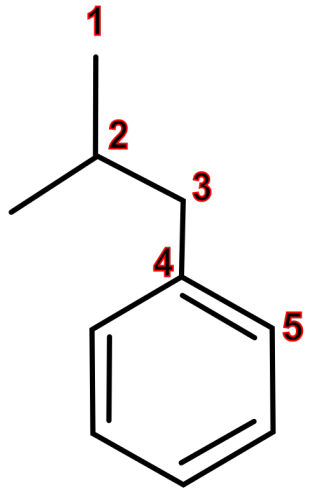
Challenges of Contemporary Group Theories

Benson Group Theory:

• The idea is to decompose molecular properties (ΔH_f, S°, 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.

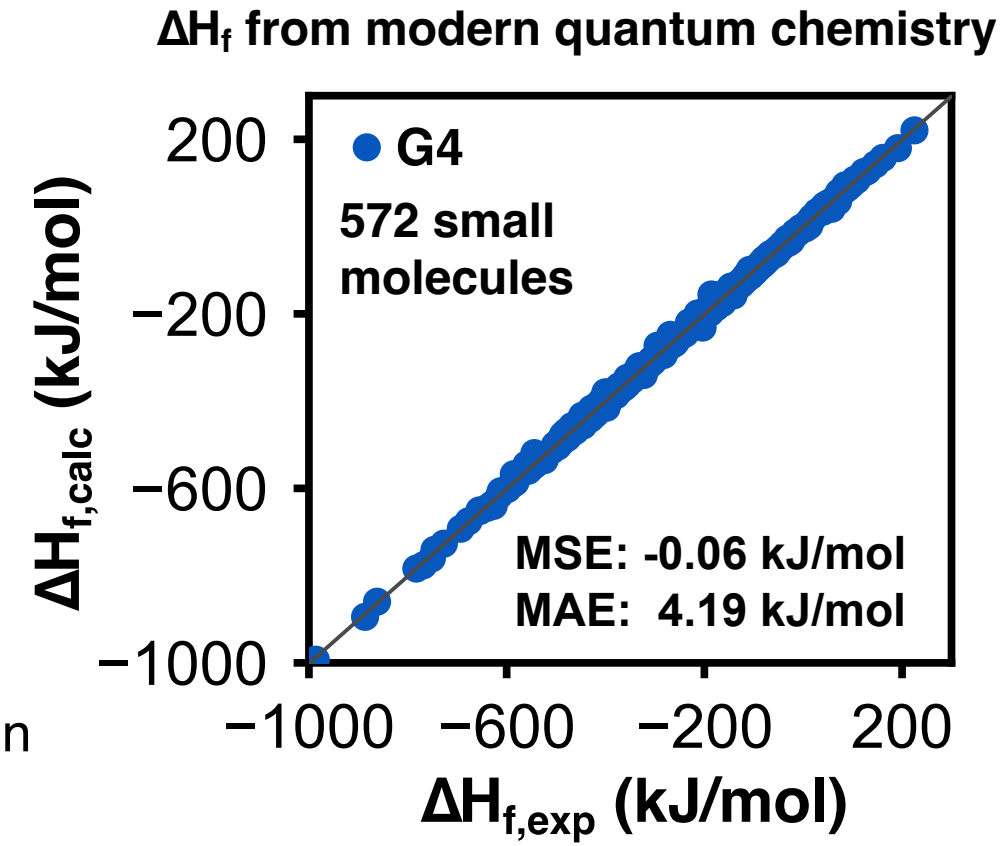
• **Extensibility:** because of the provenance and specificity problems, it isn't possible to develop new groups in a consistent way.

Challenges of Contemporary Group Theories

Benson Group Theory:

- The idea is to decompose molecular properties (ΔH_f, S°, 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.

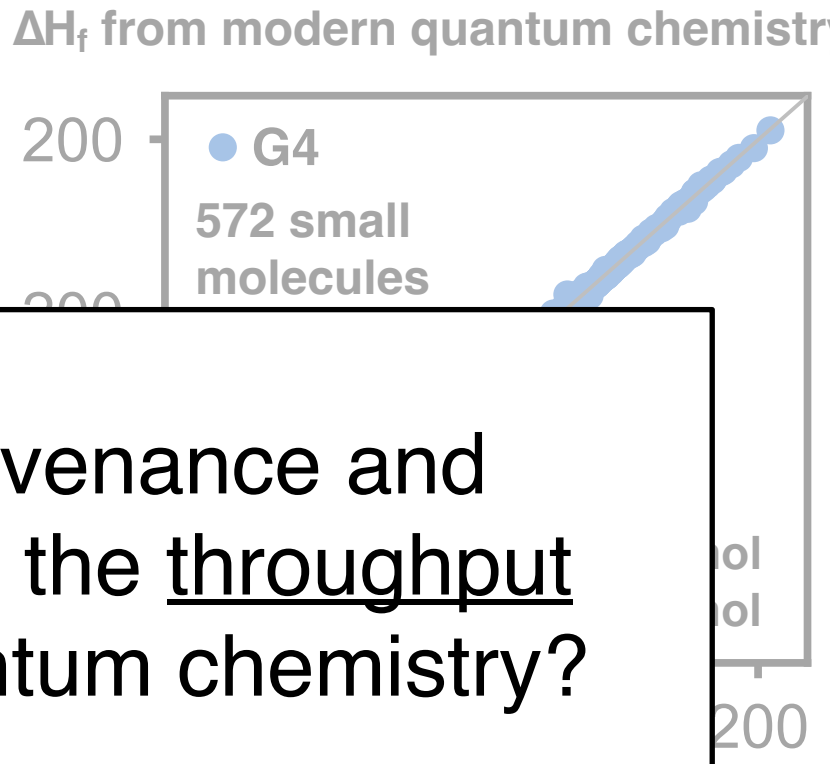
- Grou
on trus
data, a

Prob

- **Spec**
formal

- **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.



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

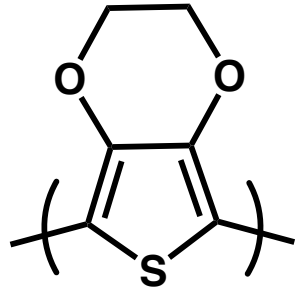
Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

TAFFI Component Increment Theory (TCIT)

The fundamental idea

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

TCIT is a component theory (2-bond specific)



graph/structure equivalence

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

TAFFI

S	0	1	0	0	1	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	0	1	0
C	1	0	0	1	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0

Adjacency matrix for PEDOT monomer

P2SAC Publications

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

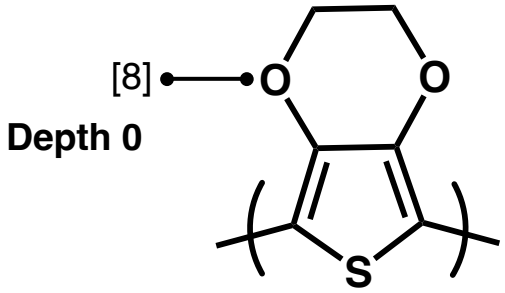
Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

TAFFI Component Increment Theory (TCIT)

The fundamental idea

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

TCIT is a component theory (2-bond specific)



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

graph/structure equivalence

S	0	1	0	0	1	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	0	1	0	0
C	1	0	0	1	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0

Adjacency matrix for PEDOT monomer

P2SAC Publications

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

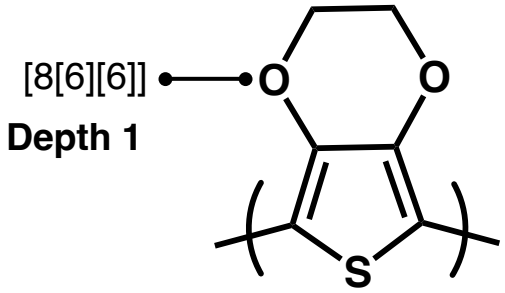
Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

TAFFI Component Increment Theory (TCIT)

The fundamental idea

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

TCIT is a component theory (2-bond specific)



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

graph/structure equivalence

S	0	1	0	0	1	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	0	1	0
C	1	0	0	1	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0

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)

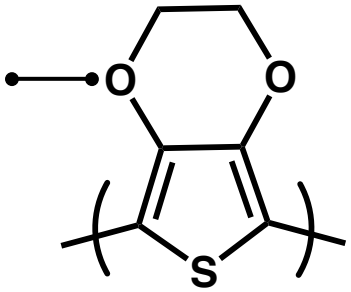
The fundamental idea

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

TCIT is a component theory (2-bond specific)

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

Depth 2



graph/structure equivalence

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

S	0	1	0	0	1	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	1	0	0
C	1	0	0	1	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0

Adjacency matrix for PEDOT monomer

P2SAC Publications

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

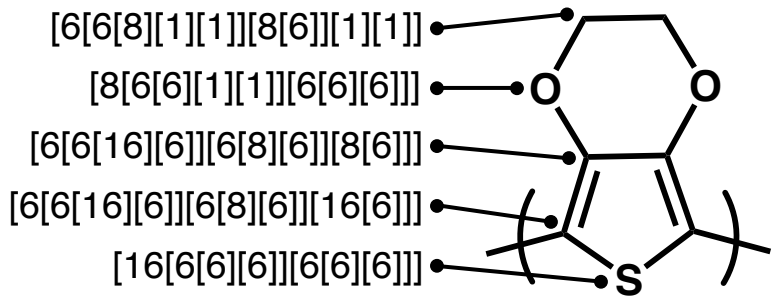
Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

TAFFI Component Increment Theory (TCIT)

The fundamental idea

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

TCIT is a component theory (2-bond specific)



graph/structure equivalence

Topology Automated Force Field Interactions

TAFFI

	S	C	C	C	C	O	C	C	H	H	O	H	H
S	0	1	0	0	1	0	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	0	1	0	0
C	1	0	0	1	0	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0	0

Adjacency matrix for PEDOT monomer

P2SAC Publications

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

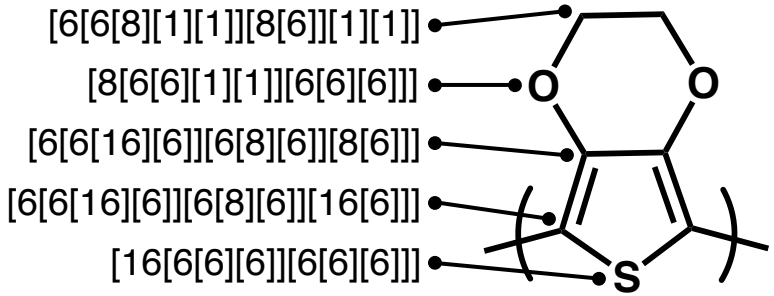
Seo, B.; Lin, Z.-Y.; Zhao, Q.; Webb, M. A.; Savoie, B. M. Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields. *J. Chem. Inf. Model.* **2021**, 61 (10), 5013–5027. https://doi.org/10.1021/acs.jcim.1c00491.

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 component theory (2-bond specific)



graph/structure equivalence

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

	S	C	C	C	C	O	C	C	H	H	O	H	H
S	0	1	0	0	1	0	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	0	0	0	0	0
C	0	0	1	0	1	0	0	0	0	0	1	0	0
C	1	0	0	1	0	0	0	0	0	0	0	0	0
O	0	0	1	0	0	0	1	0	0	0	0	0	0
C	0	0	0	0	0	1	0	1	1	1	0	0	0
C	0	0	0	0	0	0	1	0	0	0	1	1	1
H	0	0	0	0	0	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0	0
O	0	0	0	1	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0	0

Adjacency matrix for PEDOT monomer

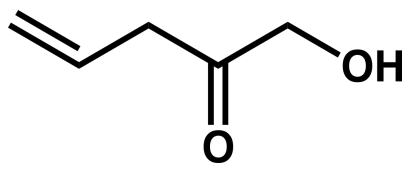
Zhao, Q.; Savoie, B. M.; "Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory". *J. Chem. Info. Model.* **2020**, 60, 2199-2207

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

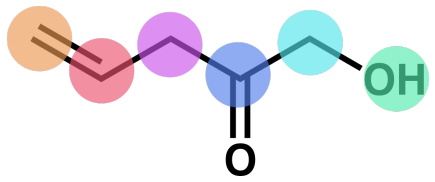


1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

Graphical Decomposition of Model Compounds

identify components*

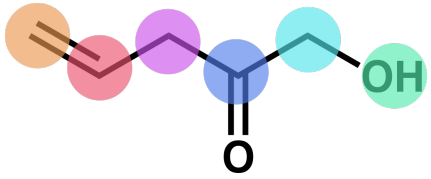


1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

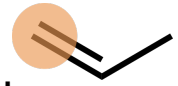
Graphical Decomposition of Model Compounds

identify components*



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

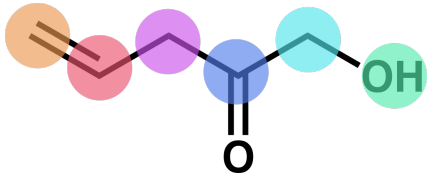
Recursively generate smallest acyclic model compounds



How will we select molecules for parameterizing TCIT components?

Graphical Decomposition of Model Compounds

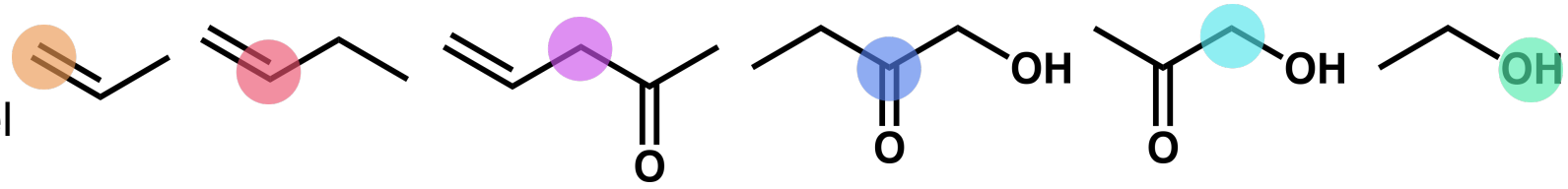
identify components*



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

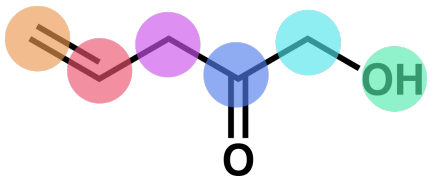
How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds



Graphical Decomposition of Model Compounds

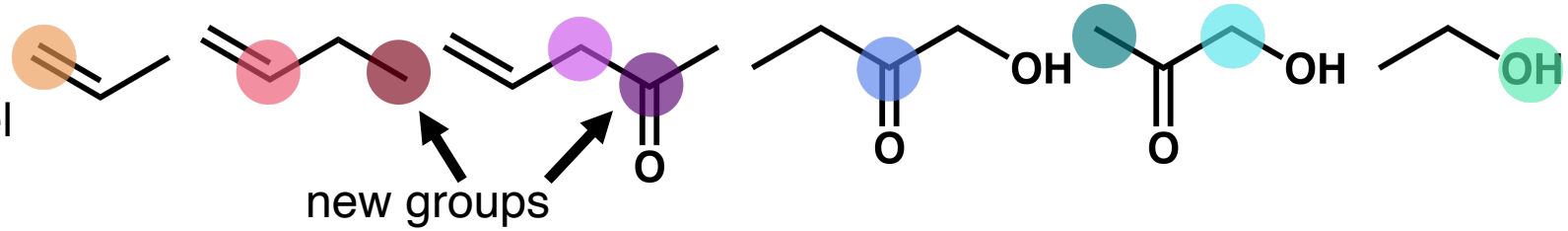
identify components*



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

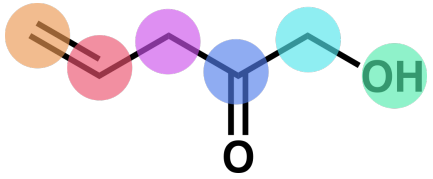
How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds



Graphical Decomposition of Model Compounds

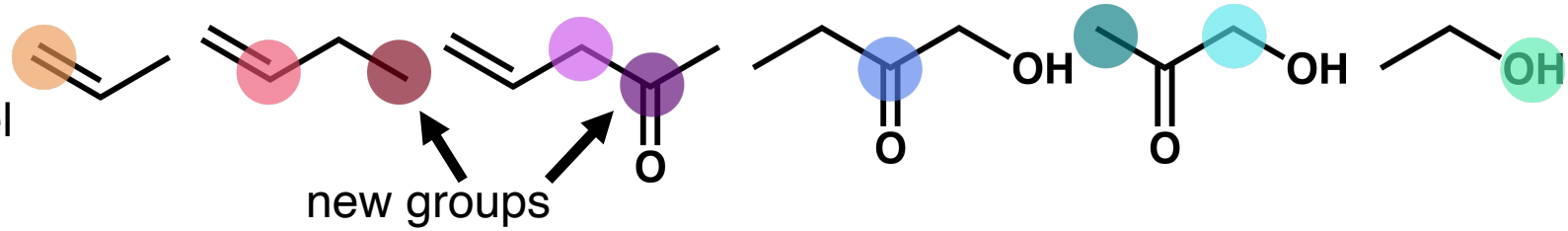
identify components*



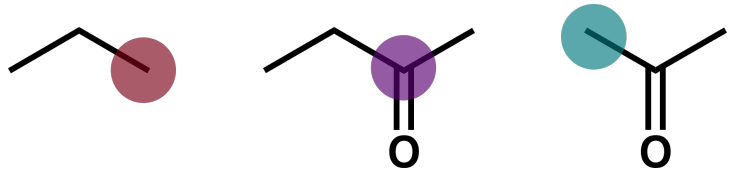
1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

Recursively generate smallest acyclic model compounds

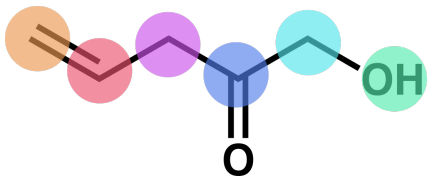


Resolve dependencies



Graphical Decomposition of Model Compounds

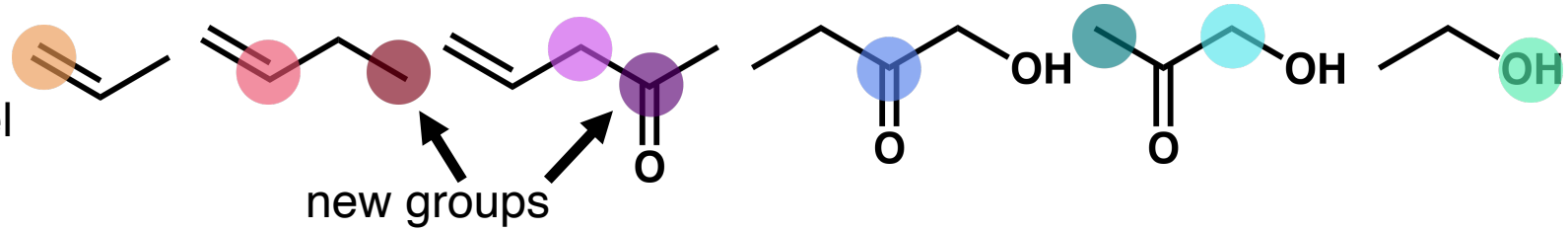
identify components*



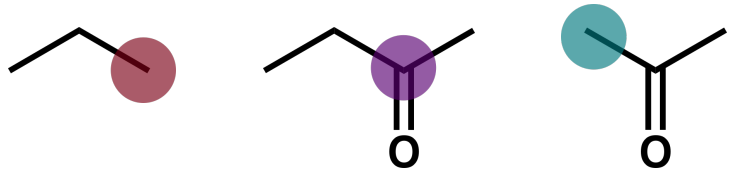
1-hydroxy-pent-2-ene-2-one

How will we select molecules for parameterizing TCIT components?

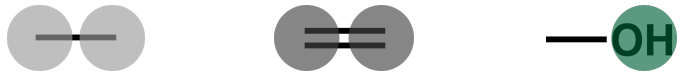
Recursively generate smallest acyclic model compounds



Resolve dependencies

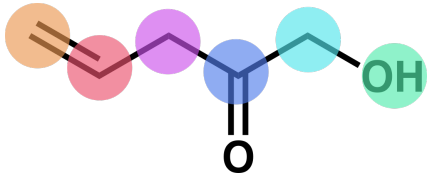


Resolve rank deficiency with elementary constraints



Graphical Decomposition of Model Compounds

Prediction target:



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

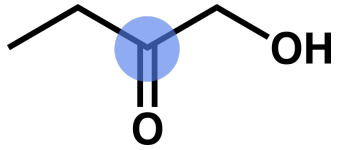
$\Delta H_{f,G4}$ = -259.9 kJ/mol
$\Delta H_{f,TCIT}$ = -259.3 kJ/mol

no experimental data

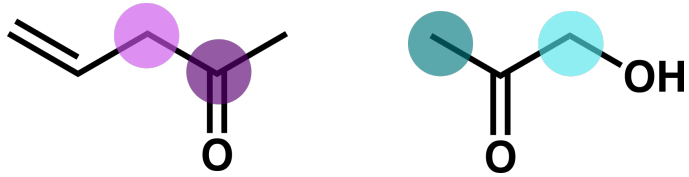
Topologically sort dependency graph

(Automatically handled by TCIT software)

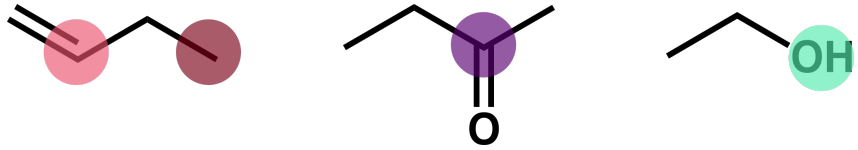
Gen 4:



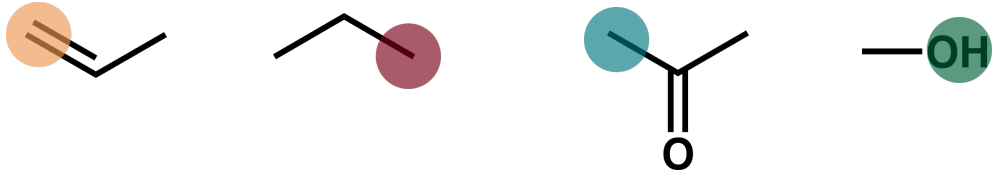
Gen 3:



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.

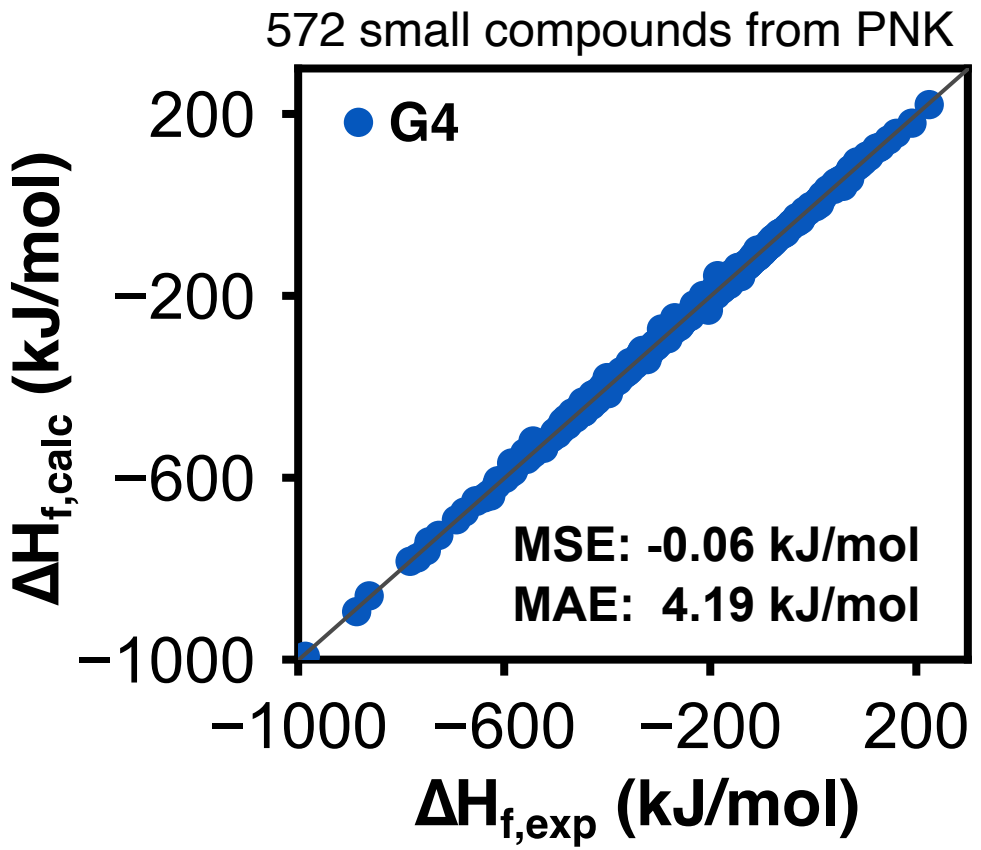
Benchmarking $\Delta H_{f,gas}$ Predictions Against the PNK Dataset

- 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

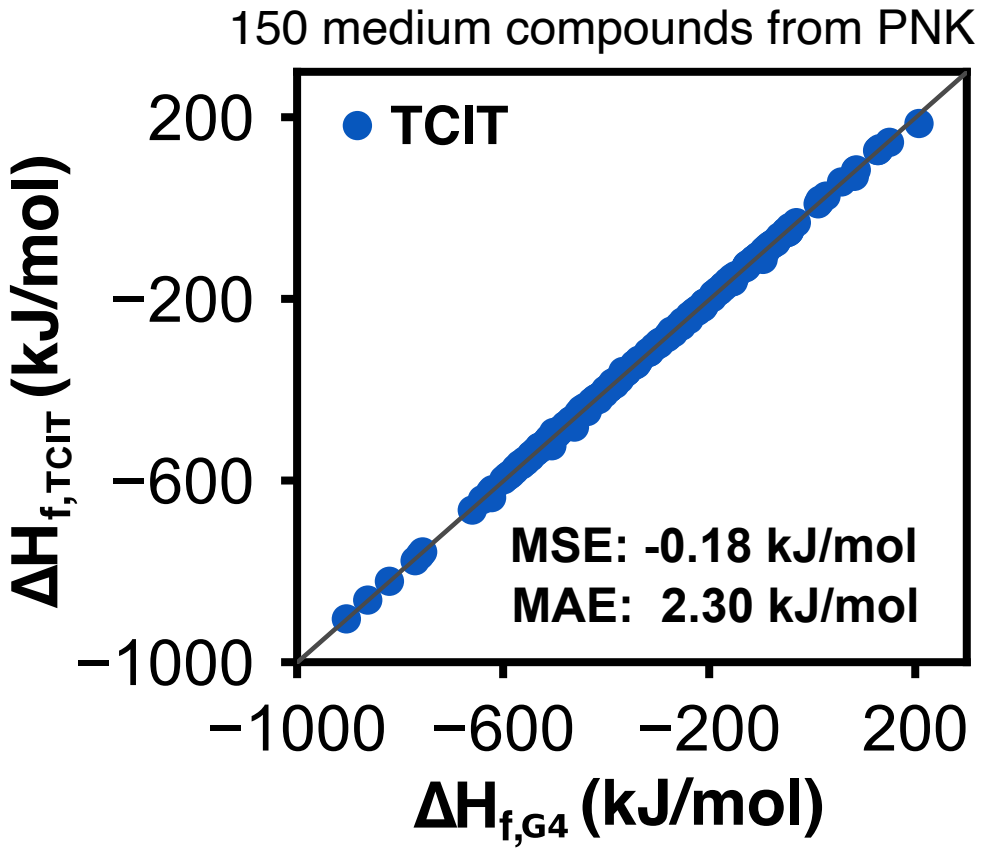
- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

(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]

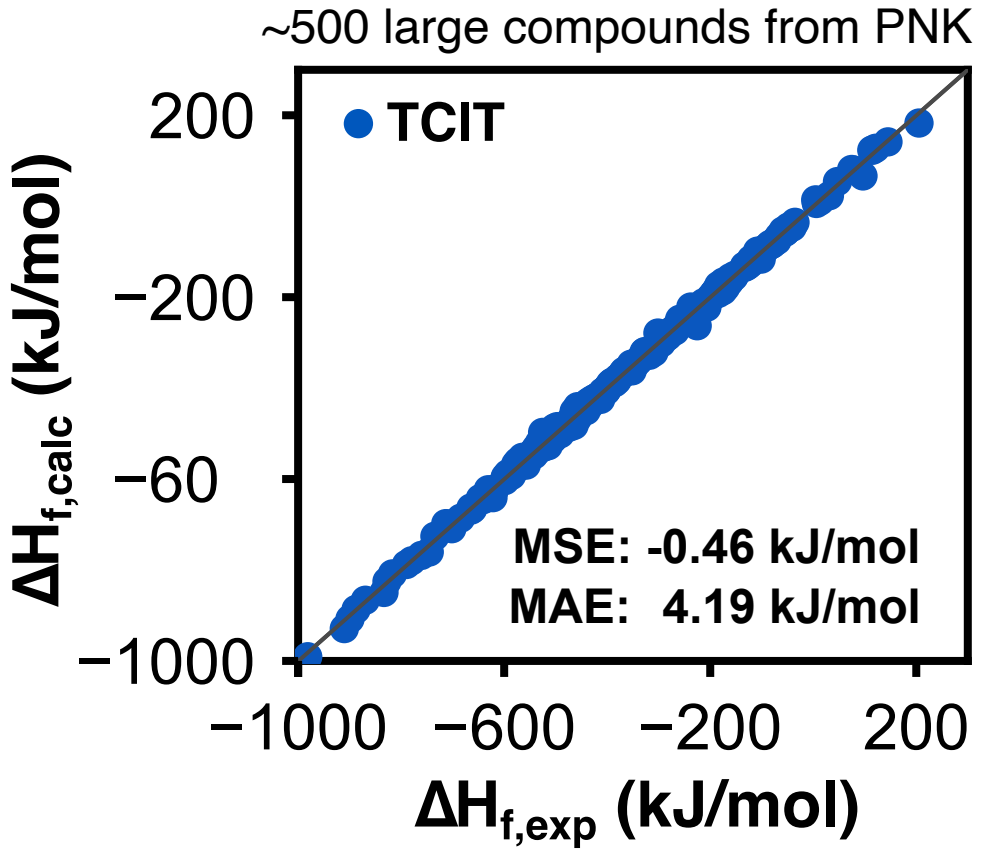
(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

- 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]

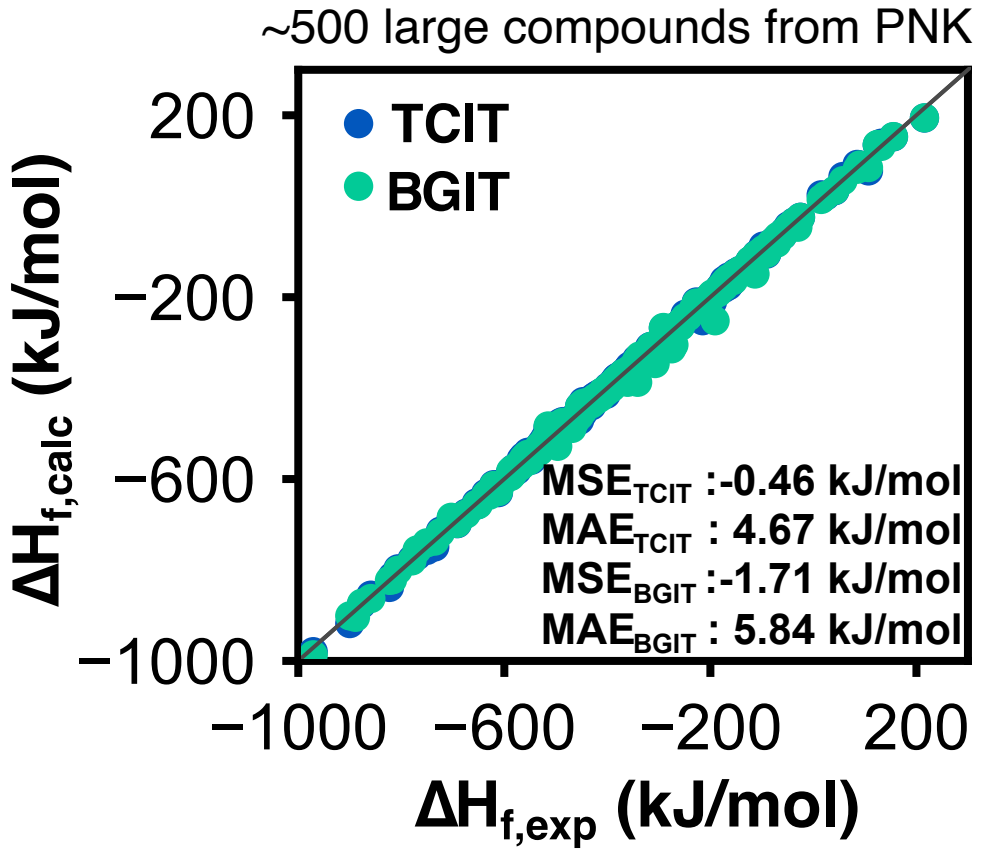
(1) J. B. Pedley, R. D. Naylor, S. P. Kirby "Thermochemical Data of Organic Compounds" 2nd ed. 1986

- 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.



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

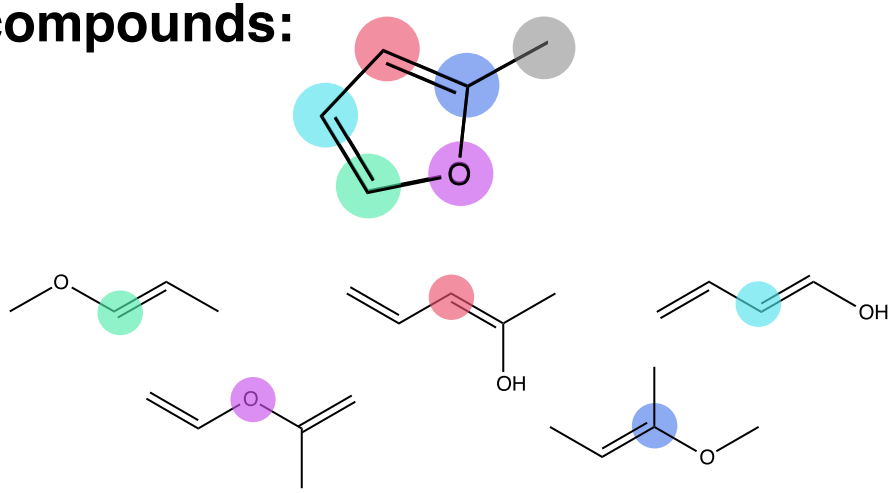
TCIT shows comparable performance to BGIT/CHETAH but is derived exclusively from extensible G4 data.

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:

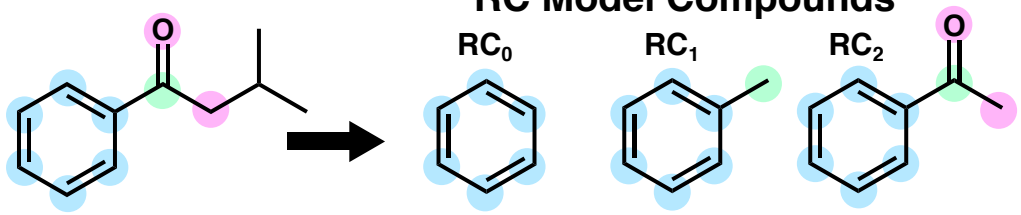


2. Add ring correction (RC) to final prediction:

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

Technical Developments

RC Model Compounds



○ : Depth 0 **○ : Depth 1** **○ : Depth 2**

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

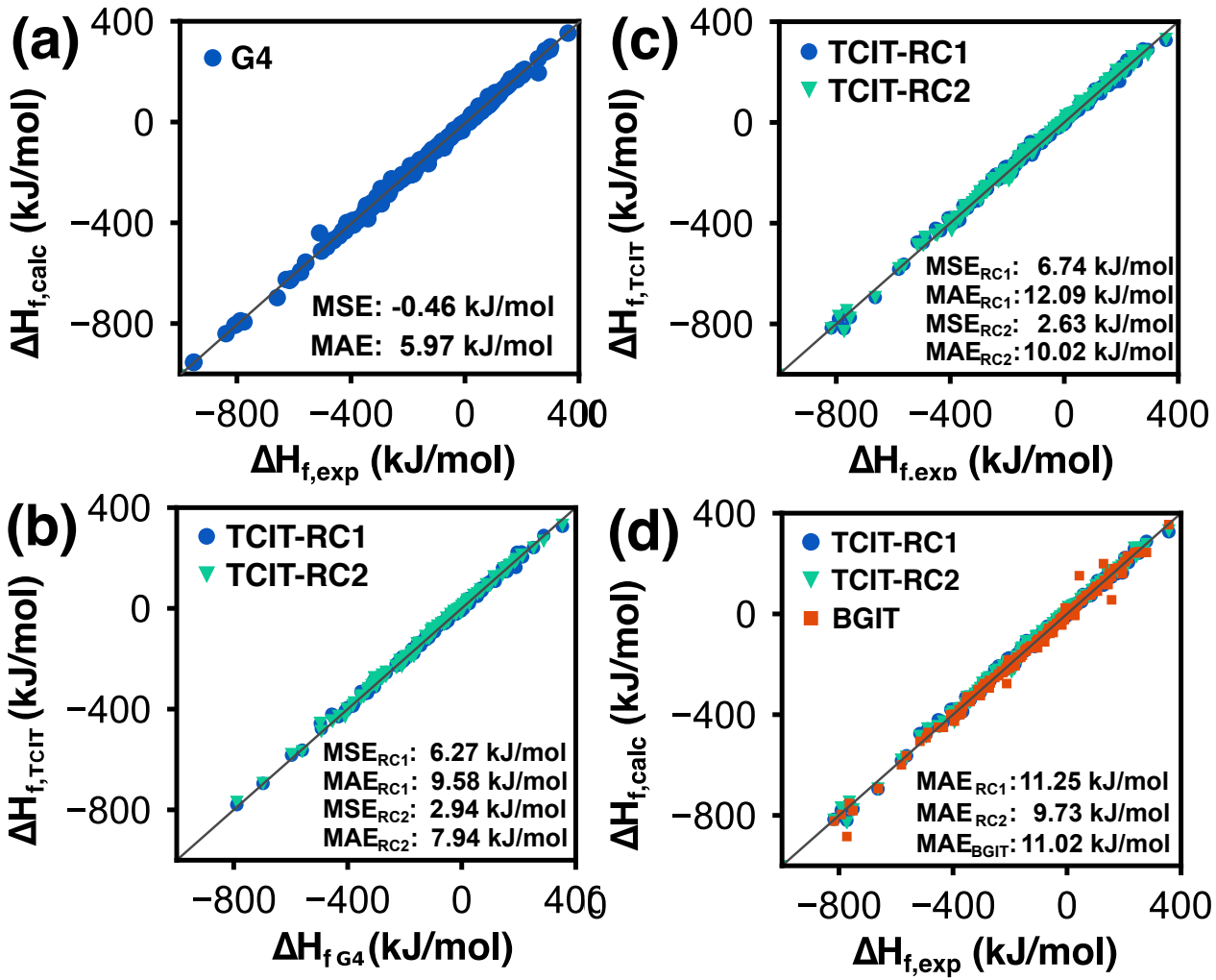
(a) G4 errors are marginally larger for ring-containing compounds but still very accurate

(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

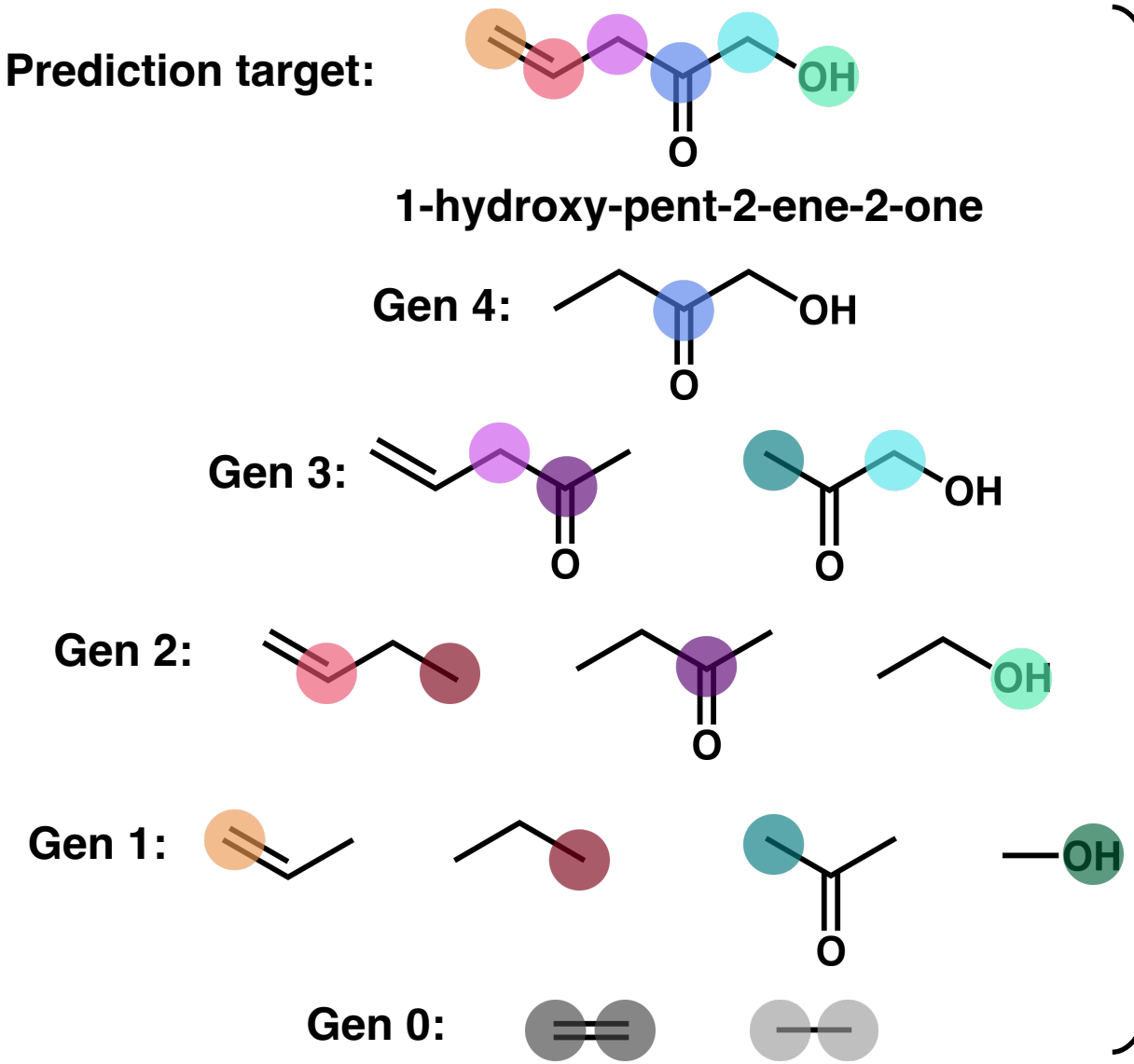
(d) The TCIT-R2 model outperforms BGIT on the large molecule benchmark while being extensible. Significantly, these compounds are within BGIT's training data.

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



BGIT cannot make predictions for ~2% of PNK compounds

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?

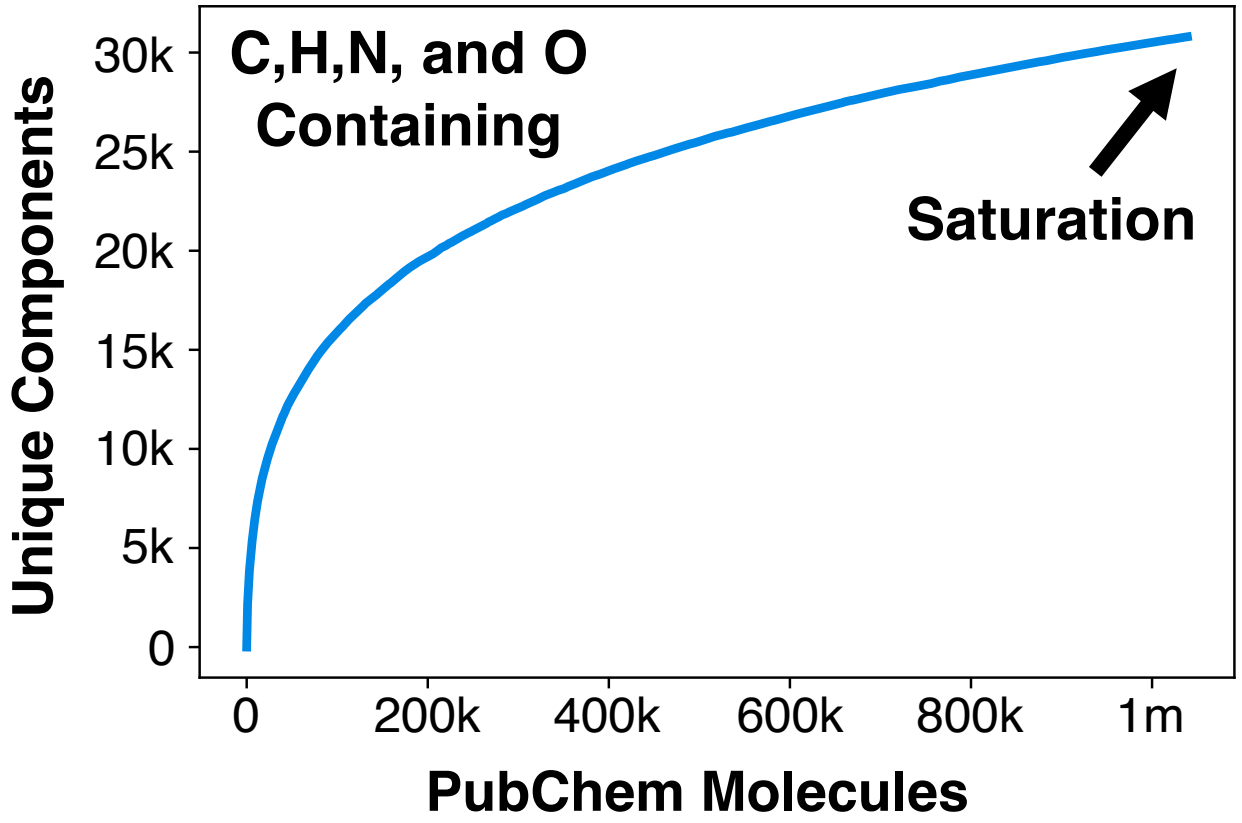
How many components are required to predict the ΔH_f of **all** (physically relevant) organic molecules?

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

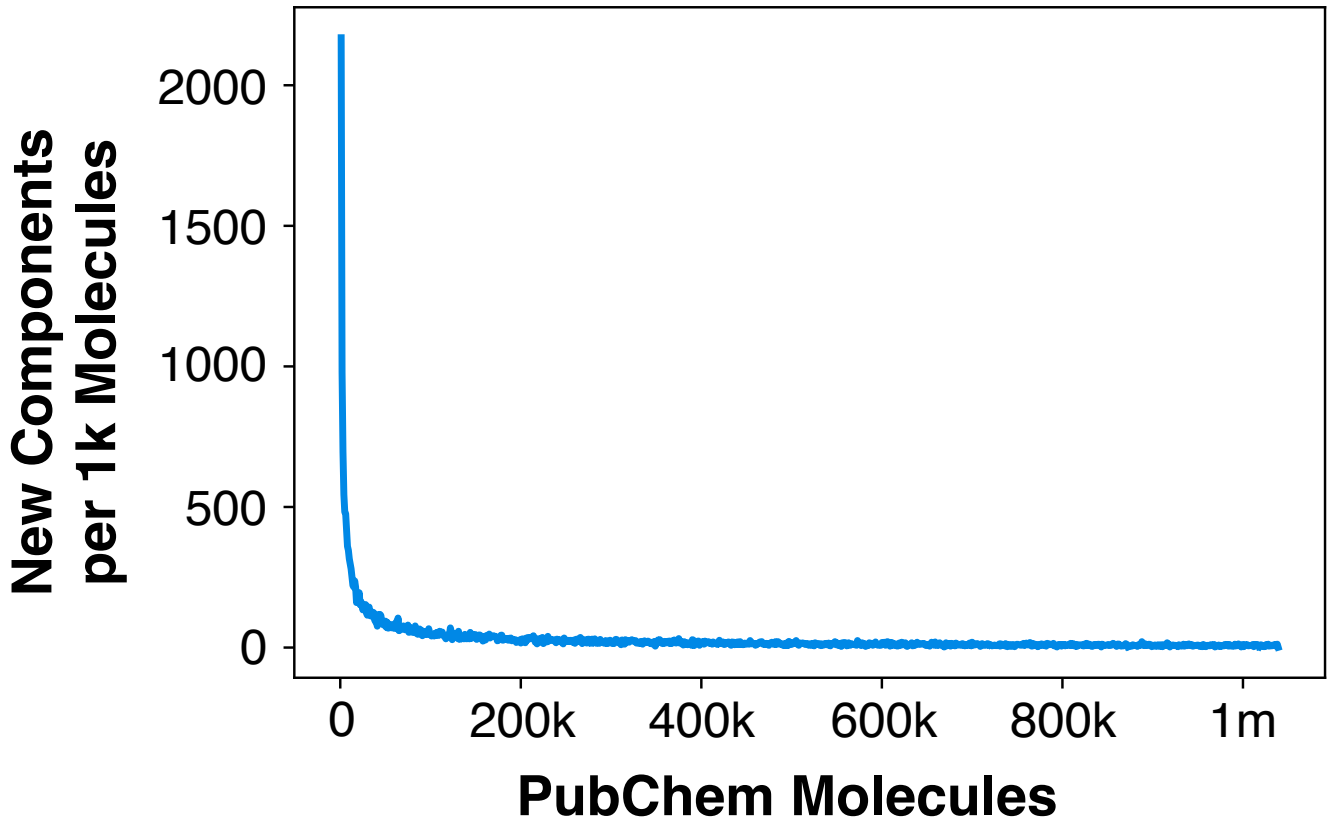


Treating PubChem as a Model of Organic Chemical Space

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.

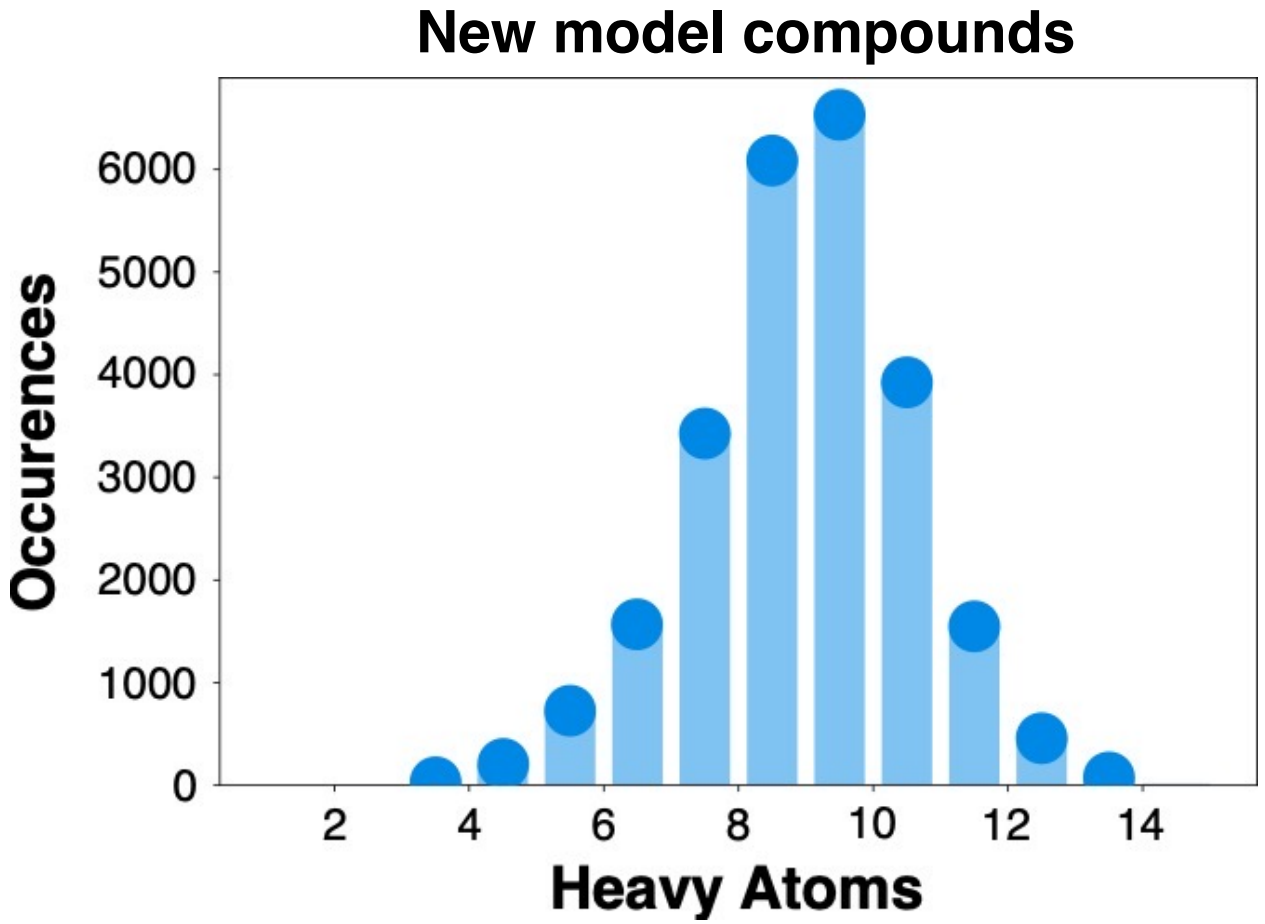


Treating PubChem as a Model of Organic Chemical Space

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.

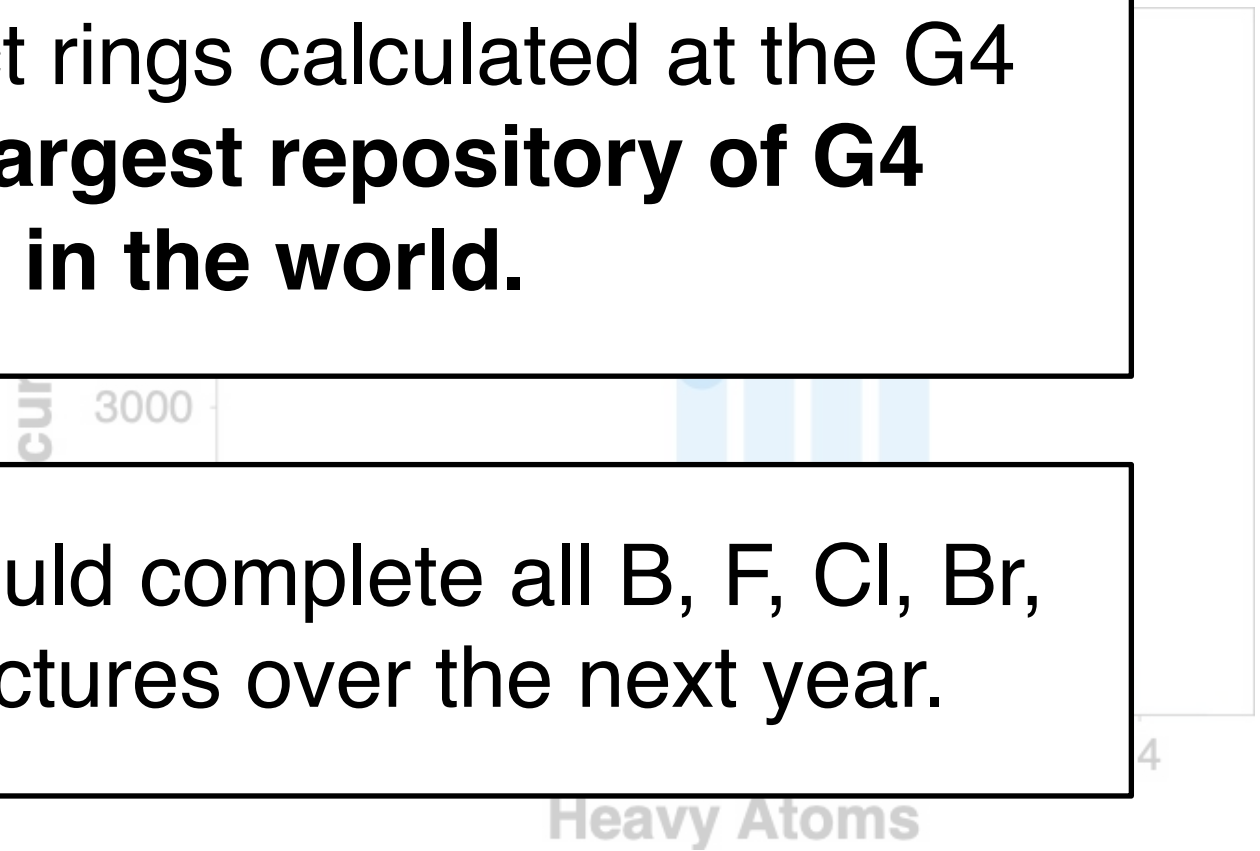


Treating PubChem as a Model of Organic Chemical Space

PubChem is a repository of chemical properties that contains many millions of organic

By the end of the summer, we will have 60k model compounds and 2k distinct rings calculated at the G4 level. **This will be the largest repository of G4 calculations in the world.**

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



The Reaction Prediction Problem

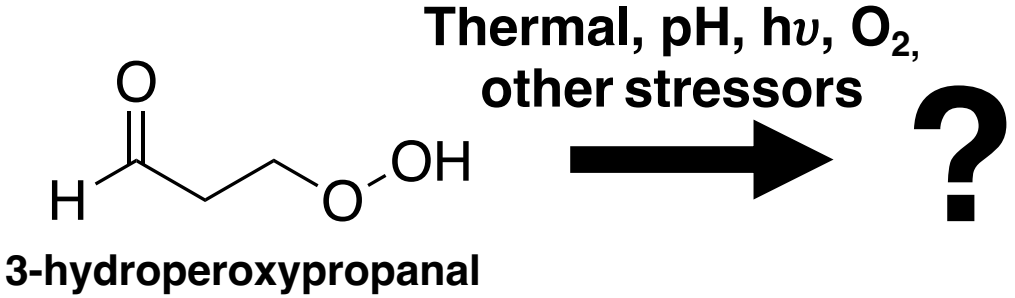
A → B : When we know the reactants and products, mature quantum chemistry tools exist to characterize transition states and establish pathways

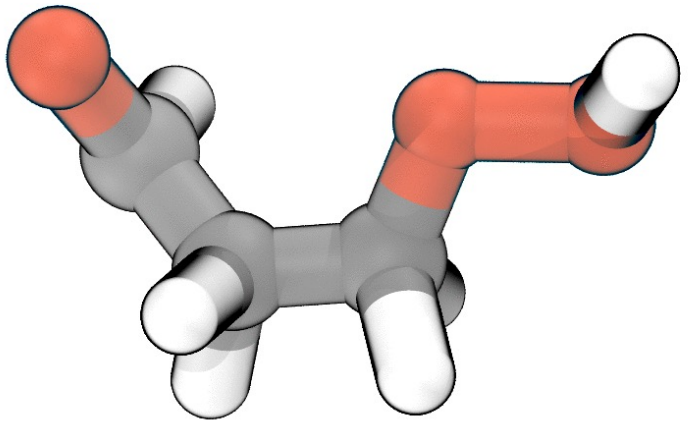
A → ? : For degradation reactions, plausible reactions are often unknown.

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.



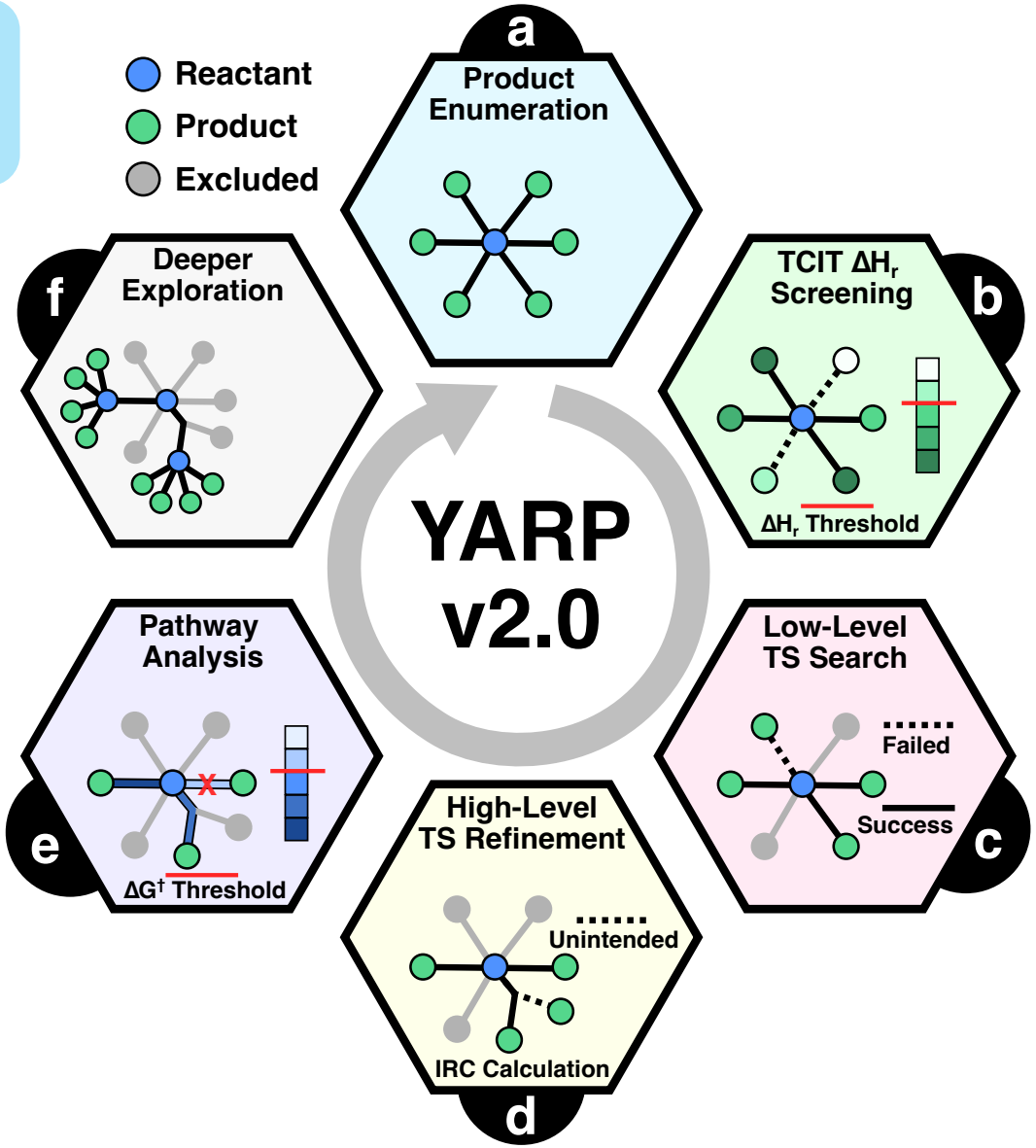


Yet Another Reaction Program (YARP)

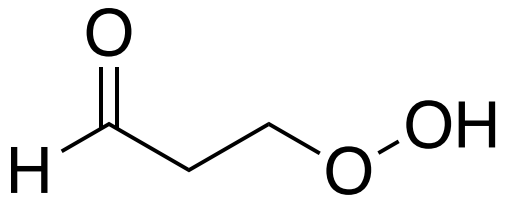
Idea: Turn the **A→?** problem into tractable (and parallelizable) **A→B** problems.

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 quantum chemistry and ML make it worthwhile to revisit this problem.

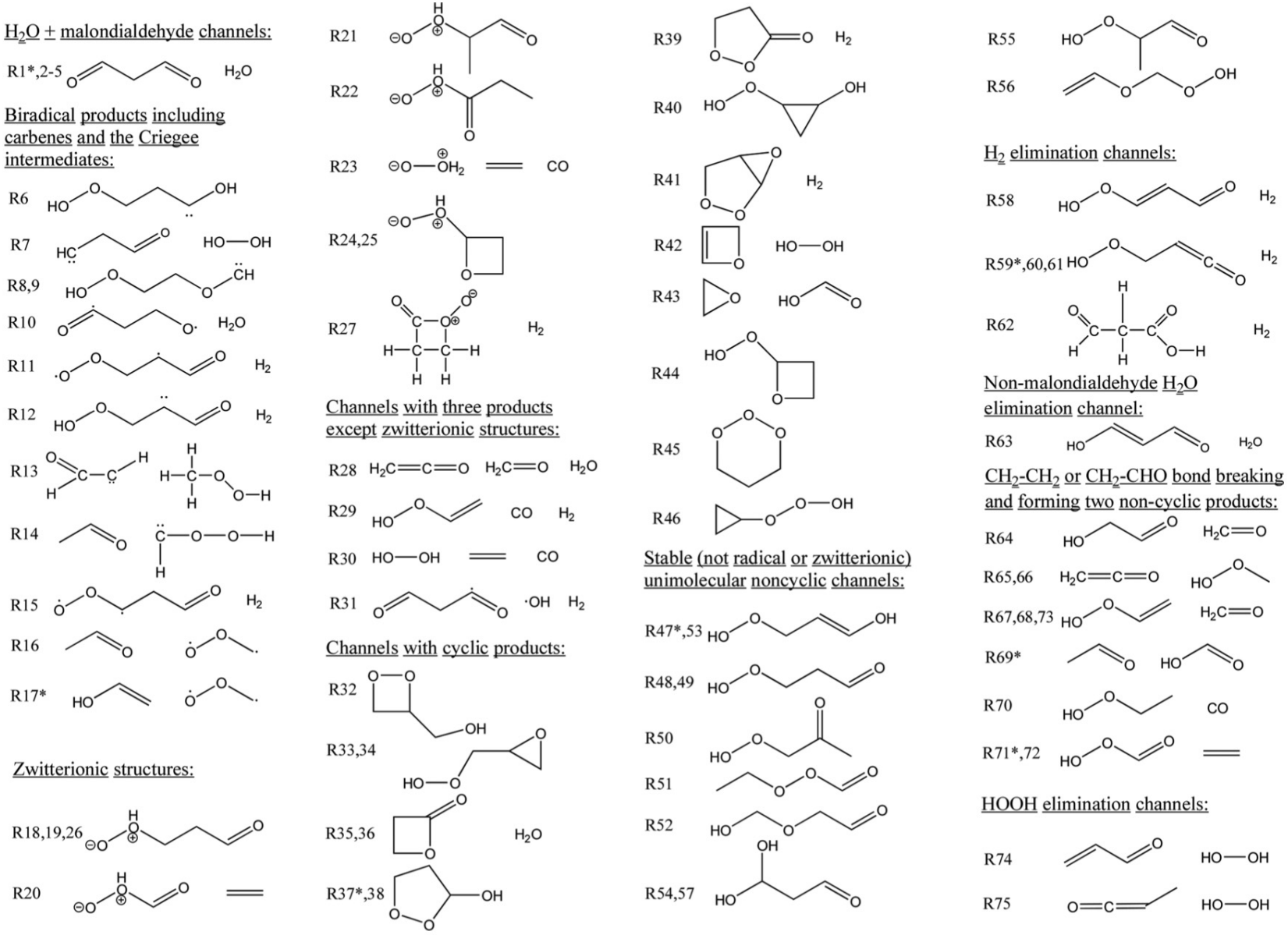


Testing YARP on a Unimolecular Decomposition Problem

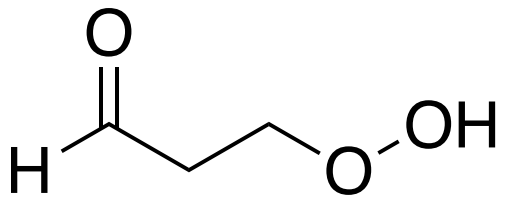


The 3-hydroperoxypropanal reaction network out to b4f4 was recently published as a benchmark for 5 reaction discovery methods.

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

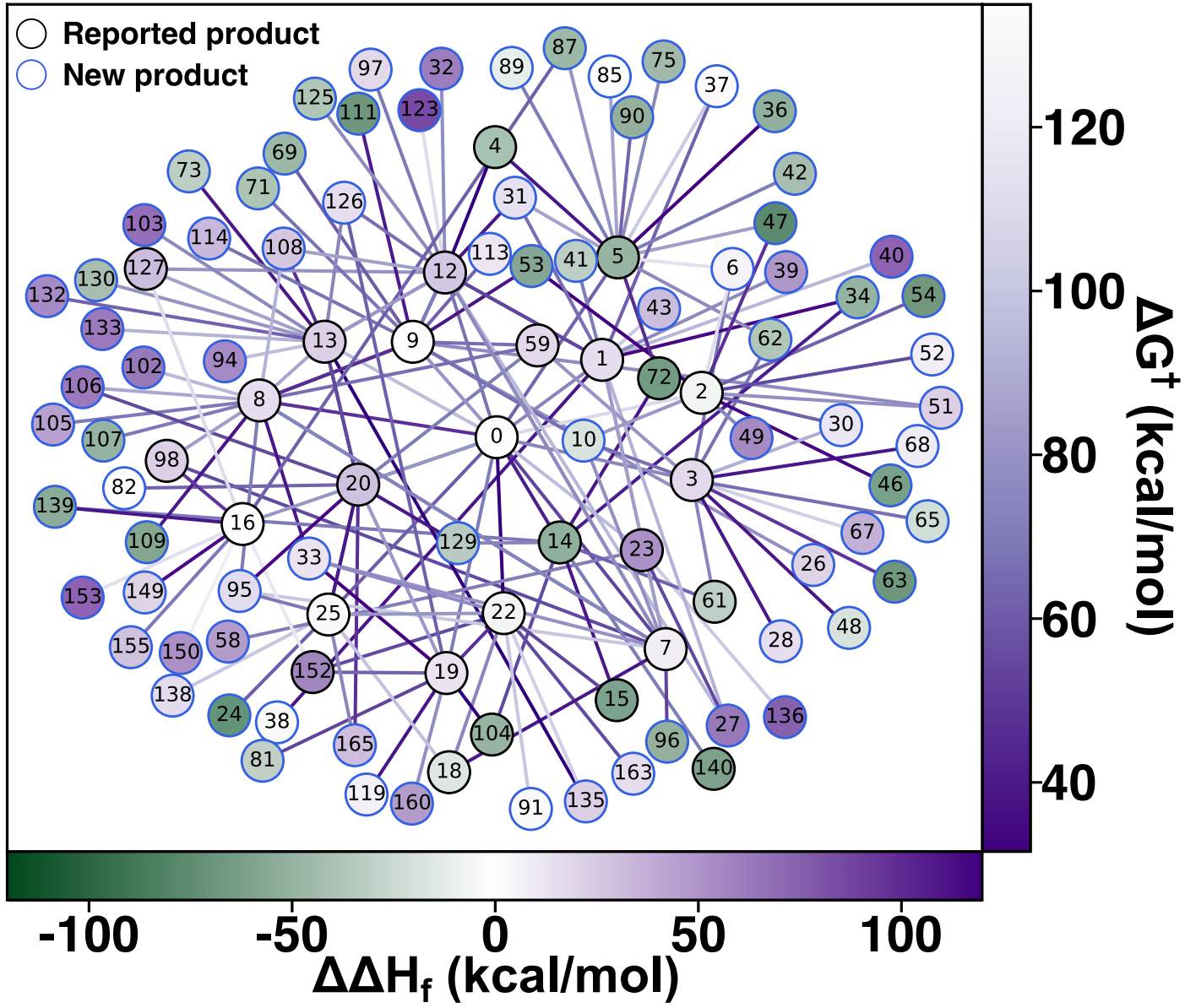
75 intended reactions discovered

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.

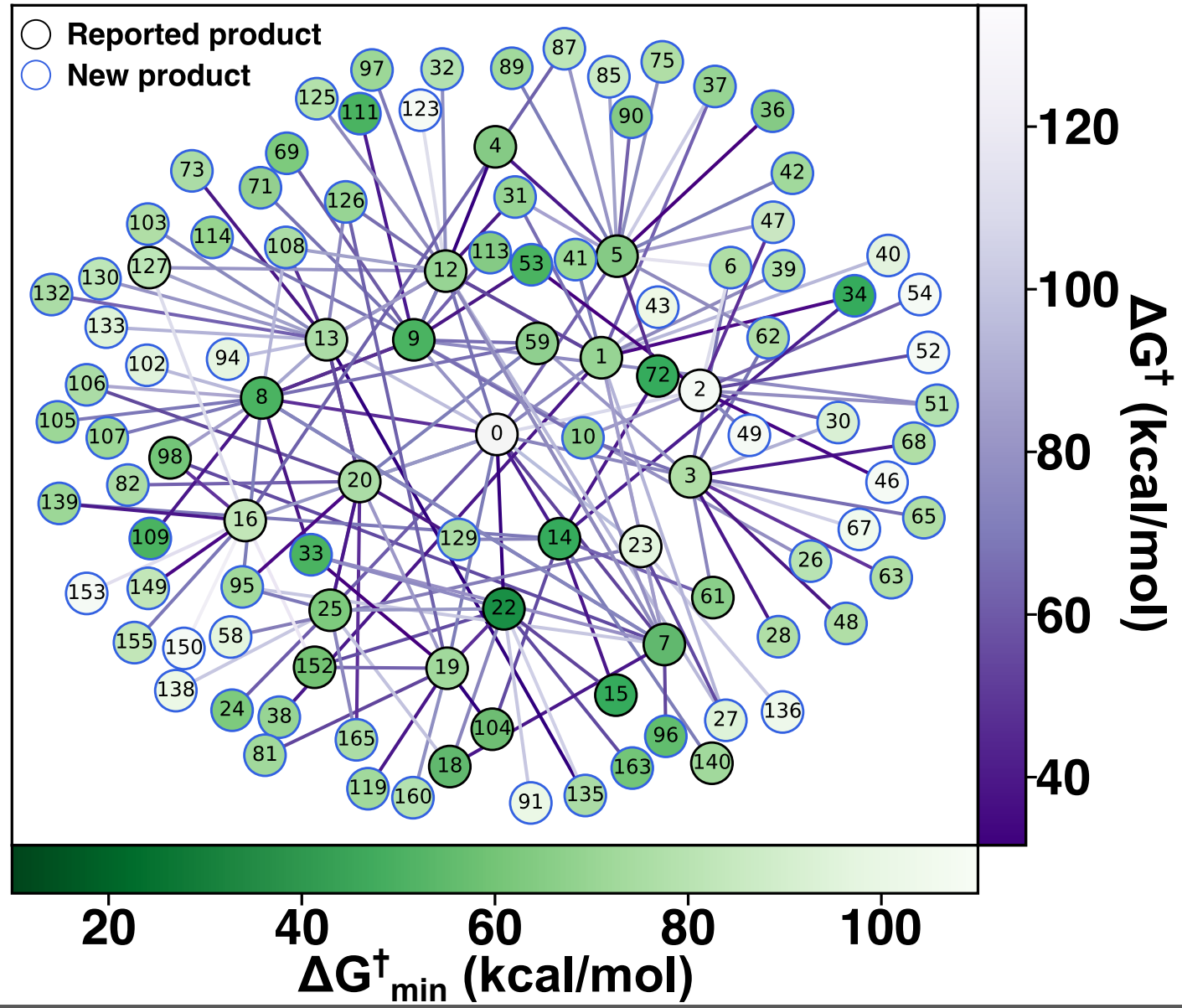
YARP finds **all known products** of this thermal decomposition network, as well as new products (77), and new reactions (157).



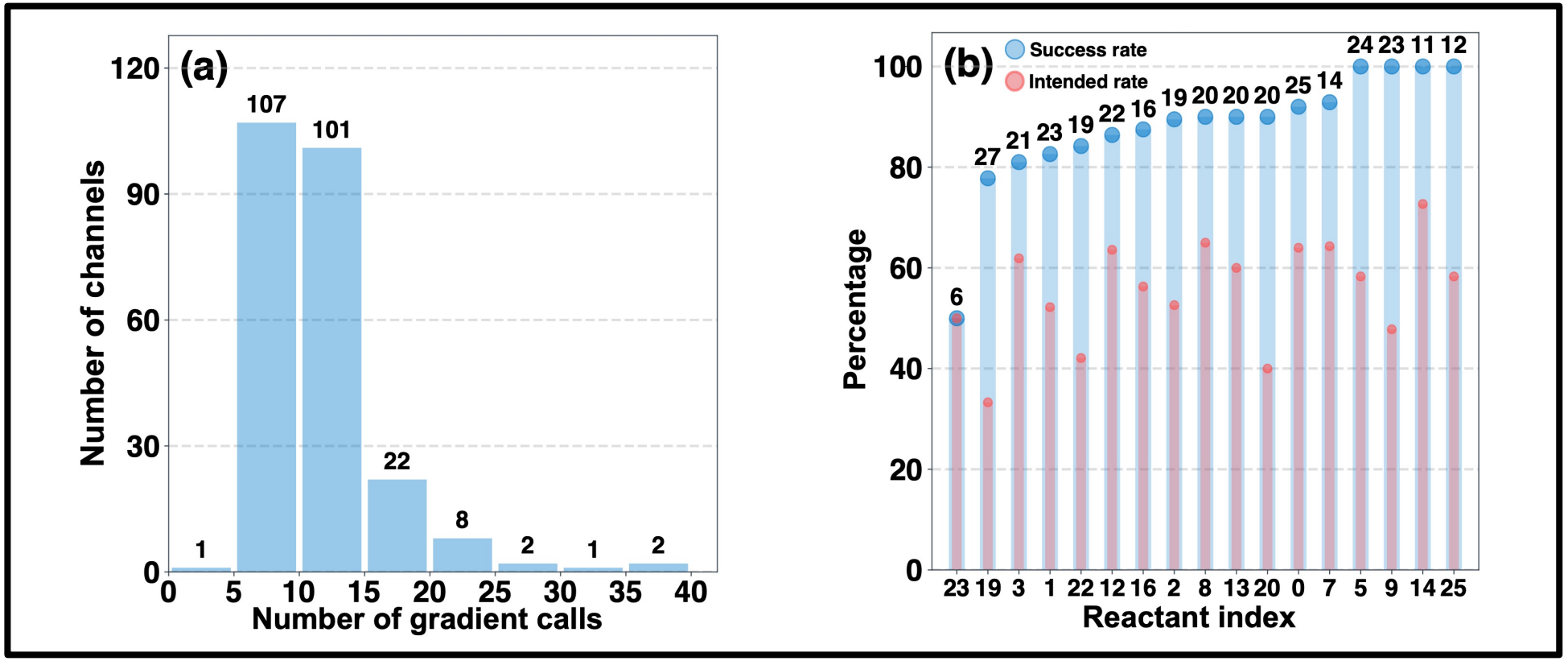
3-Hydroperoxypropanal - Reaction Network

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.

Network Exploration 2: Glucose Pyrolysis

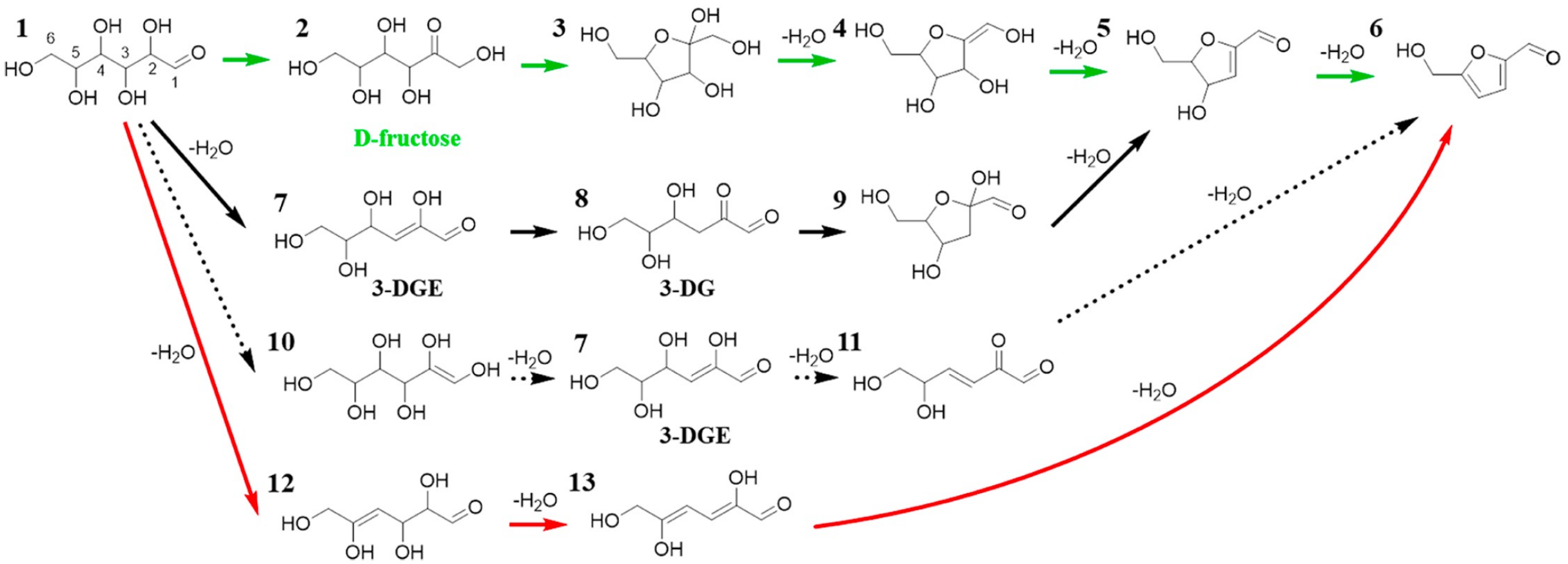
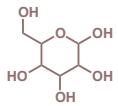


Figure 1. Proposed pathways in literature from glucose to HMF, namely the fructose path (green), 3-DG paths (black and black dotted), and direct path (red). The molecules are indicated by numbers and some key molecules are named as follows: **1**. D-glucose; **2**. D-fructose; **3**. D-fructofuranose; **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



Depth 1:

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

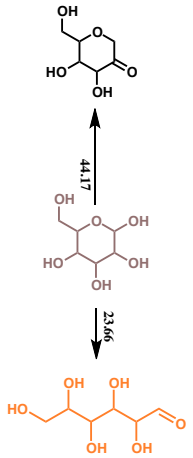
At each iteration:

(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:

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

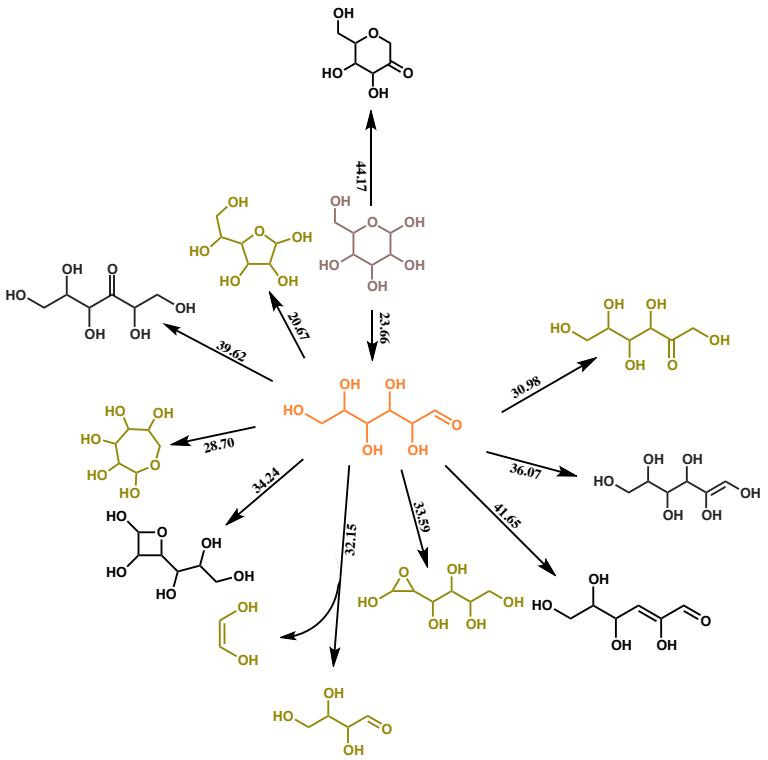
At each iteration:

(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

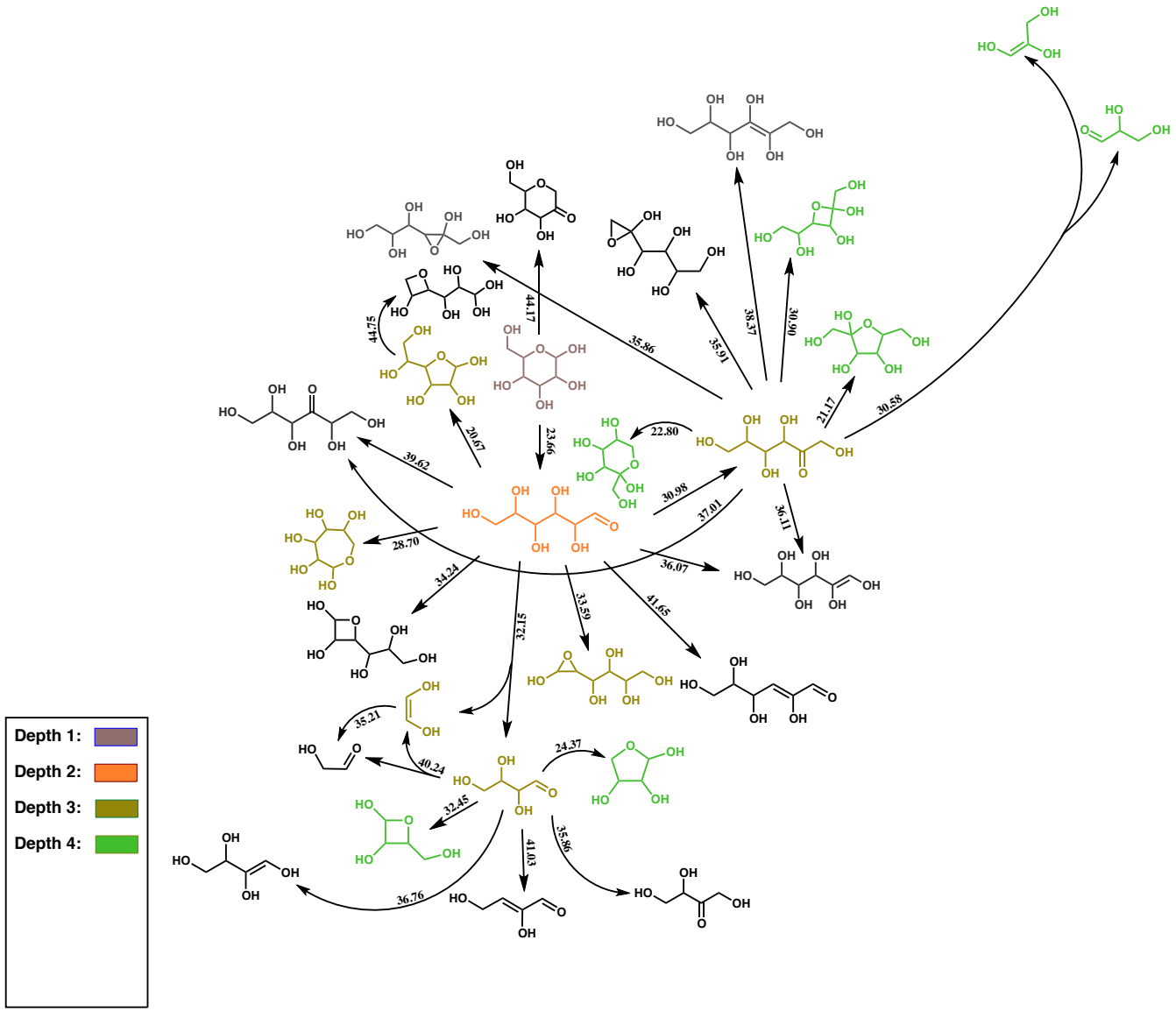
At each iteration:

(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

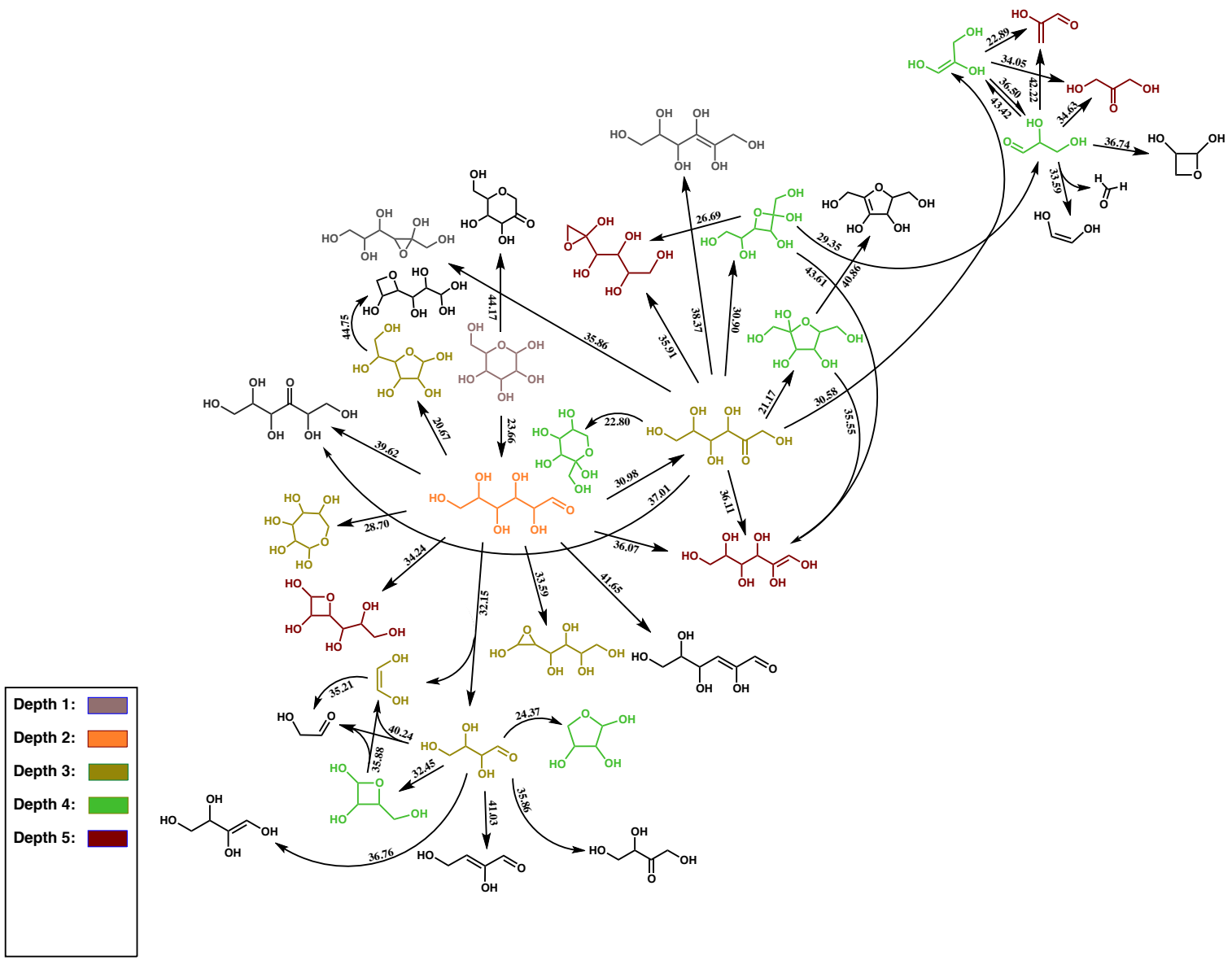
At each iteration:

(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

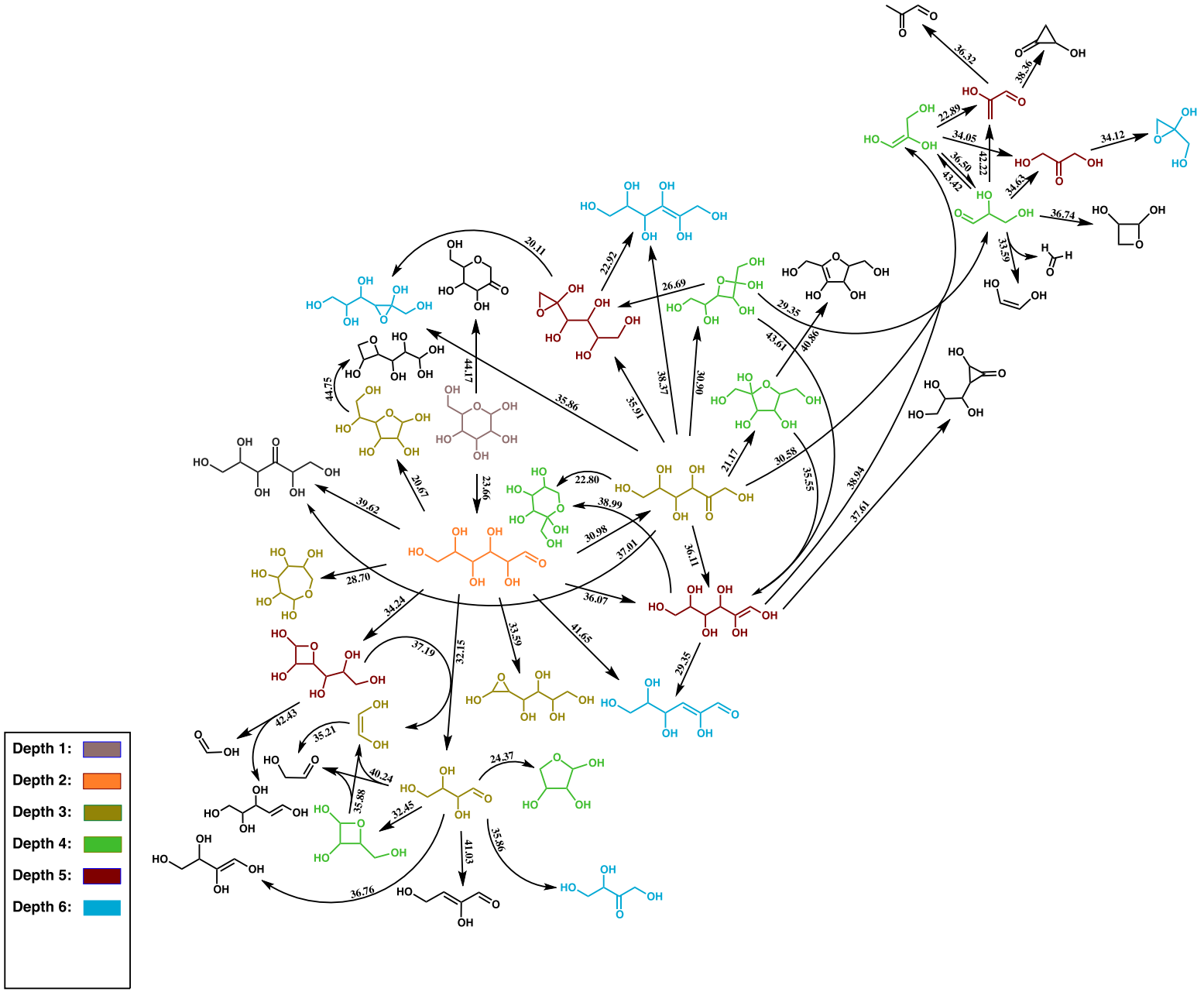
At each iteration:

(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

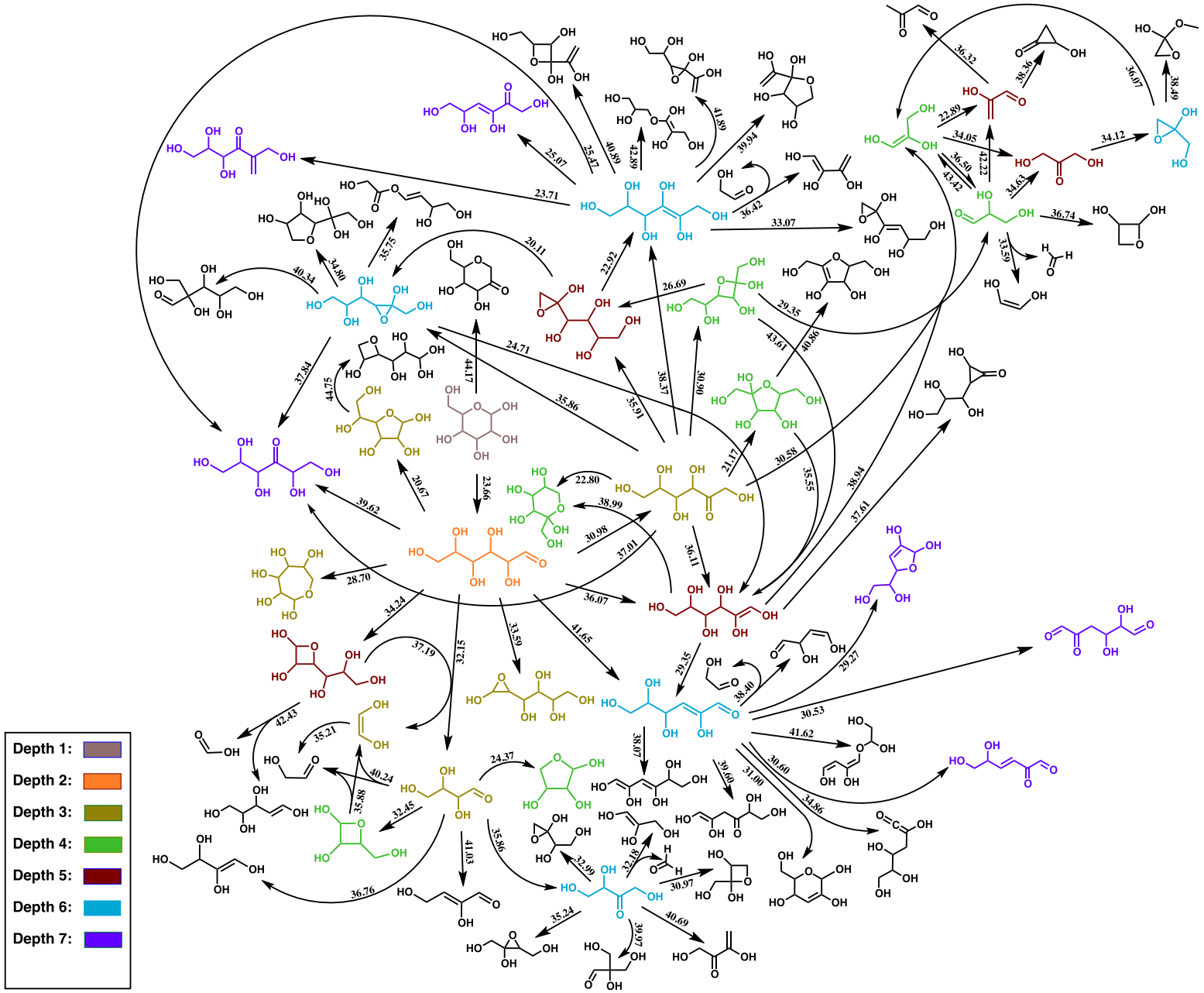
At each iteration:

(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

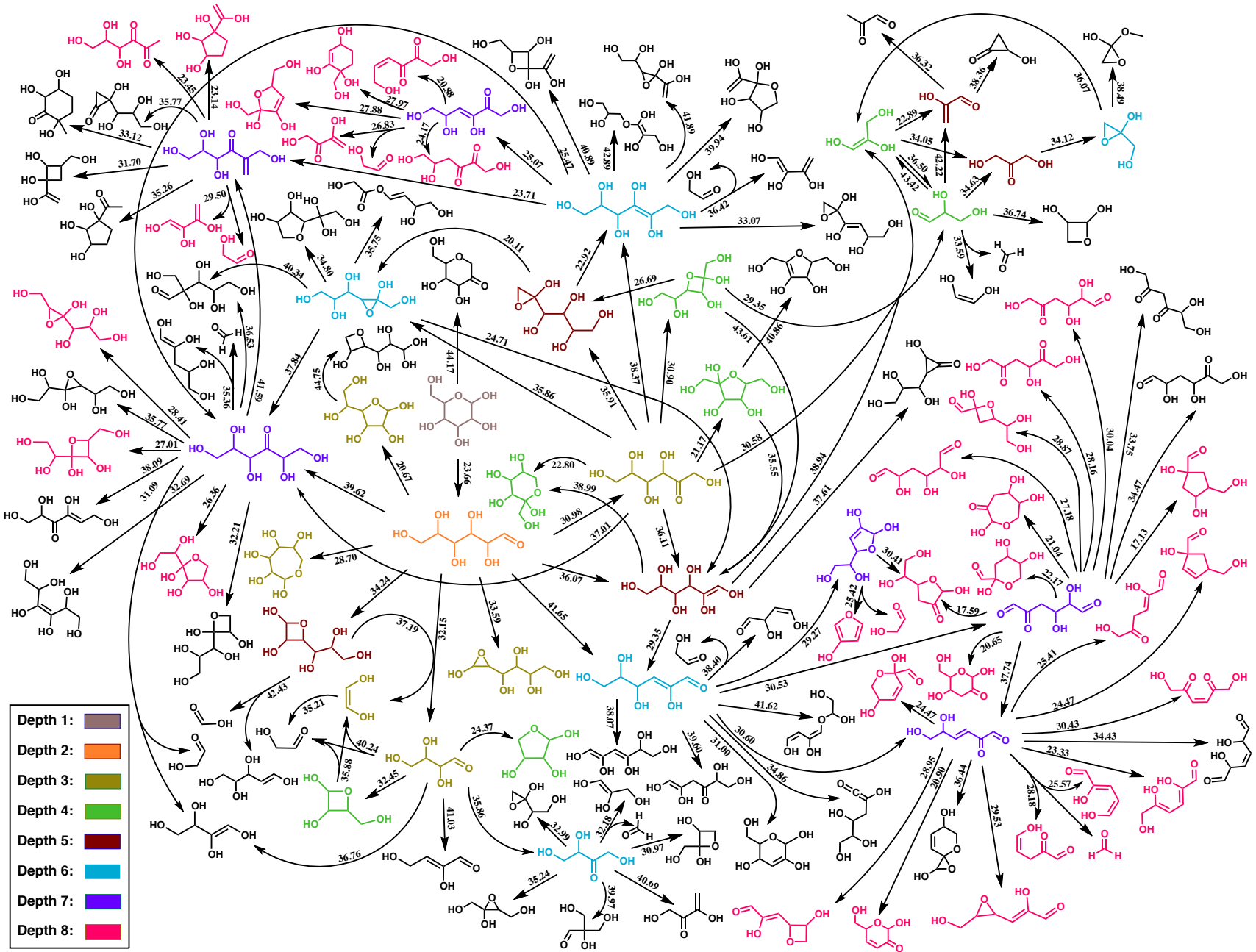
At each iteration:

(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

Outlook

The throughput enabled by YARP creates many new opportunities:

(i) Broader reaction discovery → Lowe Dataset as a discovery testbed

(ii) Generating positive and negative exemplary reaction datasets

(iii) Exploring deeper networks (e.g., materials degradation, catalysis)

In the process safety space, it seems like predicting kinetics in addition to thermochemistry could be useful.

Qiyuan, Z.; Savoie, B. M. “Simultaneously Improving Reaction Coverage and Computational Cost in Automated Reaction Prediction Tasks.” Nature Computational Science **2021**, 1, 479-490.

Outlook and Acknowledgements

Students: Qiyuan Zhao, Tyler Pasut

Project Accomplishments:

- Implemented a fully-consistent 2-bond (i.e., component) increment theory based on G4 data
- Automated model compound generation and fitting algorithms.
- Built a database infrastructure for reusing calculations and parameter fitting.
- Developed a ring-correction for TCIT to improve performance on conjugated and non-benzene structures.
- Extended TCIT to condensed phases and new thermodynamic properties and radicals.

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

P2SAC
Purdue Process Safety & Assurance Center