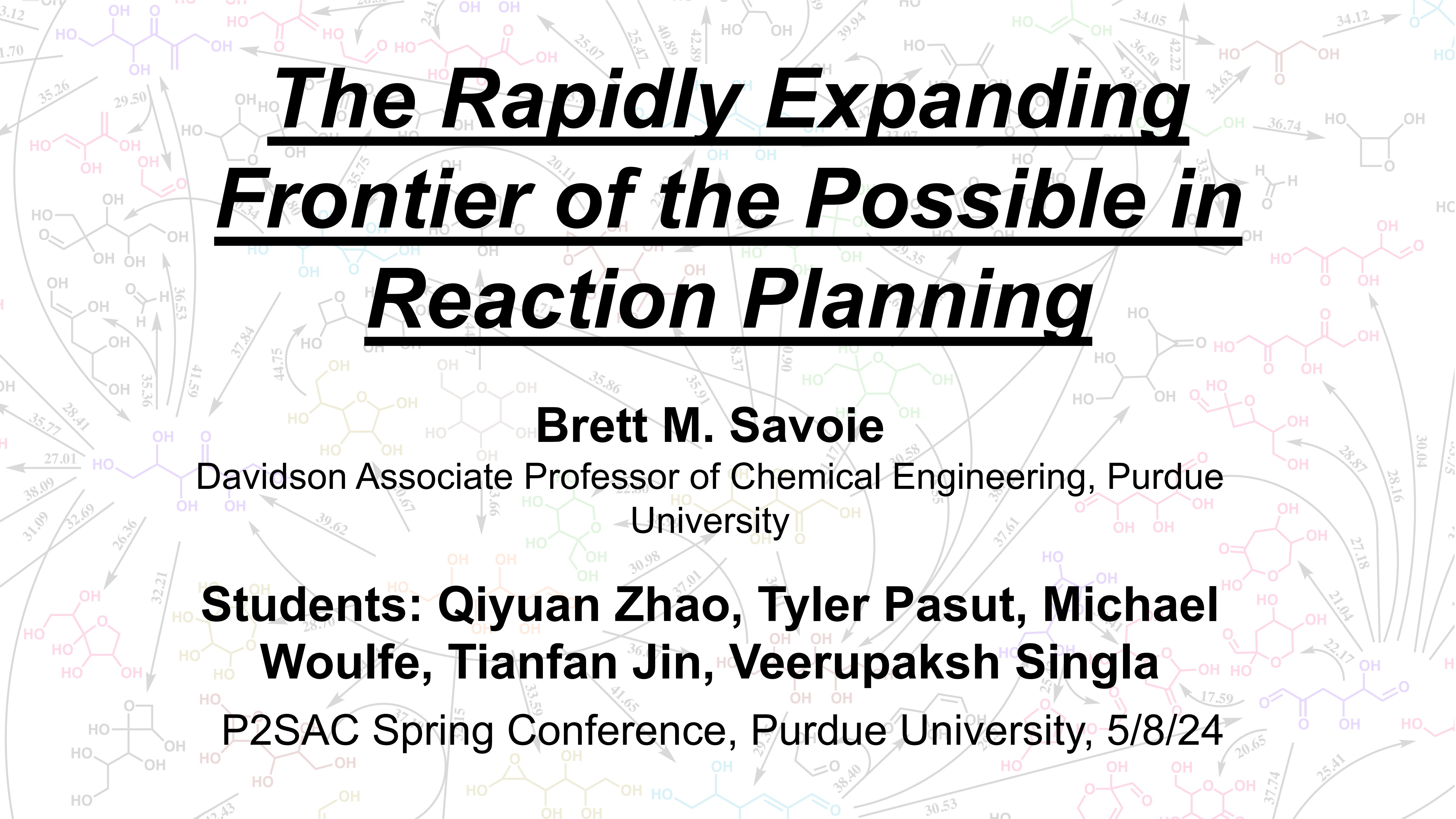


The Rapidly Expanding Frontier of the Possible in Reaction Planning

Brett M. Savoie

Davidson Associate Professor of Chemical Engineering, Purdue University

Students: Qiyuan Zhao, Tyler Pasut, Michael Woulfe, Tianfan Jin, Veerupaksh Singla

P2SAC Spring Conference, Purdue University, 5/8/24

Challenges of Contemporary Group Theories

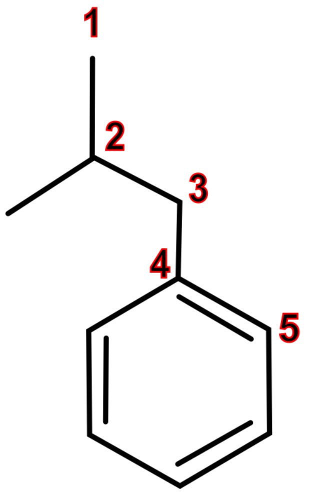
Benson Group Theory:

- 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.

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.

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

Challenges of Contemporary Group Theories

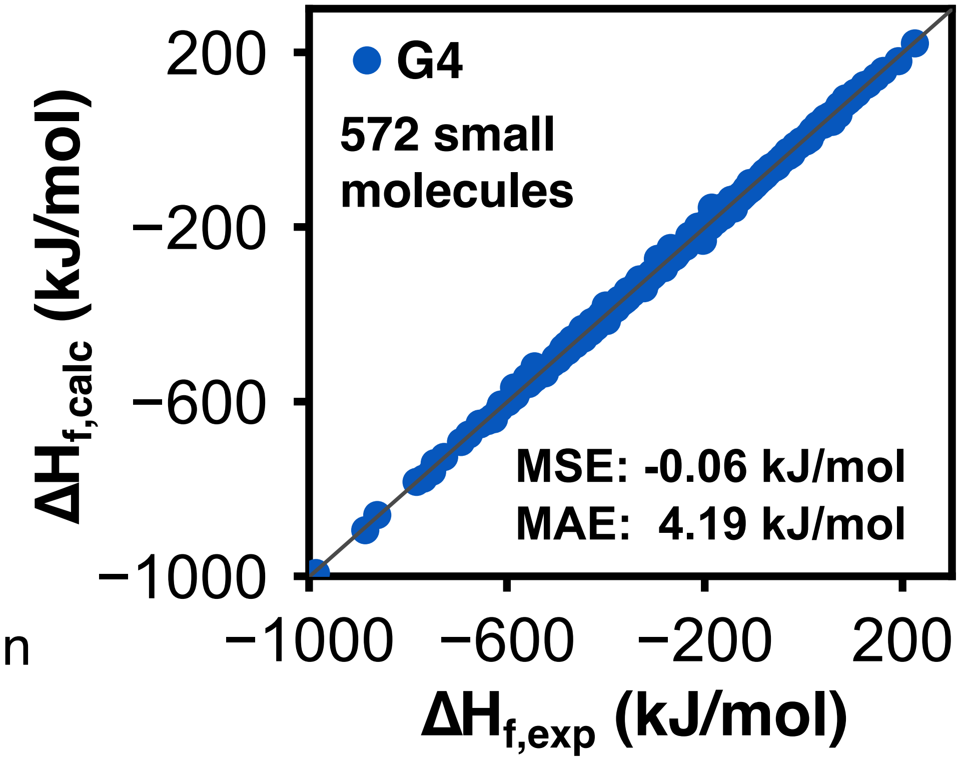
Benson Group Theory:

- 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.

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.

ΔH_f **from 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

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.

- Grou
on trus
data, a

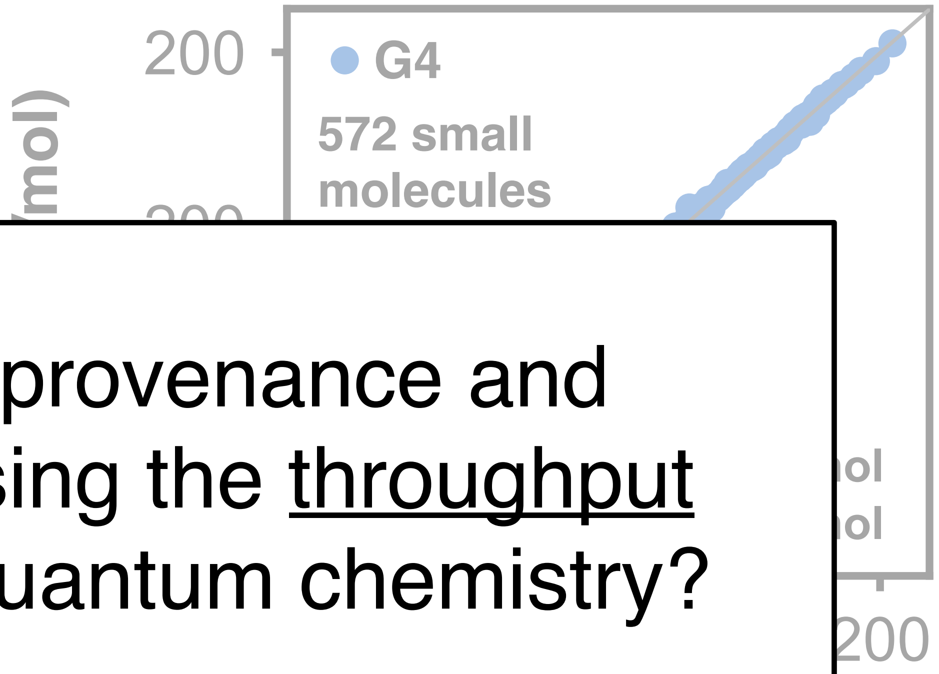
Prob

- **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.

ΔH_f from modern quantum chemistry



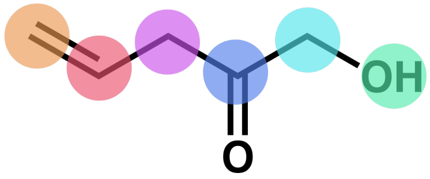
Zhao, Q.; Savoie, B. M.; Enthalpy of Formation Prediction via a fully Self-Consistent Component Increment Theory. *J. Chem. Info. Model.* **2020**, 60, 2199-2207

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

Graphical Decomposition of Model Compounds

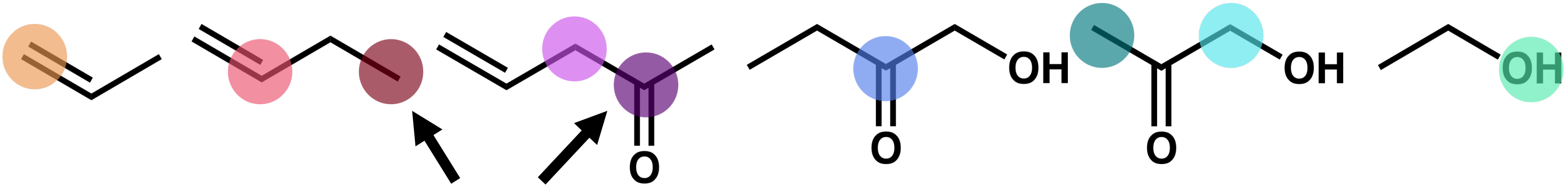
How will we select molecules for parameterizing TCIT components?

identify components*



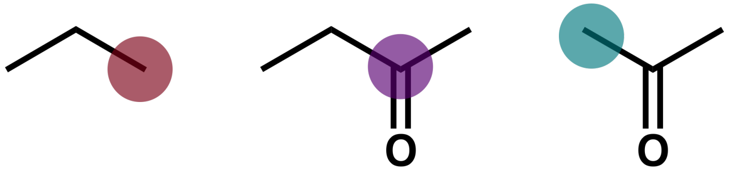
1-hydroxy-pent-2-ene-2-one

Recursively generate smallest acyclic model compounds

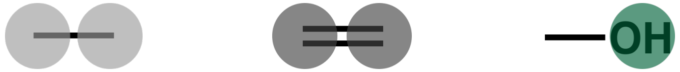


new groups

Resolve dependencies

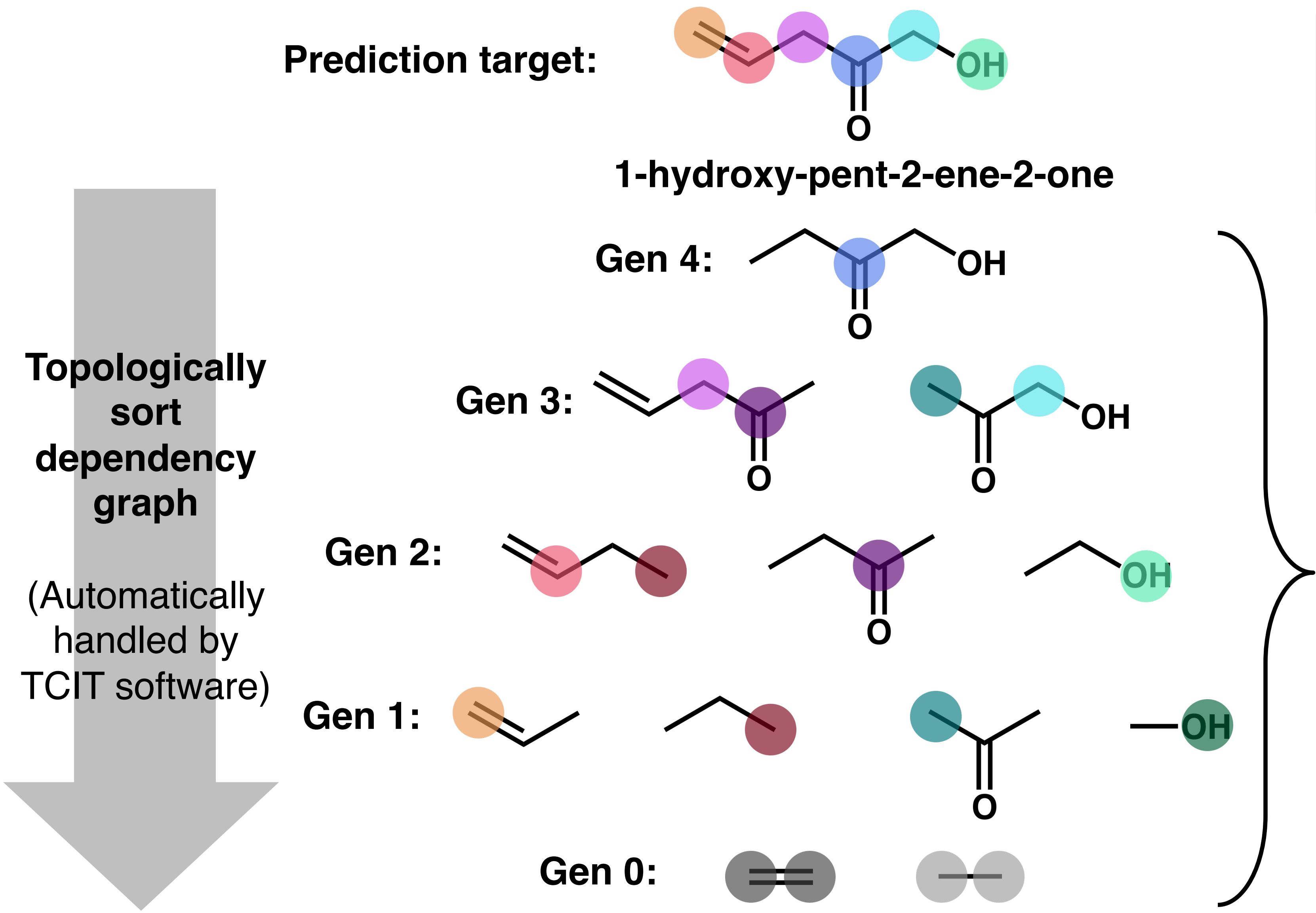


Resolve rank deficiency with elementary constraints



Zhao, Q.; Savoie, B. M. *J. Chem. Info. Model.* 2020, 60, 2199-2207.

Graphical Decomposition of Model Compounds



Topologically sort dependency graph

(Automatically handled by TCIT software)

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

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

no experimental data

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

Zhao, Q.; Savoie, B. M. *J. Chem. Info. Model.* 2020, 60, 2199-2207.

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.

Zhao, Q.; Savoie, B. M. *J. Chem. Info. Model.* 2020, 60, 2199-2207.

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

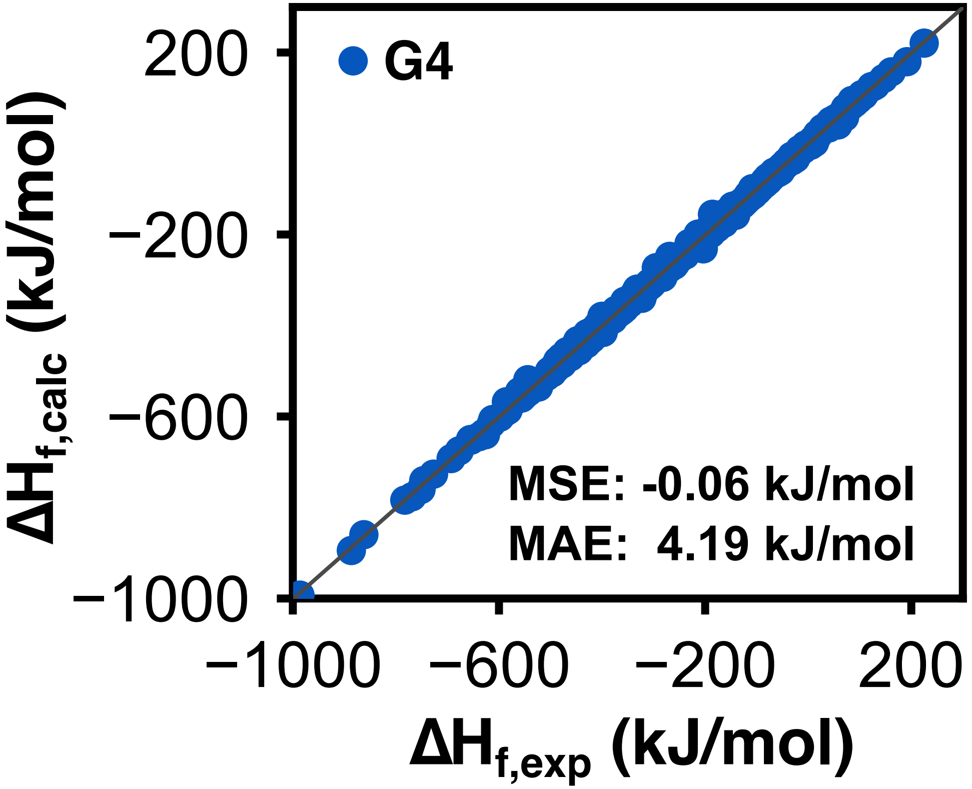
- Initial benchmarking set consists of ~1100 **linear** C,H, and O containing compounds from PNK[1]

[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.

572 small 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

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]

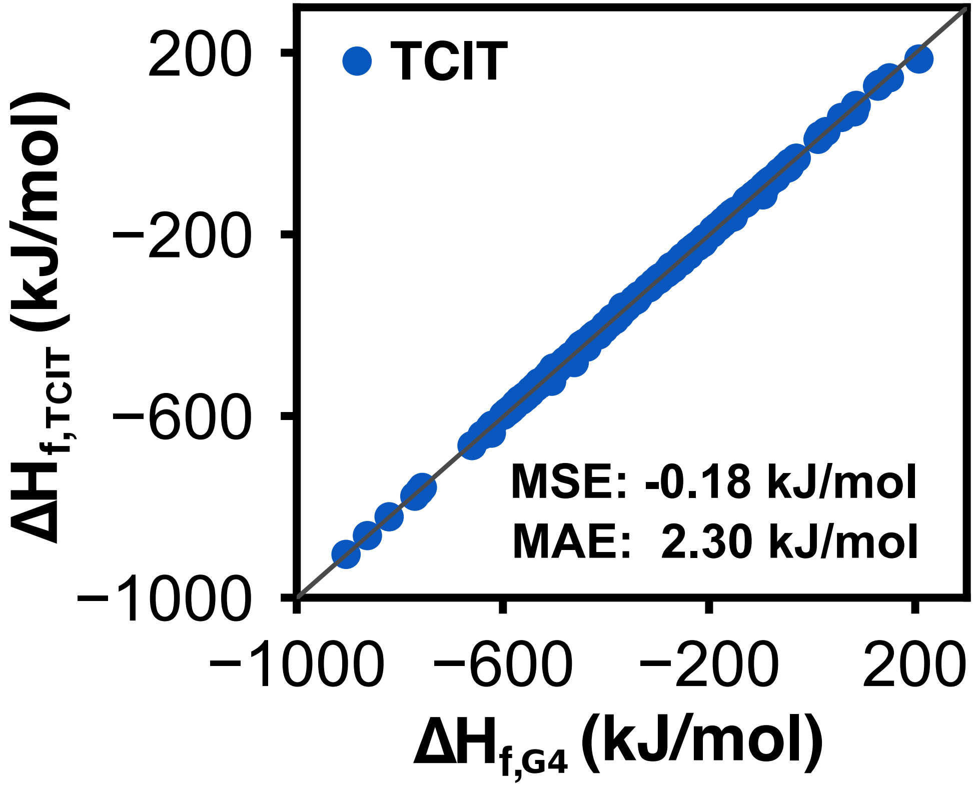
[1] J. B. Pedley, R. D. Naylor, S. P. Kirby “Thermochemical Data of Organic Compounds” 2nd ed. 1986

- 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.

150 medium 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

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

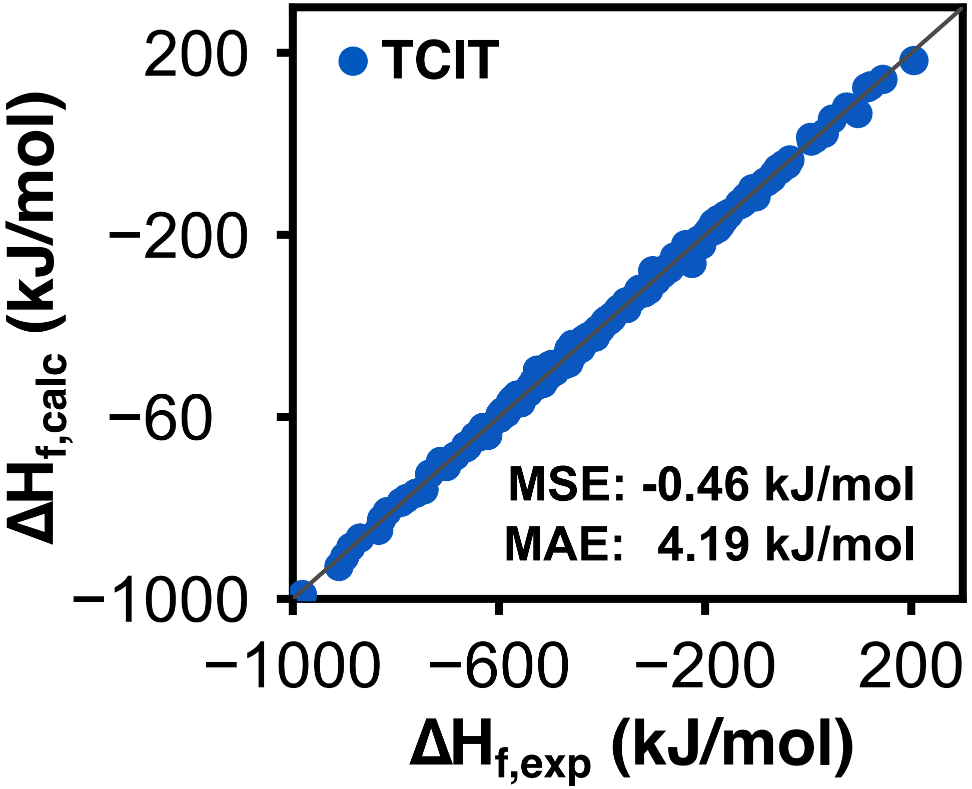
- PNK is a core dataset for fitting Benson groups

- ~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

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

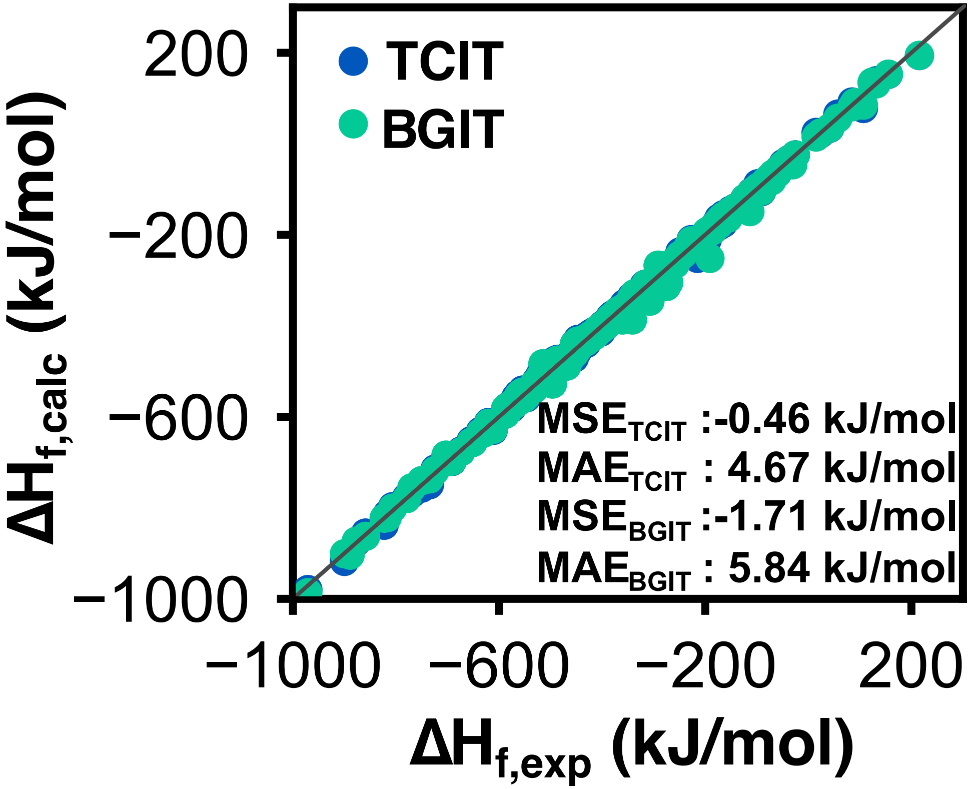
- PNK is a core dataset for fitting Benson groups

- ~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

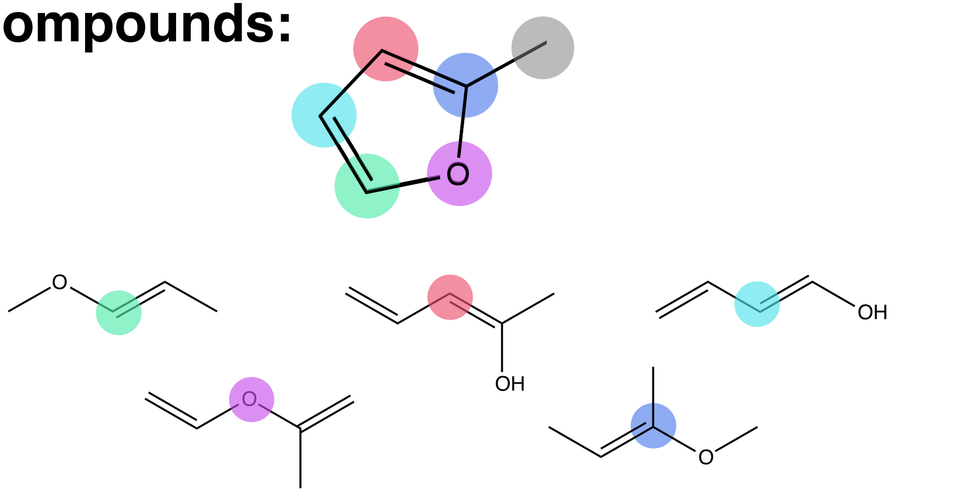
TCIT shows comparable performance to BGIT/CHETAH but is derived exclusively from extensible G4 data.

Extension to Ring-Containing Molecules

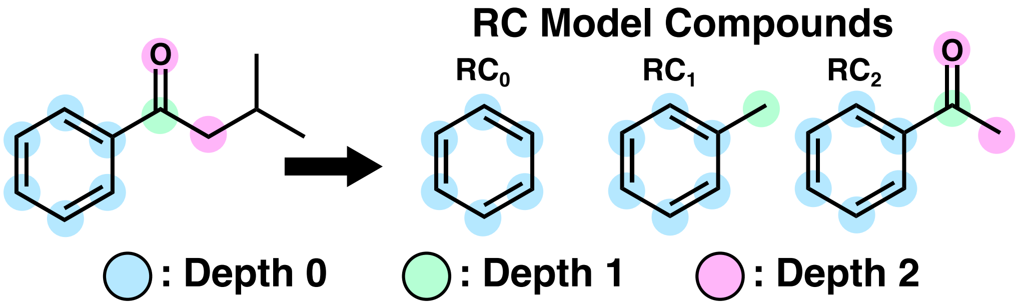
- Ring-containing molecules have additional strain and/or conjugation corrections that exacerbate the extensibility issues of Benson Theory.

- 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:



Technical Developments



Method 1: Use RC$_1$ based model parameterized to G4 data.

Method 2: Use graph-NN to predict RC$_0$-RC$_2$

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

$\mathrm{RC} = H_f(ring) -$ H_f(🔴)- H_f(🔵)- H_f(🟢)- H_f(🟣)- H_f(🔵)-H_f(⚪)

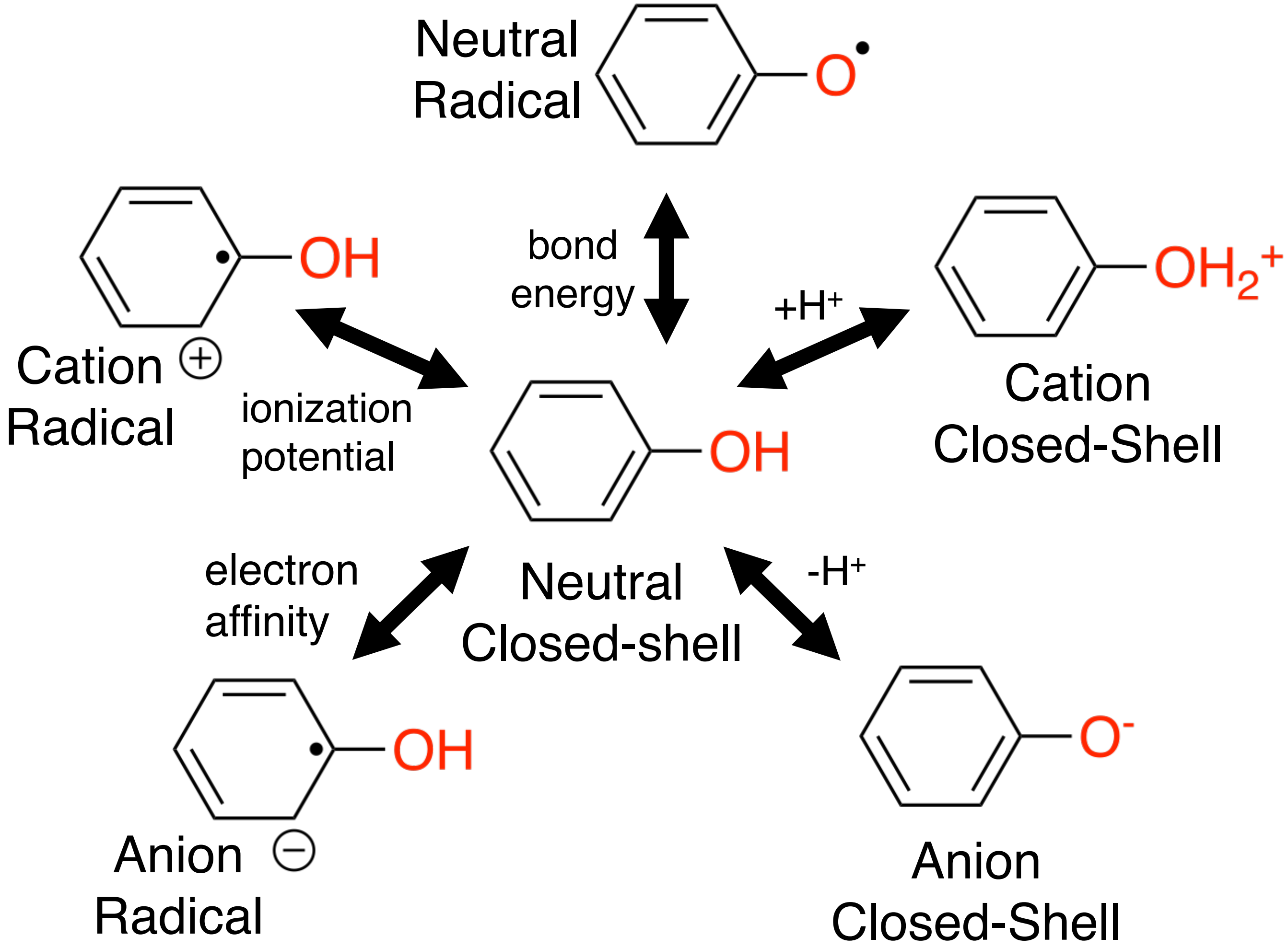
Zhao, Q.; Iovaanc, N.; Savoie, B. M. *J. Chem. Info. Model.* 2021, 61, 5013-5027

Extending TCIT to Radicals and Ions

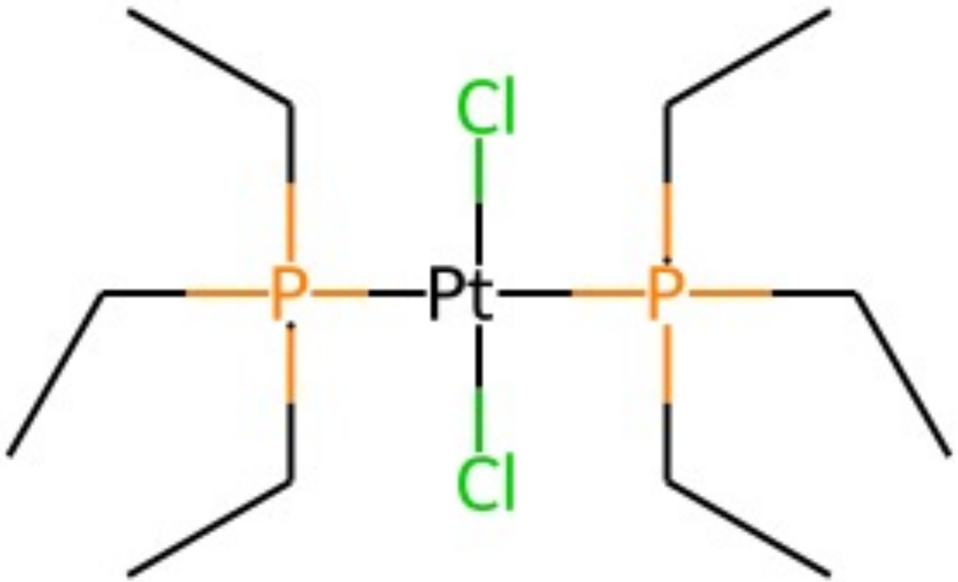
A recurring question is when will TCIT support predictions on **radicals** and **ions**?

TCIT already covers neutral close-shell species, so these extensions require us only to predict the difference between the target and the **nearest closed-shell neutral**.

This amounts to developing models to predict IP/EA/+H⁺/-H⁺



Extending TCIT to Organometallics



High level calculations on transition metals are going to get expensive, so we want to avoid calculating components directly

$$\Delta H_f^\circ = \sum_i C_{\mathrm{TCIT,i}} + \sum_{i \in M-X} \mathrm{BDE_i}$$

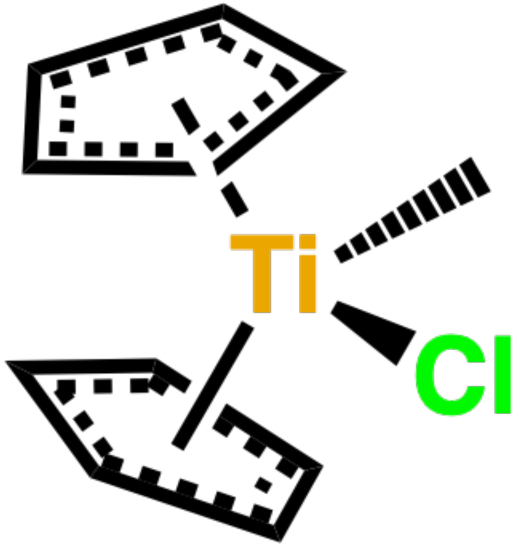
Idea: Only use TCIT for the ligands and approximate the remainder with bond energies

Species	ΔHf Contribution (kcal/mol)	Source
$PtCl_2$ (g)	26.41	Schafer 1975
Pt-P 1 (B3LYP)	-39.91	Craciun 2010
Pt-P 2 (B3LYP)	-34.81	Craciun 2010
PEt_3 (TCIT)	-35.21 * 2	
Predicted	-118.73	
trans-$Pt(PEt_3)_2Cl_2$	-118.3 ± 1.9	Takhin 1984

Sanity checks work surprisingly well

(Note: $PtCl_2$ availability probably makes it better than it ought to be**)**

Extending TCIT to Organometallics



Species	ΔHf (kcal/mol)	Source
Cp^- (TCIT/G4)	19.9 (x2 for compound)	TCIT/G4
Ti-Cl ($TiCl_4$/4)	-43.6	Calhorda 1986
Ti-Cp ($Ti(Cp)_2Cl_2$)	-7.9 (x2 for compound)	Calhorda 1986
$CH3^-$ (TCIT/G4)	32.41	TCIT/G4
$Ti\text{-}CH_3$ ($Ti(Cp)_2(CH_3)_2$)	-38.6	Calhorda 1986
Predicted	-25.8	
$Ti(Cp)_2(Cl)(CH_3)$	-29.8 +/- 3.0	Calhorda 1986

High level calculations on transition metals are going to get expensive, so we want to avoid calculating components directly

$$\Delta H_f^\circ = \sum_{\mathrm{i}} C_{\mathrm{TCIT,i}} + \sum_{i \in M-X} \mathrm{BDE_i}$$

Idea: Only use TCIT for the ligands and approximate the remainder with bond energies

Sanity checks work surprisingly well

(Note: $PtCl_2$ availability probably makes it better than it ought to be**)**

Extending TCIT to Organometallics

High level calculations on transition metals

to

y

E_i

approximate the remainder with bond energies

TCIT now contains all CAVs necessary to predict ΔH_f of 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.

ply

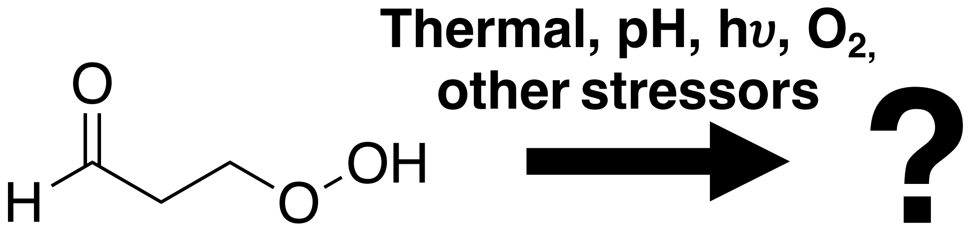
makes it better than it ought to be)

Species	ΔHf (kcal/mol)	Source
Cp^- (T		
Ti-Cl (		
Ti-Cp (		
$CH3^-$ (		
Ti-CH_3		
Predicted	-25.8	
Ti$(Cp)_2$(Cl)(CH_3)	-29.8 +/- 3.0	Calhorda 1986

The Reaction Prediction Problem

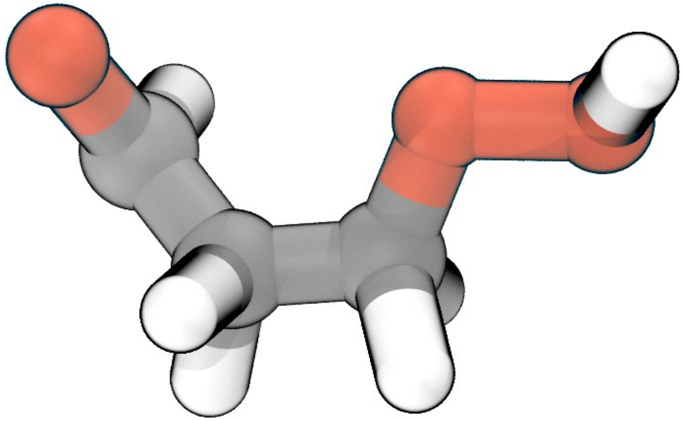
A → B : When we know the reactants and products, mature quantum chemistry tools exist to characterize transition states and establish pathways

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



Thermal, pH, hυ, O_2, other stressors

3-hydroperoxypropanal

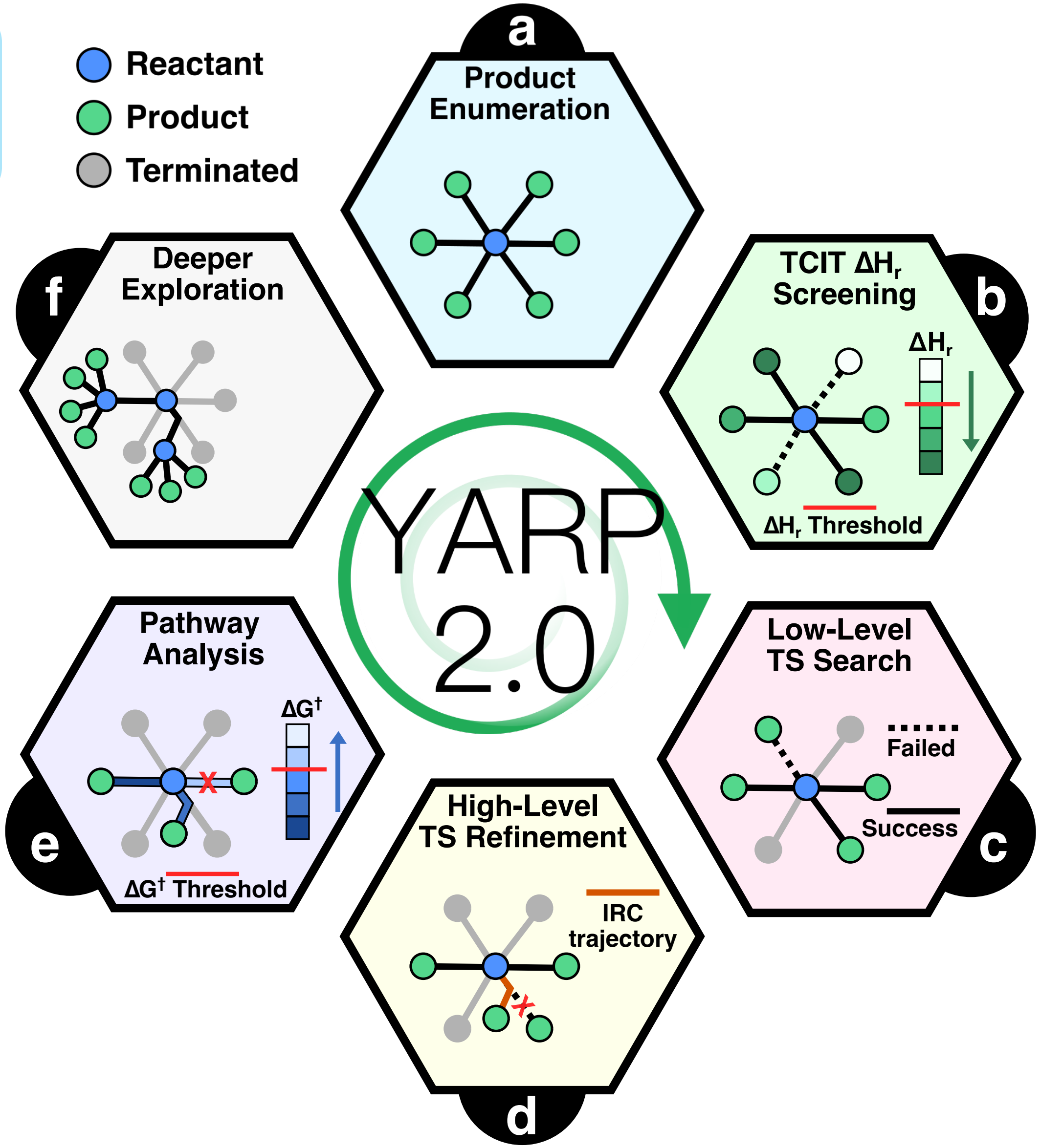


Yet Another Reaction Program (YARP)

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.
- Solving the **A→?** problem is the prerequisite for reaction network prediction.

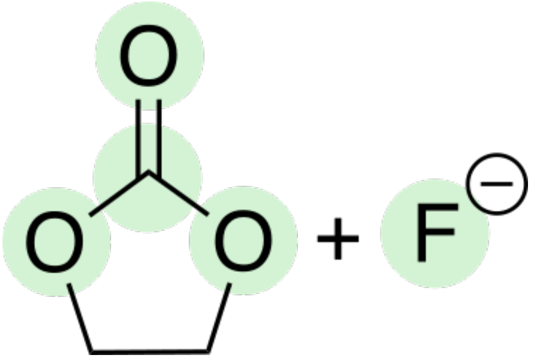


Zhao, Q.; Savoie, B. M. *Nature Comput. Sci.* 2021 ; Zhao, Q.; Savoie, B. M. *Angew. Chemie* 2022.

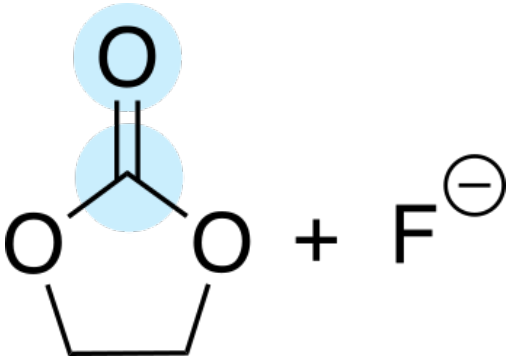
YARP: Elementary Reaction Step(s)

Polar and pericyclic organic reactions are decomposed into elementary electron donor and acceptor reactions with concomitant σ-bond breaks

bnfn will refer to σ-bond changes, π-bonds are allowed to arbitrarily rearrange.

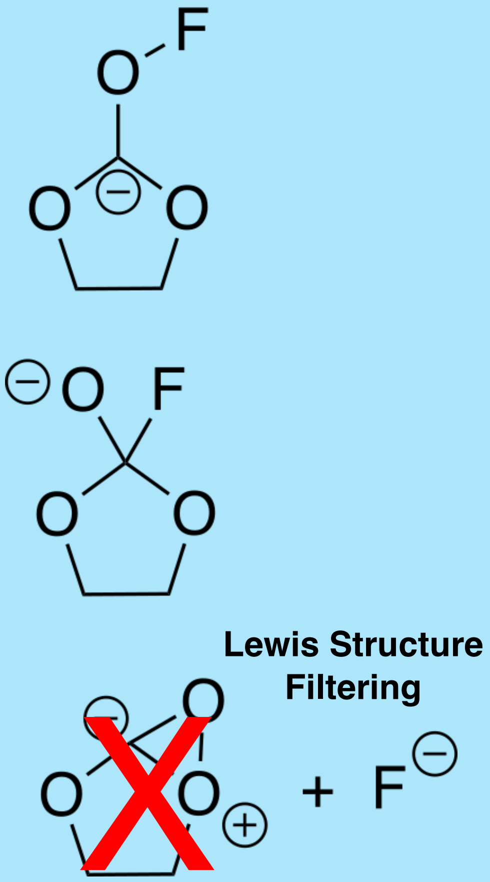


Lone-Pair Donors



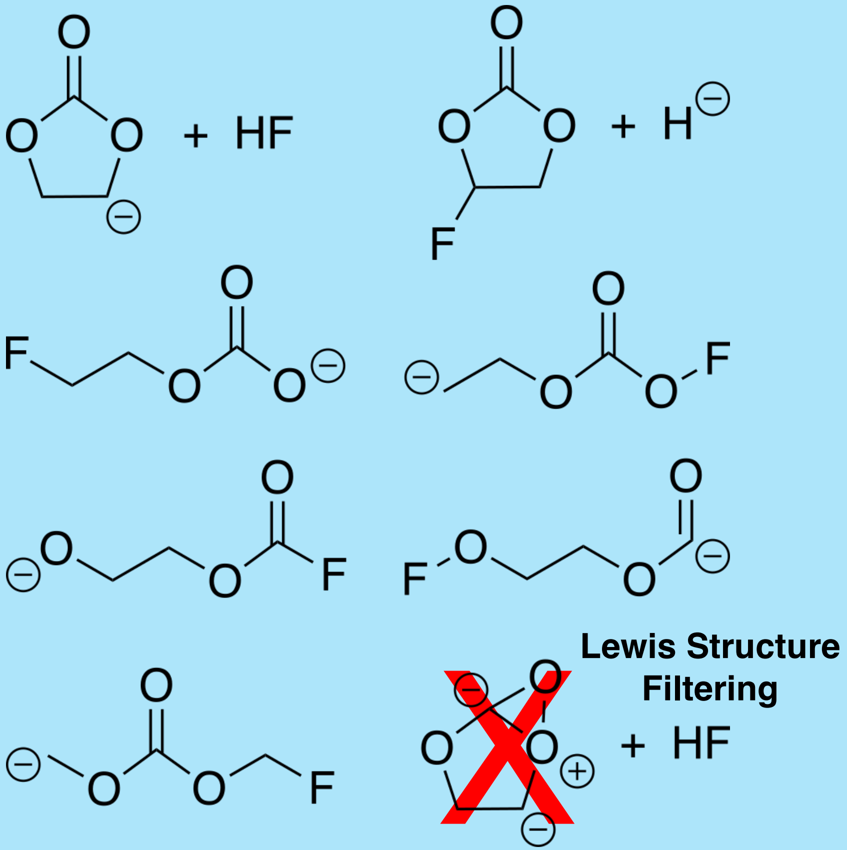
Lone-Pair Acceptors

Form 1 Products



Lewis Structure Filtering

Break 1 Form 1 Products



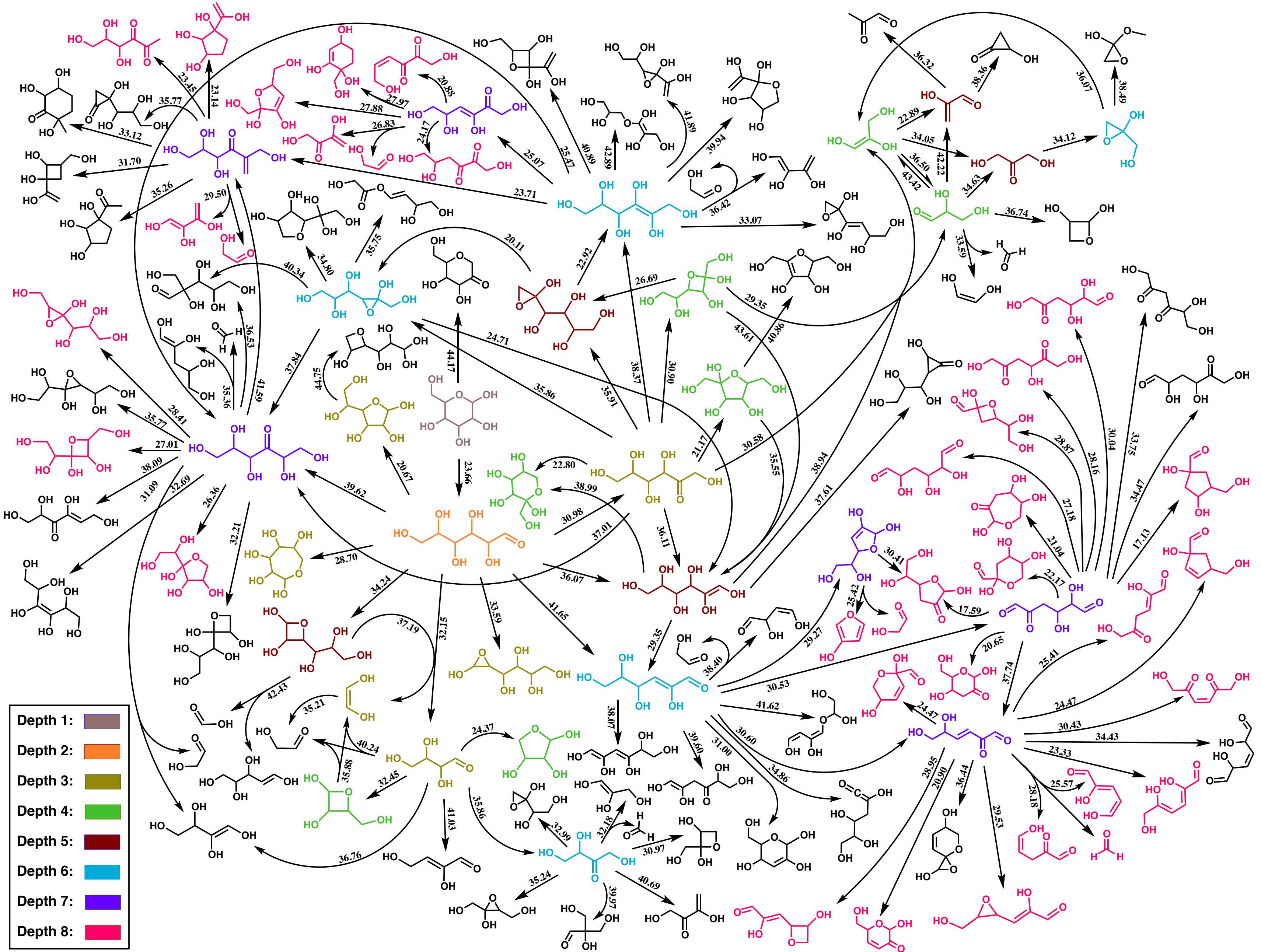
Lewis Structure Filtering

+ 28 others

Hsu, H. H., Zhao, Q., Savoie, B. M. *In Preparation*

β-D-Glucose Pyrolysis Network Exploration

Zhao, Q.; Savoie, B. M. Proc. Nat. Acad. Sci. 120, 2023



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

At each iteration:

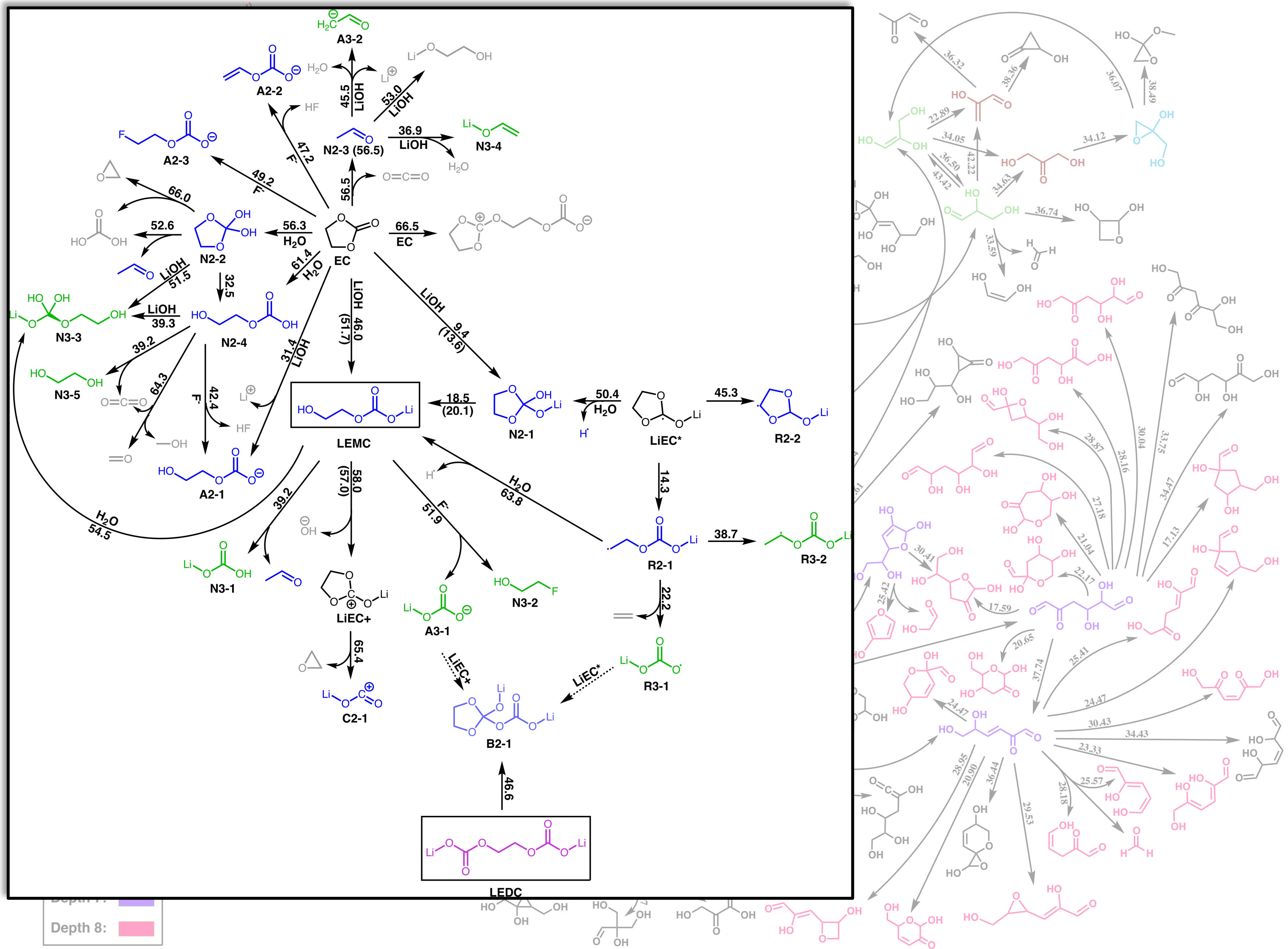
(1) all b2f2 reactions are explored for active nodes.

(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

β-D-Glucose Pyrolysis Network Exploration

Zhao, Q.; Savoie, B. M. Proc. Nat. Acad. Sci. 120, 2023



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

At each iteration:

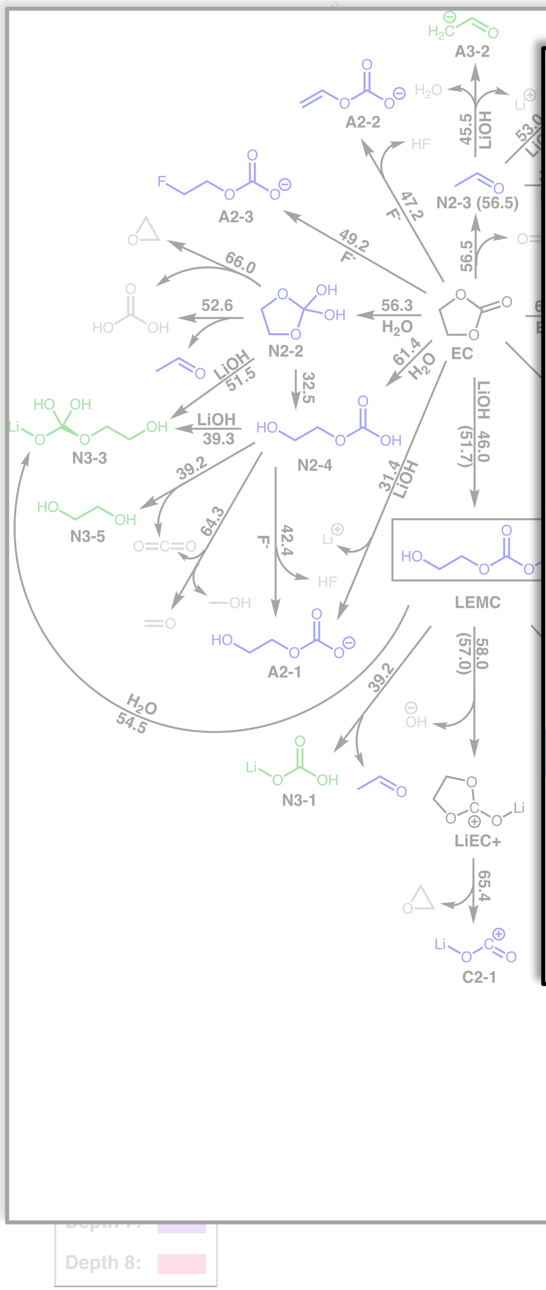
(1) all b2f2 reactions are explored for active nodes.

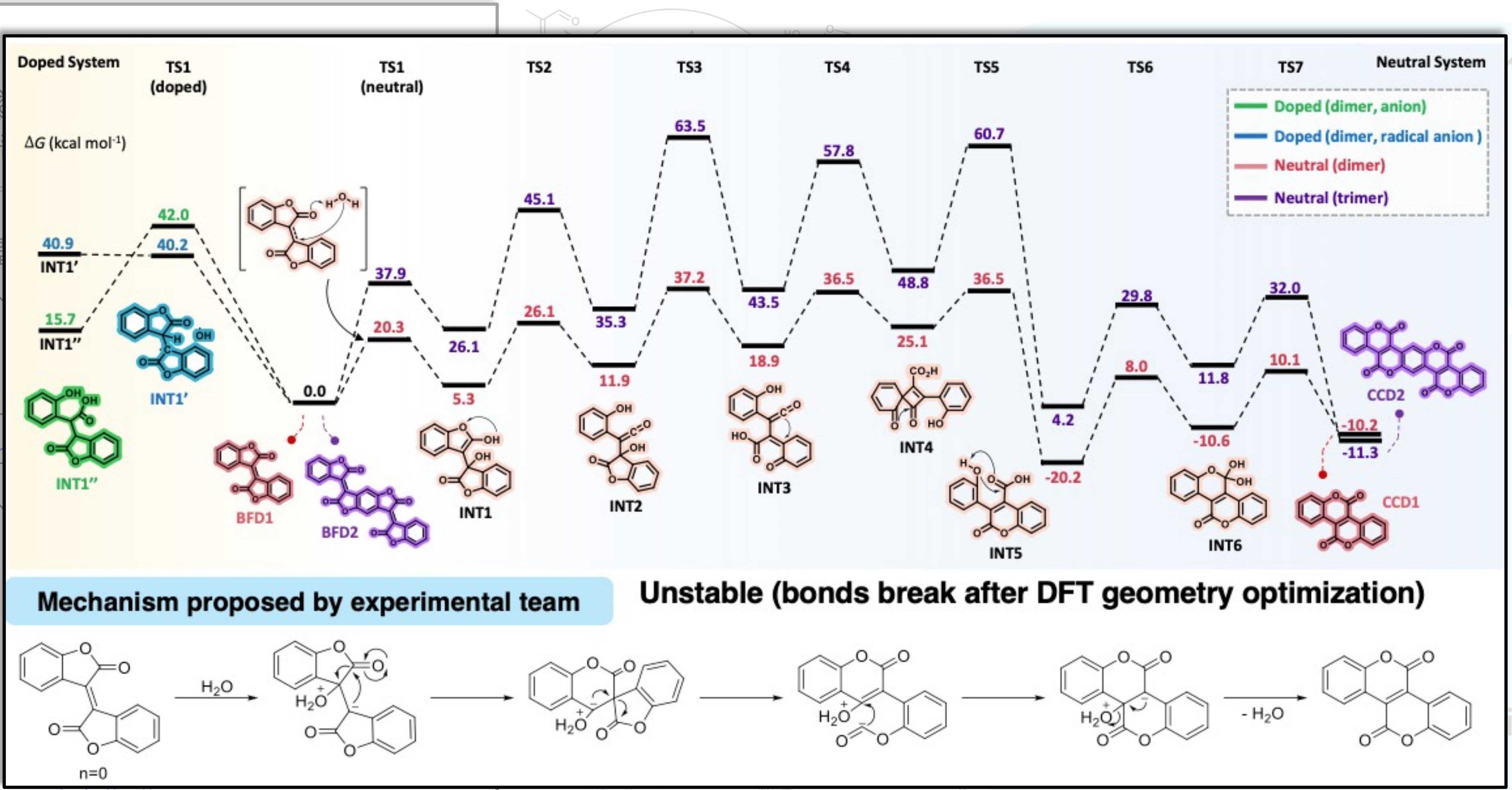
(2) Active nodes are determined by the minimum barrier to a given product (with a window)

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

β-D-Glucose Pyrolysis Network Exploration

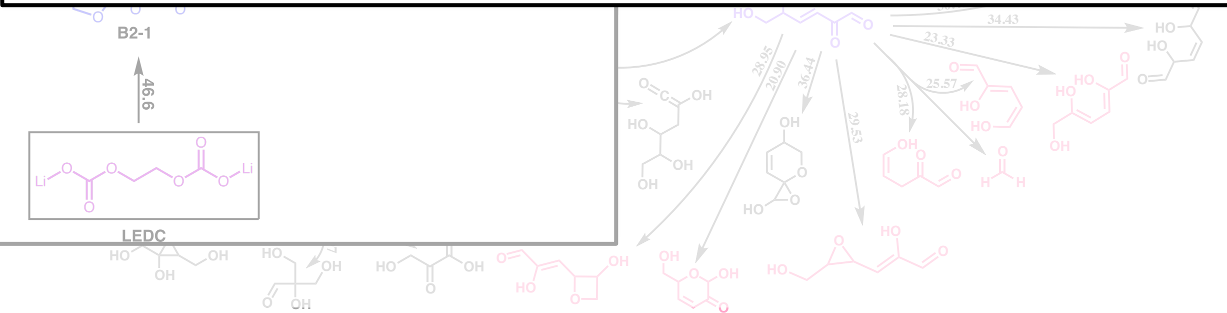
Zhao, Q.; Savoie, B. M. Proc. Nat. Acad. Sci. 120, 2023





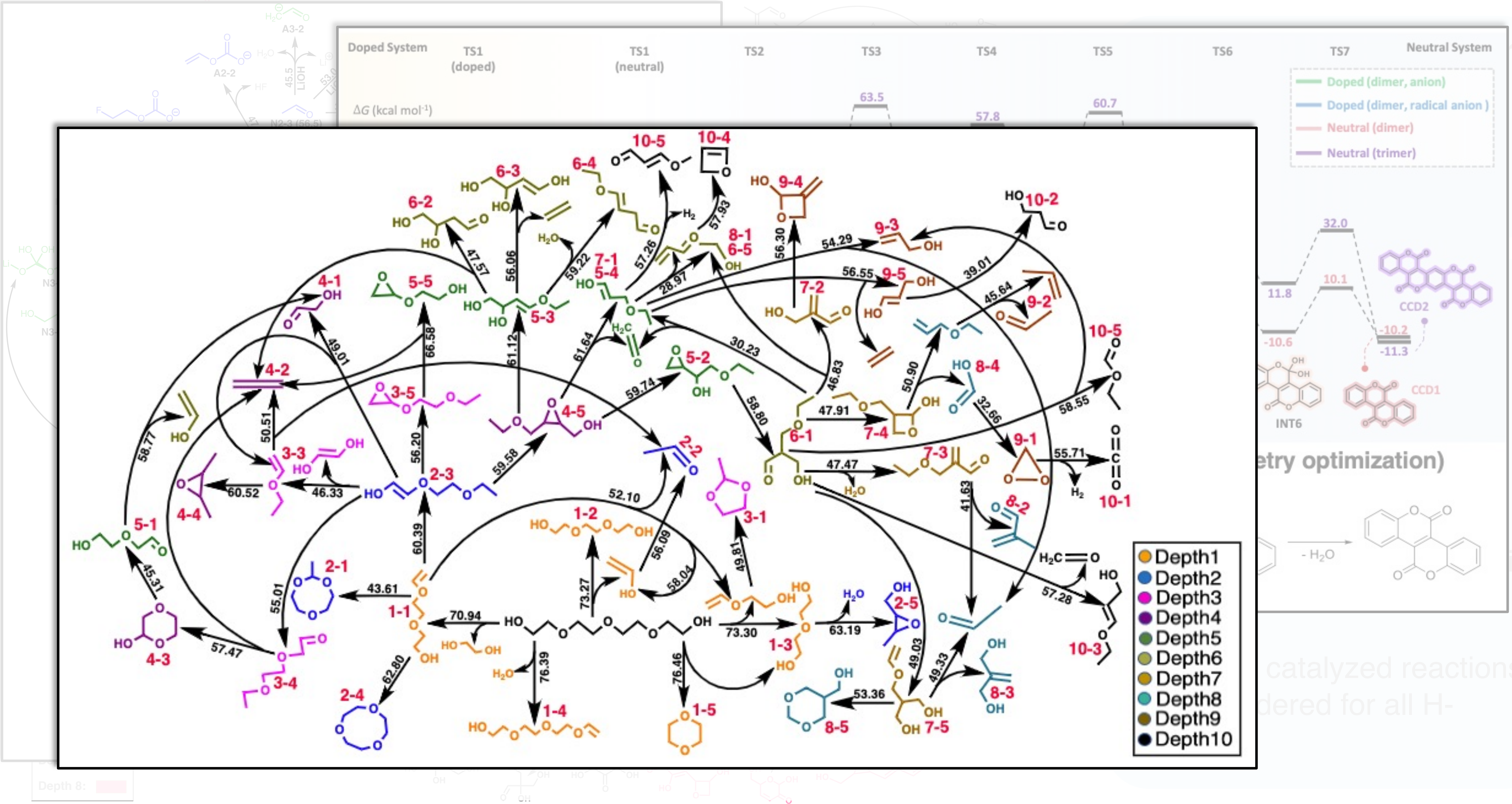
Mechanism proposed by experimental team

Unstable (bonds break after DFT geometry optimization)

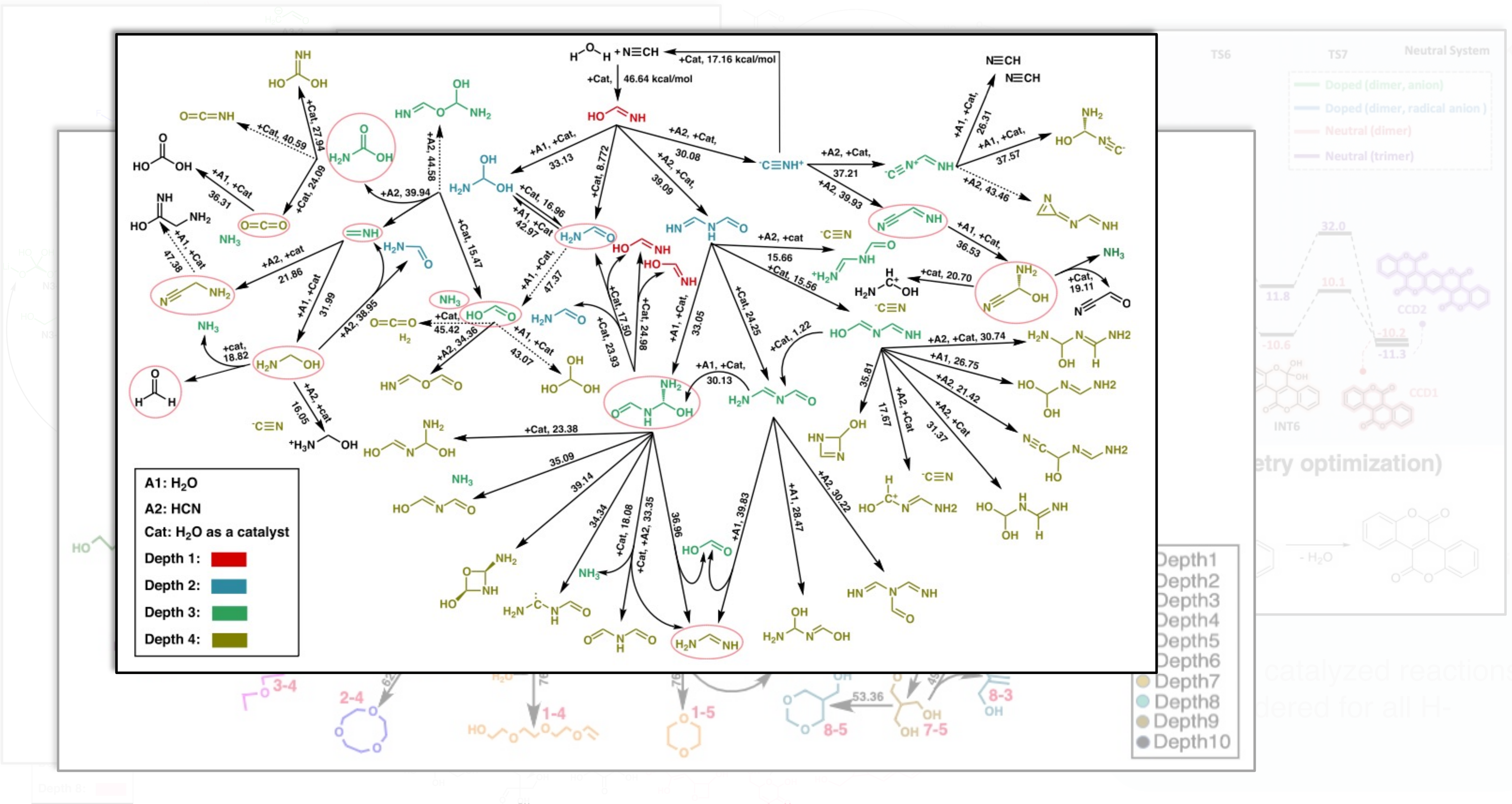


(3) Water catalyzed reactions are considered for all H-transfers

β-D-Glucose Pyrolysis Network Exploration

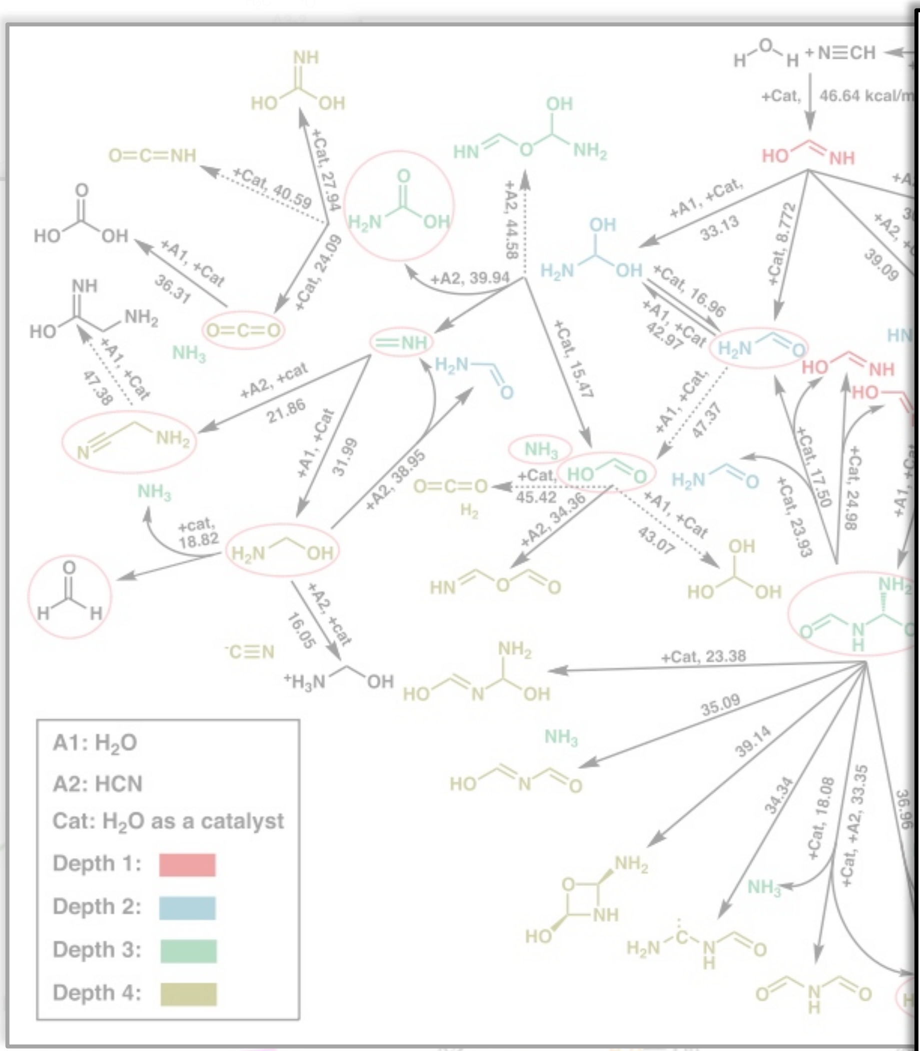


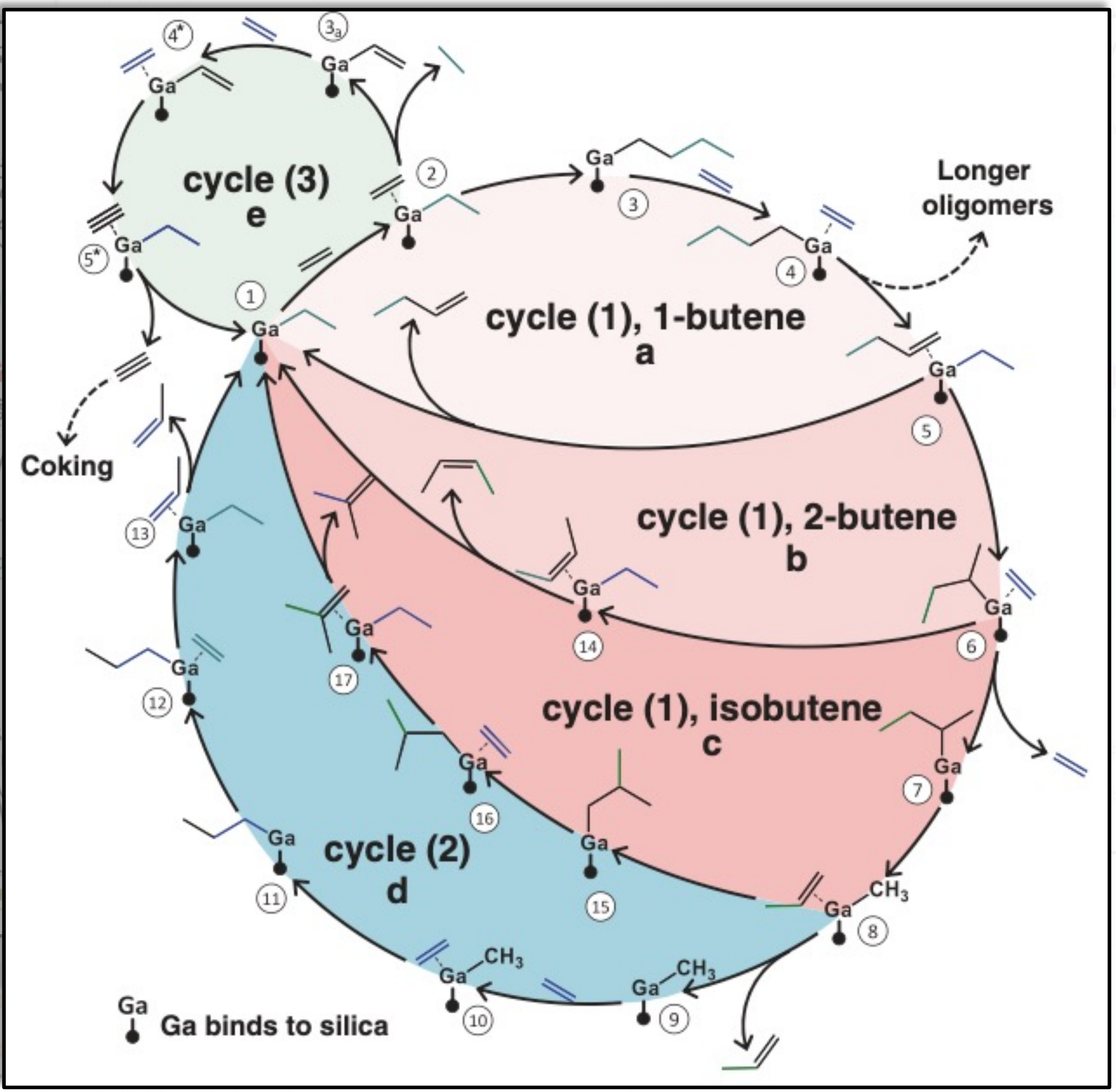
β-D-Glucose Pyrolysis Network Exploration



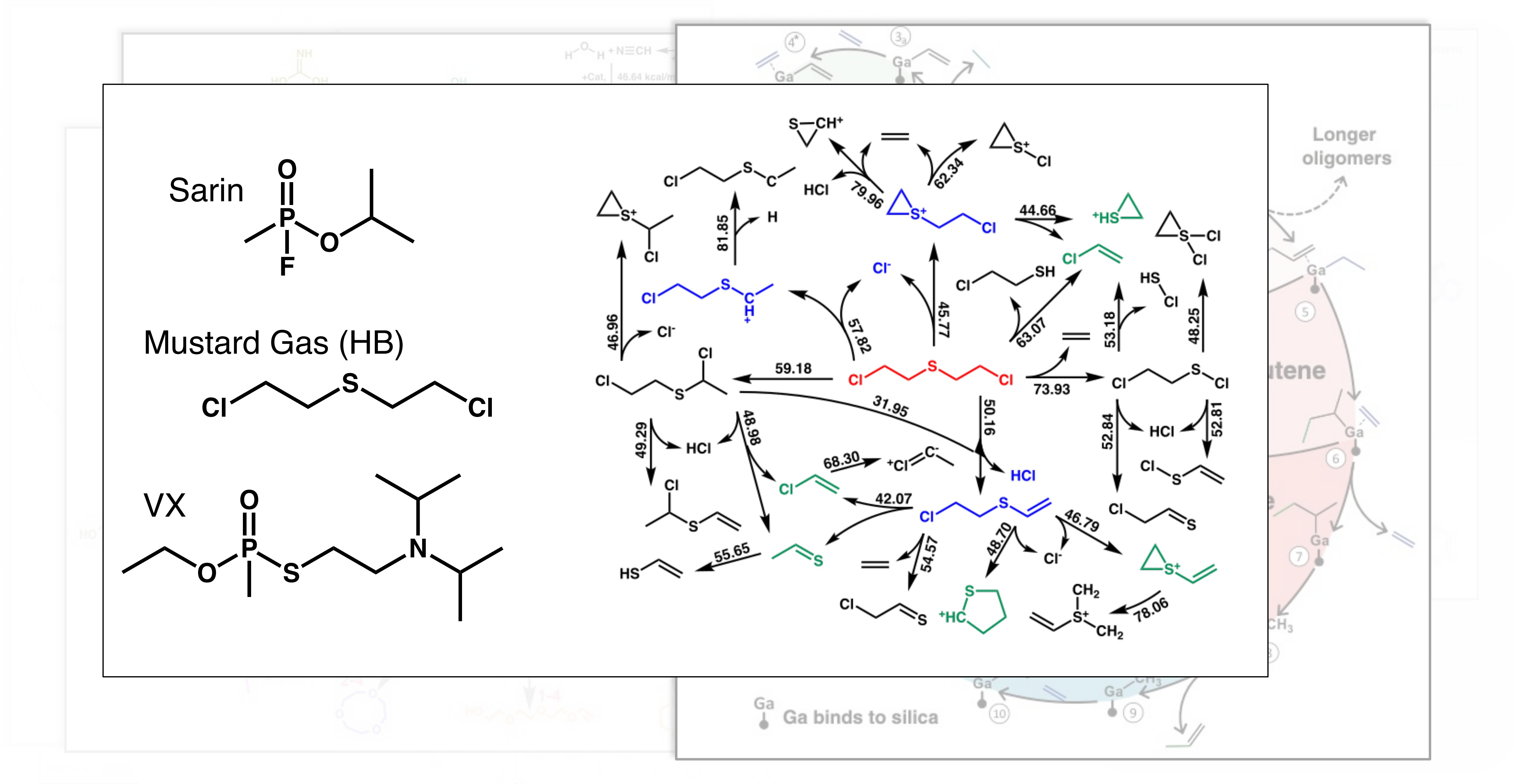
Zhao, Q.; Savoie, B. M. Proc. Nat. Acad. Sci. 120, 2023

β-D-Glucose Pyrolysis Network Exploration





β-D-Glucose Pyrolysis Network Exploration

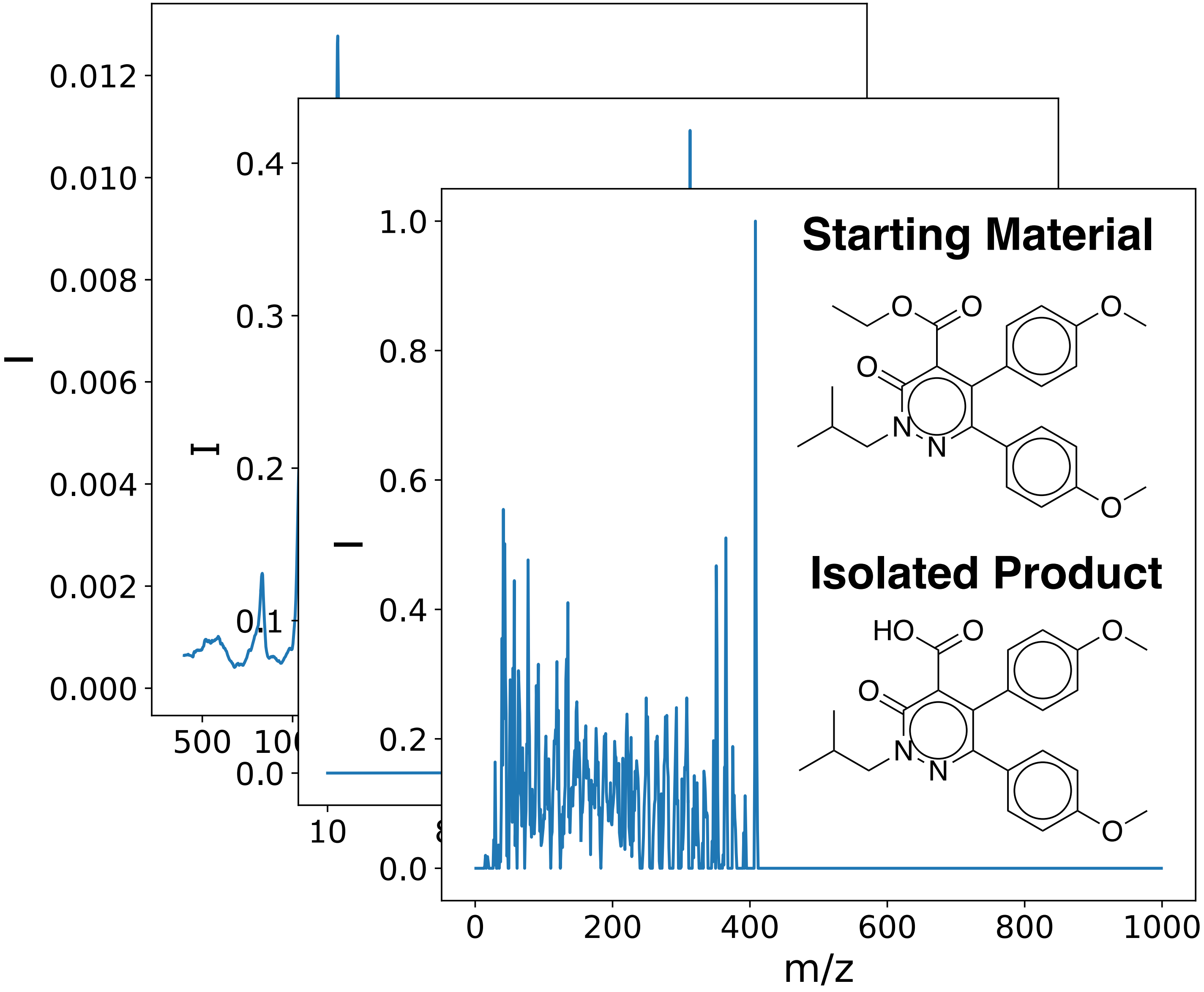


(Pop Quiz) The Product Identification Problem

Suppose we've isolated an unknown degradant and I start offering you spectra.

How long until you can unequivocally identify the structure?

What if I told you the starting material?



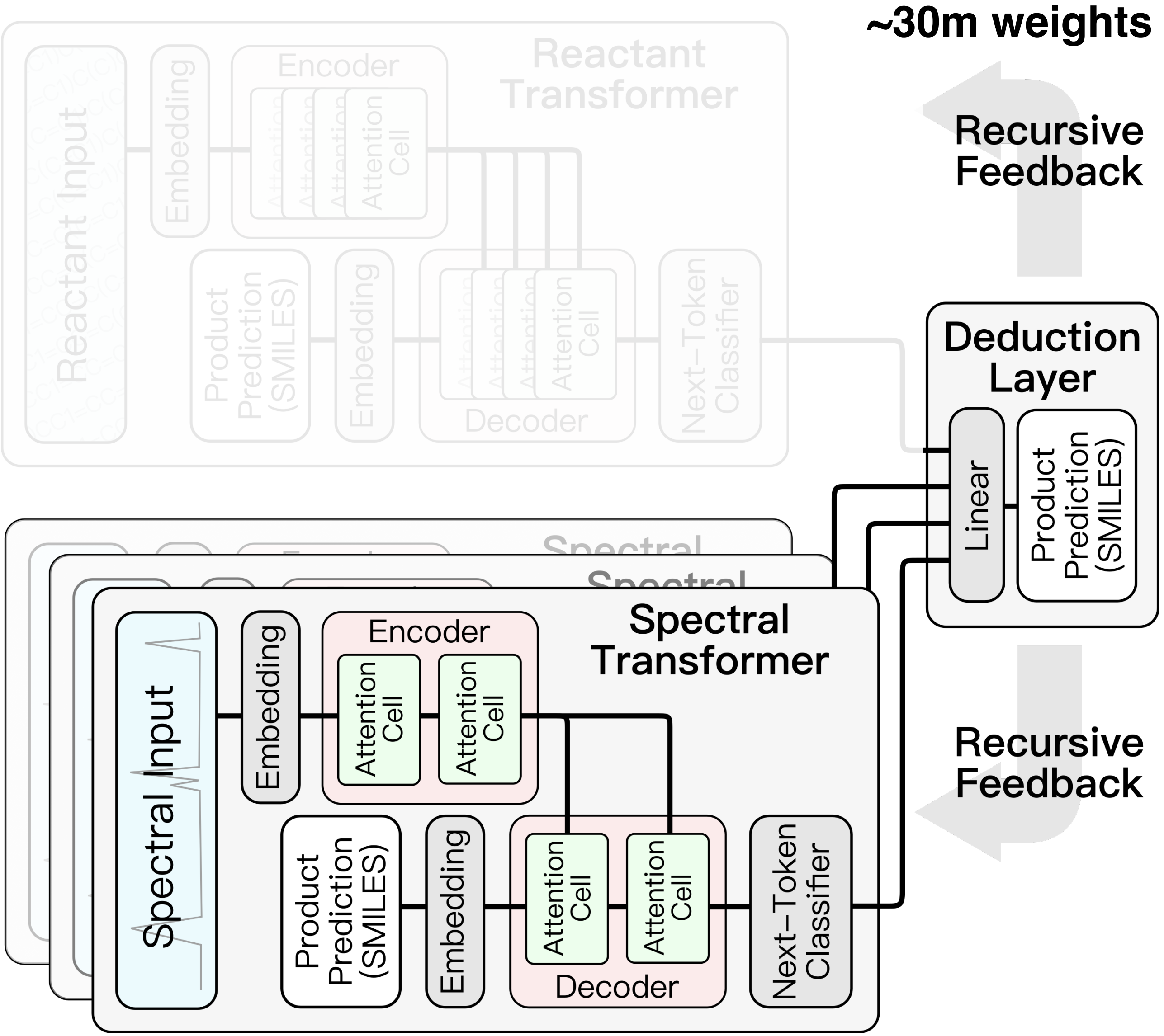
McGill, C et al. JCIM 2021, 2594-2609 (Simulated IR) ; Wei, J. N. et al. ACS Cent. Sci. 2019, 700-708 (Simulated EI-MS)

Leveraging Mixture of Experts Transformer Architectures

We developed this architecture to emulate the deductive process that experts use during product identification

The **product identification** architecture, corresponding to the full network, accepts information about the **reactants** and **spectral information of products** to predict the corresponding product identity.

The **spectrum to structure** (**StS**) identification architecture corresponds to just the bottom half. This architecture doesn't use any knowledge of reactants.



Jin, T., Zhao, Q., Savoie, B.M., Chemrxiv, doi:10.26434/chemrxiv-2023-l6lzp

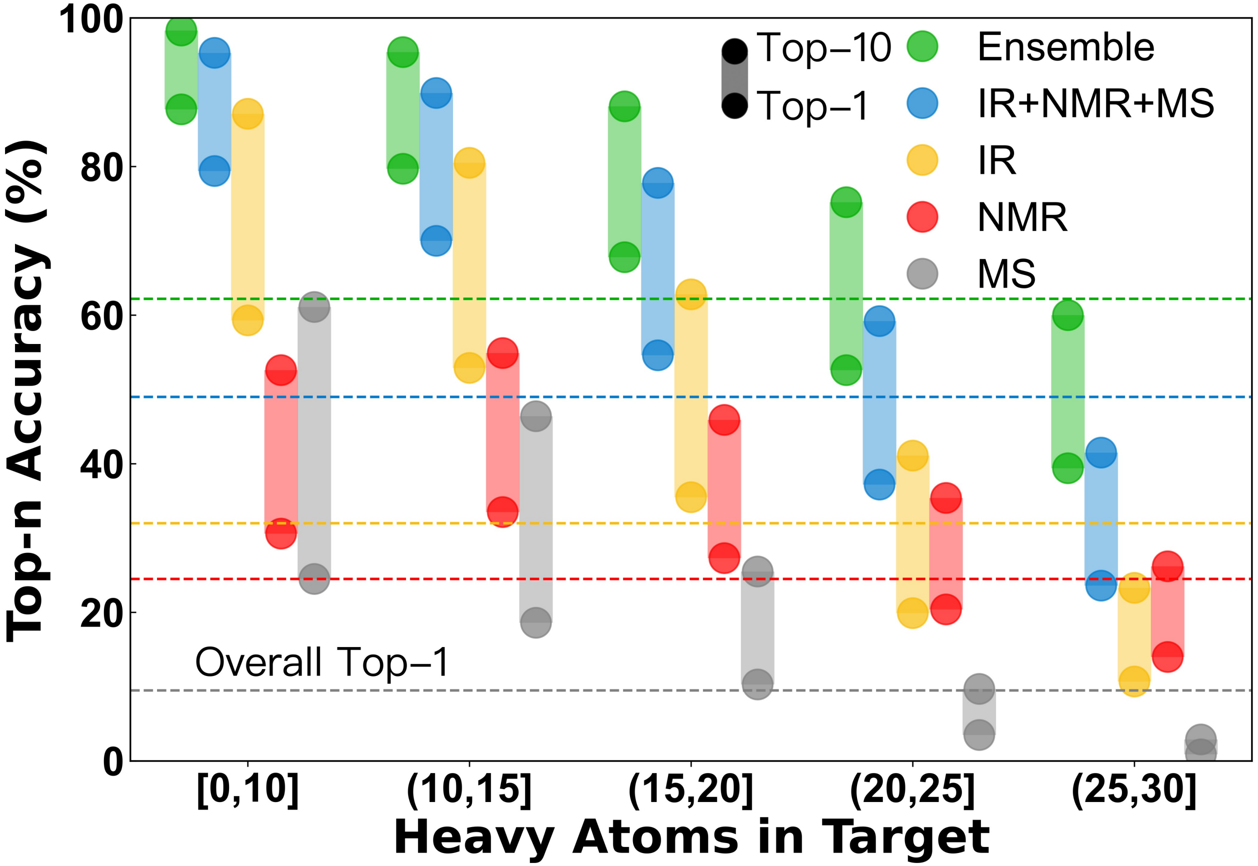
Is there sufficient information for the StS task?

1H-NMR, IR, and EI-MS spectra were simulated for 305,493 USPTO species and 524,860 from PubChem.

Sufficient information is probable for small molecules: **~85% top-1**

Potential deductive bottleneck for large molecules: **~38% top-1**

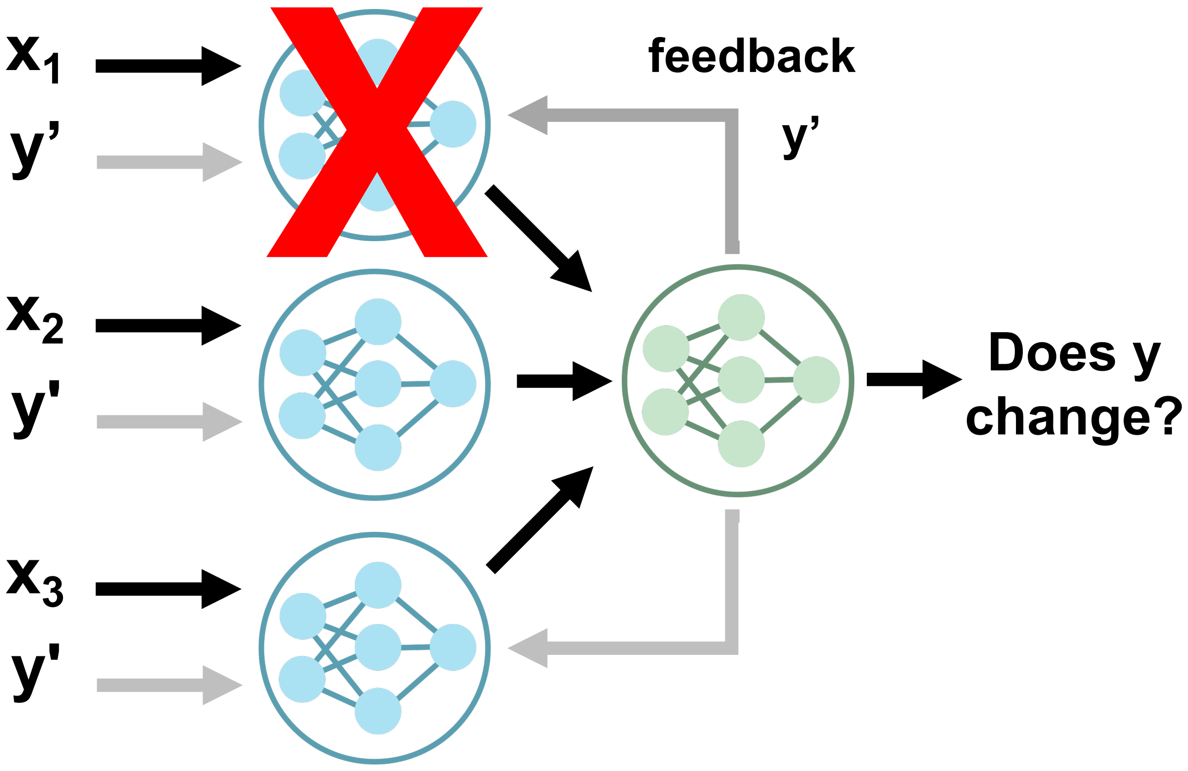
Top-10 predictions obtained with beam-search. Performance is shown on a random 10% testing split.

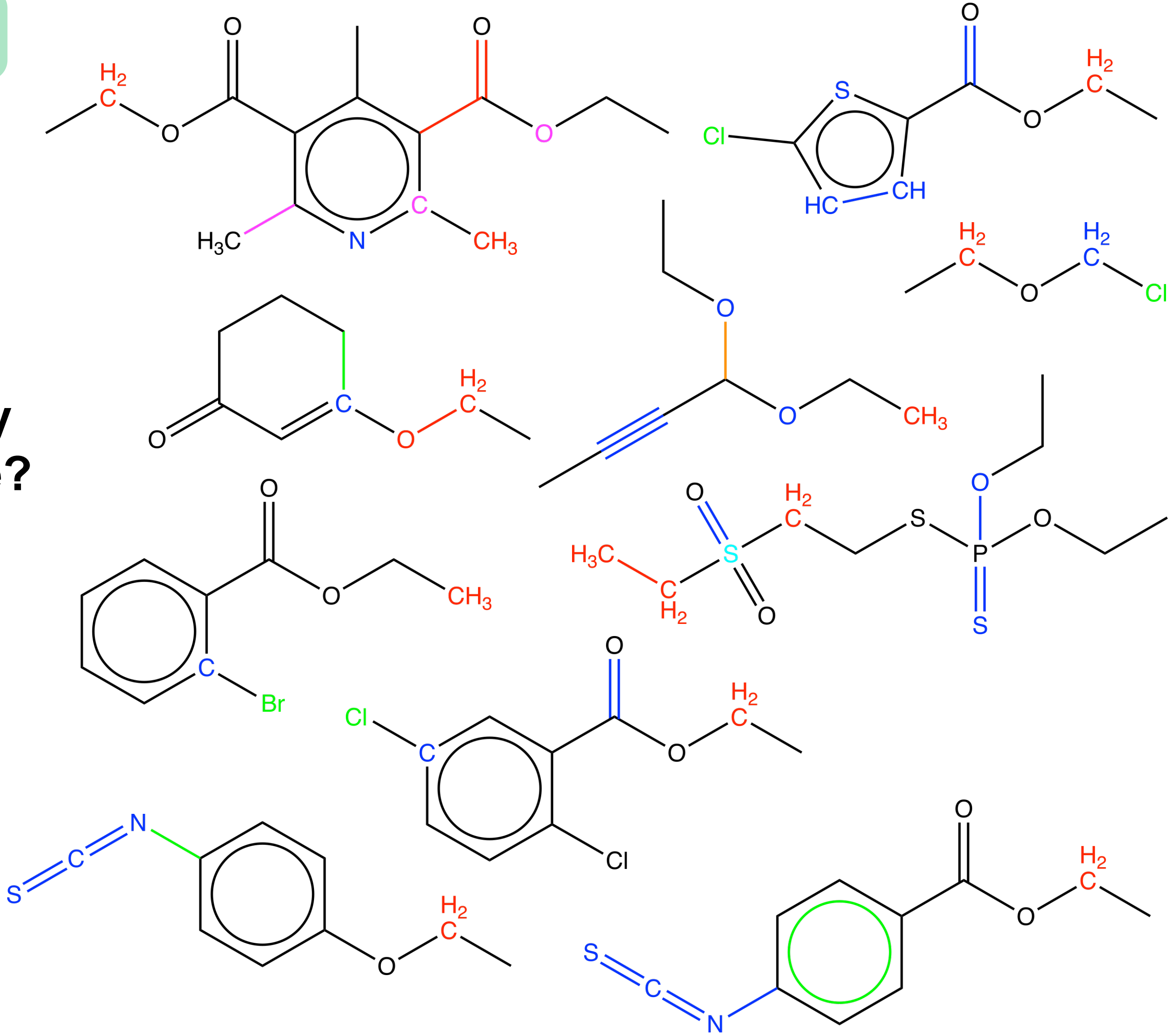


Jin, Schofield, Savoie, *In Preparation*

Direct Evidence of Dynamic Deduction by these Models

Visualizing Per Token Spectral Decisiveness





Color Legend: 1H-NMR IR EI-MS

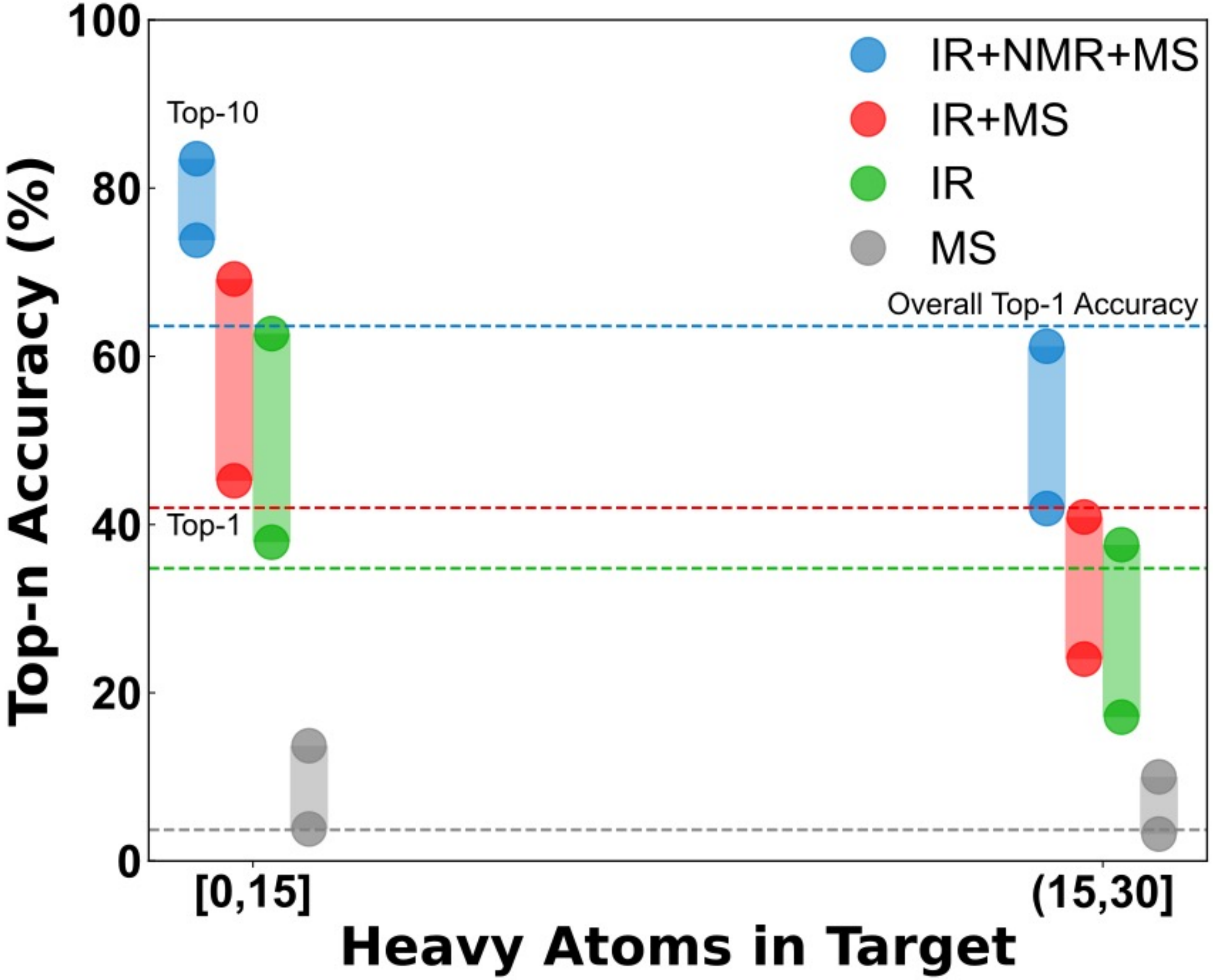
Jin, Schofield, Savoie, *In Preparation*

Does the StS Model Transfer to Real Spectra?

We used 5544 molecules from NIST with experimental IR/H-NMR/EI-MS* as an external test for the model

None of the NIST molecules were in the training data, and all training was on simulated not experimental spectra.

External NIST Testing Set



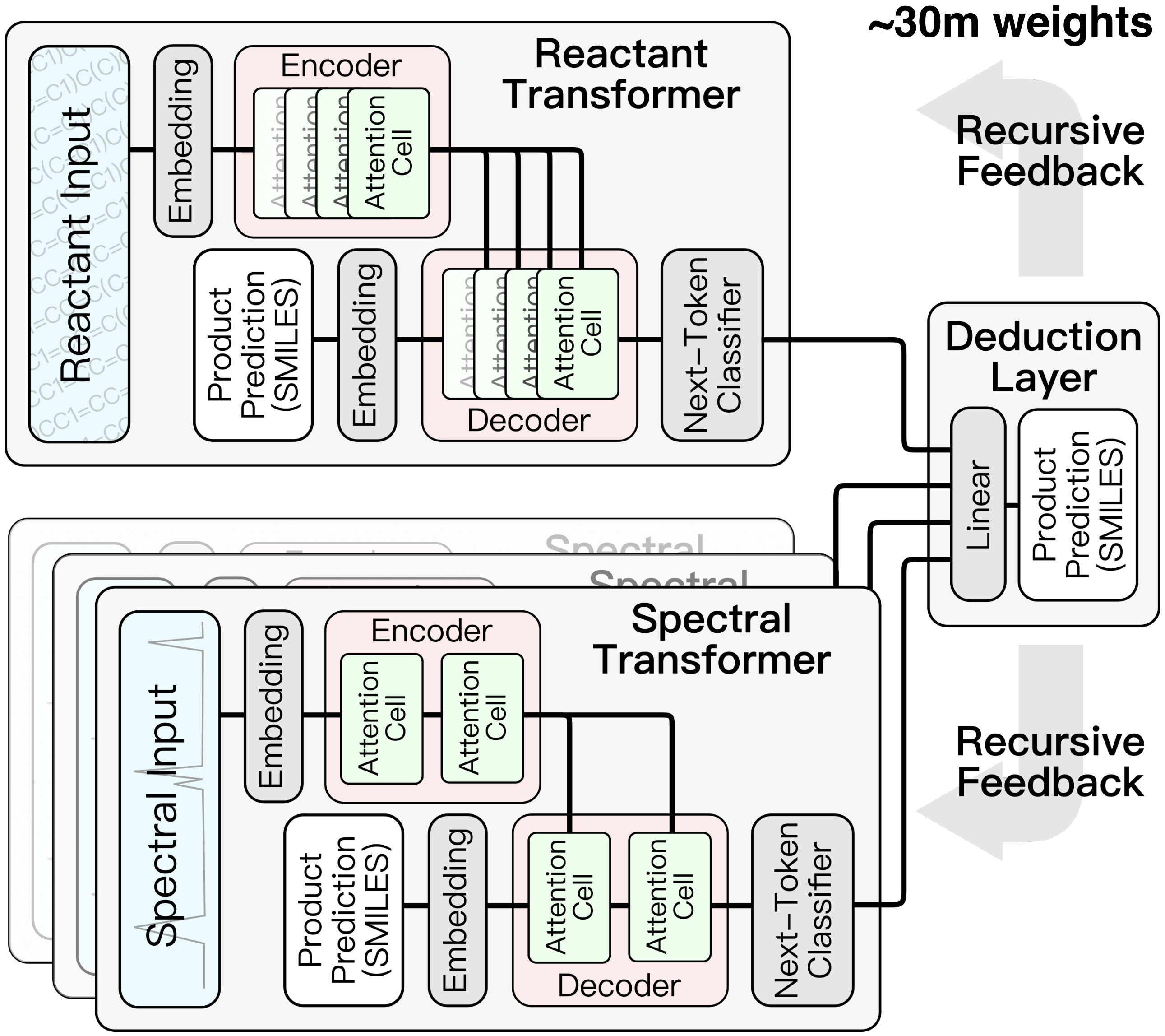
*EI-MS was still simulated in these cases.

Leveraging Mixture of Experts Transformer Architectures

We developed this architecture to emulate the deductive process that experts use during product identification

The **product identification** architecture, corresponding to the full network, accepts information about the **reactants** and **spectral information of products** to predict the corresponding product identity.

The **spectrum to structure** (**StS**) identification architecture corresponds to just the bottom half. This architecture doesn't use any knowledge of reactants.

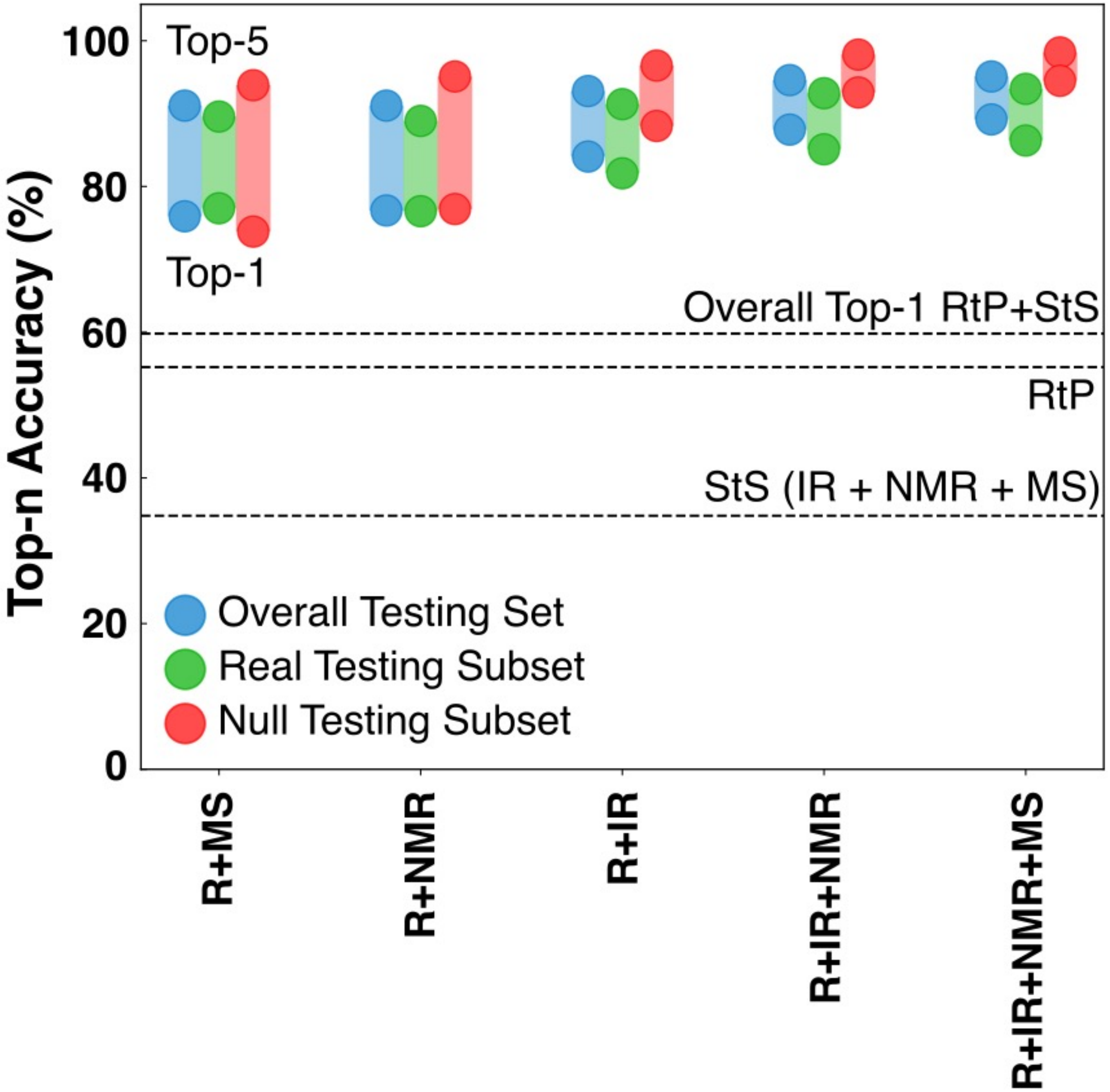


Jin, T., Zhao, Q., Savoie, B.M., Chemrxiv, doi:10.26434/chemrxiv-2023-l6lzp

Adding Reactant Info Qualitatively Improves Performance

Testing these Reactant + Spectra models on product identification:

1. 466,330 Reactions from D. Lowe
2. 2:1 Split between real products and starting material
3. 80:10:10 Train:Val:Test
4. Simulated IR/1H-NMR/EI-MS for all 305,493 species

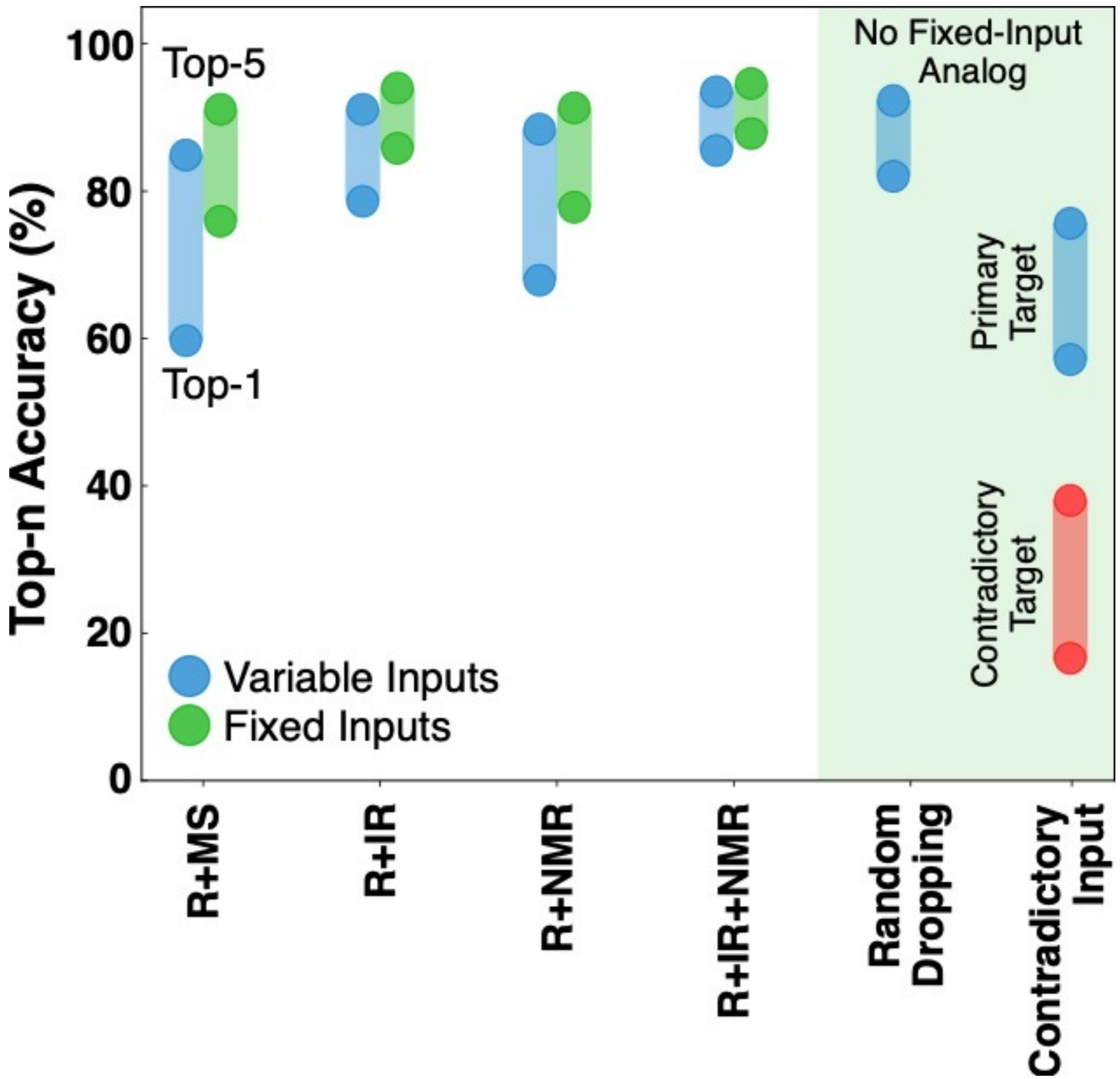


Jin, T., Zhao, Q., Savoie, B.M., Chemrxiv, doi:10.26434/chemrxiv-2023-l6lzp

Can the Mixture of Expert Models Handle Contradictory Information?

A second aspect of deductive behavior is to reason with partial, or even, contradictory inputs.

To test the performance under contradictory information we supplied a contradictory spectrum to one transformer (at random)

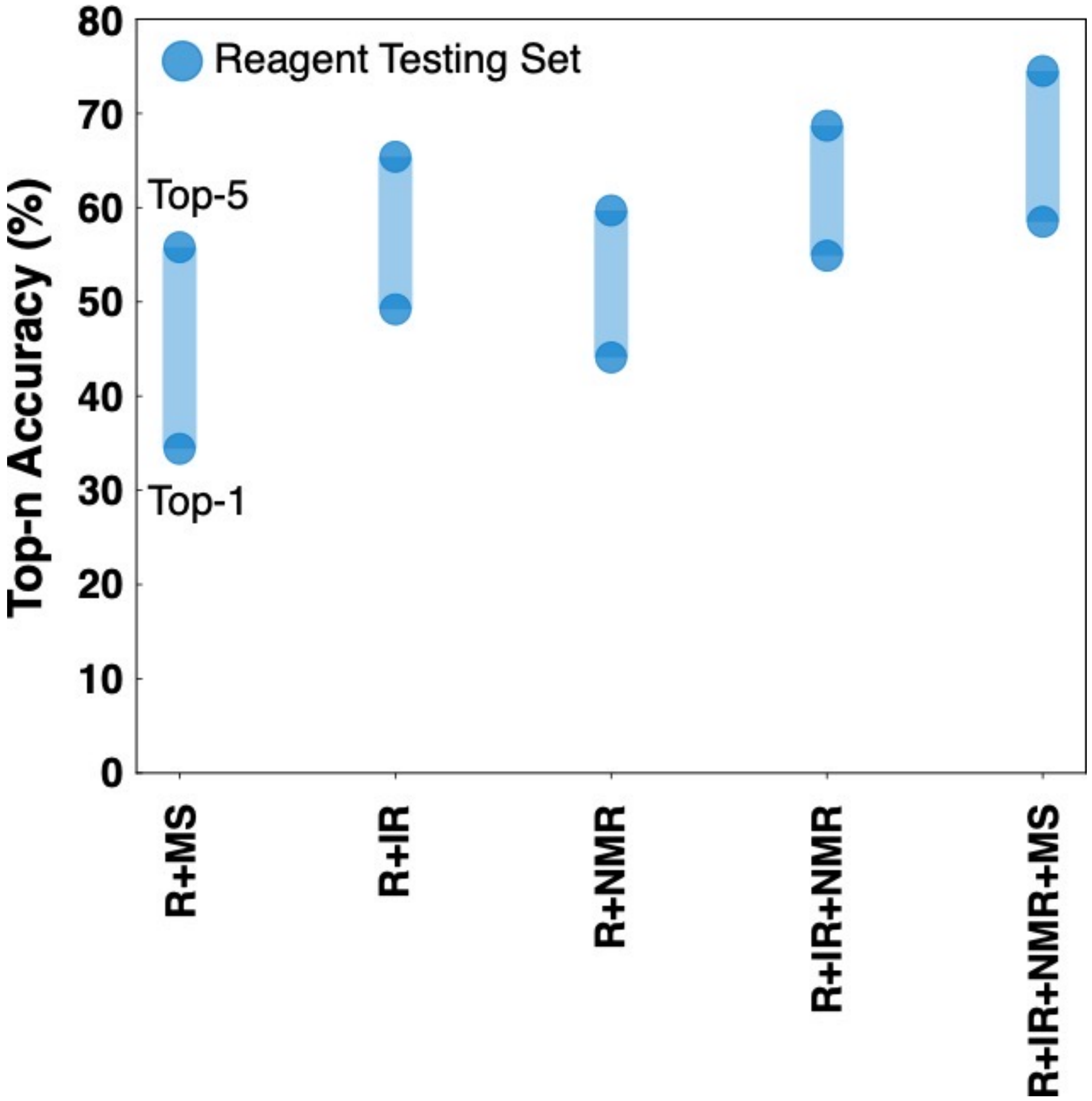


Jin, T., Zhao, Q., Savoie, B.M., Chemrxiv, doi:10.26434/chemrxiv-2023-l6lzp

Can the Mixture of Expert Models Perform Out-of-Distribution?

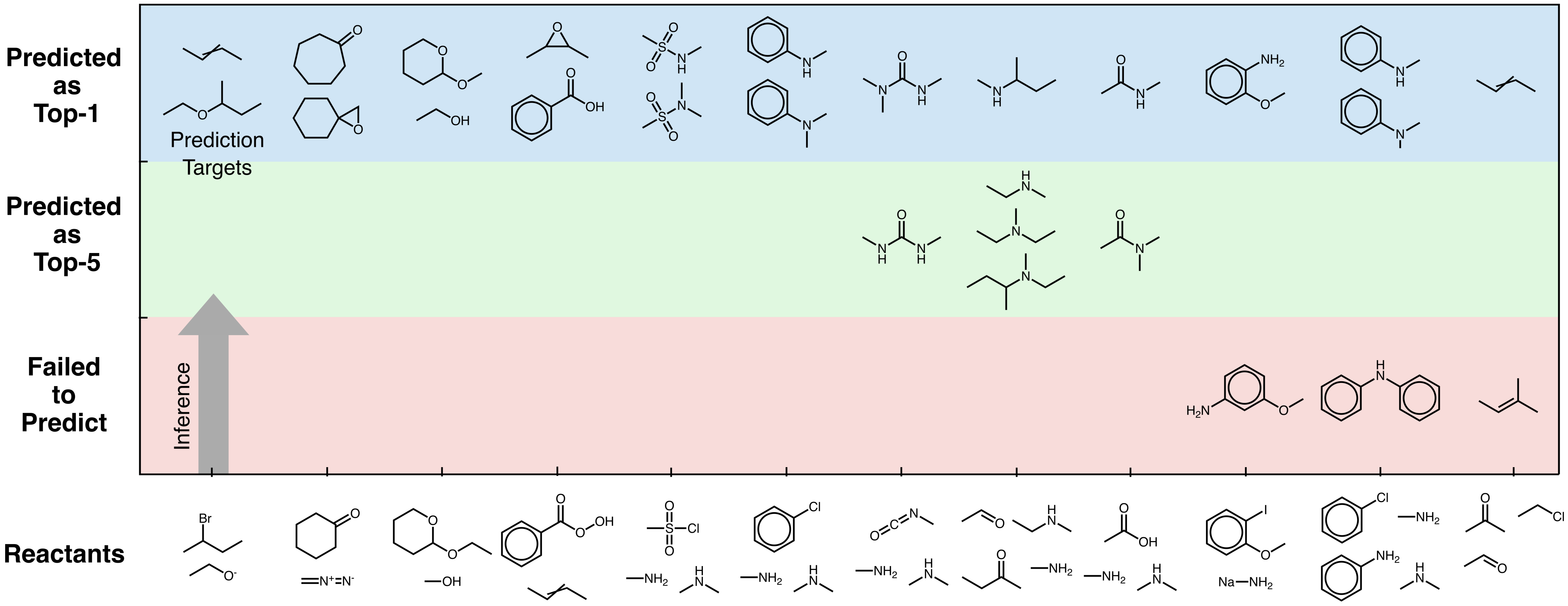
Reagent identification is an untrained task that has the same inputs as product/null characterization

We curated ~1k reagents involved in ~3k reactions from the USPTO dataset. None of these species were seen as prediction targets during training.



Jin, T., Zhao, Q., Savoie, B.M., Chemrxiv, doi:10.26434/chemrxiv-2023-l6lzp

Can the Mixture of Expert Models Perform Out-of-Distribution?

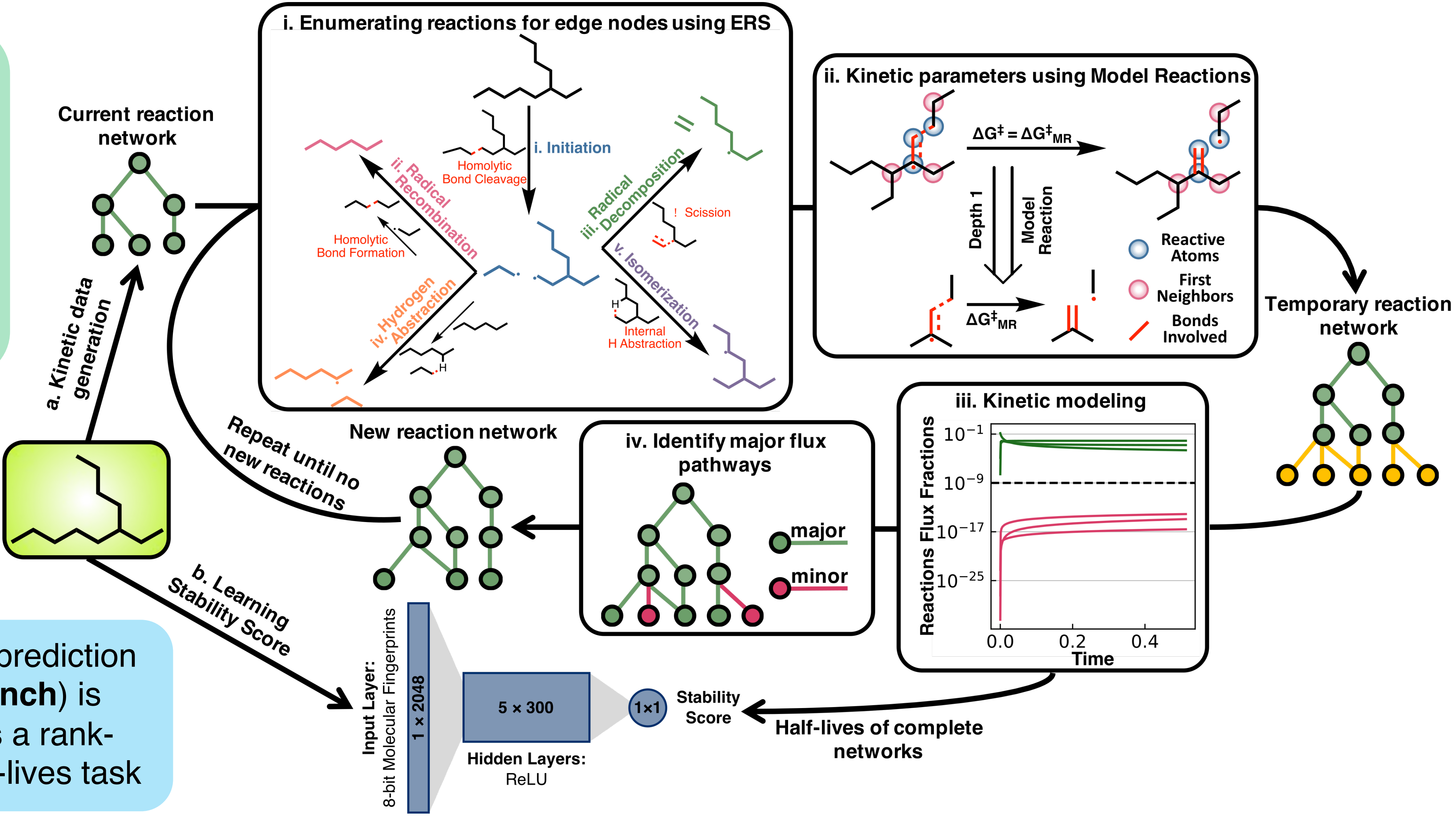


Minor product identification was tested using 18 reactants with 50 distinct products from Grossman's textbook and Hartenfeller et al.

Hartenfeller, Renner, et al. JCIM, 11, 2011

Actionable Design Information: Thermal Stability Scores

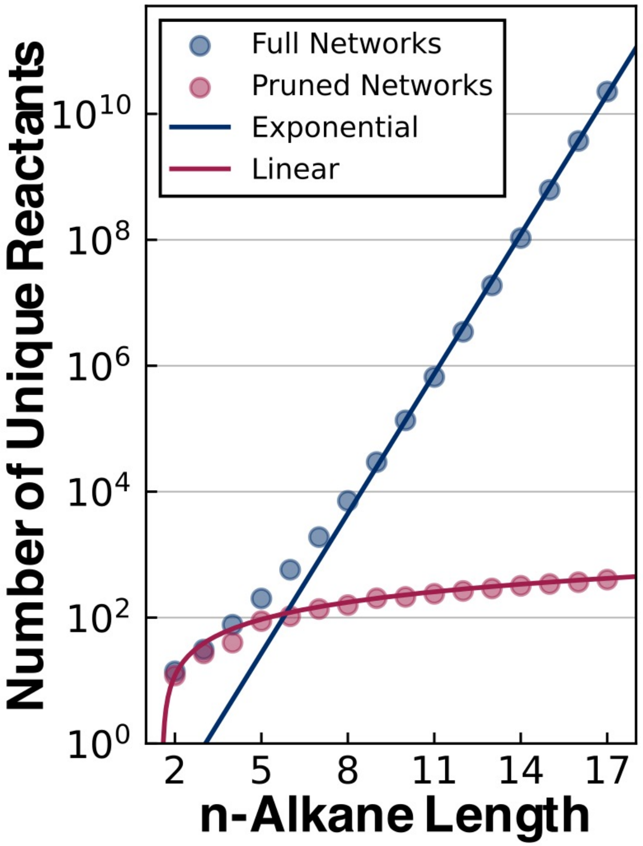
Computational half-life generation (**top-branch**) is expensive and incompatible with on-the-fly design needs.

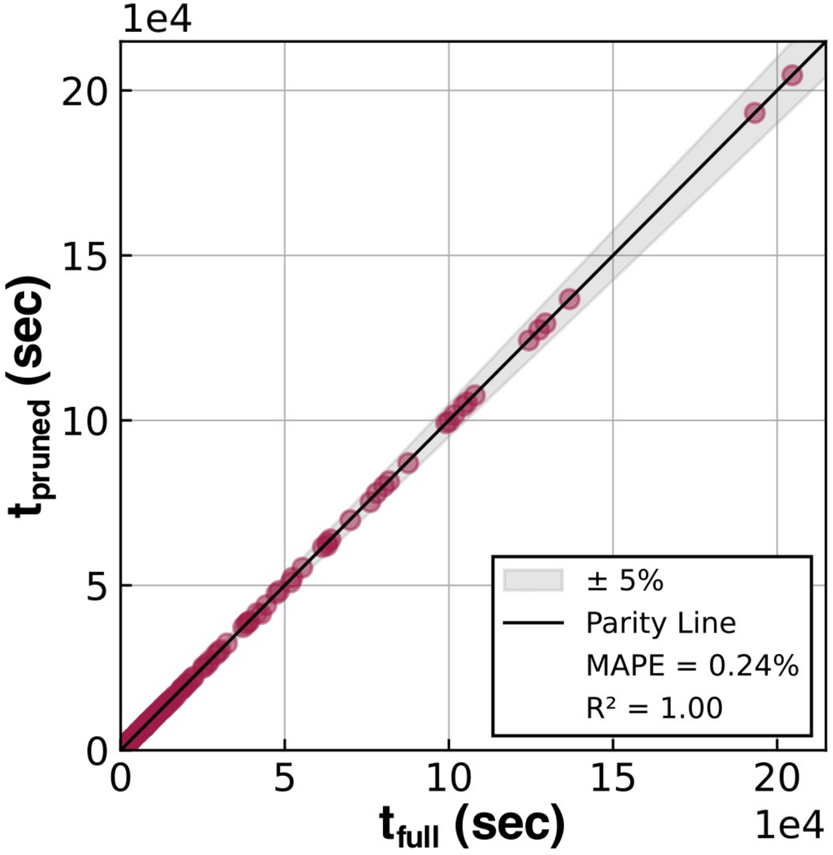


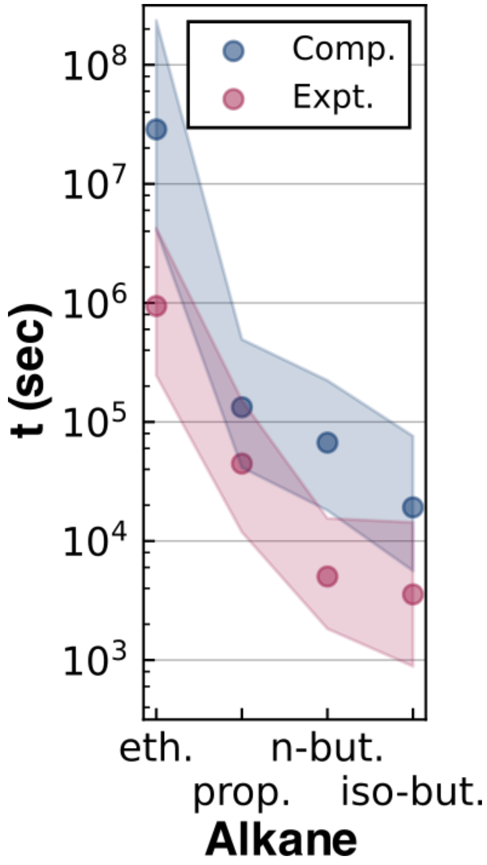
Stability score prediction (**bottom branch**) is formulated as a rank-ordering of half-lives task

Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Half-Life Simulations are Non-Trivial



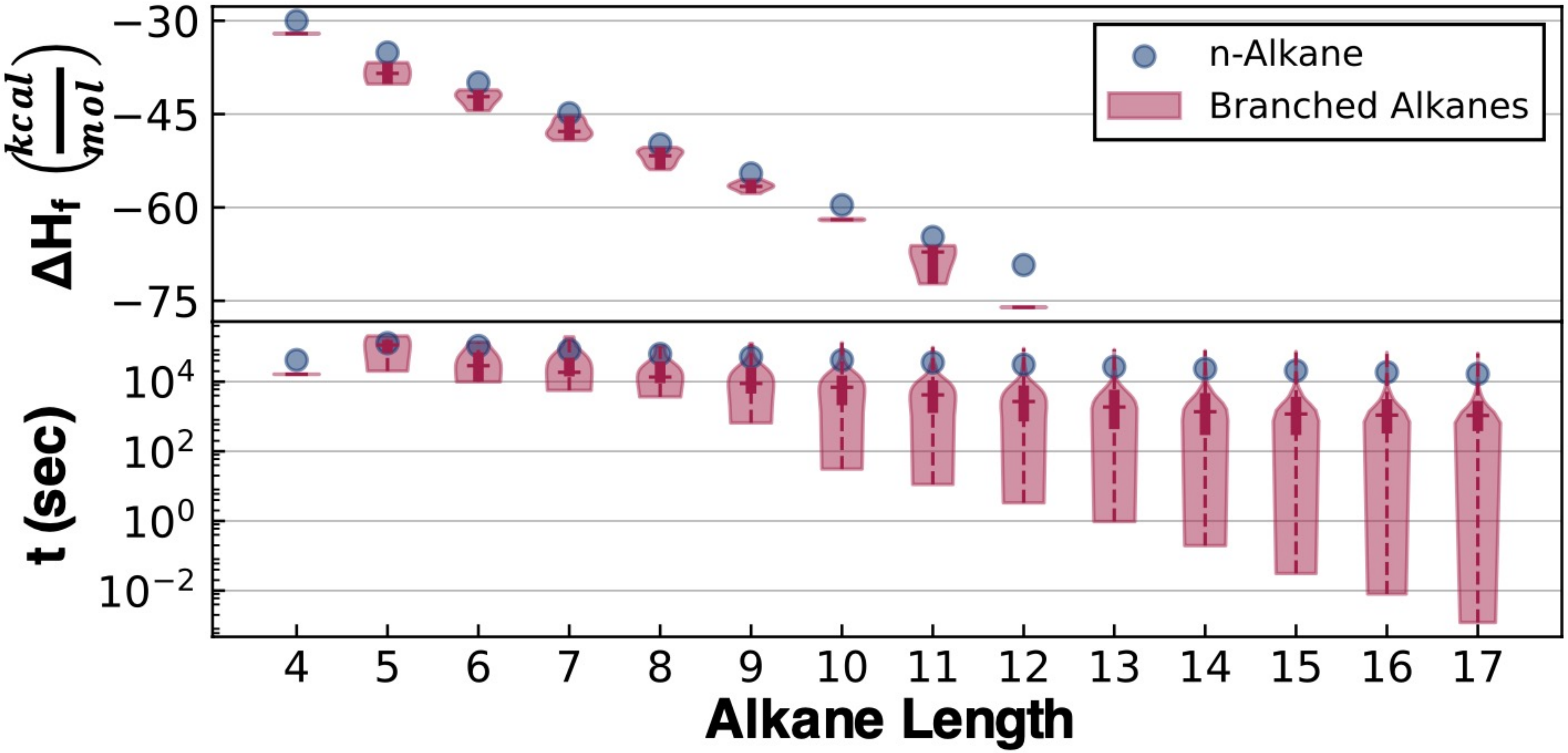




The half-lives of ~32k alkanes with varying topology were simulated under pyrolysis conditions as a surrogate for thermal stability

Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Kinetics Matter for Stability



Heats of formation are completely non-predictive for relative stability (even for the simplest organic materials class).

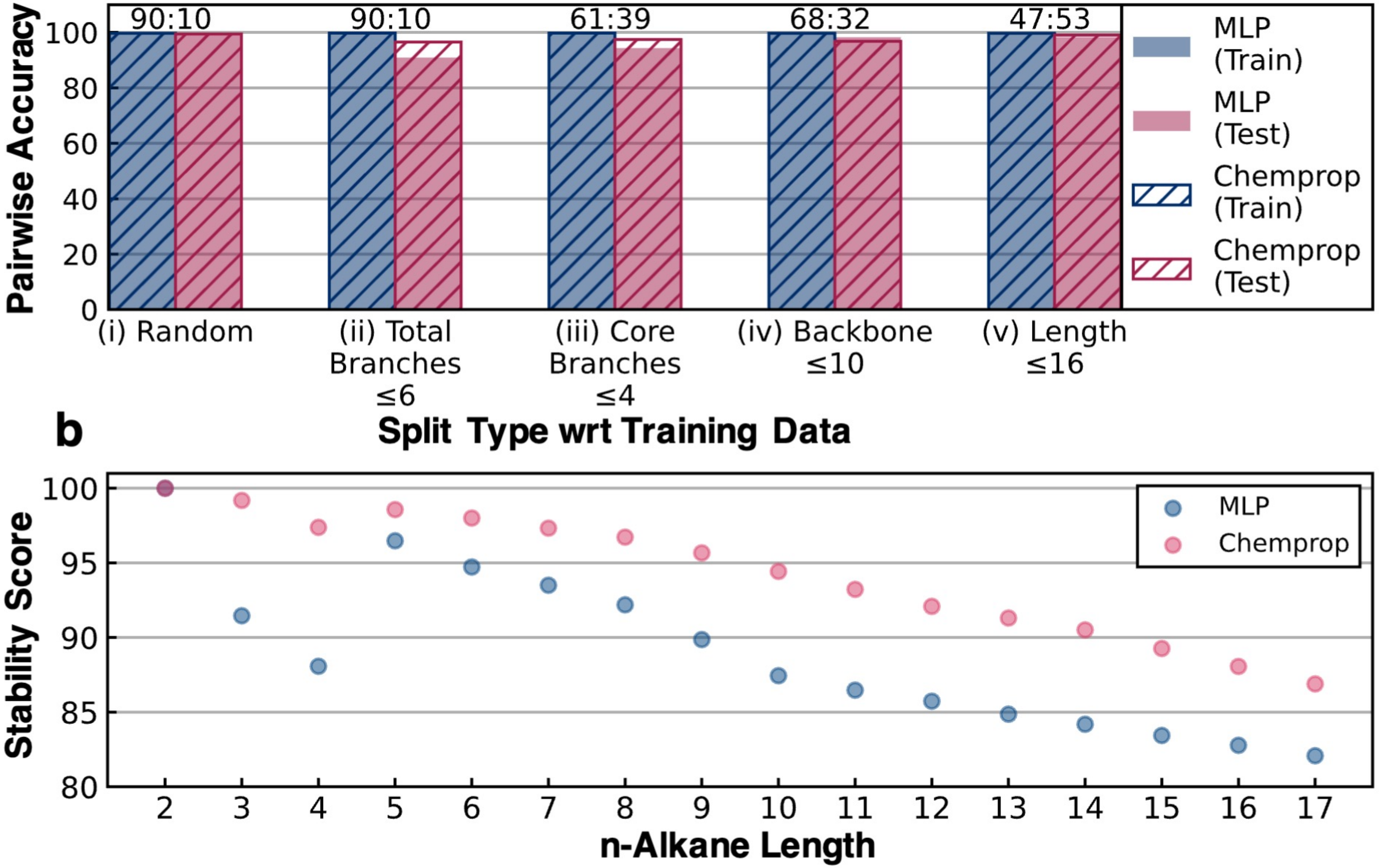
Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Kinetics Matter for Stability

The stability score is trained to predict the rank order of the log half-life using a hinge-loss function

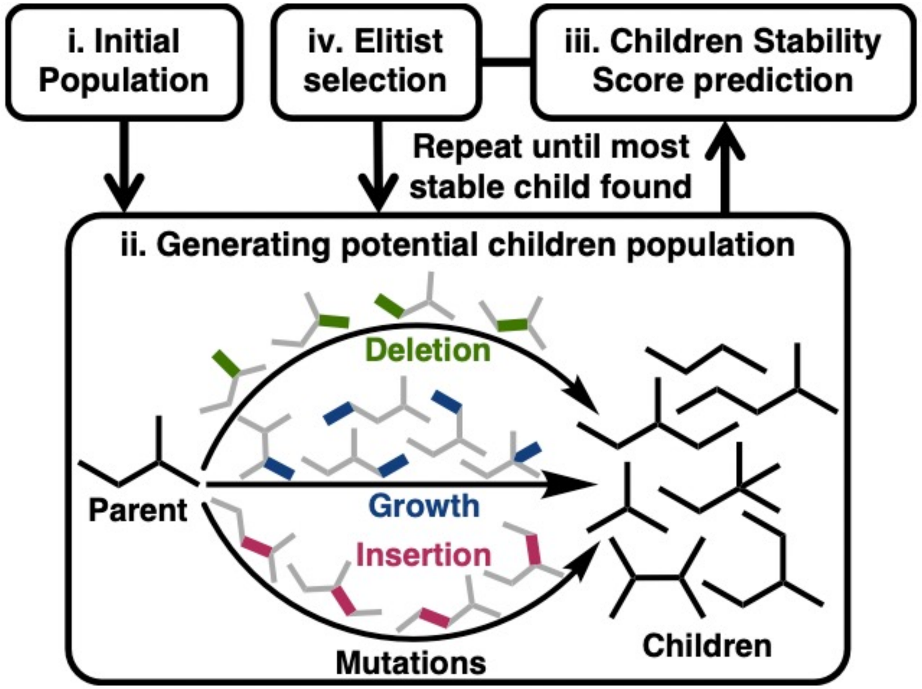
The relative stability is really easy to learn (basically every architecture we tried >95% accurate)

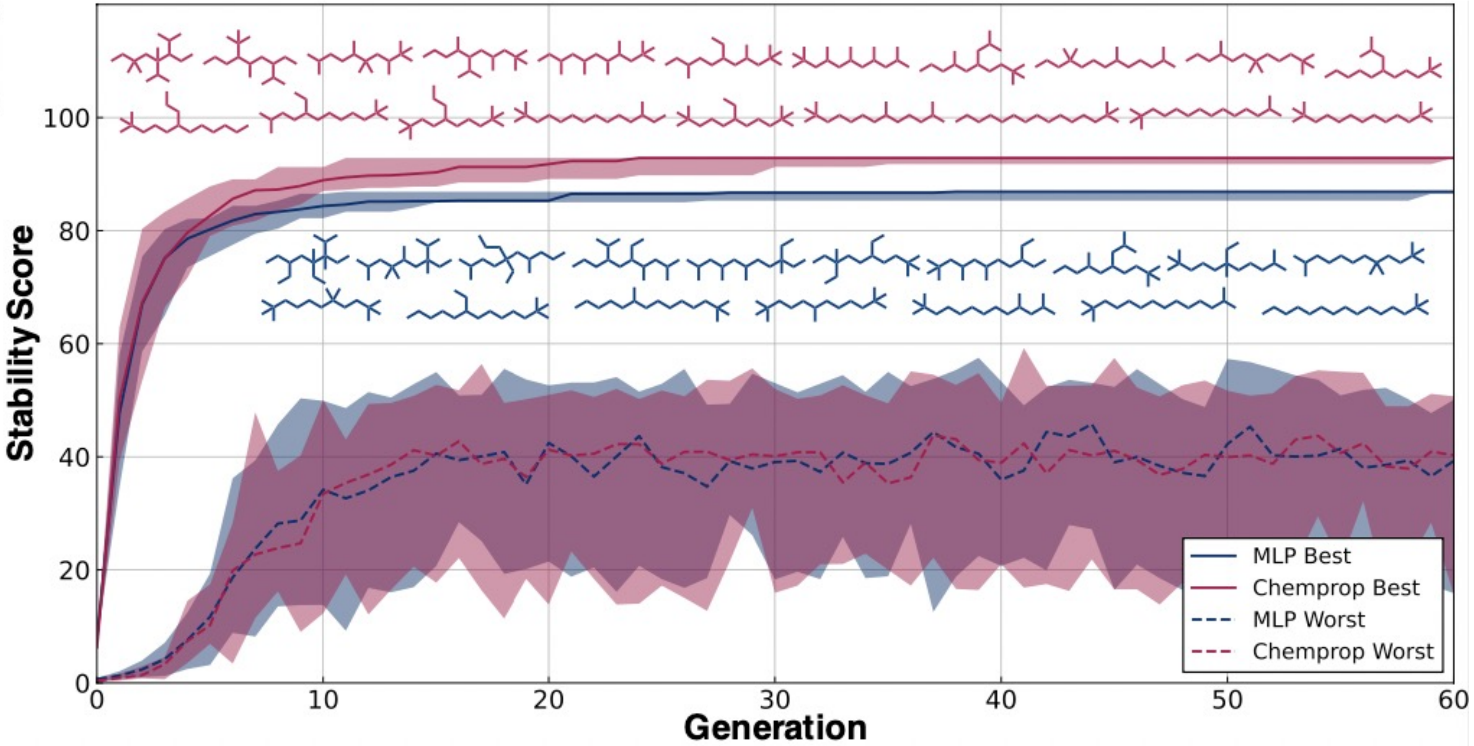
Non-trivial kinetic stability trends are learned in testing set



Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Actionable Design Information: Thermal Stability Scores


i. Initial Population
iv. Elitist selection
iii. Children Stability Score prediction
Repeat until most stable child found
ii. Generating potential children population
Deletion
Parent
Growth
Insertion
Mutations
Children

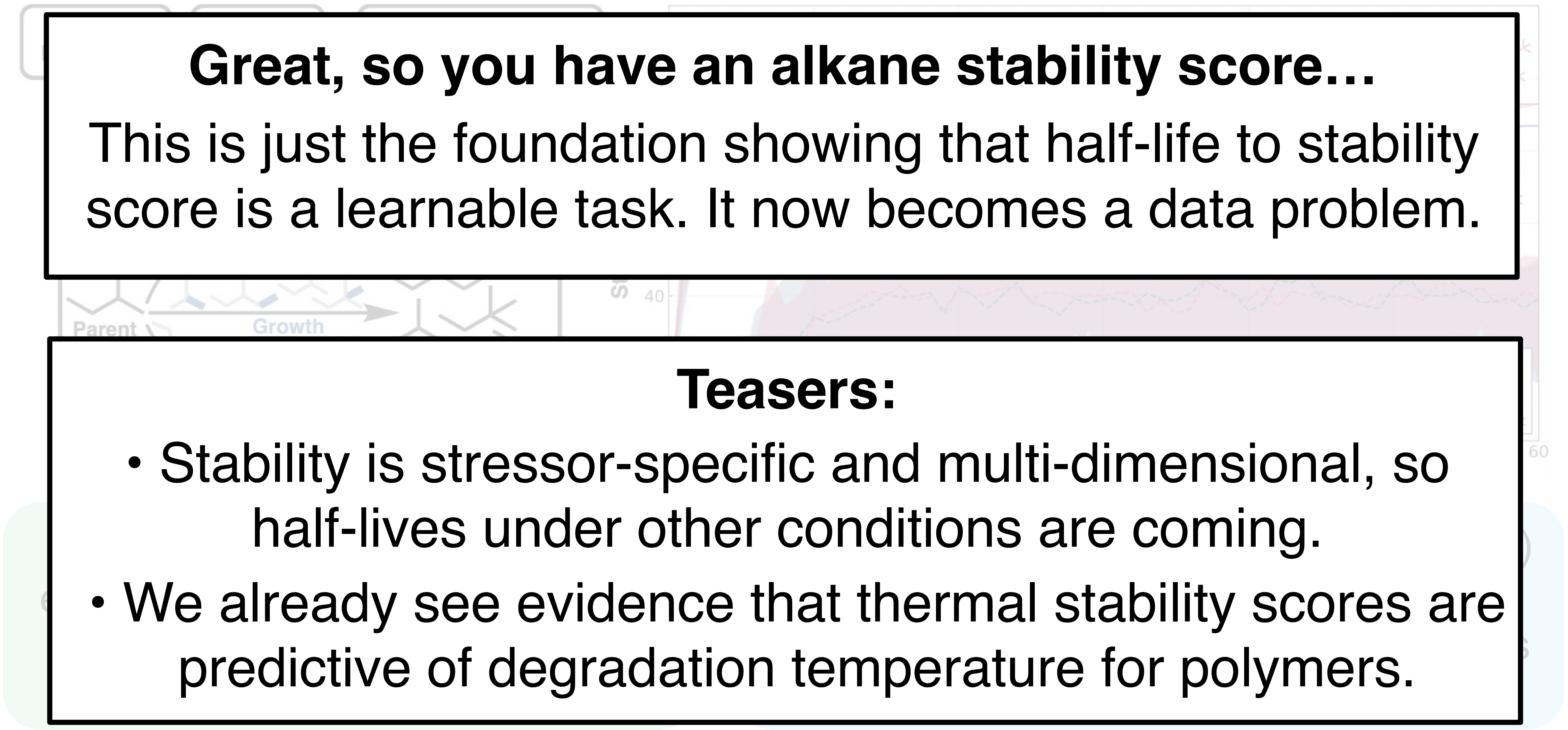


Genetic algorithms are used to explore chemical space, using the stability score as an objective function

The simple (ML) and complex (Chemprop) models are both extremely proficient at guiding the search towards stable alkanes with a minimal number of branches

Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Actionable Design Information: Thermal Stability Scores



Great, so you have an alkane stability score…

This is just the foundation showing that half-life to stability score is a learnable task. It now becomes a data problem.

Teasers:

- Stability is stressor-specific and multi-dimensional, so half-lives under other conditions are coming.
- We already see evidence that thermal stability scores are predictive of degradation temperature for polymers.

Singla, V.; Zhao, Q.; Savoie, B. M. *Chemrxiv.* 2023, https://doi.org/10.26434/chemrxiv-2023-0pjxv

Acknowledgements

Collaborators:

Ray Mentzer (TCIT; P2SAC)
Julia Laskin (Mass Spec Degradation; MURI)
Geoffrey Hutchison (U. Pitt; MURI)
Olexandr Isayev (Carnegie Mellon; MURI)
Jianguo Mei (Conjugated Polymers; MURI)
Jie Xu (ANL; Automated Characterization; MURI)

Students:

Qiyuan Zhao, Tyler Pasut, Michael Woulfe,
Tianfan Jin, Veerupaksh Singla

Funding:

P²SAC
Purdue Process Safety & Assurance Center

Office of Naval Research
ONR
Science & Technology

The Camille & Henry Dreyfus Foundation, Inc.

Army Research Office

Air Force Office of Scientific Research

NSF

P²SAC

U.S. Department of Energy

ORAU

The Camille & Henry Dreyfus Foundation, Inc.

ACS PRF