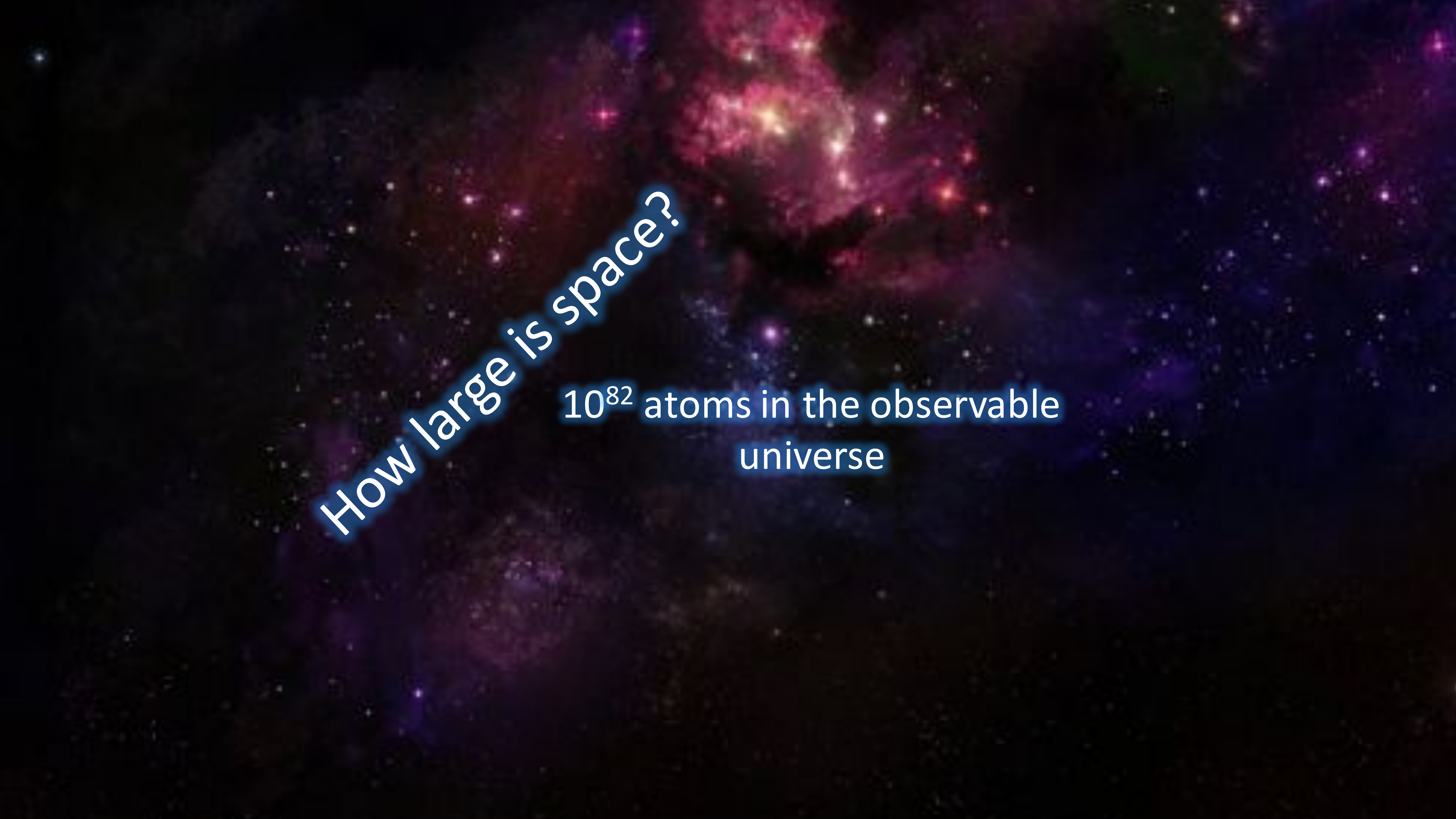
Exploring chemical space: Molecular Space Shuttle

Alán Aspuru-Guzik
Professor
Harvard University

http://aspuru.chem.harvard.edu
Twitter: A_Aspuru_Guzik
aspuru@chemistry.harvard.edu

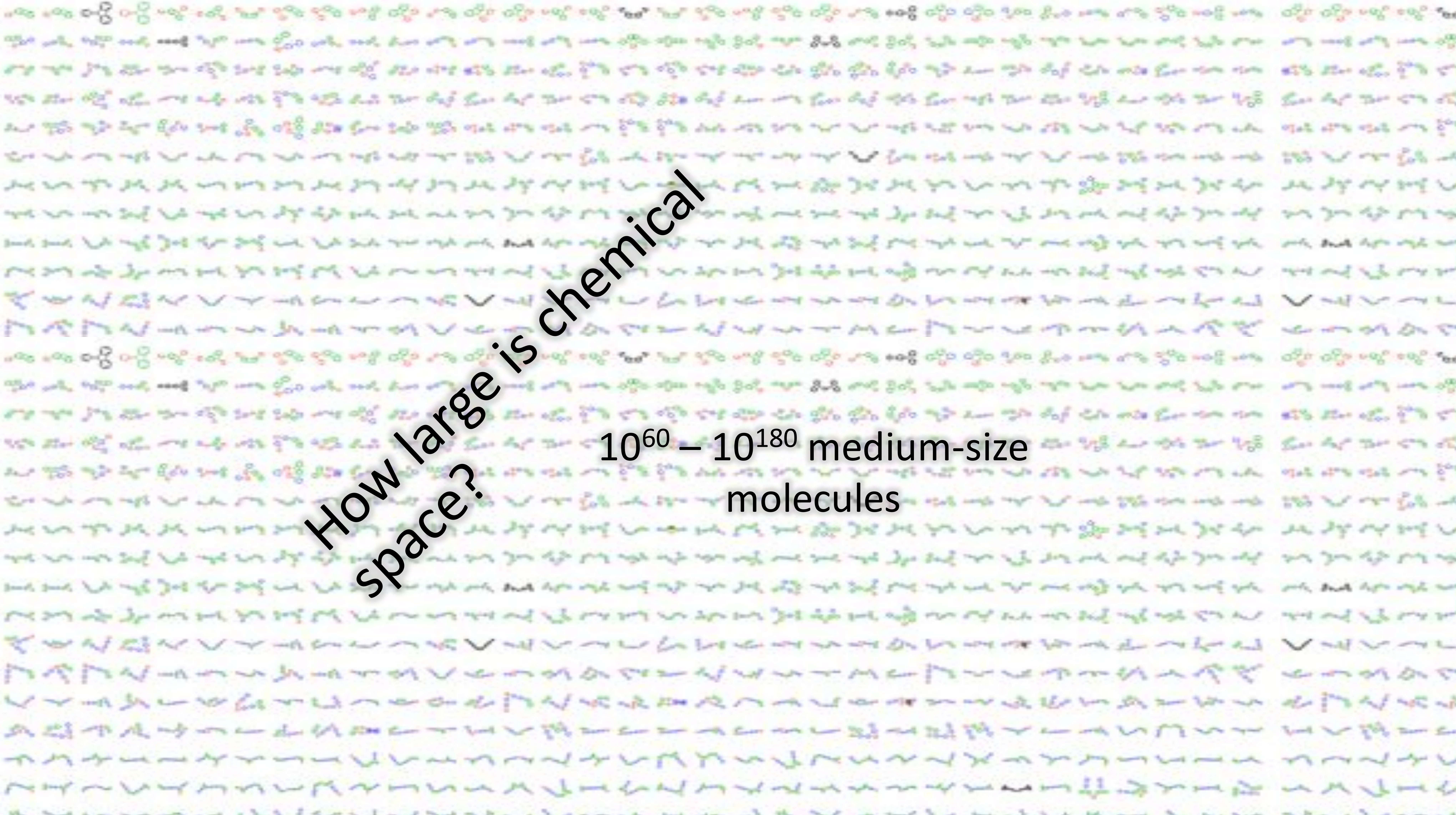


How large is space?

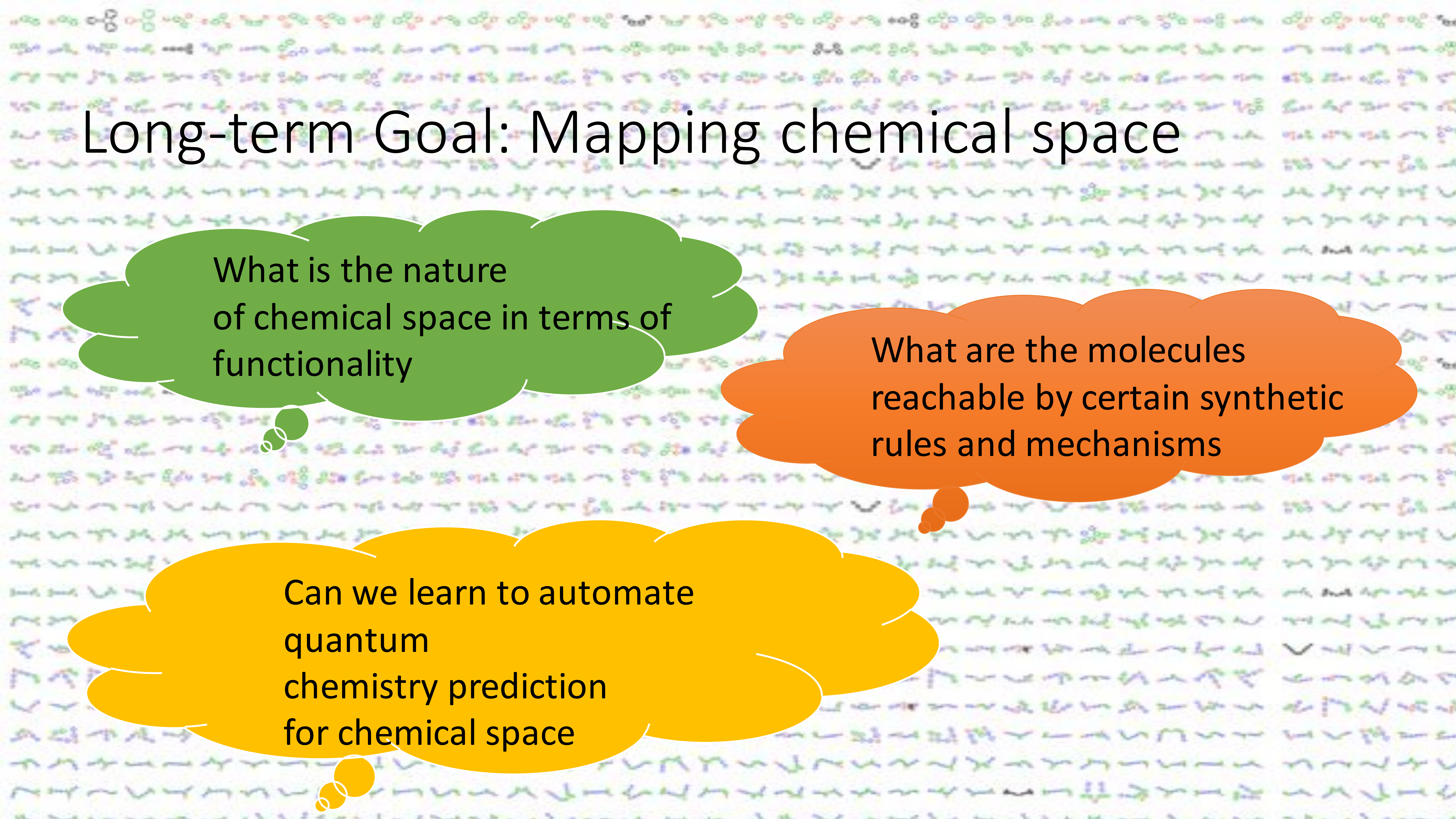
10^{82} atoms in the observable universe

How large is chemical space?

$10^{60} – 10^{180}$ medium-size molecules



Long-term Goal: Mapping chemical space

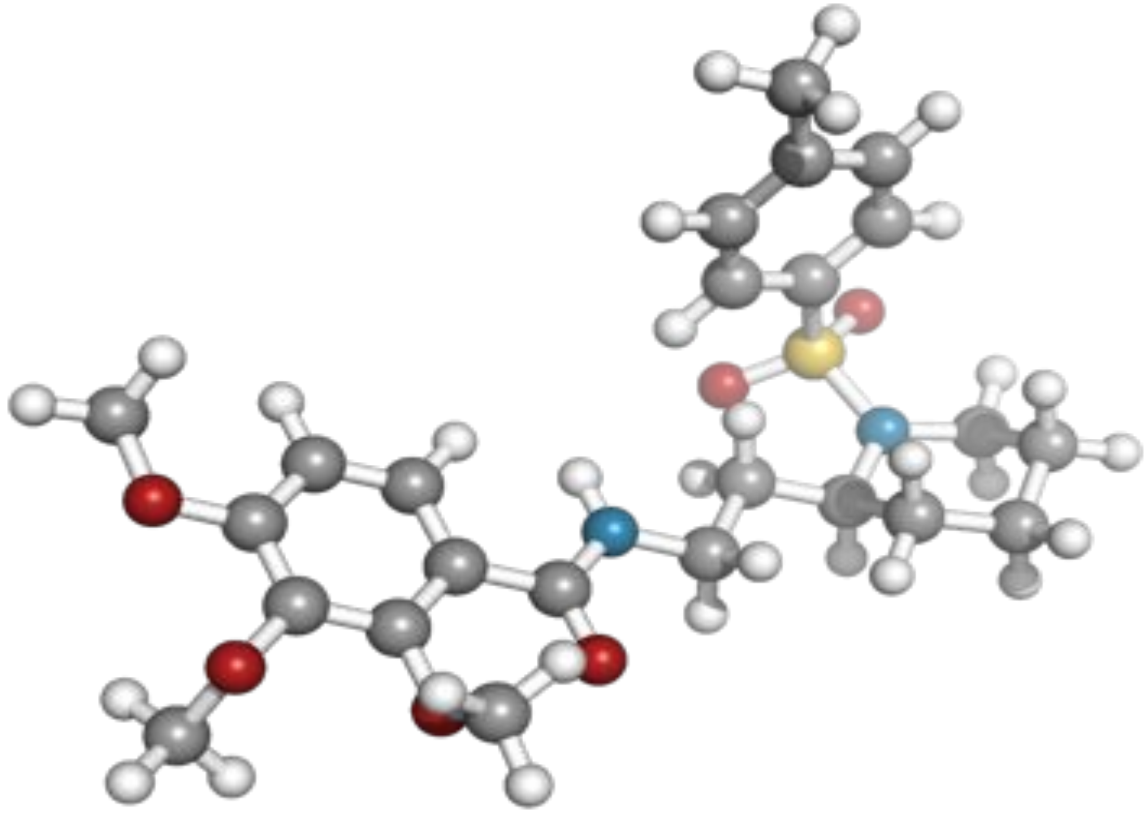


What is the nature of chemical space in terms of functionality

What are the molecules reachable by certain synthetic rules and mechanisms

Can we learn to automate quantum chemistry prediction for chemical space

Molecular screening for organic materials

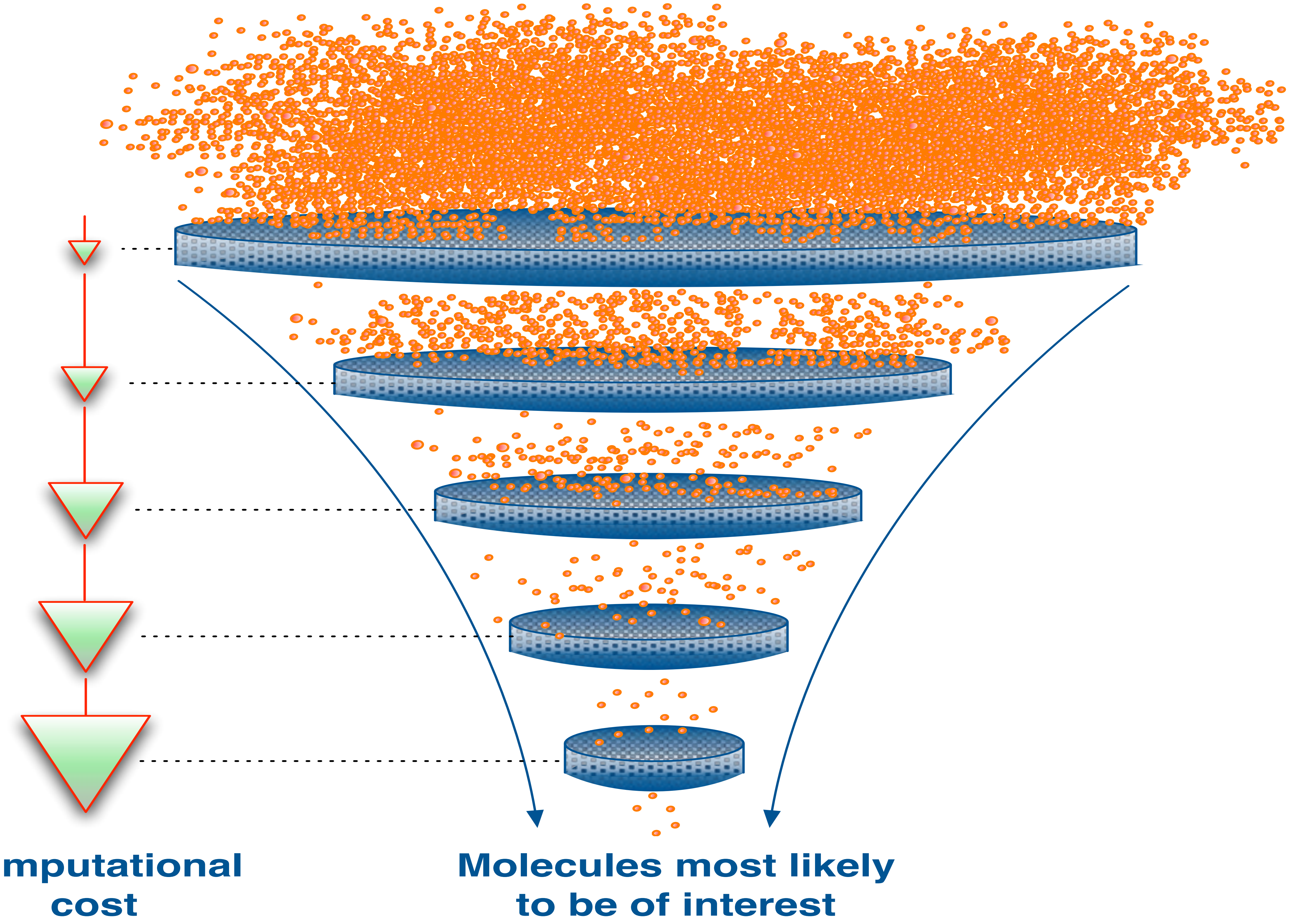


Quantum Mechanics

Machine Learning

How good is this molecule as a battery material?

From 10^{60} to 10^{6} to 10…

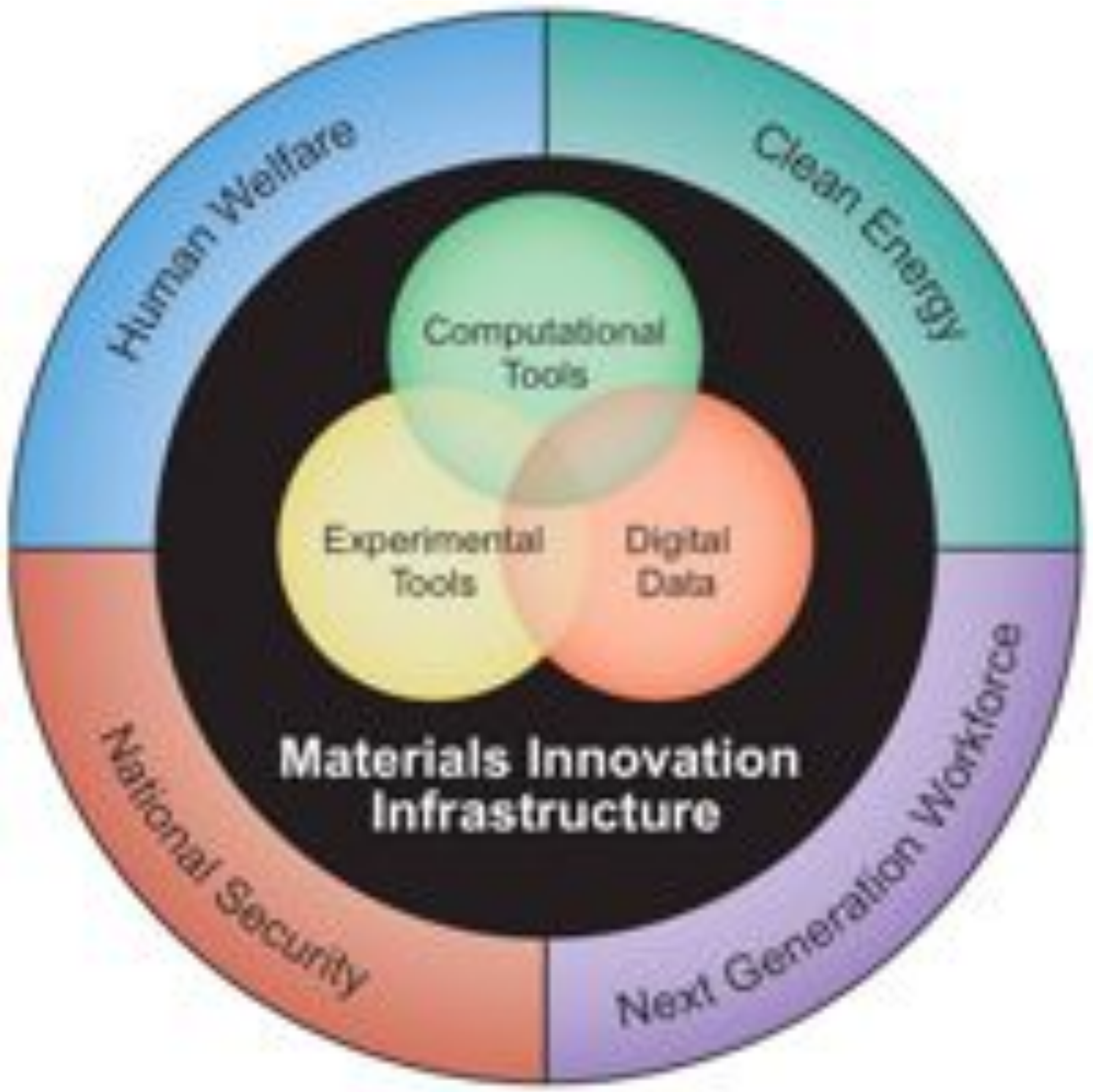


Initial library

Computational screening

Synthesis and testing

Organic materials in the larger context



Human Welfare

Clean Energy

National Security

Next Generation Workforce

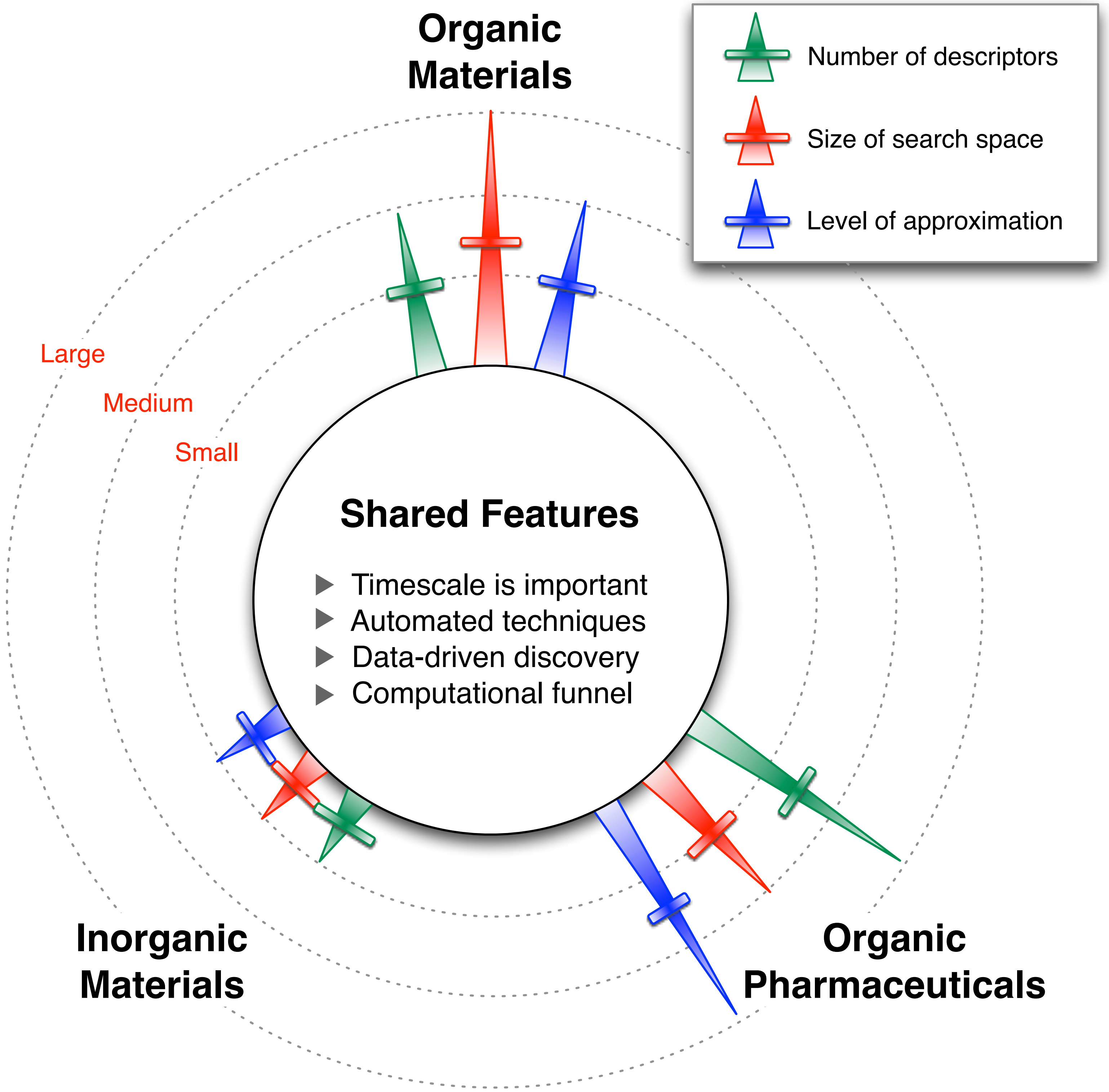
Computational Tools

Experimental Tools

Digital Data

Materials Innovation Infrastructure

US Materials Genome Initiative



Organic Materials

Inorganic Materials

Organic Pharmaceuticals

Number of descriptors

Size of search space

Level of approximation

Large

Medium

Small

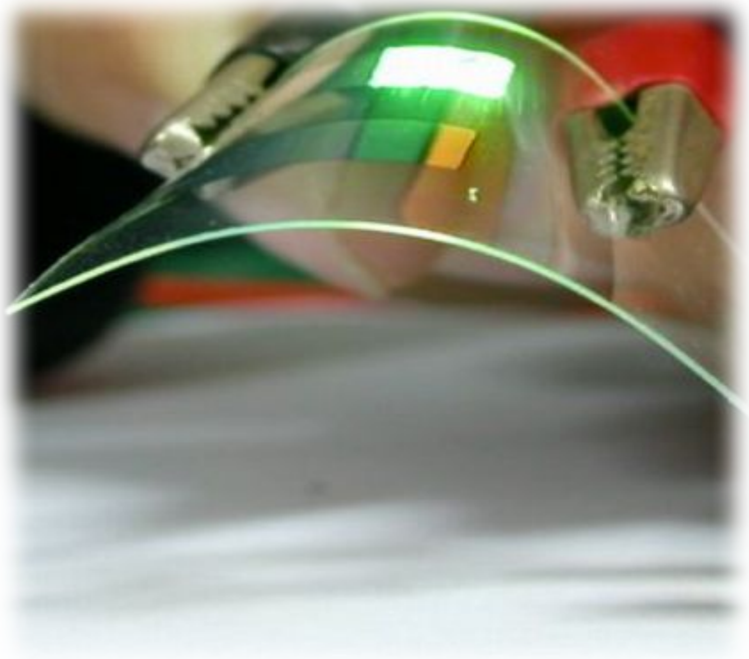
Shared Features

- Timescale is important
- Automated techniques
- Data-driven discovery
- Computational funnel

Pyzer-Knapp, et al. Ann Rev. Mat Sci. (2015)

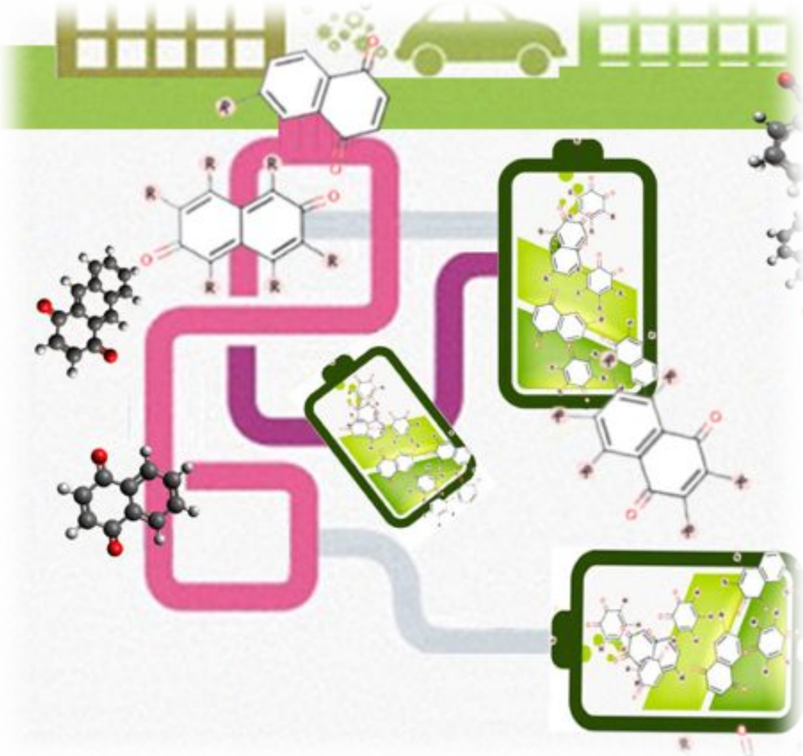
My research group's explorations of chemical space

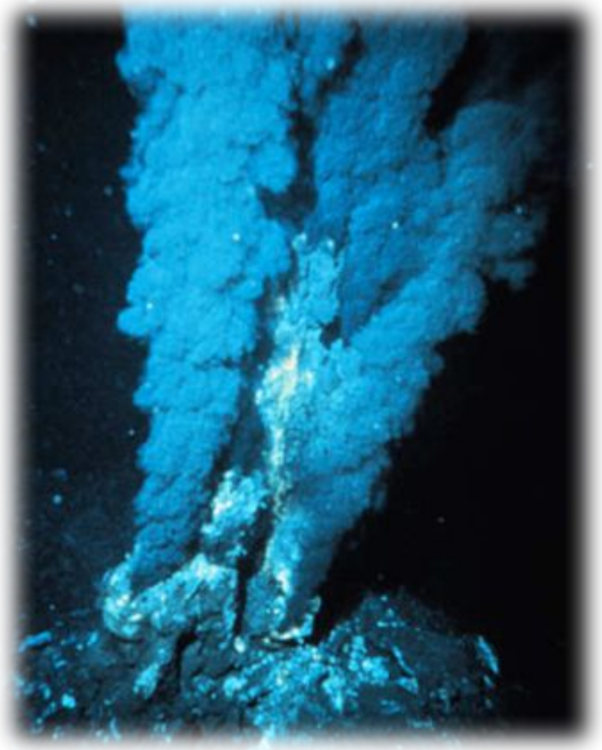
The Harvard Clean Energy Project
Generating renewable energy



Blue Organic LED
For your next gadget or TV

Organic flow batteries
Storing renewable energy



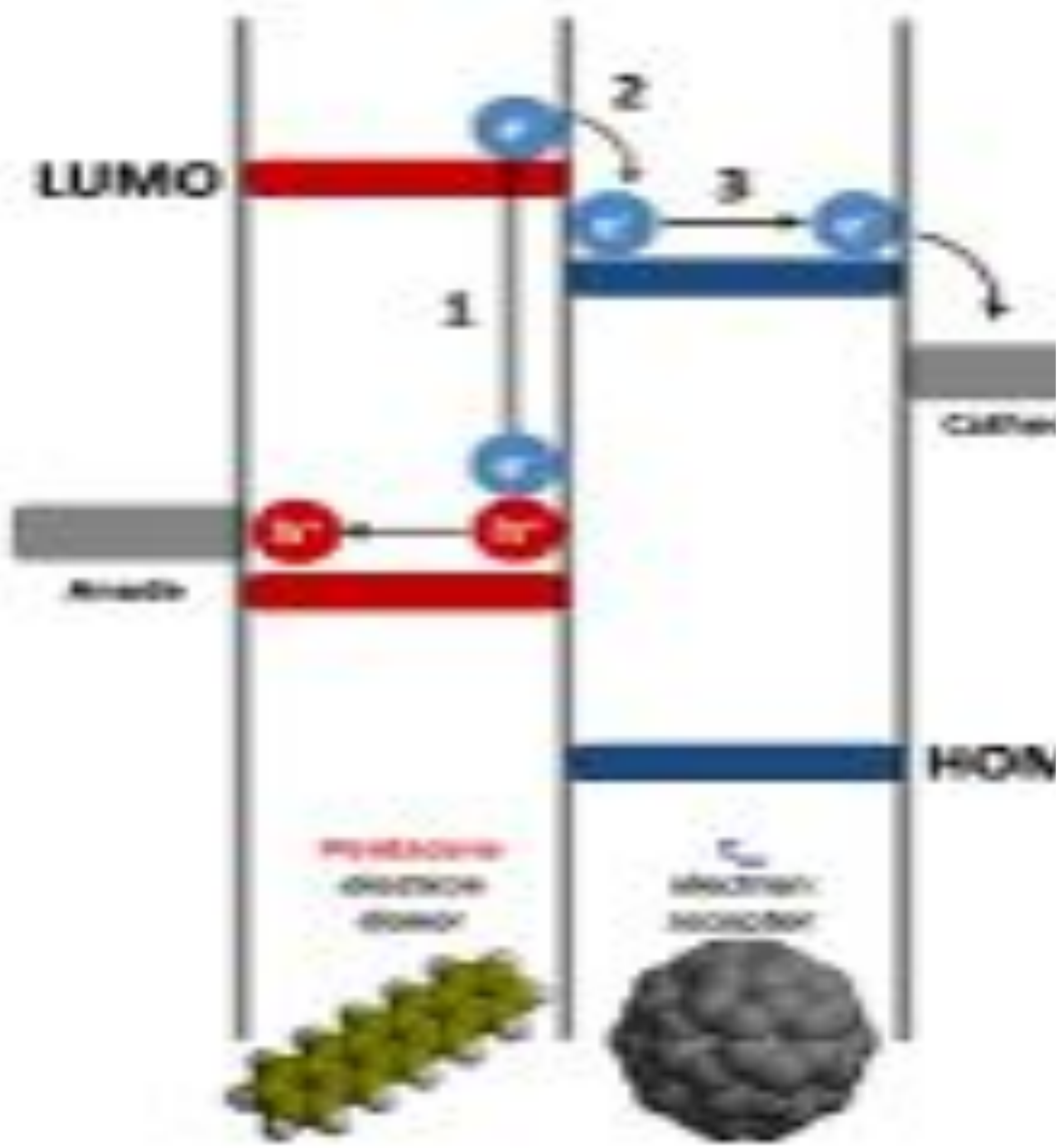


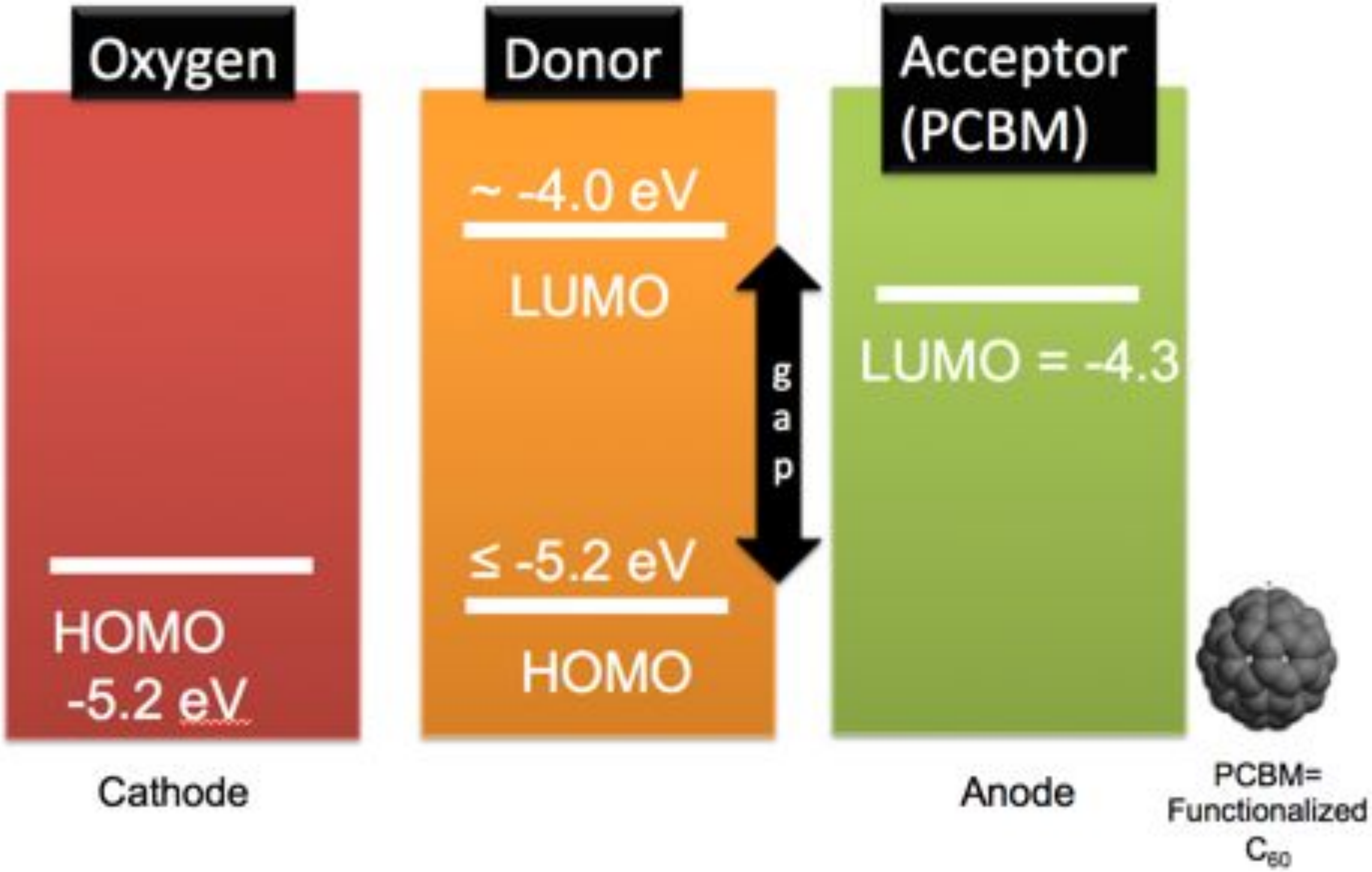
Origins of life
How life may have come about?

Project chronology and screening methodology improvements

1.0	Harvard Clean Energy Project	2006-	Distributed high-throughput	+	Experimental calibration	+	MolecularSpace database
2.0	Organic flow batteries	2013-	Tight experimental feedback cycle	+	Stability screening	+	Synthesizability at forefront
3.0	Organic Light Emitting diodes	2014-	Super-flexible molecular builder	+	Machine learning screening	+	Infinite libraries
4.0	Harvard Clean Energy Project	2014-	Thompson sampling	+	Several experimental collaborations at once	+	Molecular crystal ML and prediction

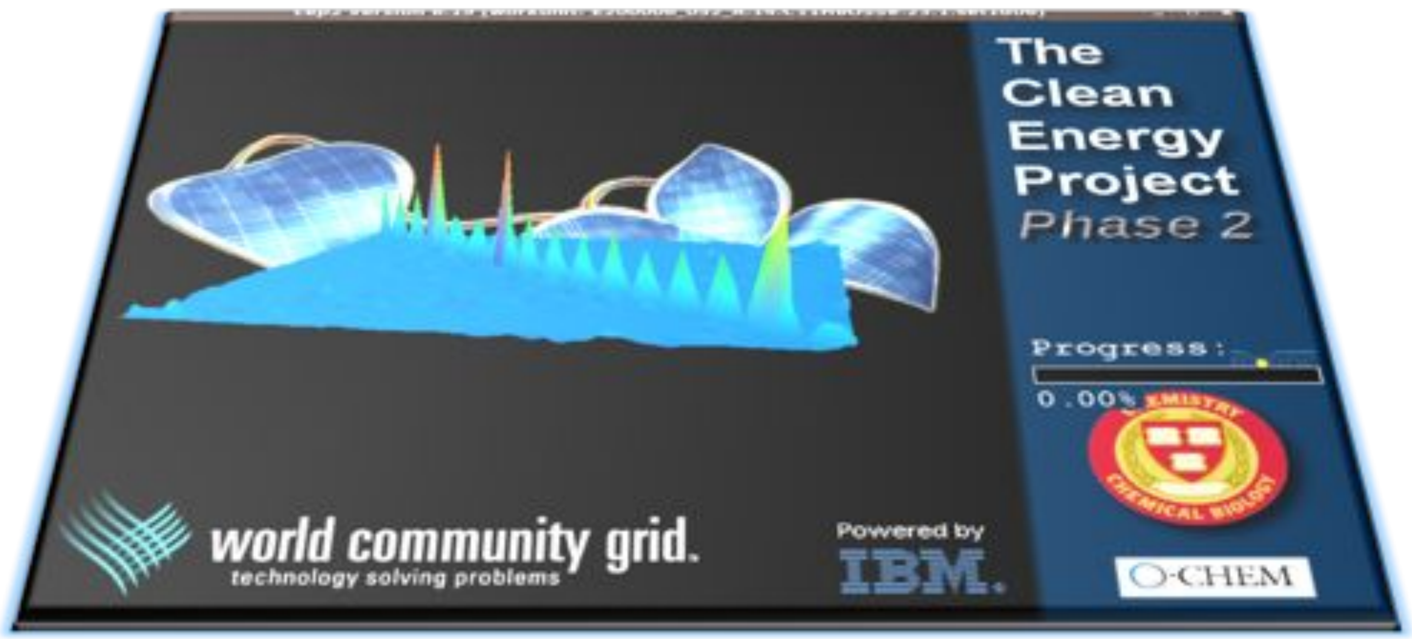
Harvard Clean Energy Project:Organic solar cells





The Harvard Clean Energy Project

The Largest quantum chemistry screening project to date



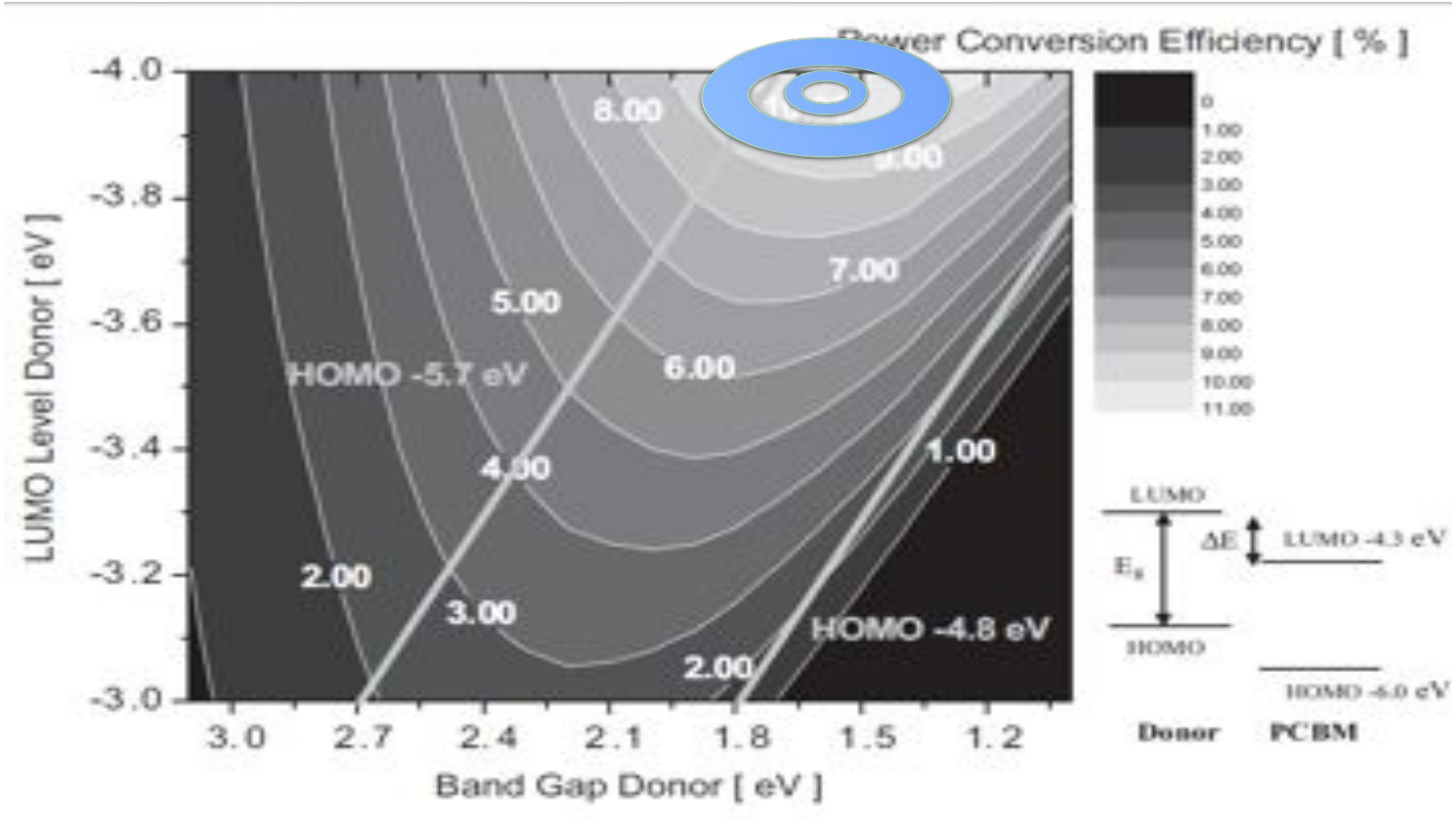
Part IV: Clean Energy Project

30,000+ CPU years have led to more than 35,000 high-performance organic photovoltaic candidates.

Collaborators: Juan Hindo (IBM) Zhenan Bao (Stanford), Johannes Hachmann (Buffalo), Alejandro Briseño (UMass), Carlos Amador (UNAM), and …

Hachmann, et. Al, J. Phys. Chem Lett. (2011), Energ. Env. Sci. (2014)

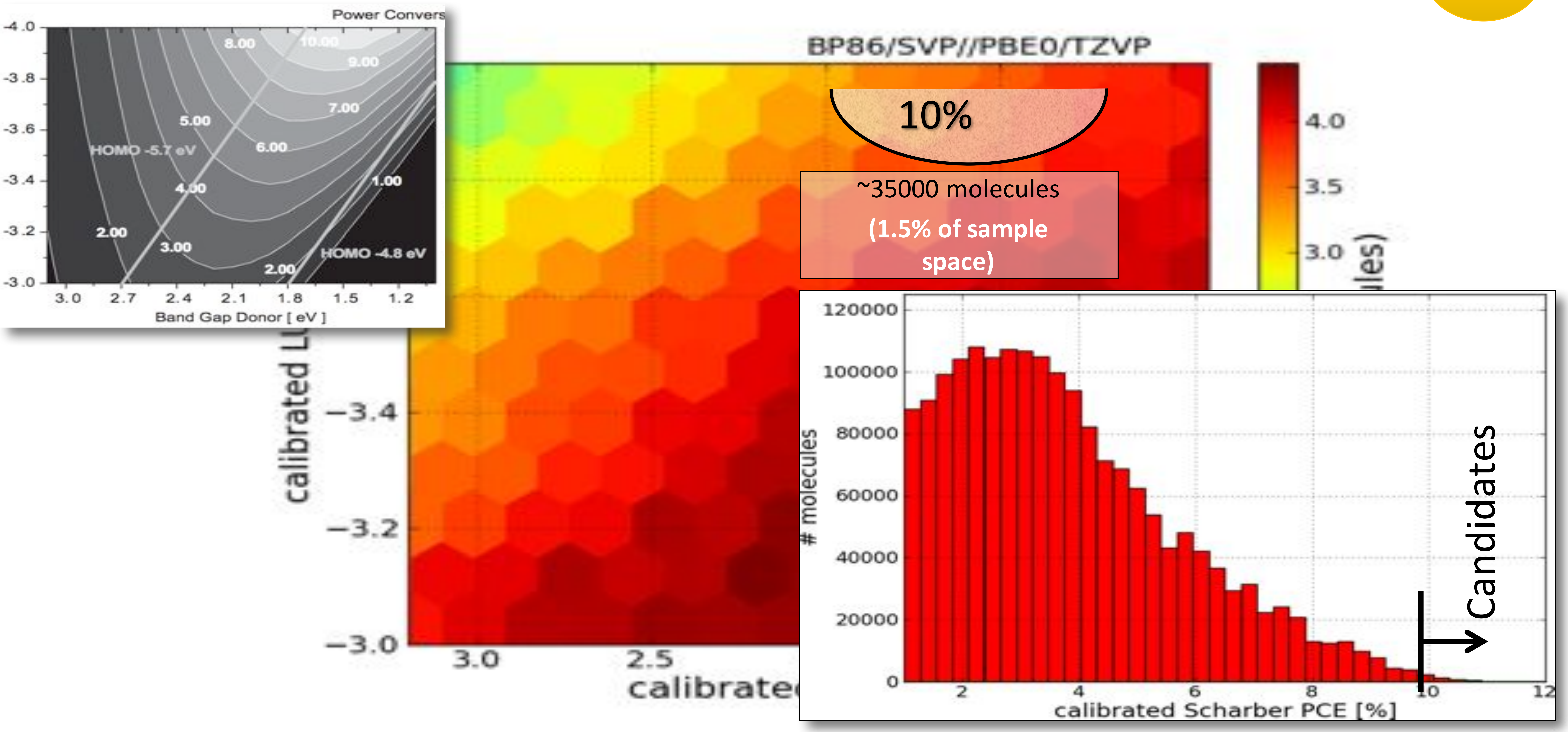
Energy Levels and Efficiency



Scharber M, Mühlbacher D, Koppe M, et al., Advanced Materials (2006)

Sifting through 2.3 million molecules

1.0



10%

~35000 molecules

(1.5% of sample space)

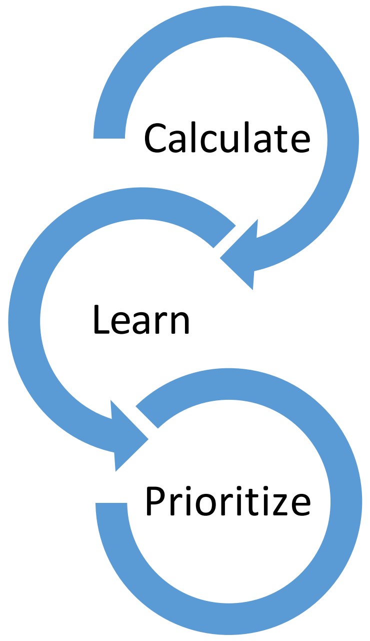
Candidates

Energy and Environmental Science, 7, 698 (2014)

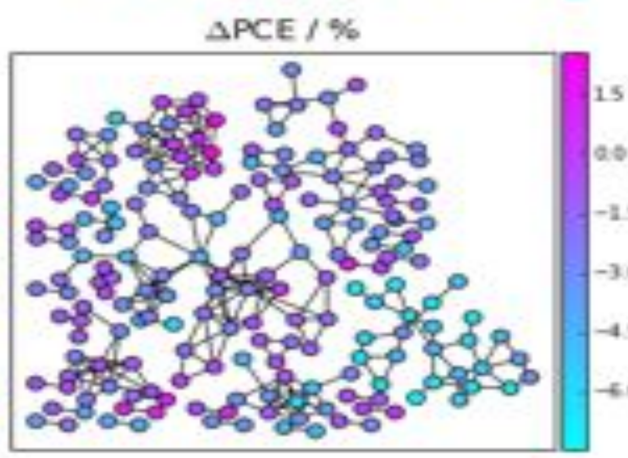
The Clean Energy Project gets an artificial intelligence boost!

4.0

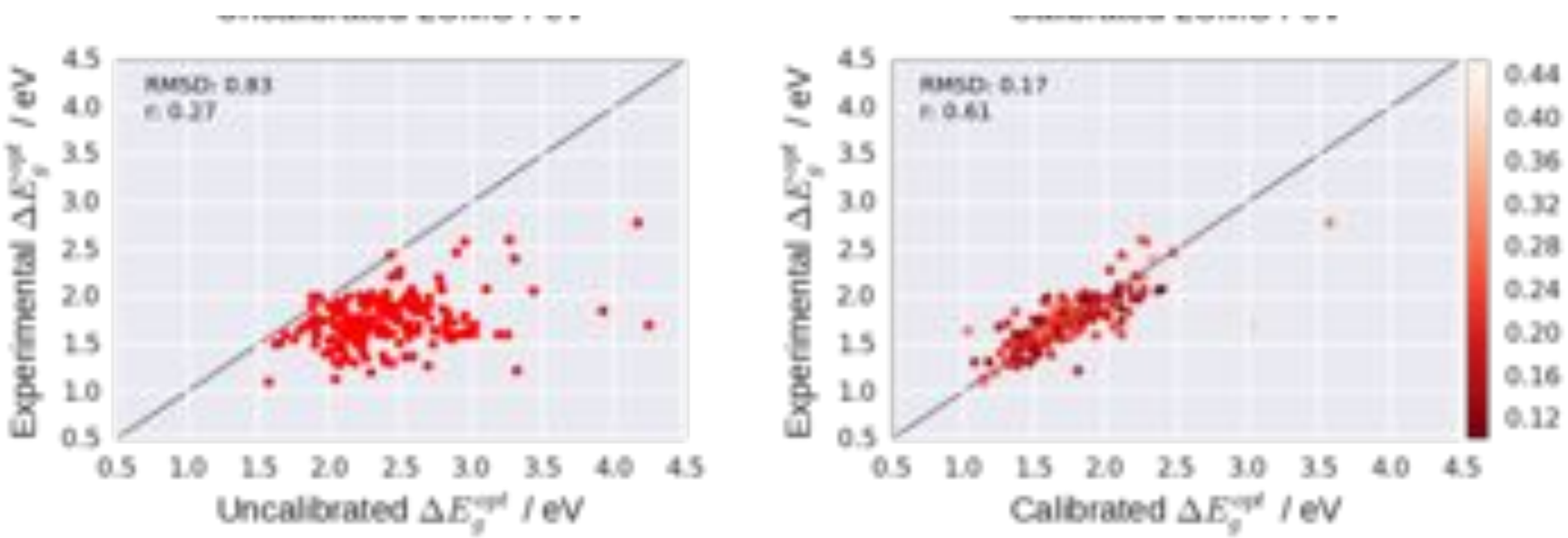
Machine learning



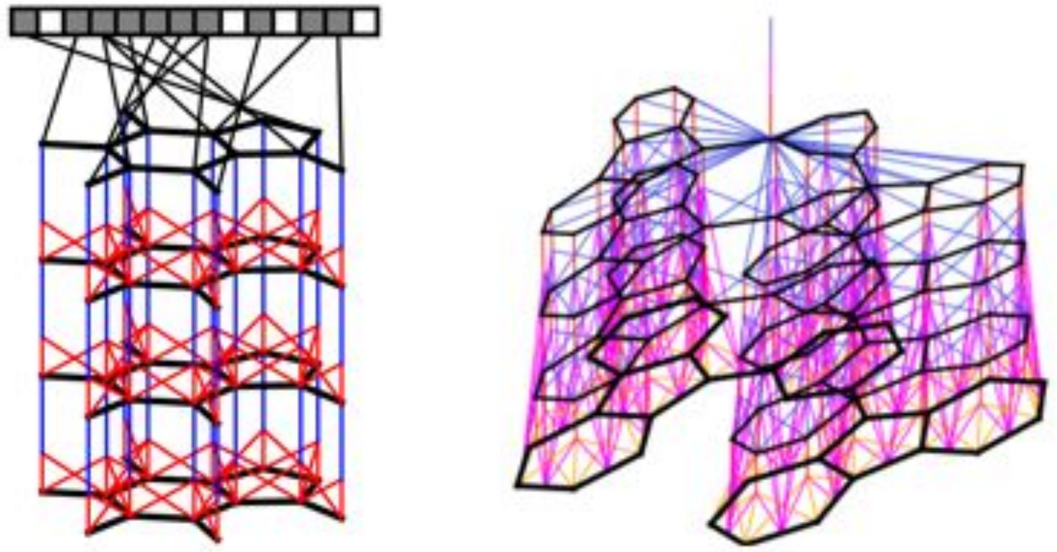
Easy to synthesize libraries



Bayesian calibration



Smart Screening Using machine learning



Neural Fingerprints

E. O. Pyzer-Knapp, et al. Advanced Functional Materials 2015

E. O. Pyzer-Knapp, et al. arXiV:1510.00388 2015

D. Duvenaud, arXiV:1509.09292 NIPS 2015

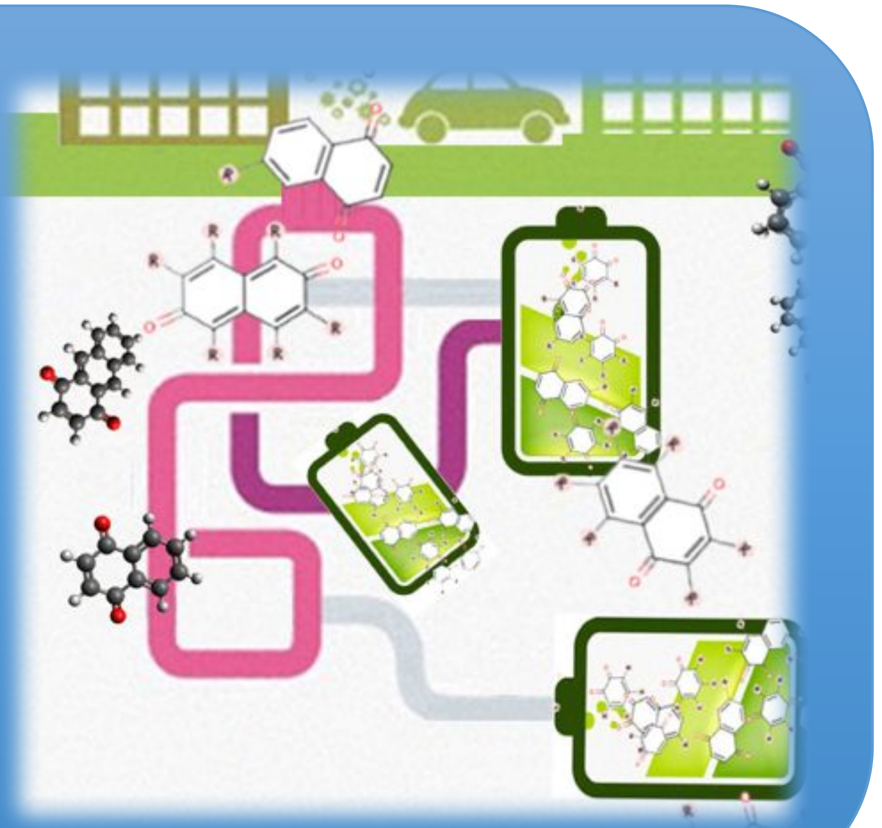
Collaborator: Ryan Adams (Harvard)

Organic Flow Batteries

The Harvard Clean Energy Project
Generating renewable energy

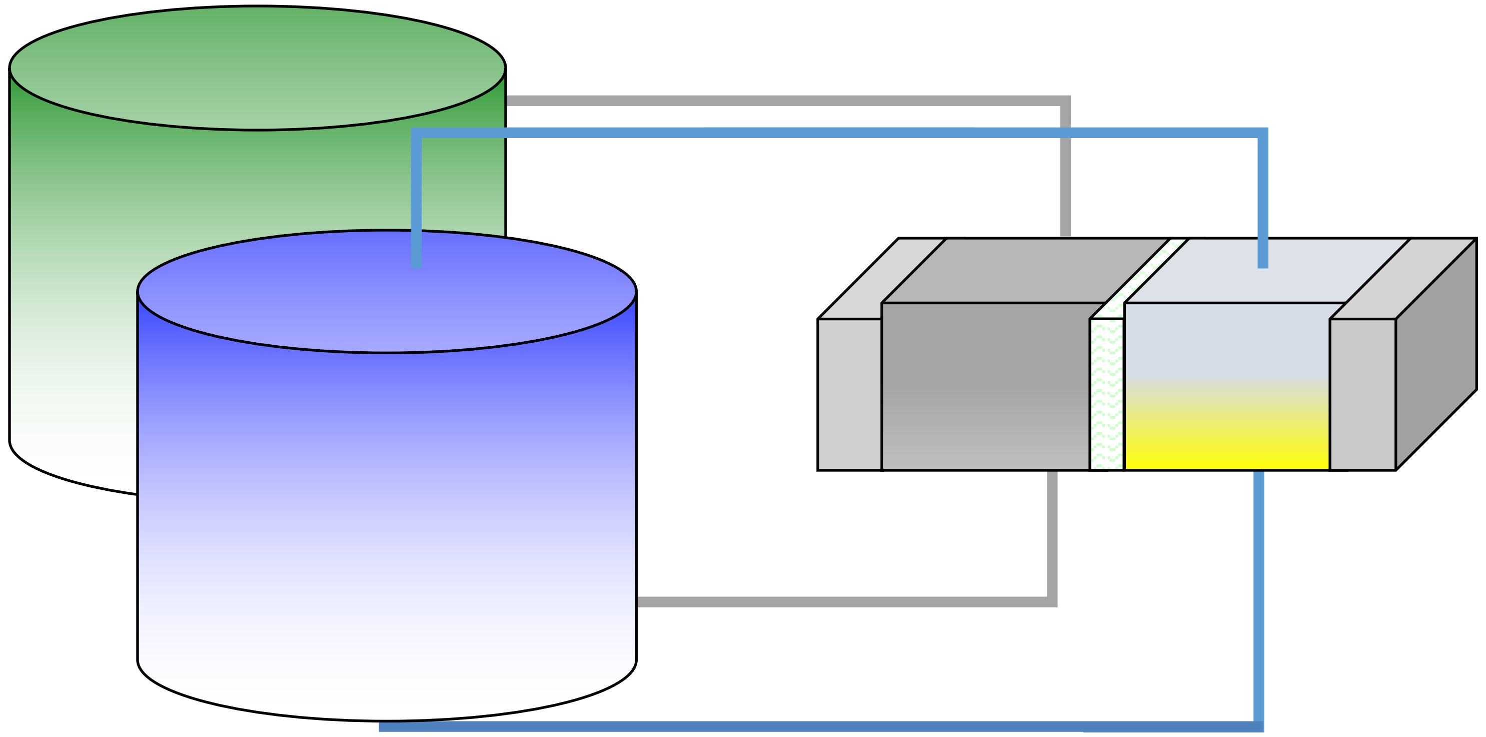
Blue Organic LED
For your next gadget or TV

Organic flow batteries
Storing renewable energy



Chemical networks
Origins of life
Organic reactions
Chemical autoencoders

Organics for storing clean energy



Organic flow batteries

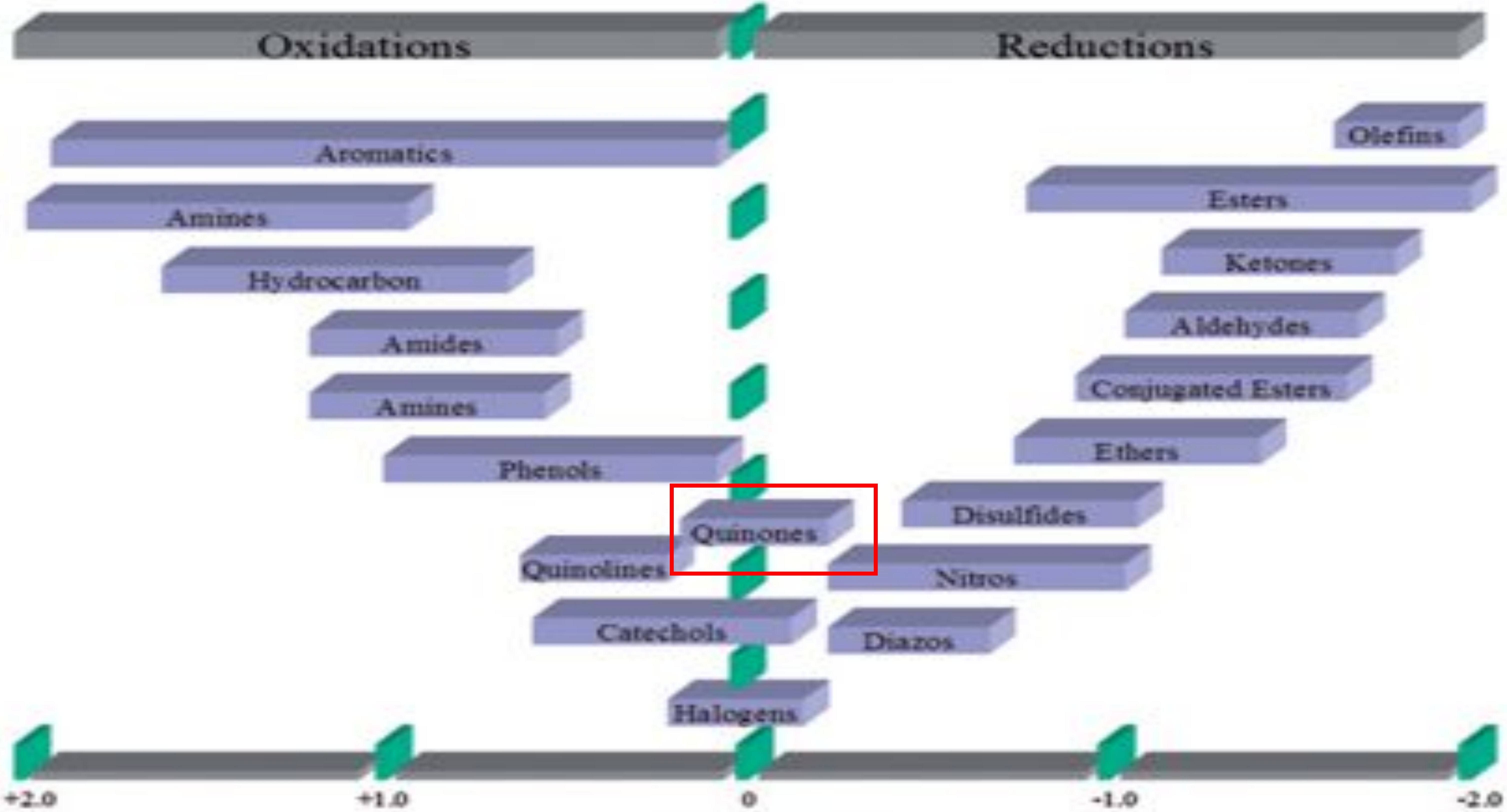
Suh, et al., Chem. Sci., 6, 2015, p. 885

Huskinson, et al., Nature, 505, 2014, p. 195

Lin, et al., Science, 349, 2015, p. 1529

Collaborators: Mike Aziz and Roy Gordon (Harvard)

Search space for redox potentials



E (V vs. Saturated Calomel Electrode)

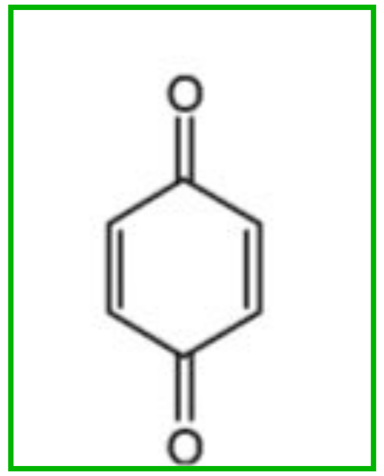
Estimated potential range of organic functional group@ 25°C

Handbook of Electrochemistry Ed. C.G. Zoski

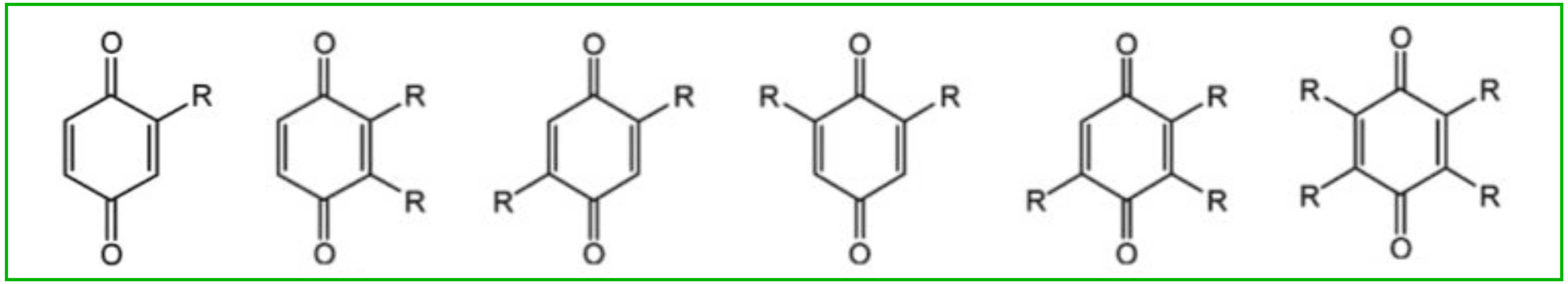
Choice for combinatorial library:
1R and fully substituted cases only

2.0

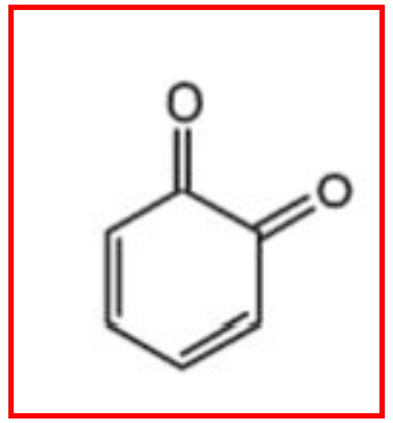
1. $N(CH_3)_2$
2. NH_2
3. OCH_3
4. OH
5. SH
6. CH_3
7. SiH_3
8. F
9. Cl
10. C_2H_3
11. CHO
12. $COOCH_3$
13. CF_3
14. CN
15. COOH
16. PO_3H_2
17. SO_3H
18. NO_2

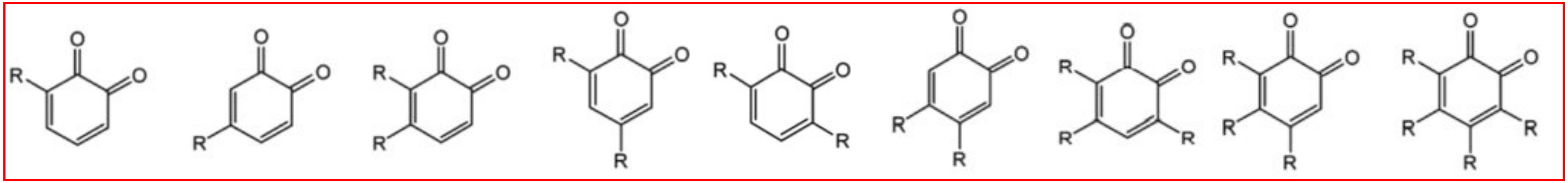


1,4-BenzoQuinones

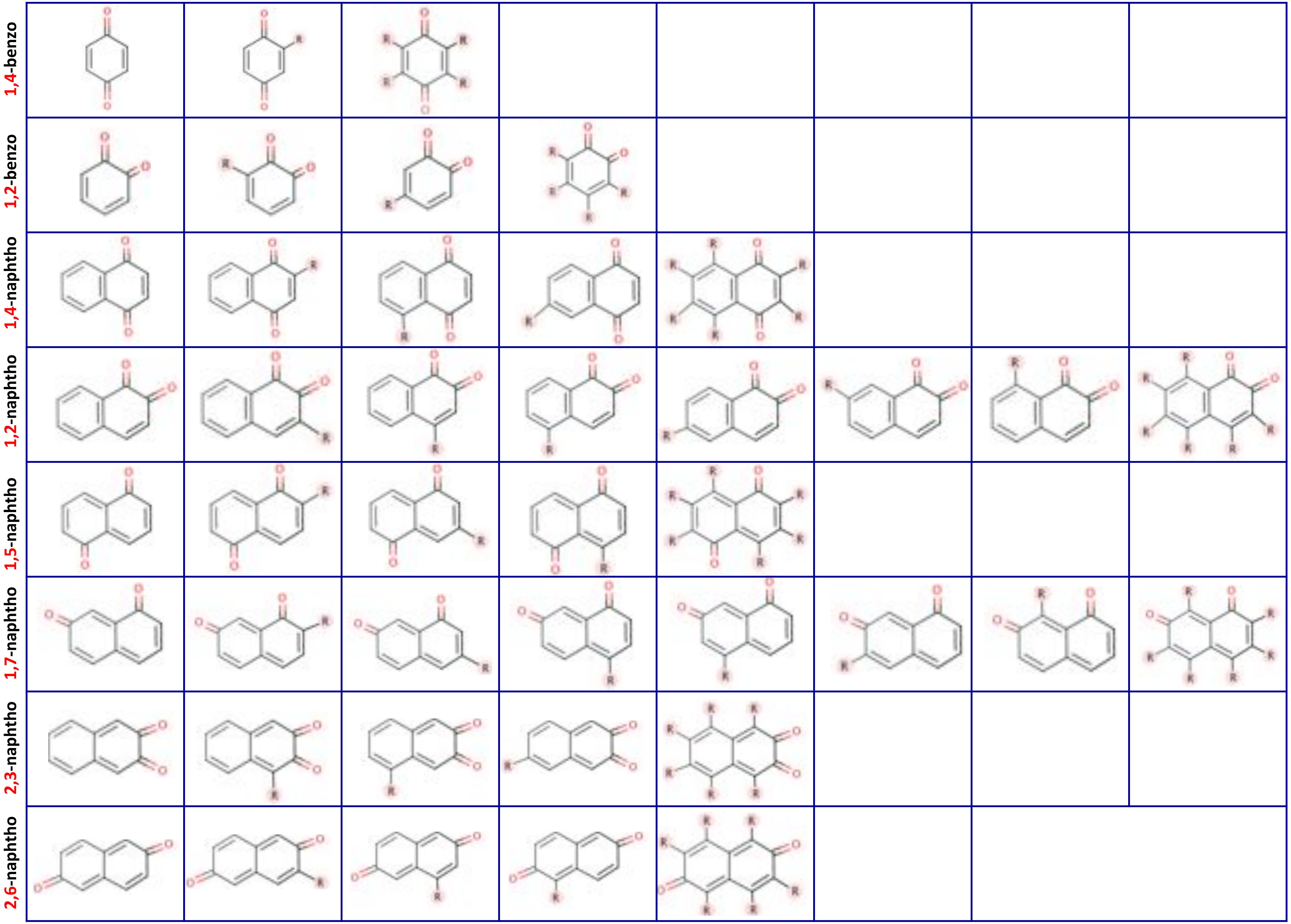


1,2-BenzoQuinones





Naphtoquinones



1,4-benzo

1,2-benzo

1,4-naphtho

1,2-naphtho

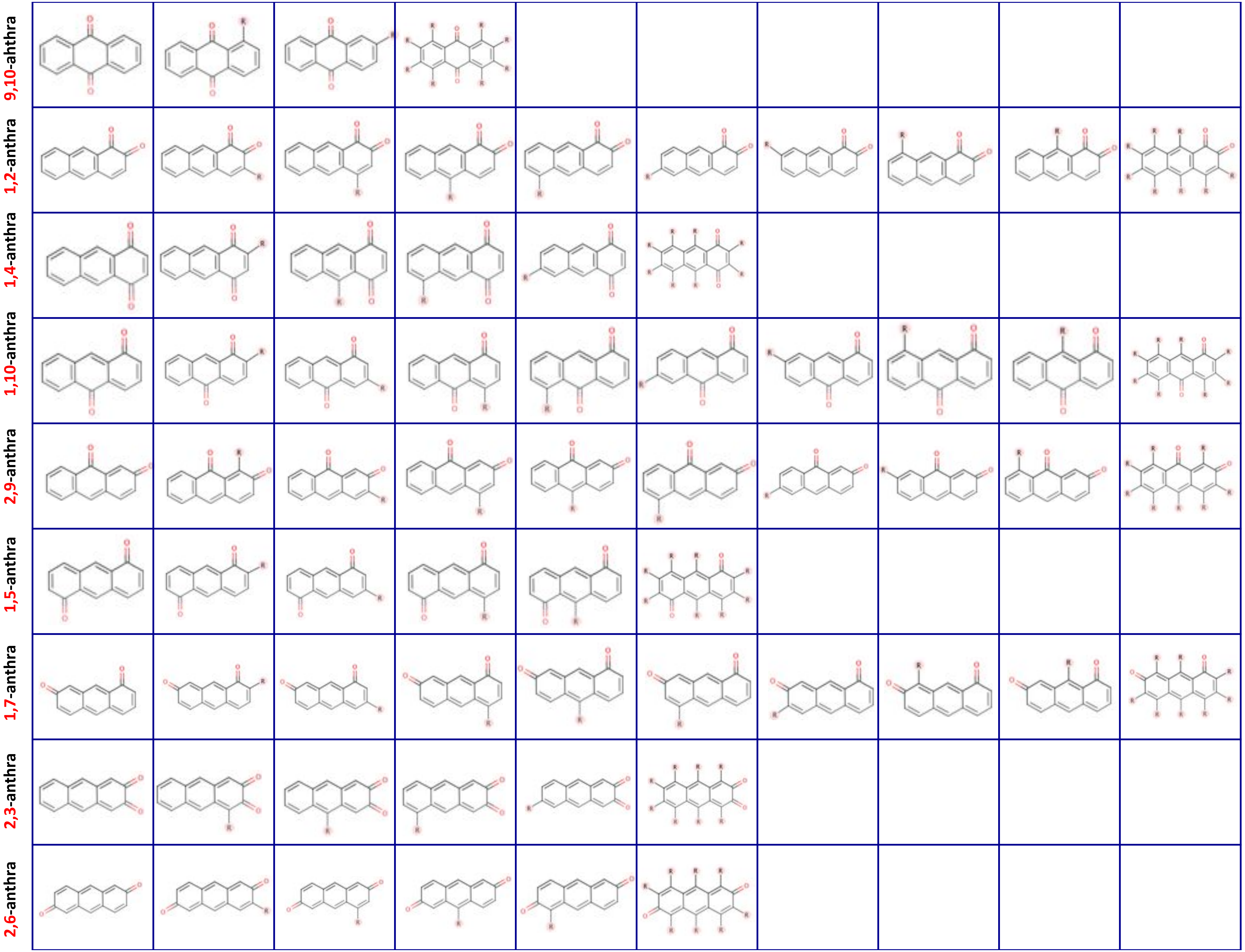
1,5-naphtho

1,7-naphtho

2,3-naphtho

2,6-naphtho

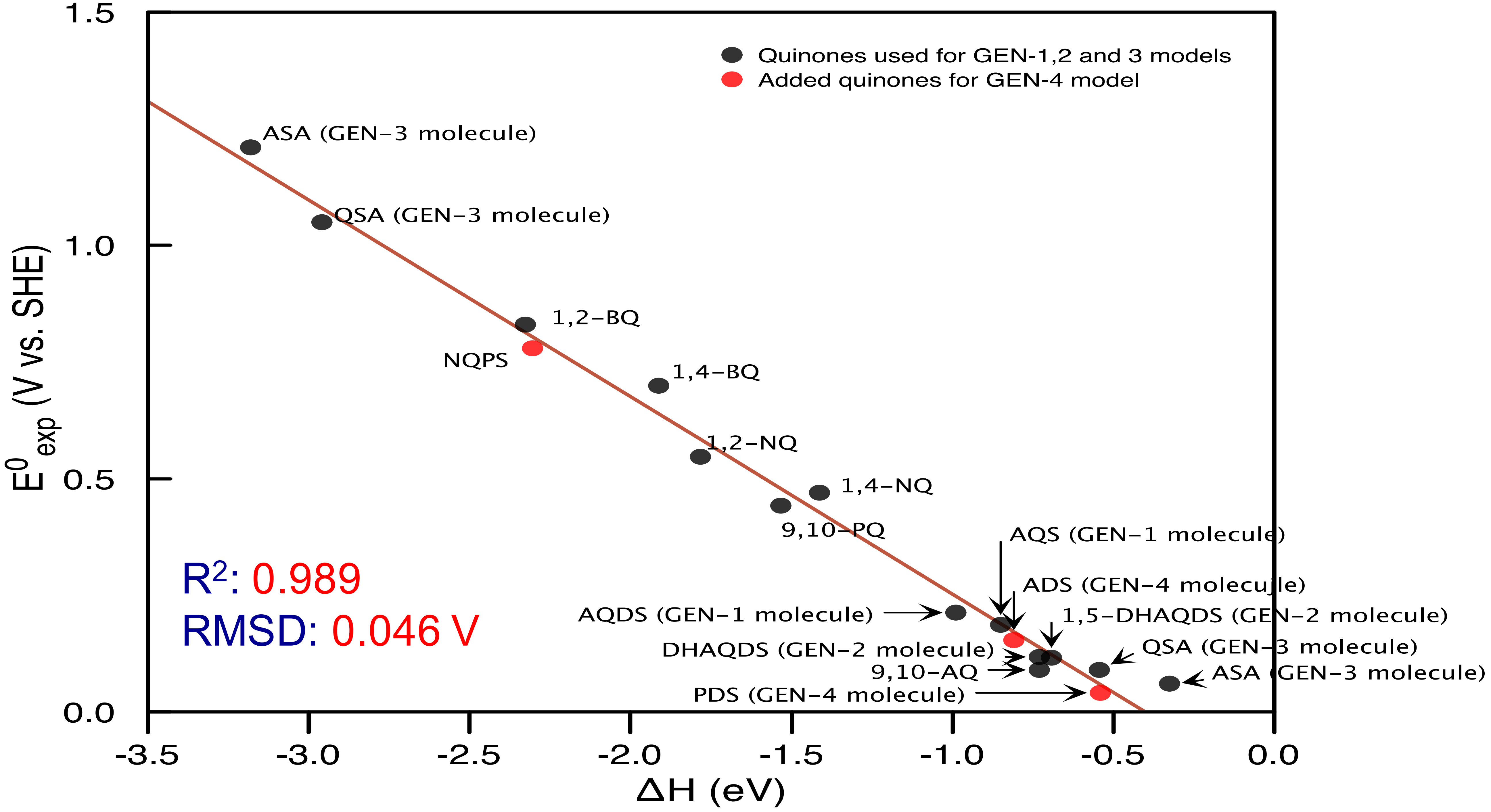
Anthraquinones



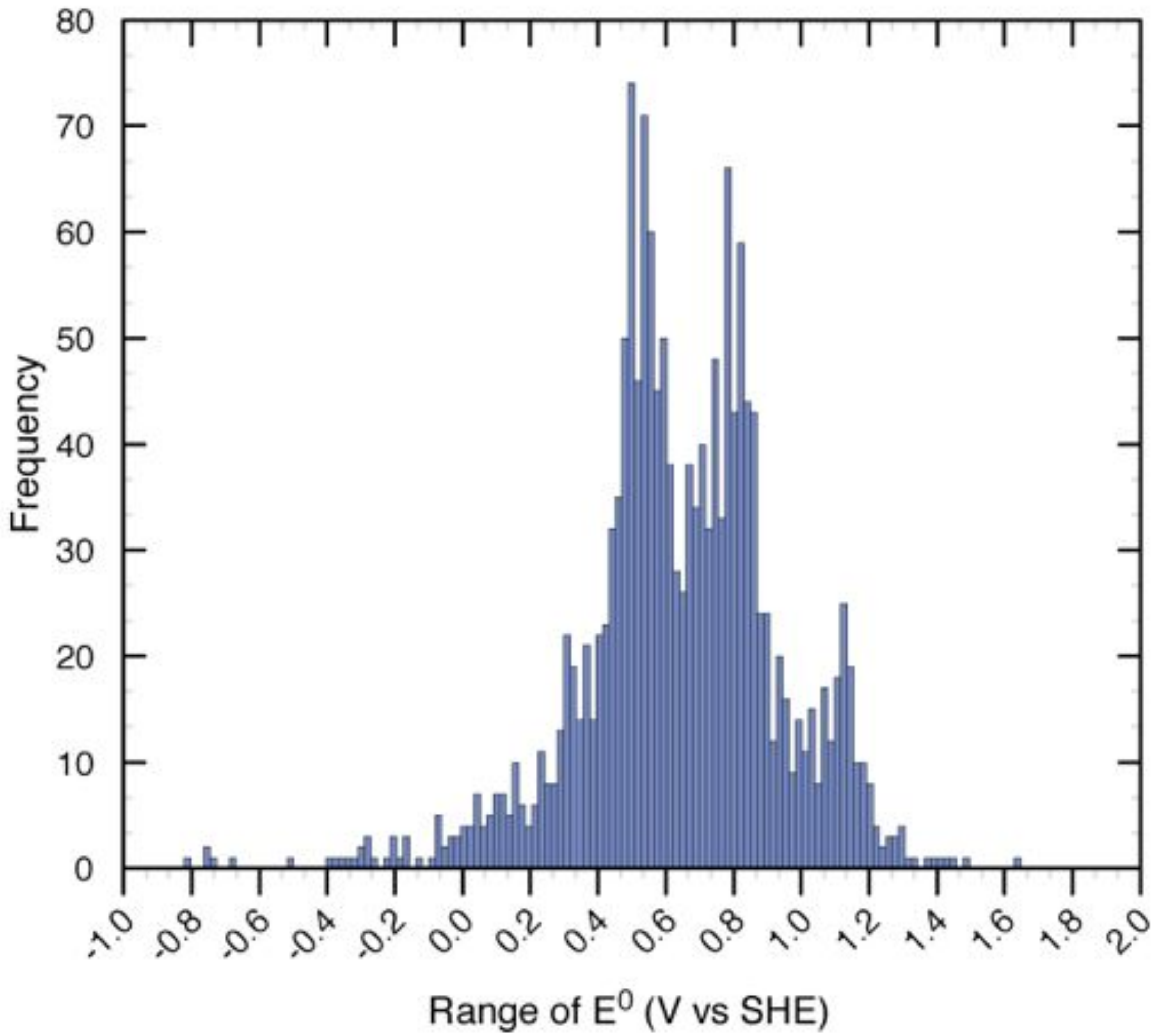
Theoretical calibration of quinone redox potentials

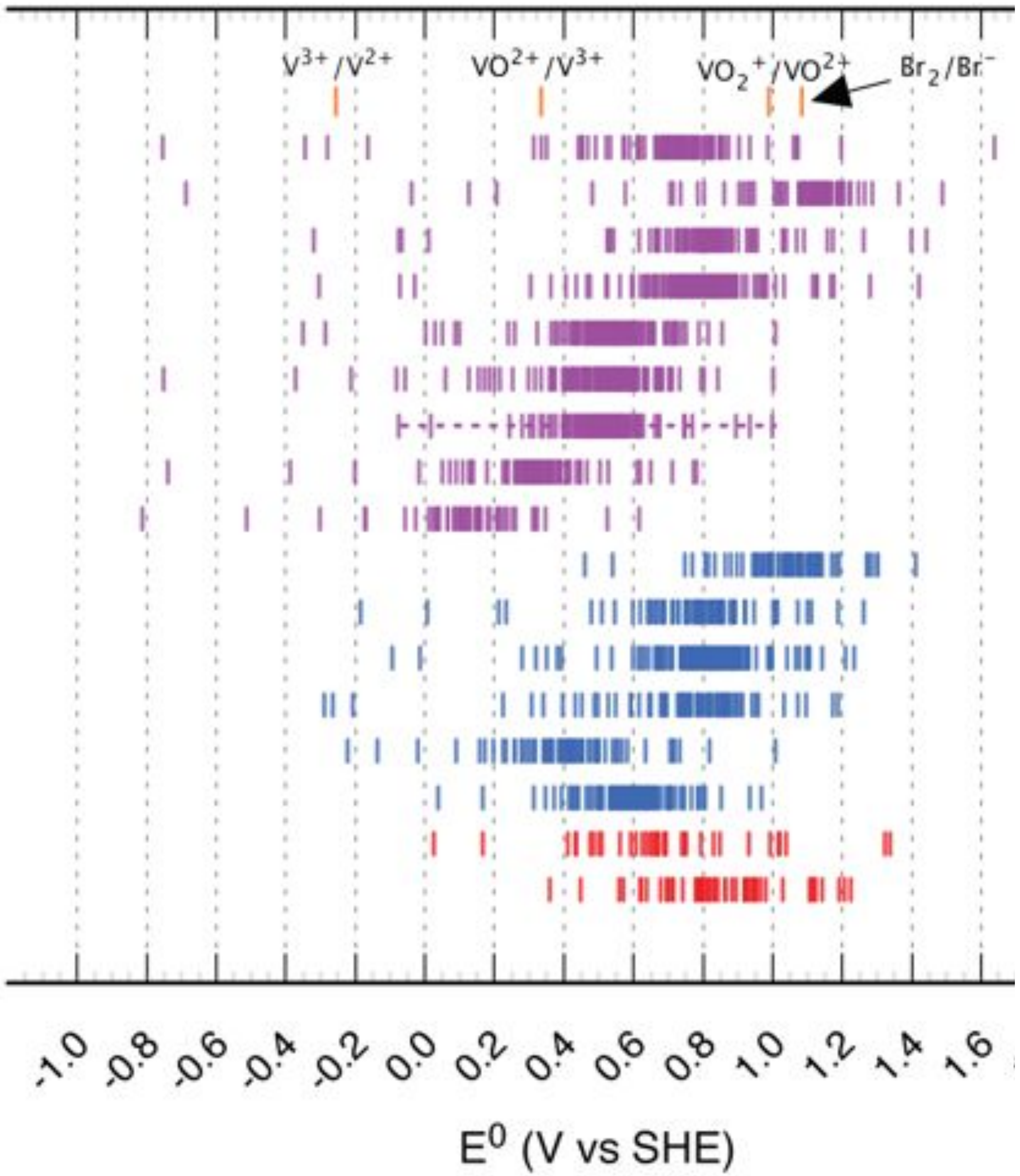
GEN-4 model

with inclusion of GEN-4 molecules (ADS and PDS) and NQPS



> 300 new candidate quinones predicted

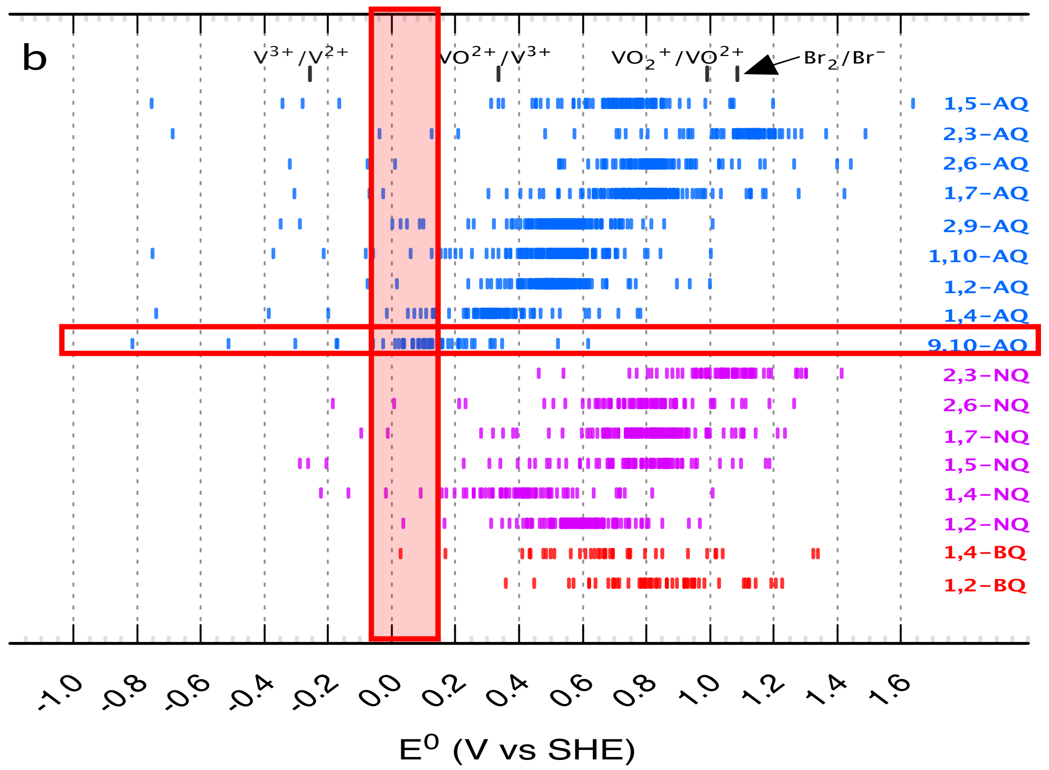


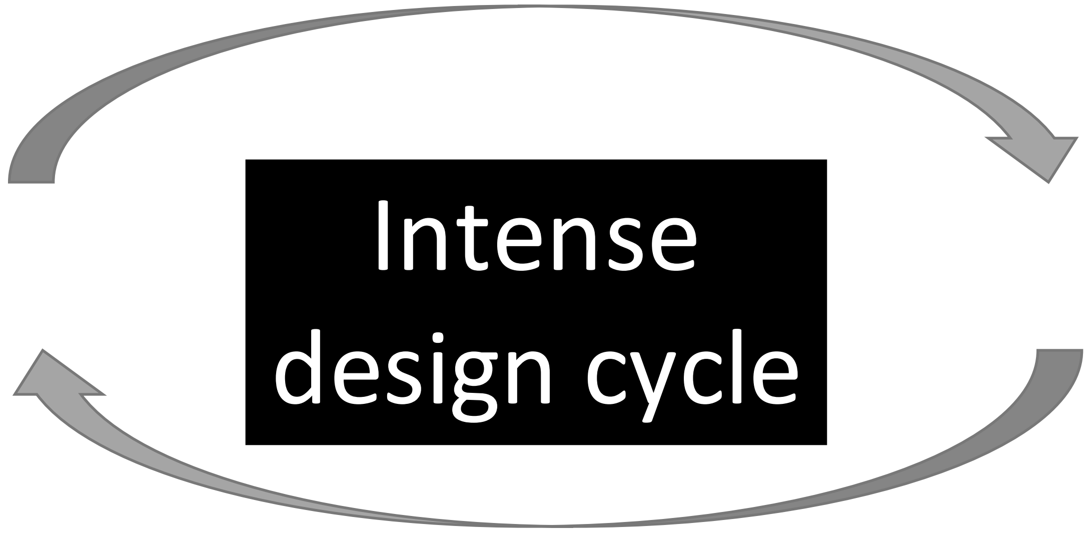


S. Er., C. Suh, M. P. Marshak, A. Aspuru-Guzik,
Chemical Science (2015)

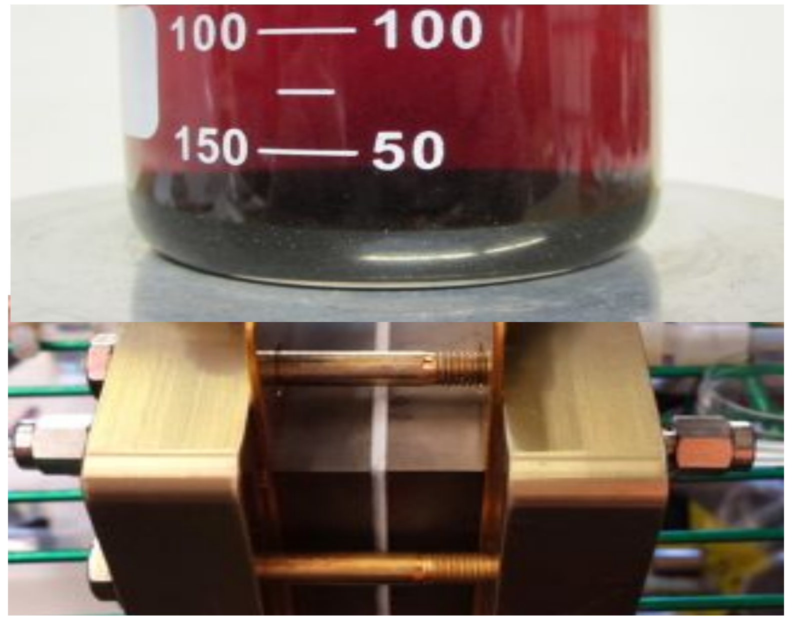
Our metal-free aqueous flow battery

2.0



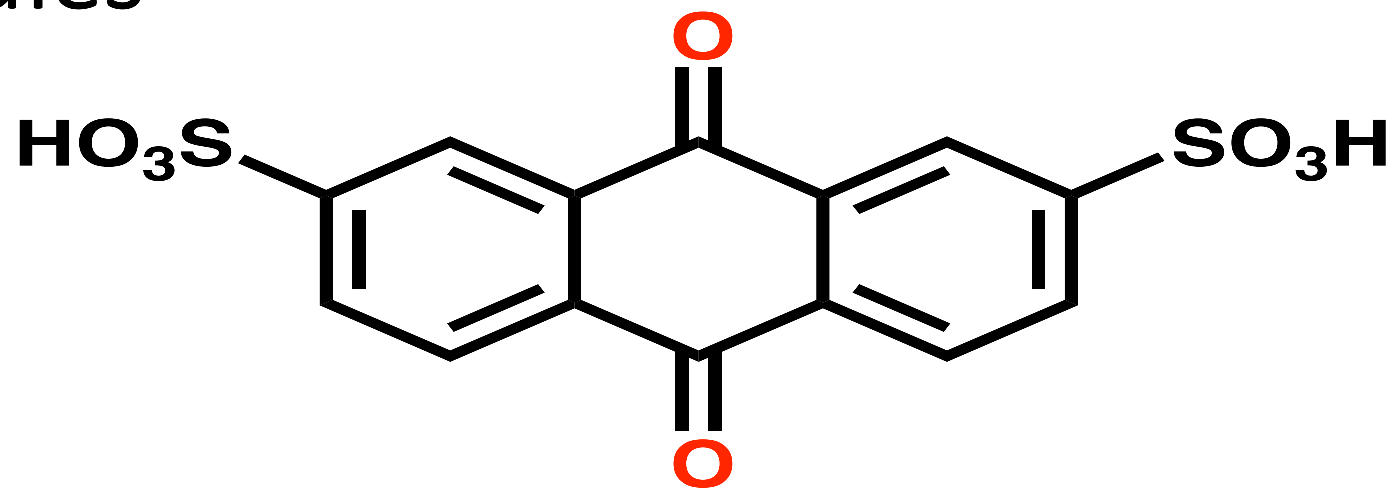


Intense design cycle



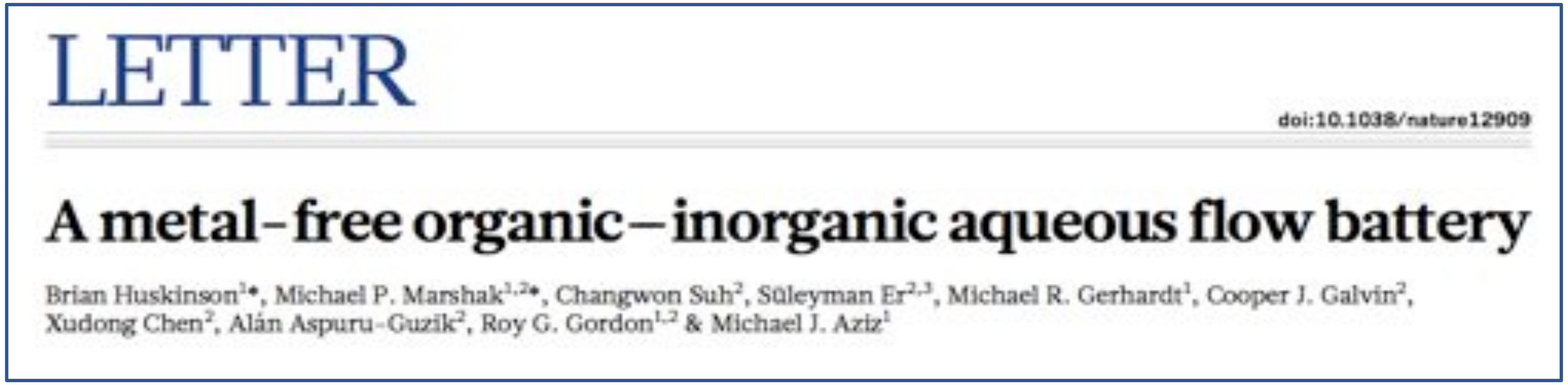
Computational screening of 10,000 quinone molecules

Synthesize molecules
Test in flow battery



Selected molecule

Theory-experiment collaboration



LETTER

doi:10.1038/nature12909

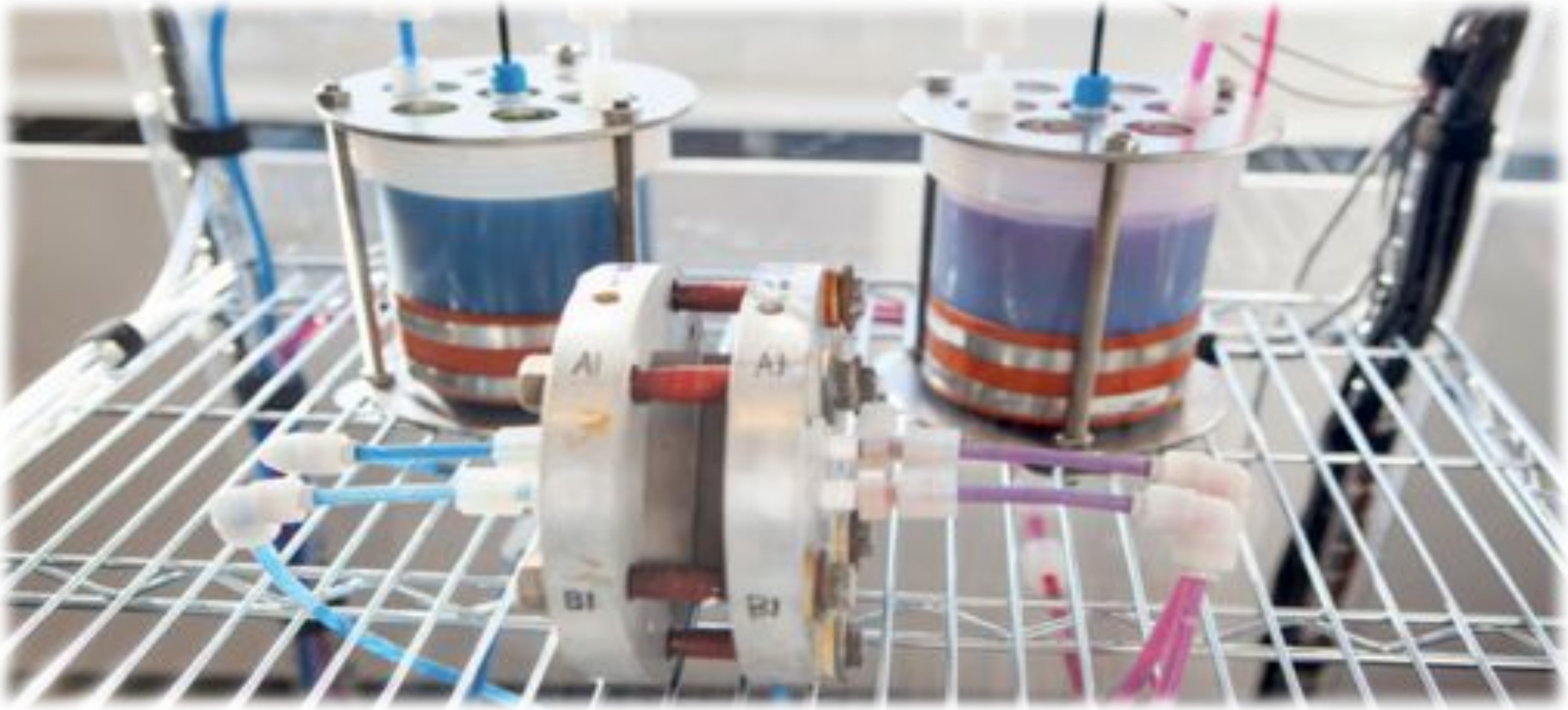
A metal–free organic–inorganic aqueous flow battery

Brian Huskinson[1]*, Michael P. Marshak[1,2]*, Changwon Suh[2], Süleyman Er[2,3], Michael R. Gerhardt[1], Cooper J. Galvin[2], Xudong Chen[2], Alán Aspuru-Guzik[2], Roy G. Gordon[1,2] & Michael J. Aziz[1]

Michael Aziz
Engineering

Roy Gordon
Chemistry

Alán Aspuru-Guzik
Chemistry



Nature, 505, 2014, p. 195

Molecular Flow Battery Data View

Blue: Stable molecule
Red: Unstable molecule

X axis: Redox Potential
Y axis: Free energy of Solvation

~ 100,000 molecules shown

Molecular Flow Battery Data View

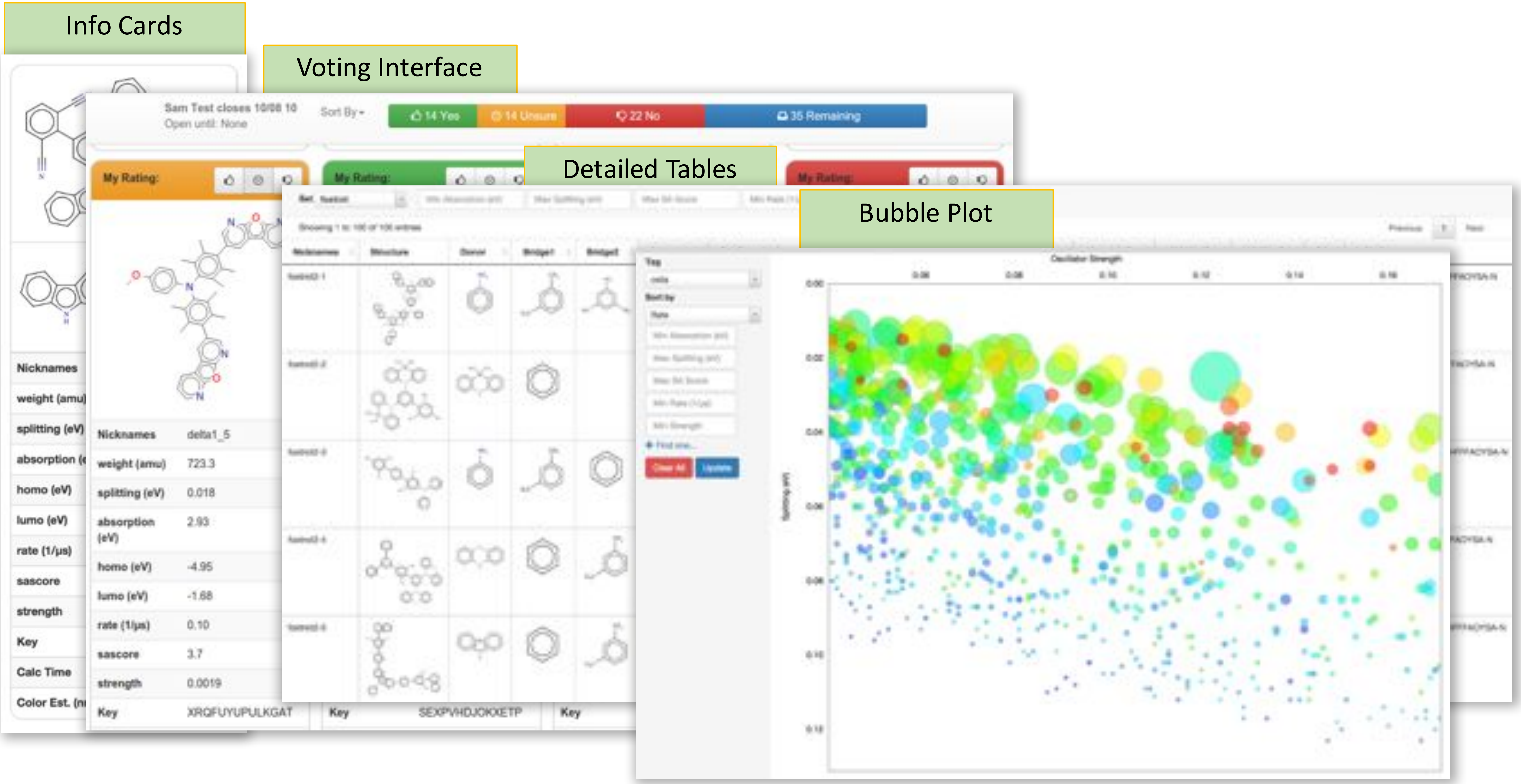
Filtering the data view

Molecular Flow Battery Data View

Baseball card view

High-throughput materials discovery process and tools

Moelcular Space Shuttle: advanced molecular discovery platform



Info Cards

Voting Interface

Detailed Tables

Bubble Plot

Sam Test closes 10/08 10
Open until: None

Sort By ▾

14 Yes | 14 Unsure | 22 No | 35 Remaining

My Rating:

Nicknames	delta1_5
weight (amu)	723.3
splitting (eV)	0.018
absorption (eV)	2.93
homo (eV)	-4.95
lumo (eV)	-1.68
rate (1/μs)	0.10
sascore	3.7
strength	0.0019
Key	XRQFUYUPULKGAT

Key SEXPVHDJOKKETP

Key

Nicknames, weight (amu), splitting (eV), absorption (eV), homo (eV), lumo (eV), rate (1/μs), sascore, strength, Key, Calc Time, Color Est. (n

Showing 1 to 100 of 106 entries

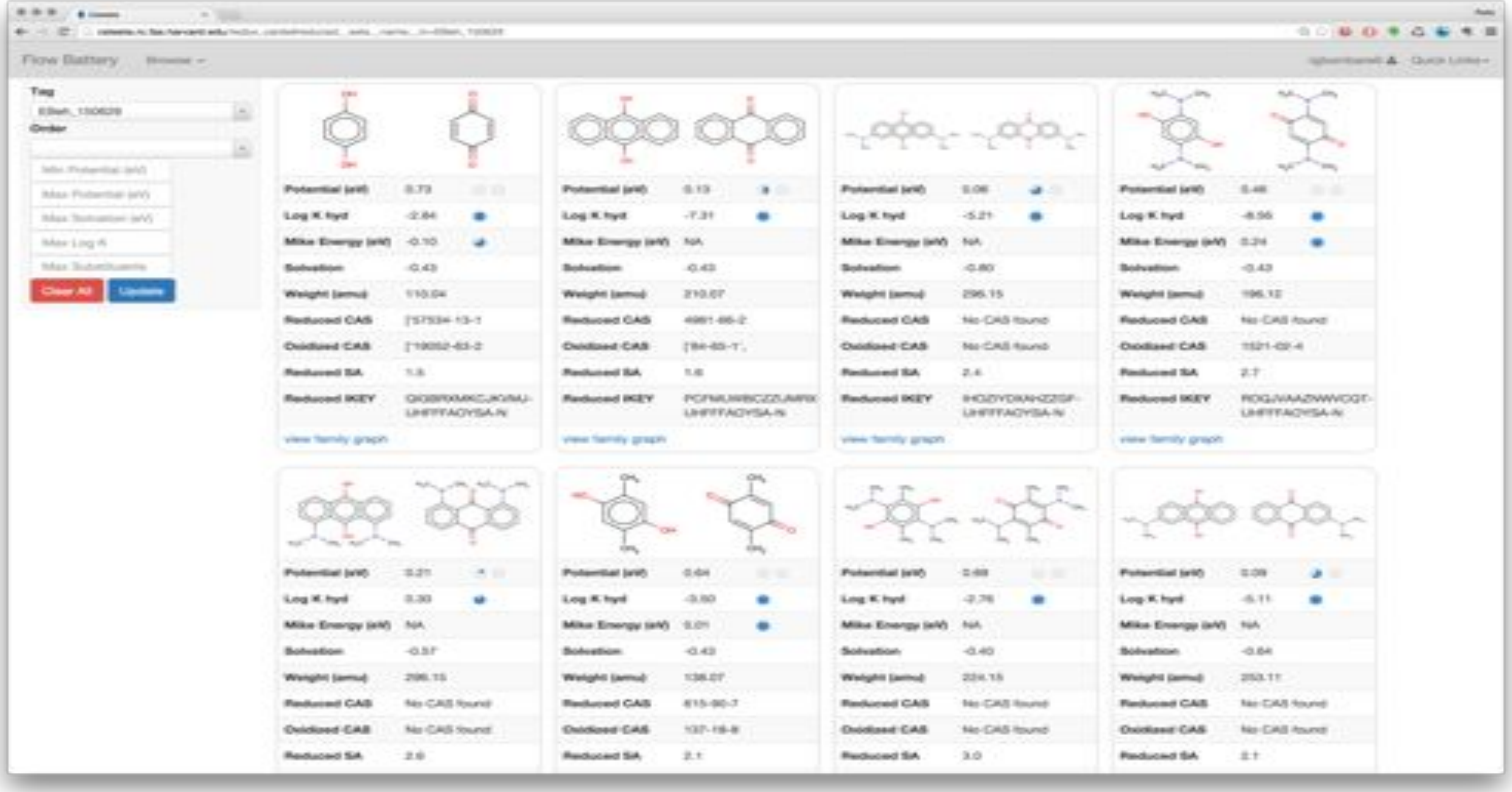
Nicknames | Structure | Donor | Bridge1 | Bridge2

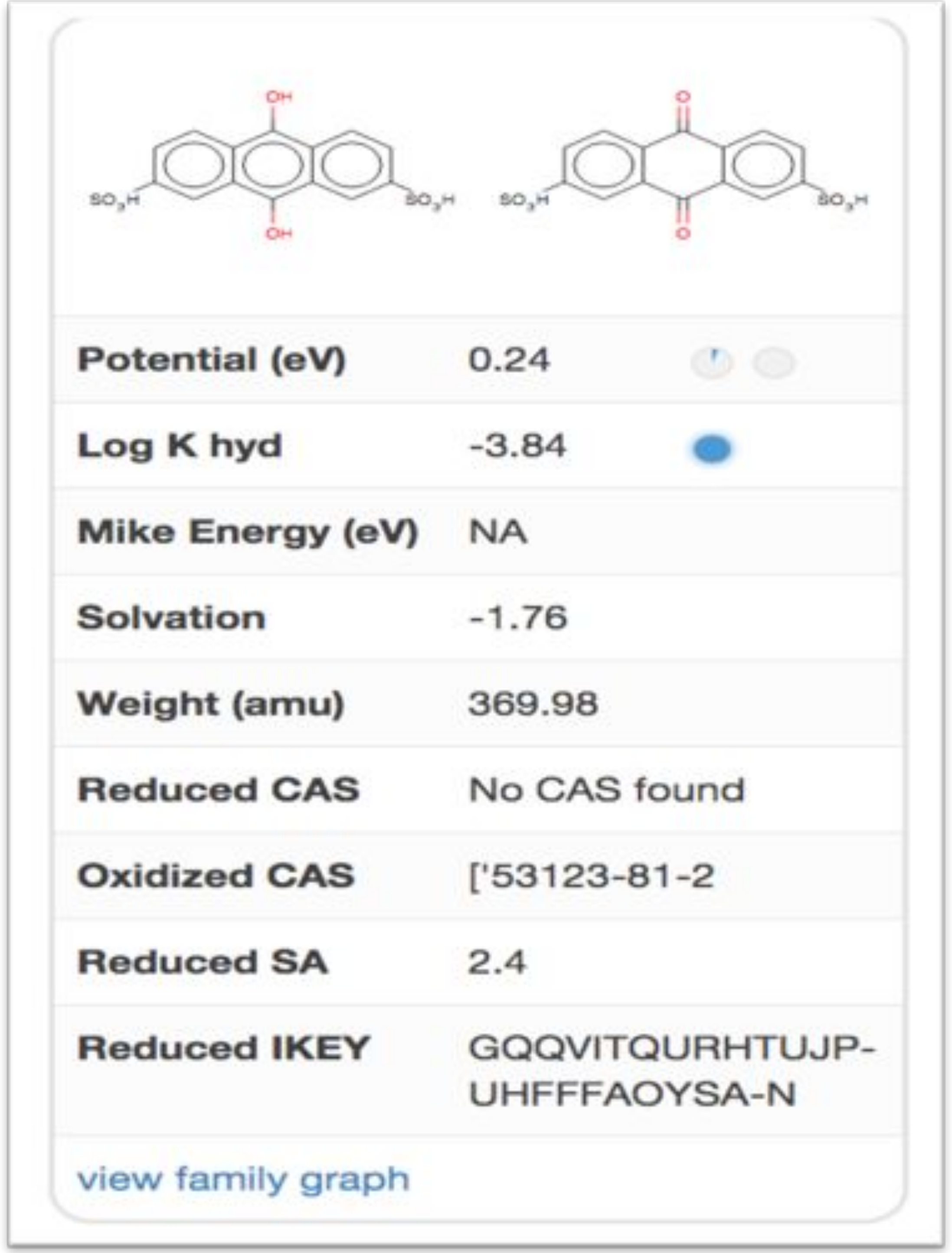
Web tools and critically enable partner communication and successful molecular discovery

Feedback tool

Database-backed web system tracks:

- ~1,000,000 machine-generated molecules
- ~1,500 (8000 including oxidation, decomposition and dissociation products)

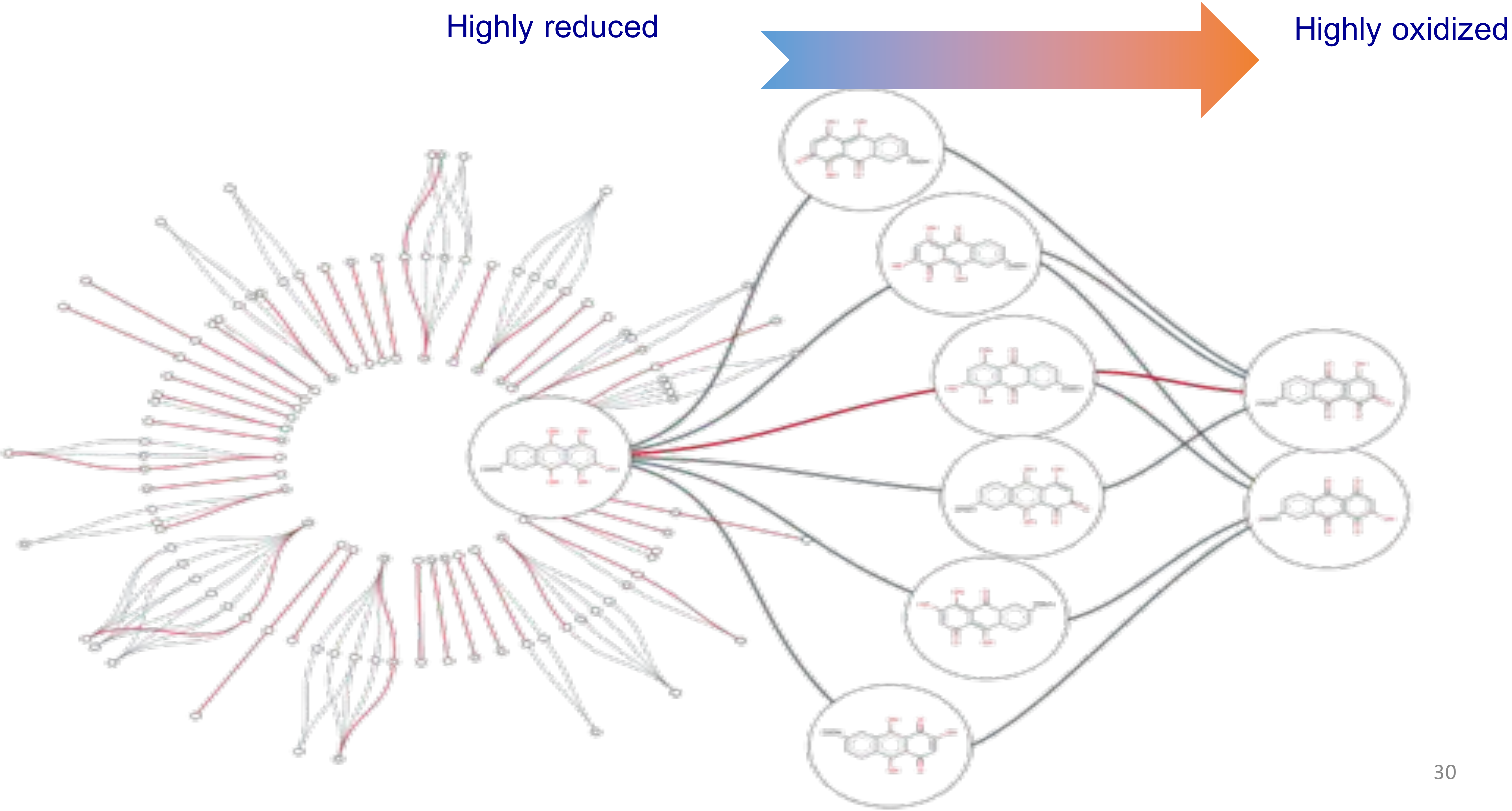




Potential (eV)	0.24
Log K hyd	-3.84
Mike Energy (eV)	NA
Solvation	-1.76
Weight (amu)	369.98
Reduced CAS	No CAS found
Oxidized CAS	['53123-81-2
Reduced SA	2.4
Reduced IKEY	GQQVITQURHTUJP-UHFFFAOYSA-N

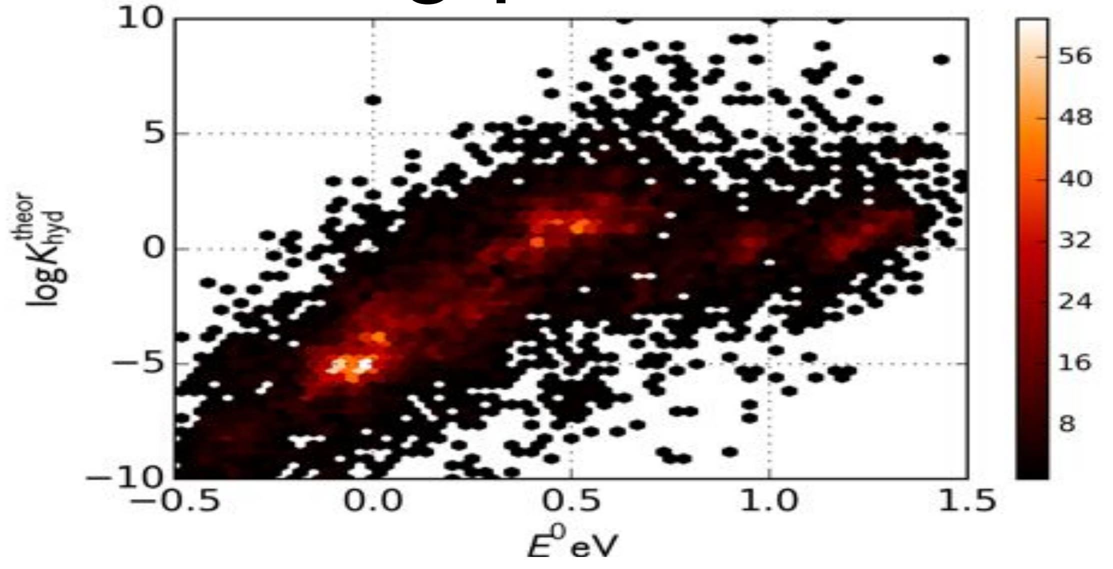
view family graph

Complex quinone redox pathways

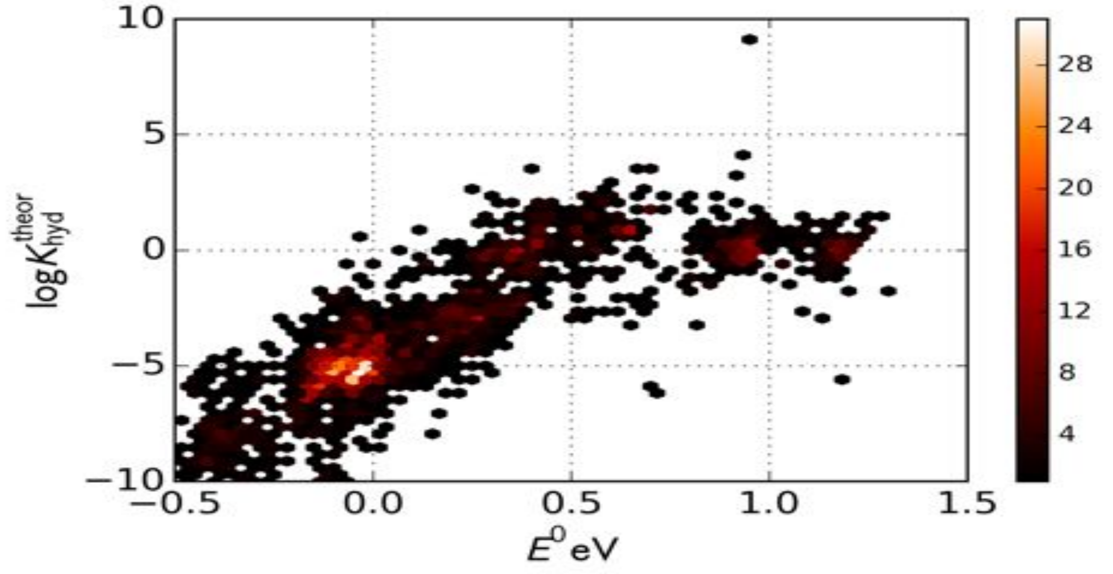


Additional current-theory work 1: quinone stability

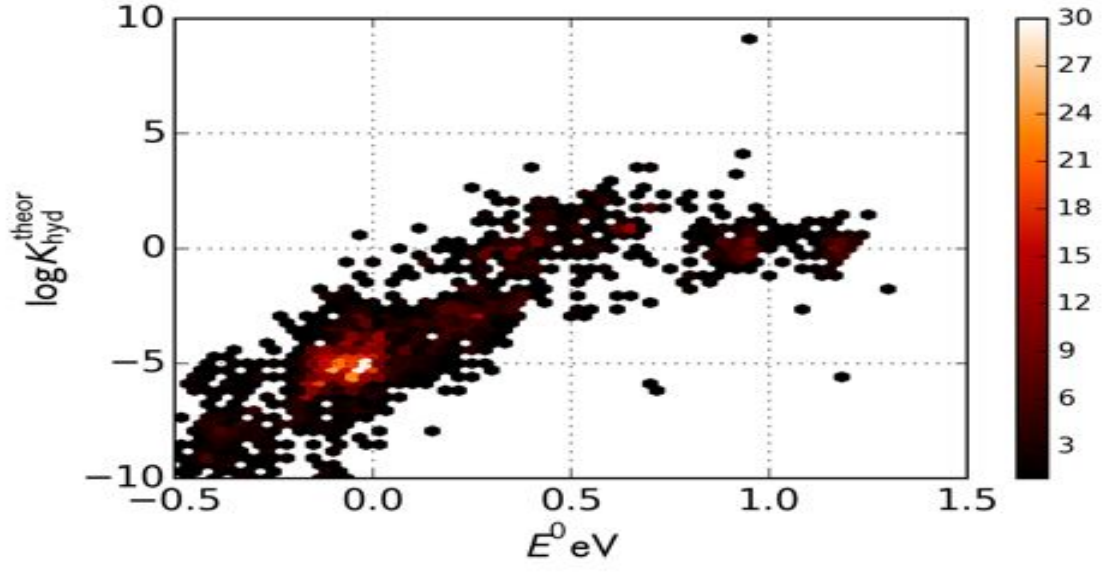
Screening procedures excluding potential Michael addition



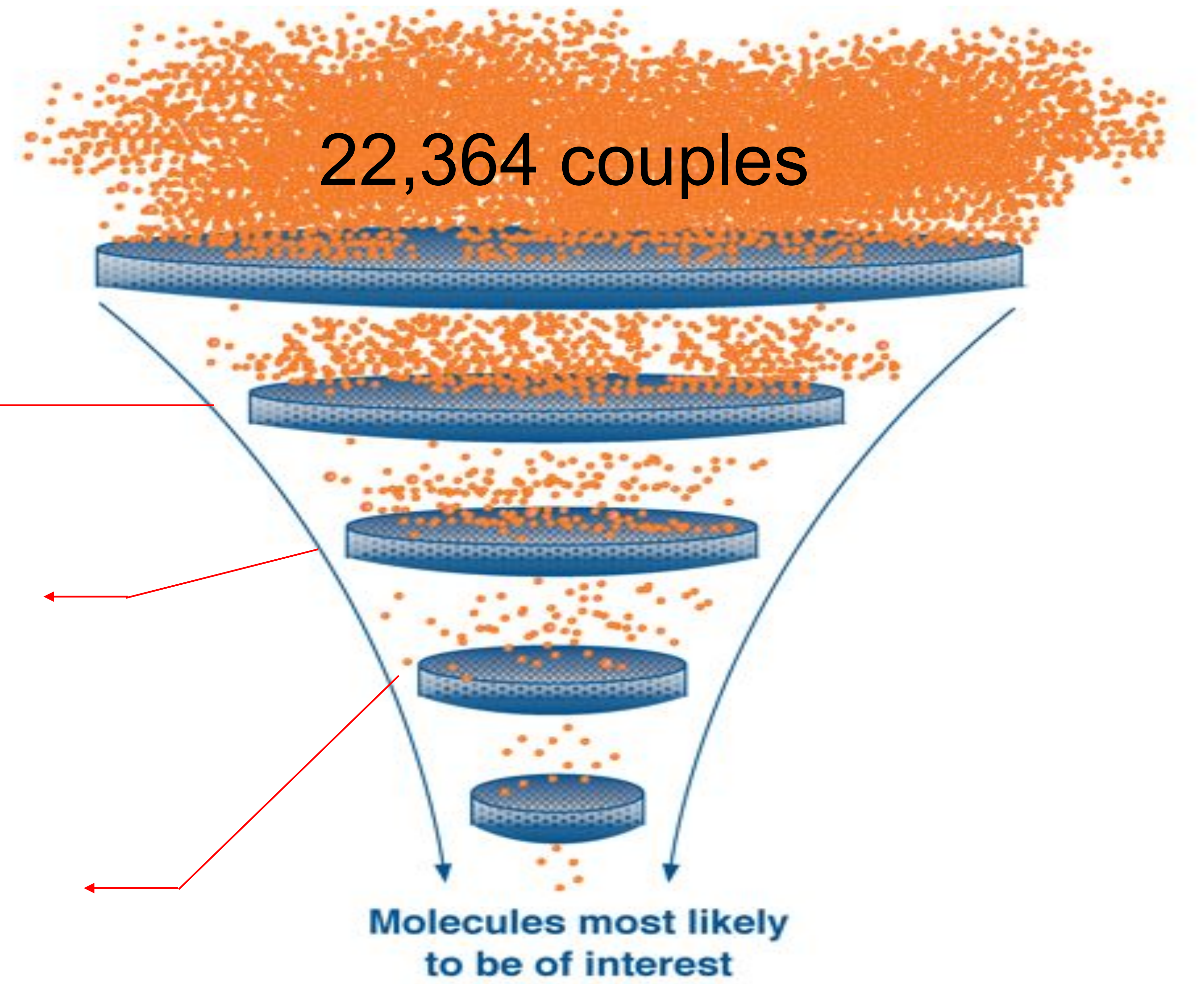
1st screening with Michael addition
9,866 couples



2nd screening:
Fewer than two R-groups
2,290 couples



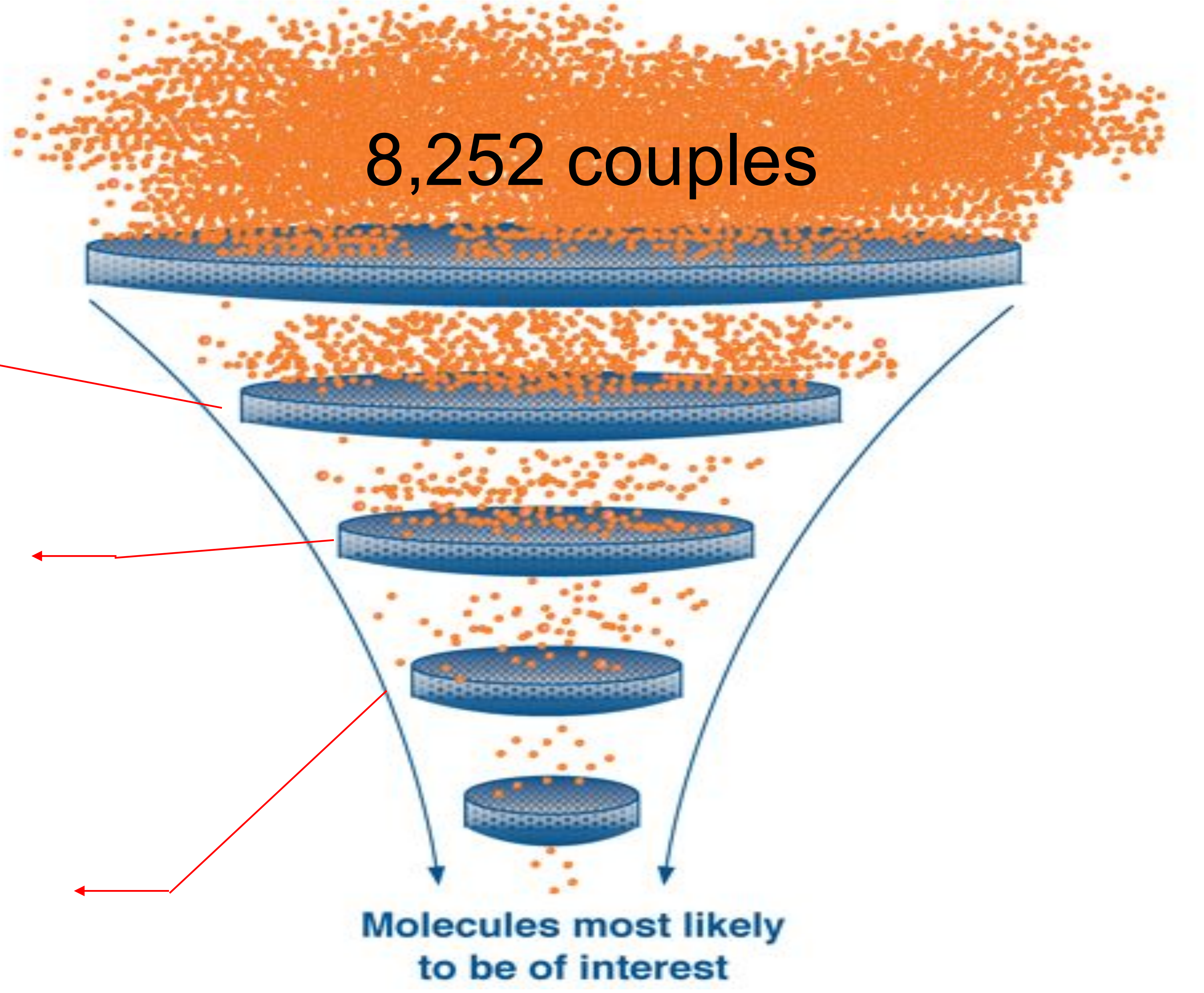
3rd screening:
Fewer than two R-groups
and good solubility
($\Delta G^0_{solv} < -0.75$ eV)
2,052 couples

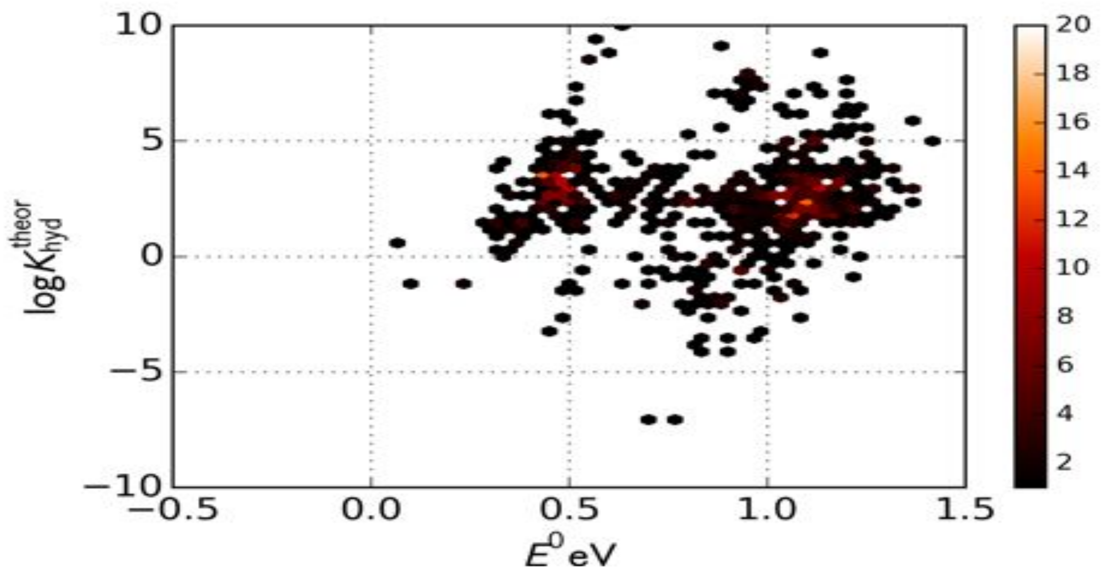


- *X-axis: Quinone redox potential (E^0_1) / Y-axis: Stability (K_{hyd})*
- Warmer colors represent higher density of molecules

Additional current-theory work 2: Second-oxidation quinones

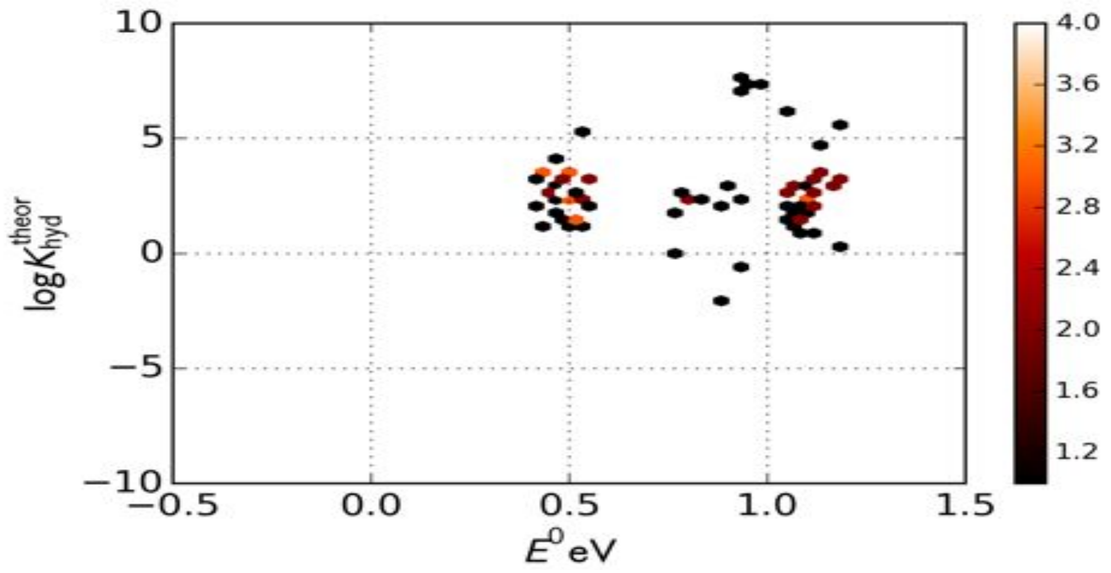
Screening Procedures with consideration of Michael addition


8,252 couples



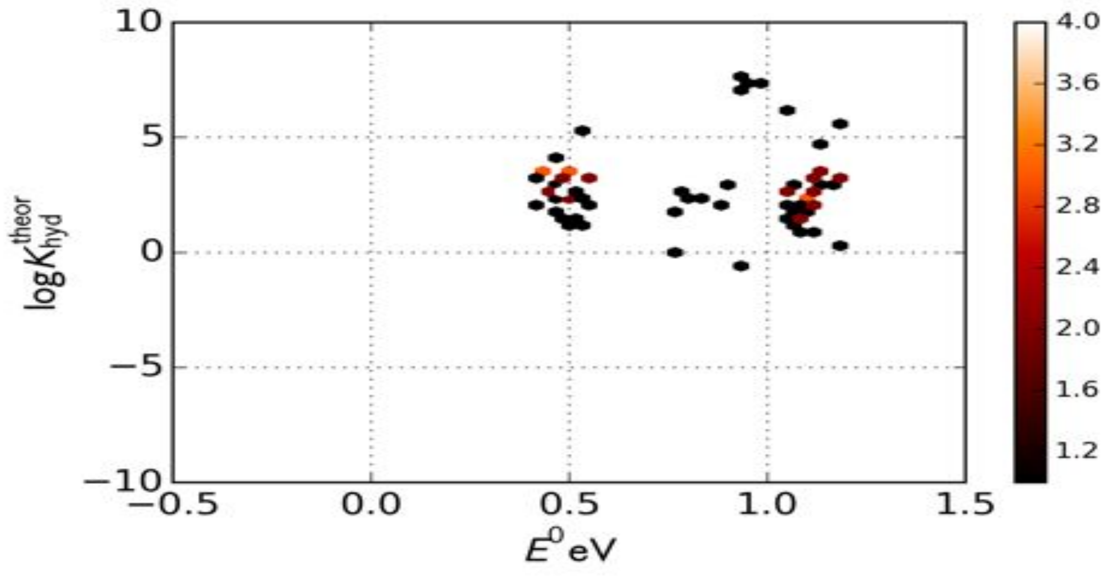
1st screening with Michael addition

762 couples



2nd screening: Fewer than two R-groups

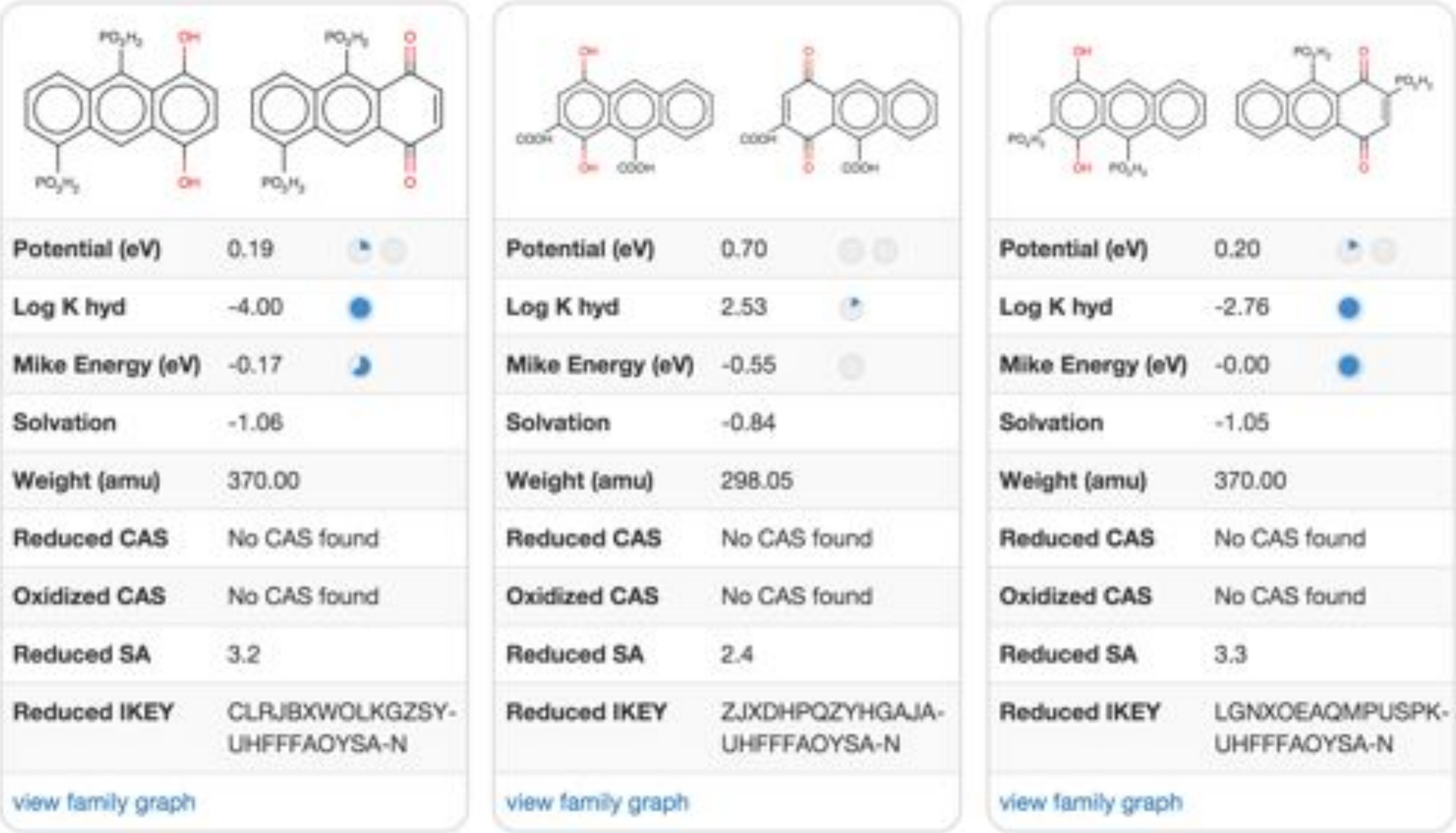
98 couples



3rd screening: Fewer than two R-groups and good solubility (ΔG^0_{solv} < -0.75 eV)

84 couples

Molecular baseball cards including stability



Potential (eV)	0.19	◔ ○
Log K hyd	-4.00	●
Mike Energy (eV)	-0.17	◕
Solvation	-1.06	
Weight (amu)	370.00	
Reduced CAS	No CAS found	
Oxidized CAS	No CAS found	
Reduced SA	3.2	
Reduced IKEY	CLRJBXWOLKGZSY-UHFFFAOYSA-N	

view family graph

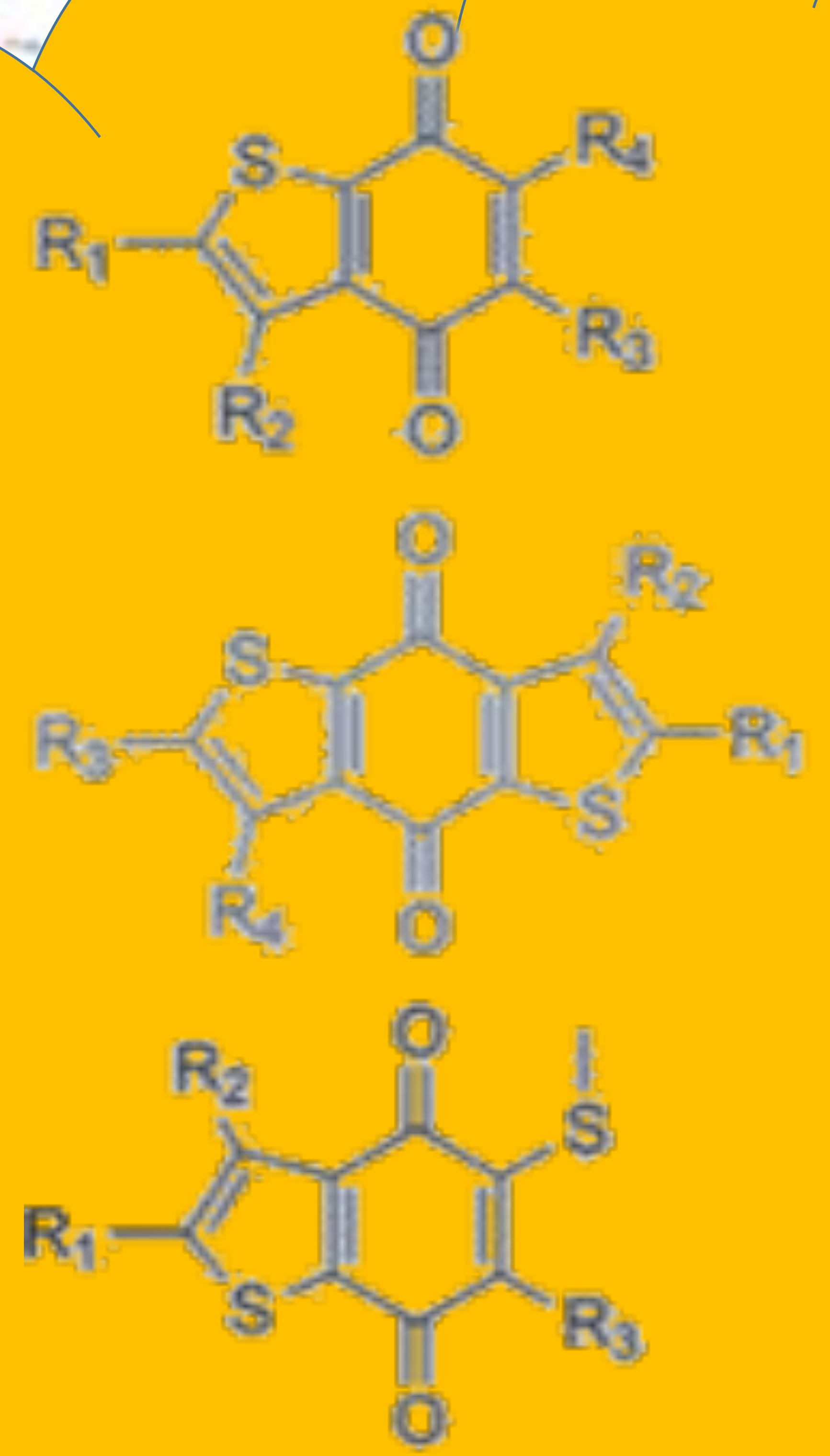
Potential (eV)	0.70	○ ○
Log K hyd	2.53	◔
Mike Energy (eV)	-0.55	○
Solvation	-0.84	
Weight (amu)	298.05	
Reduced CAS	No CAS found	
Oxidized CAS	No CAS found	
Reduced SA	2.4	
Reduced IKEY	ZJXDHPQZYHGAJA-UHFFFAOYSA-N	

view family graph

Potential (eV)	0.20	◔ ○
Log K hyd	-2.76	●
Mike Energy (eV)	-0.00	●
Solvation	-1.05	
Weight (amu)	370.00	
Reduced CAS	No CAS found	
Oxidized CAS	No CAS found	
Reduced SA	3.3	
Reduced IKEY	LGNXOEAQMPUSPK-UHFFFAOYSA-N	

view family graph

Beyond quinones

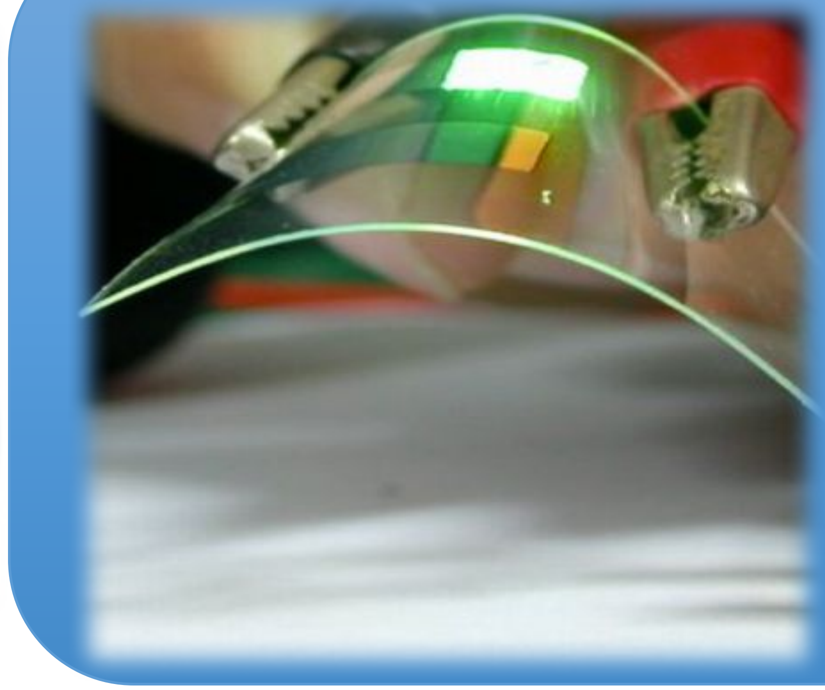


Sulfolobus archaebacteria

Pineda-Flores, et al. J. Phys. Chem. C 119 21800 (2015)

Long-lasting blue organic LED

The Harvard Clean Energy Project
Generating renewable energy



Blue Organic LED
For your next gadget or TV

Organic flow batteries
Storing renewable energy

Origins of life
How life may have come about?

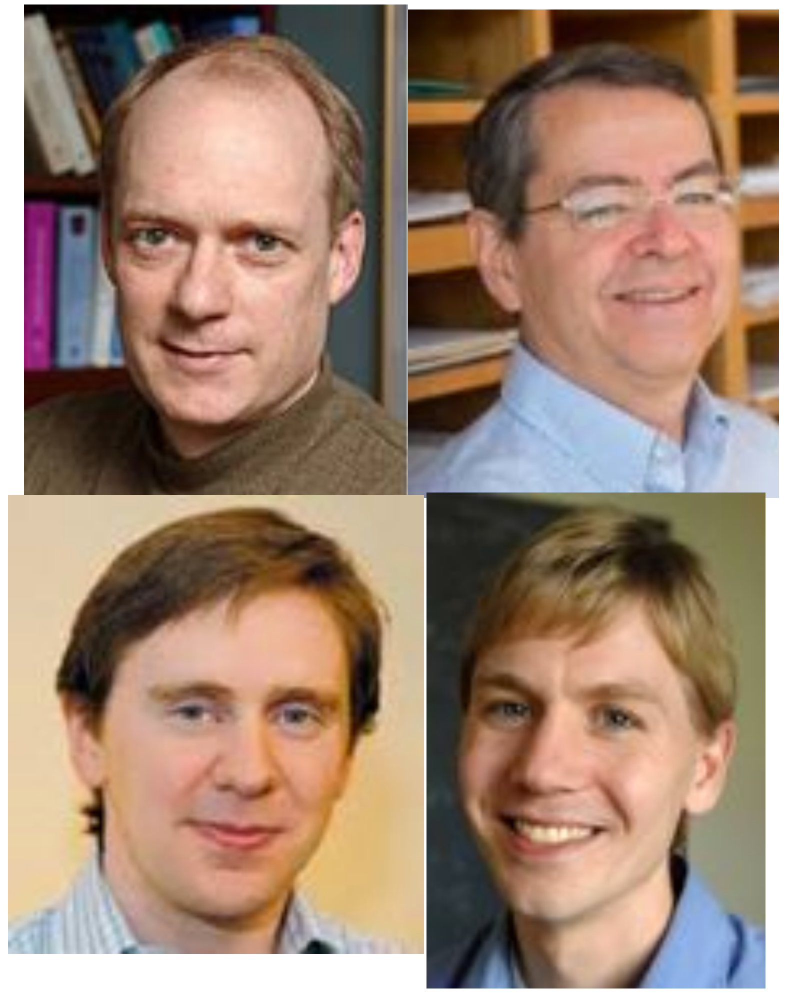
Harvard-MIT collaboration

Harvard

Ryan Adams
Machine Learning

Alán Aspuru-Guzik
High-throughput quantum chemistry

MIT



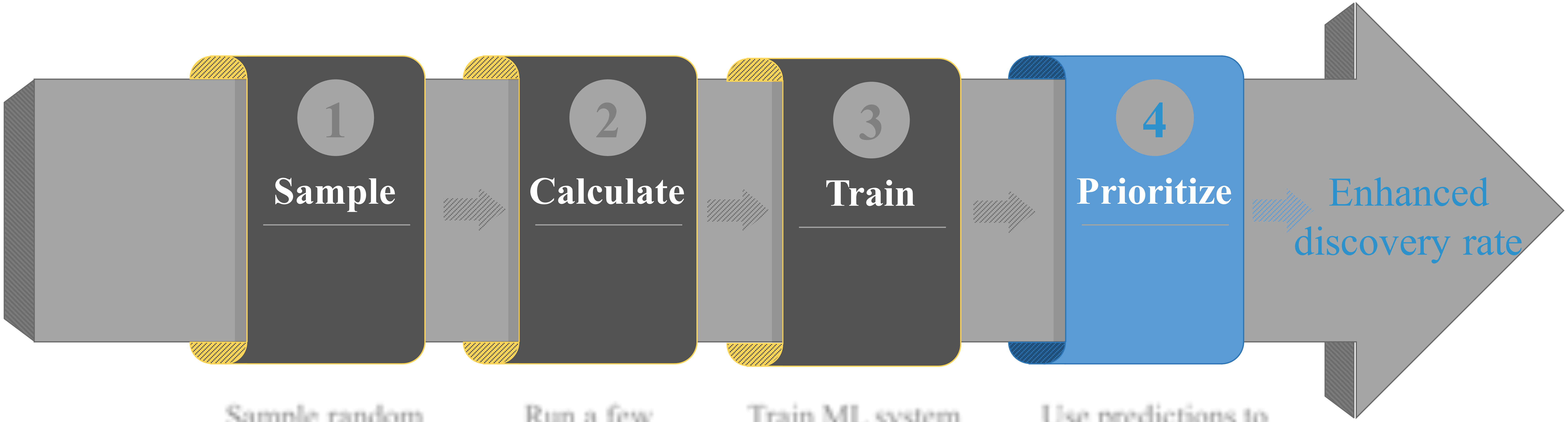
Tim Swager
Stephen Buchwald
Synthetic Chemistry

Marc Baldo
Device Engineering

Troy Van Voorhis
Microscopic theory

>450,000 molecules screened so far! ~25 synthesized and tested

Speedy screening: Machine Learning



1 Sample

2 Calculate

3 Train

4 Prioritize

Enhanced discovery rate

Sample random molecules in the DB

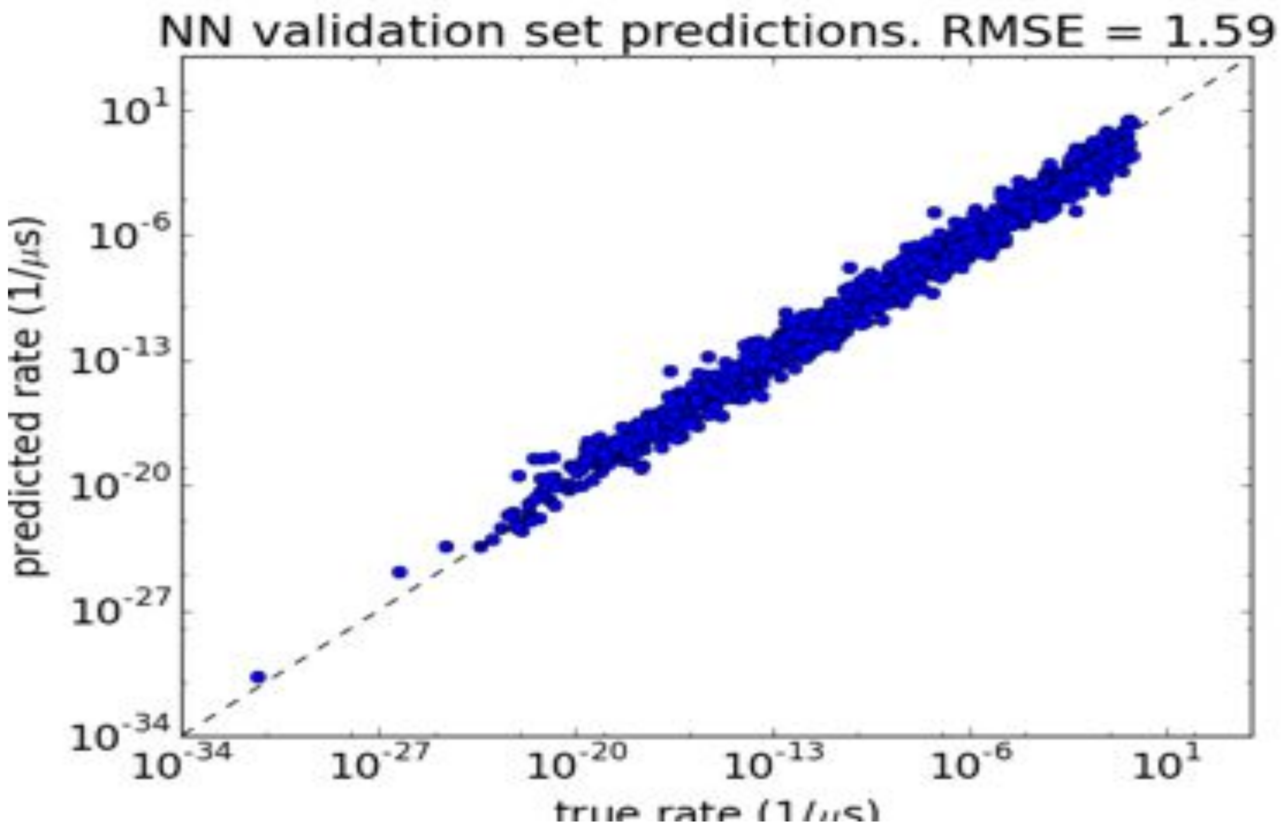
Run a few thousand calculations on those molecules

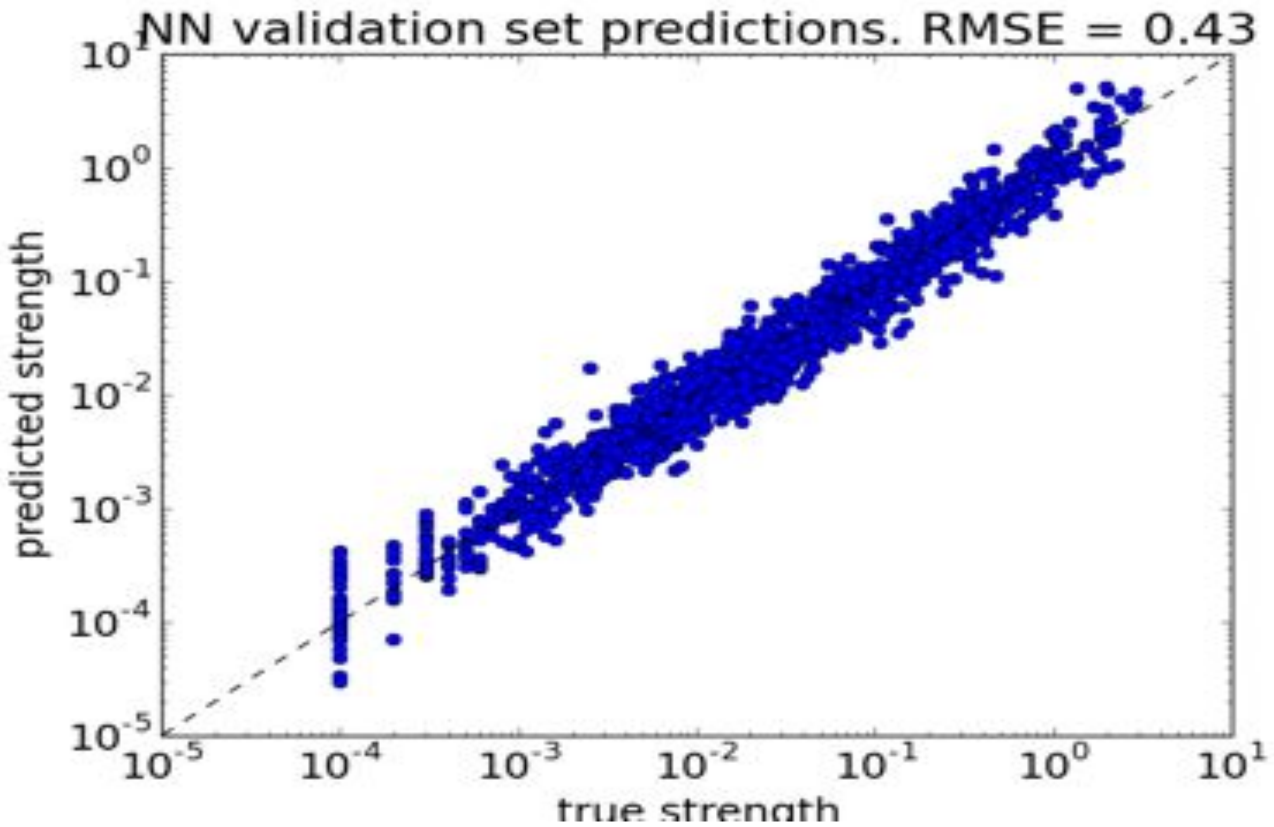
Train ML system on calcualtion results

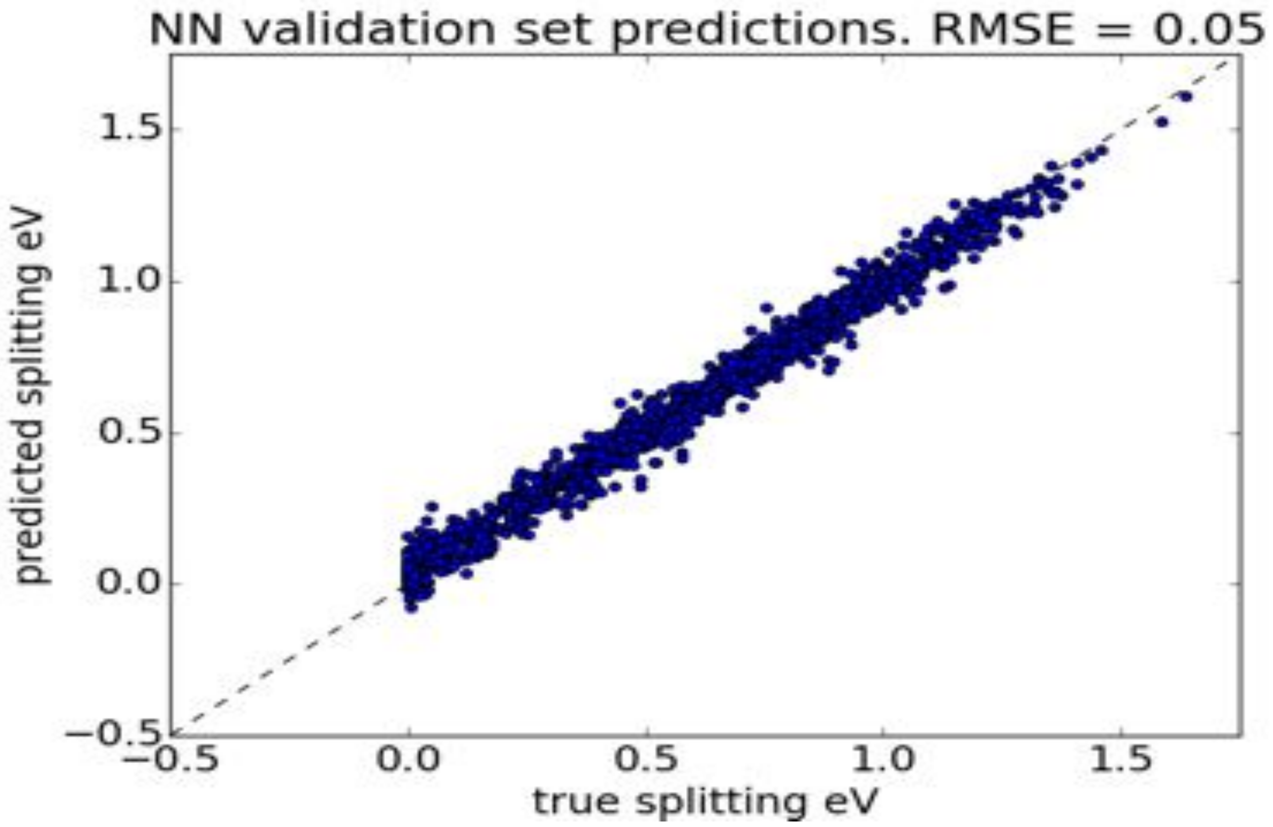
Use predictions to rank candidates and prioritized the calculations

Machine Learning

- Supervised learning algorithms
  - Neural networks for ultrafast predictions leveraging thousands of data-points.
  - Result in 10x speedup by discarding poor candidates
- Role of dimensionality
  - Chemical space is sparse but libraries are dense. Powerful interpolation
- Explore-exploit strategy


NN validation set predictions. RMSE = 1.59


NN validation set predictions. RMSE = 0.43

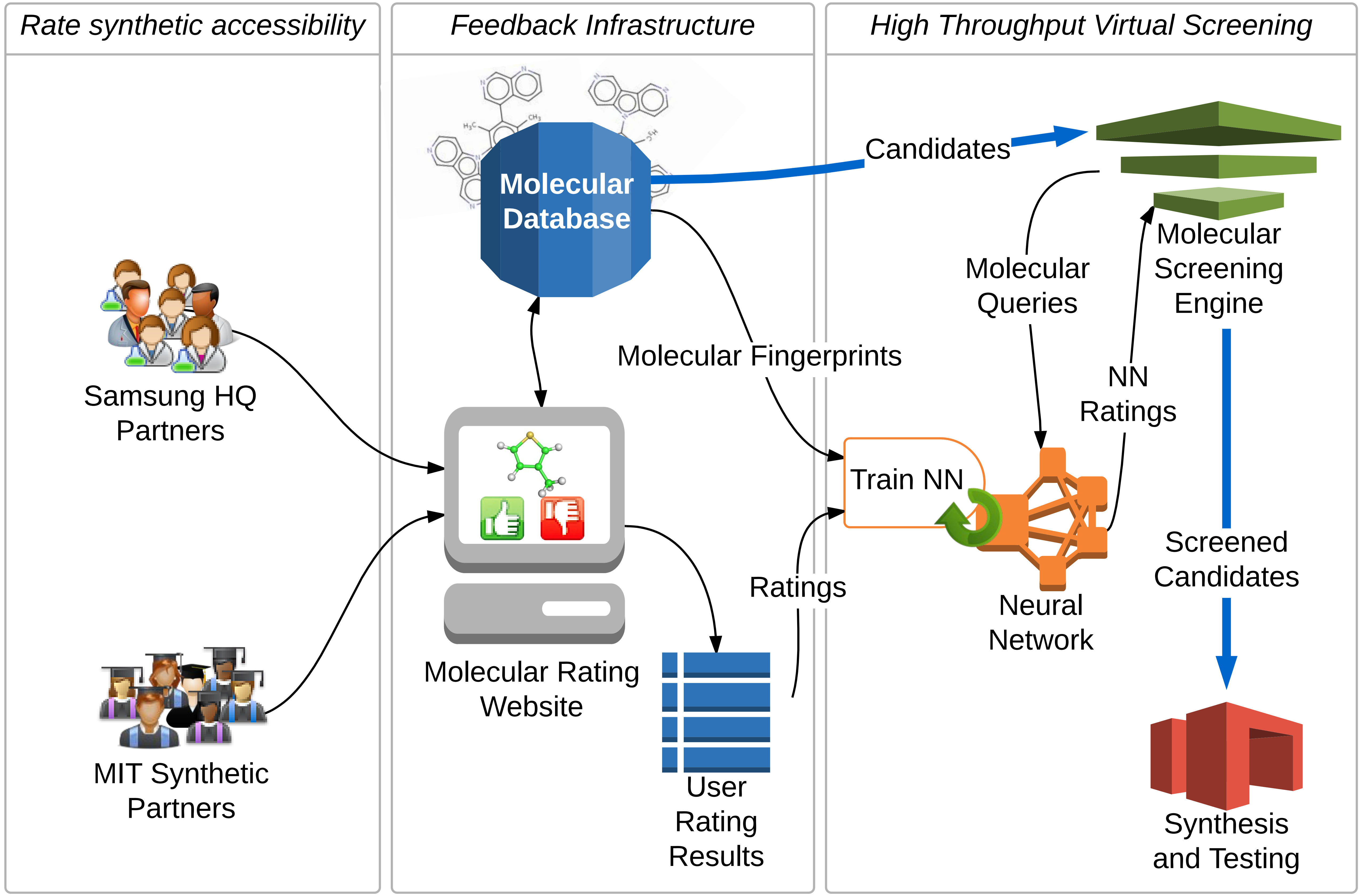

NN validation set predictions. RMSE = 0.05

Selecting molecules is like dating.

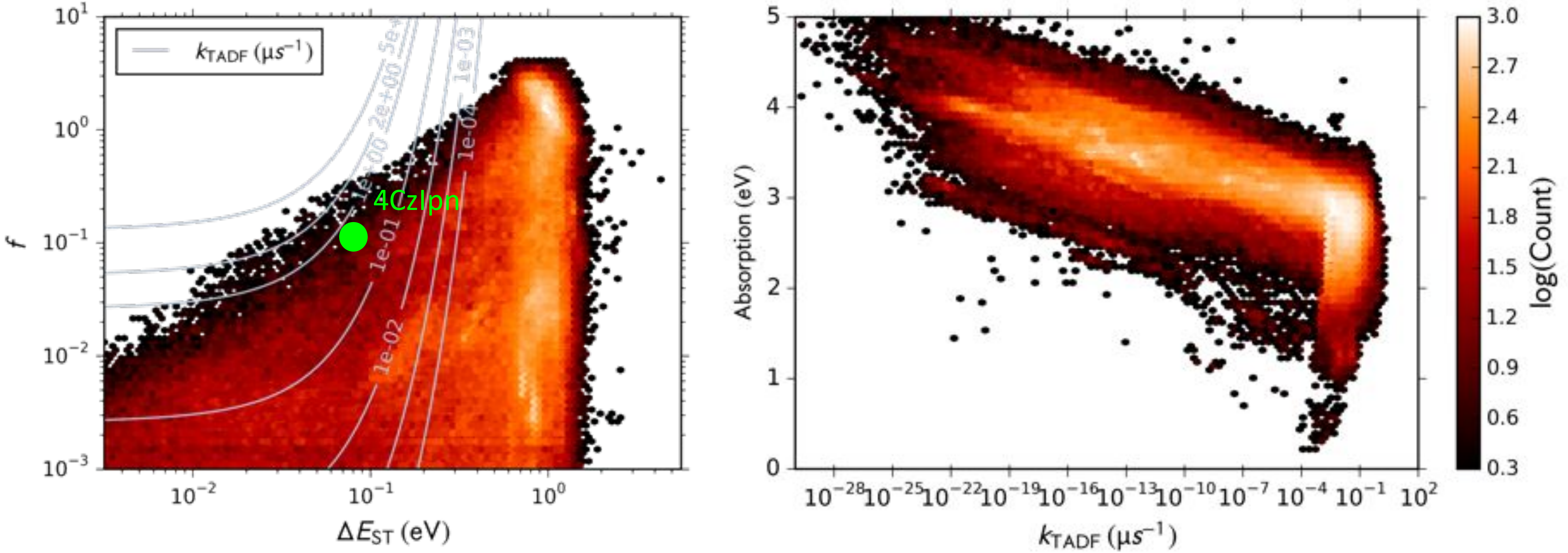
Organic LED Screening

Synthetic accessibility voting tool

Neural Net Training Workflow



Data mining 500,000 quantum calculations

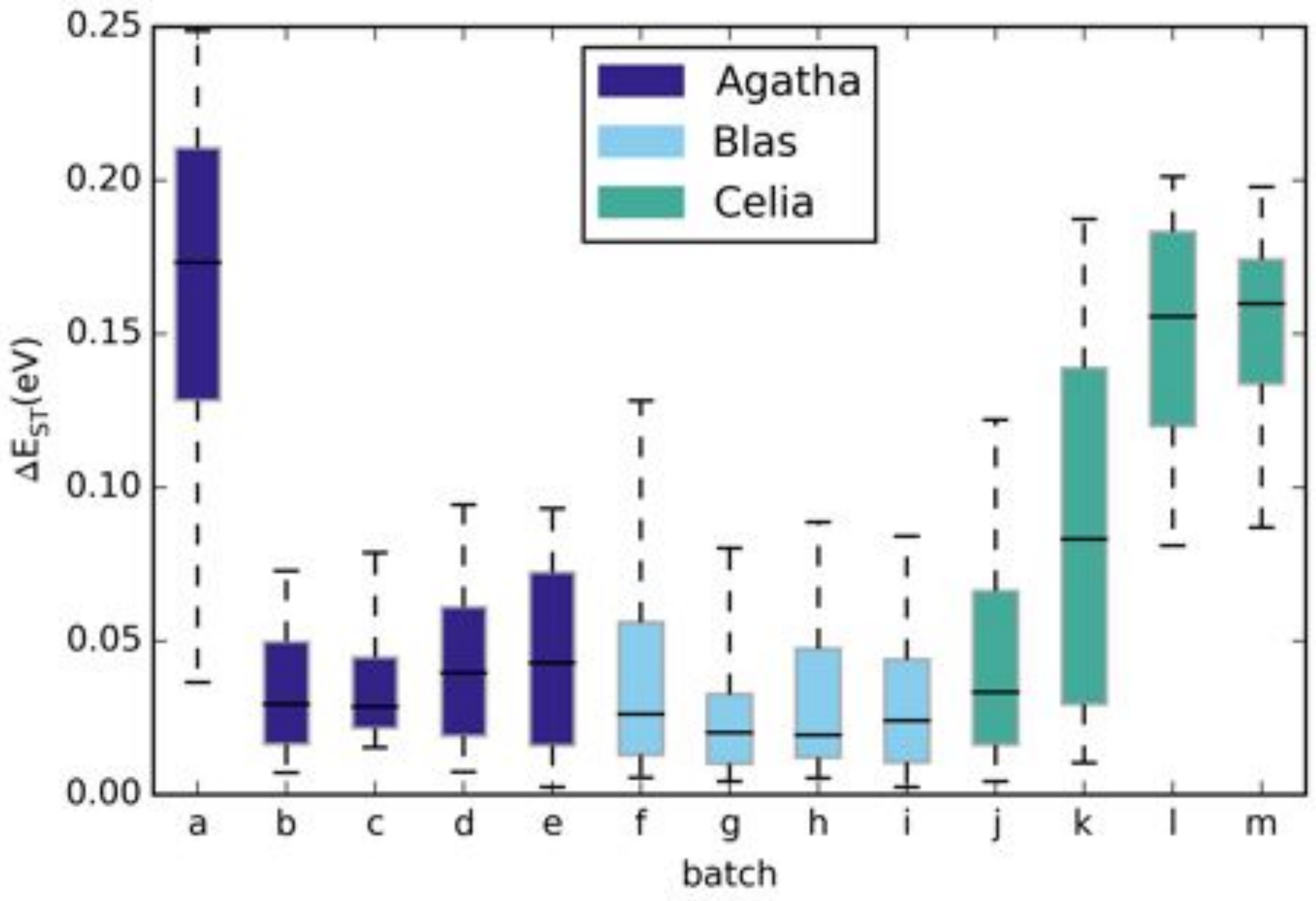


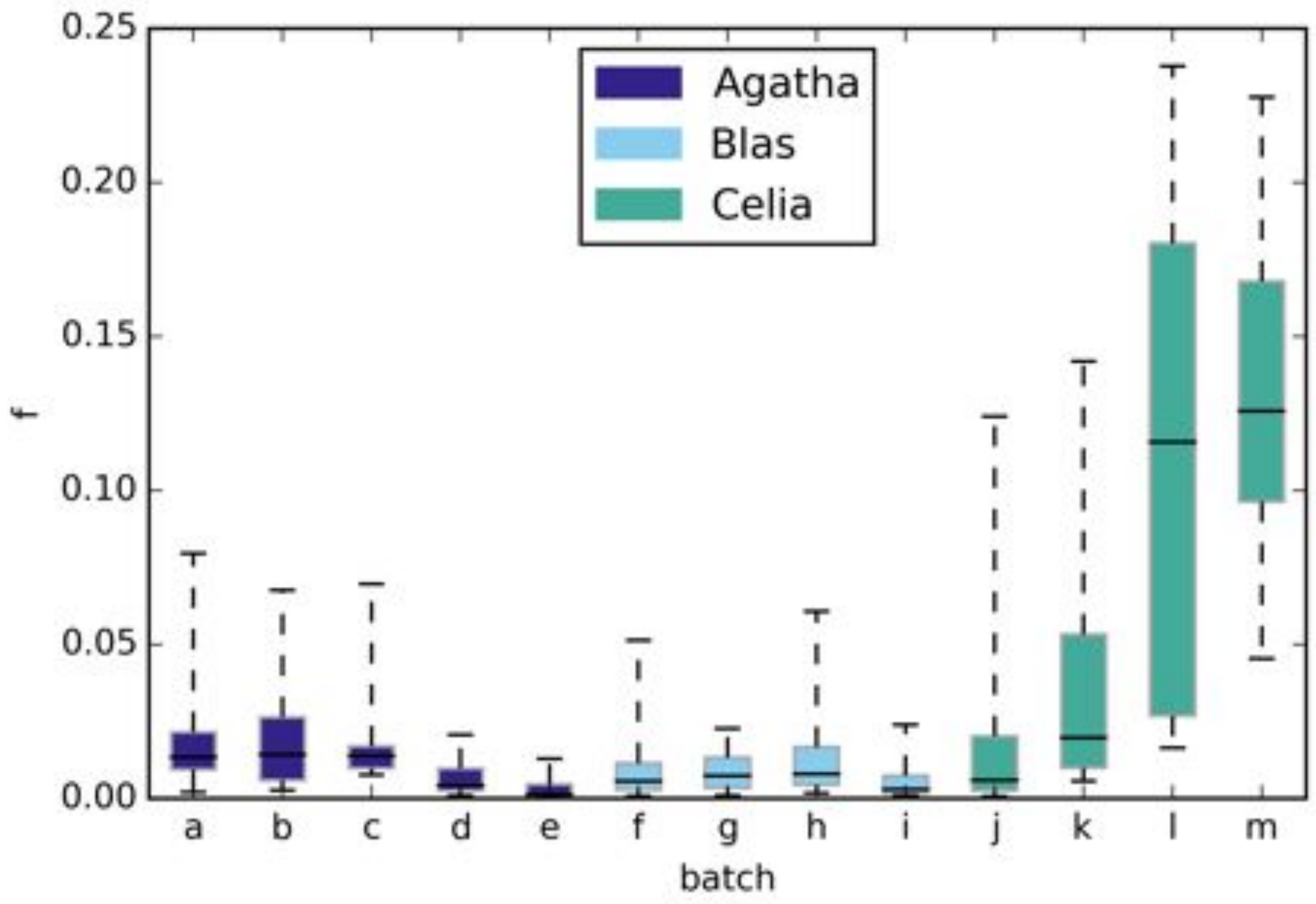
R. Gomez-Bombarelli, et al. Submitted (2015)

Batches

Batches

Selection of 100-200 molecules for experimentalists to browse in a contained way.
Usually explore some chemical *family*, using ancestry from database.
Need to confirm novelty *post hoc*: sometimes re-discover known molecules.





22.5%

Rafael Gómez-Bombarelli, Jorge Aguilera-Iparaguirre, Tim Hirzel
Martin BloodAdams, Baldo, Swager groups, Samsung IT

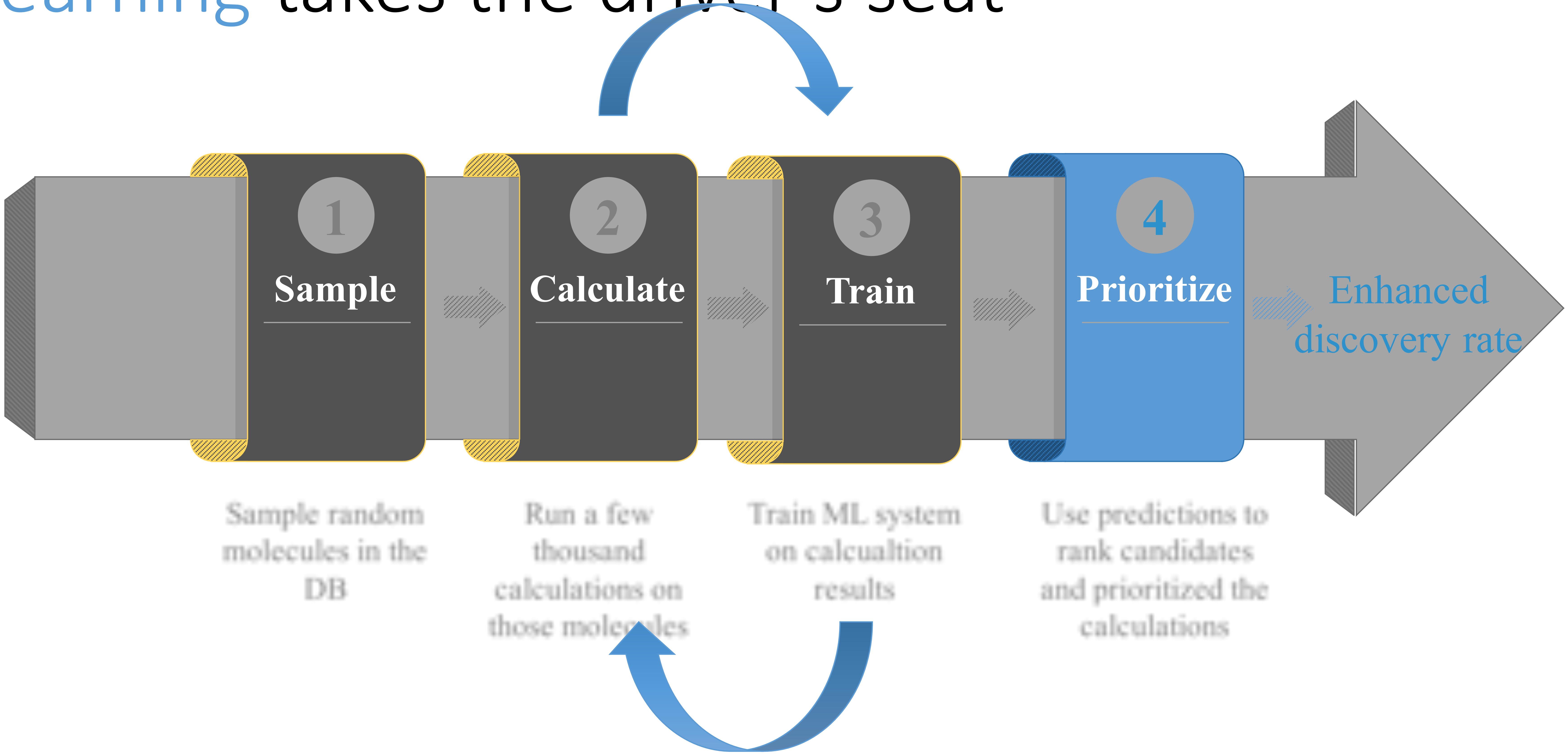
R. Gomez-Bombarelli, et al. Submitted (2015)

Key breakthroughs in efficiency: Strength

Name	S_0 splitting	T_1 splitting	S_0 strength	T_1 strength	EQE(%)
4CzIPN	0.124	0.101	0.063	0.049	20
Foxtrot1-21	0.015	0.031	0.003	0.000	20
Hotel1-38	0.017	0.046	0.008	0.012	7
Julie2-16-1	0.104	0.145	0.124	0.186	22
Lima17-36	0.179	0.187	0.257	0.240	17*

- A small gap is crucial for TADF behavior
- We need also need a big fluorescence
- We have managed to control both for great overall efficiency

Screening billions of molecules: Machine learning takes the driver’s seat



Sample random molecules in the DB

Run a few thousand calculations on those molecules

Train ML system on calcualtion results

Use predictions to rank candidates and prioritized the calculations

To design something really well you have to get it. You have to really grok what it's all about. It takes a passionate commitment to really thoroughly understand something. Chew it up, not quickly swallow it. Most people don't take time to do that.

Aspuru-Guzik group

http://aspuru.chem.harvard.edu
Twitter: A_Aspuru_Guzik
aspuru@chemistry.harvard.edu

Sponsors: **DOE BES, ARPA-E, Samsung, NSF**, ARO, ONR, AFOSR, Samsung, Sloan Foundation, Camille and Henry Dreyfus Foundation, DTRA, DARPA