Exploring chemical space: Molecular Space Shuttle

Alán Aspuru-Guzik
Professor
Harvard University

http://aspuru.chem.harvard.edu
Twitter: A_Asputru_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

How good is this molecule as a battery material?

Quantum Mechanics

Machine Learning
From $10^{60}$ to $10^6$ to $10$...

Initial library

Computational screening

Synthesis and testing

Computational cost

Molecules most likely to be of interest
Organic materials

in the larger context

US Materials Genome Initiative

Shared Features

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

Organic Materials

Inorganic Materials

Organic Pharmaceuticals

- Number of descriptors
- Size of search space
- Level of approximation

My research group’s explorations of chemical space

**The Harvard Clean Energy Project**
Generating renewable energy

**Organic flow batteries**
Storing renewable energy

**Blue Organic LED**
For your next gadget or TV

**Origins of life**
How life may have come about?
# Project chronology and screening methodology improvements

<table>
<thead>
<tr>
<th>Version</th>
<th>Project Description</th>
<th>Timeline</th>
<th>Key Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Harvard Clean Energy Project</td>
<td>2006-2014</td>
<td>Distributed high-throughput + Experimental calibration + MolecularSpace database</td>
</tr>
<tr>
<td>2.0</td>
<td>Organic flow batteries</td>
<td>2013-2014</td>
<td>Tight experimental feedback cycle + Stability screening + Synthesizability at forefront</td>
</tr>
<tr>
<td>4.0</td>
<td>Harvard Clean Energy Project</td>
<td>2014-2014</td>
<td>Thompson sampling + Several experimental collaborations at once + Molecular crystal ML and prediction</td>
</tr>
</tbody>
</table>
Harvard Clean Energy Project: Organic solar cells
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 ...

Energy Levels and Efficiency

Sifting through 2.3 million molecules

~35000 molecules
(1.5% of sample space)

Energy and Environmental Science, 7, 698 (2014)
The Clean Energy Project gets an artificial intelligence boost!

Machine learning

Calculate

Learn

Prioritize

Bayesian calibration

Smart Screening Using machine learning

Easy to synthesize libraries

Neural Fingerprints

Collaborator: Ryan Adams (Harvard)

Organic Flow Batteries

The Harvard Clean Energy Project
Generating renewable energy

Blue Organic LED
For your next gadget or TV

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

1,4-Benzooquinones

1. N(CH₃)₂
2. NH₂
3. OCH₃
4. OH
5. SH
6. CH₃
7. SiH₃
8. F
9. Cl
10. C₂H₃
11. CHO
12. COOCH₃
13. CF₃
14. CN
15. COOH
16. PO₃H₂
17. SO₃H
18. NO₂

1,2-Benzooquinones
We have finished calculations of ΔHs for RPs. Naphtoquinones
<table>
<thead>
<tr>
<th>Anthraquinones</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,10-anthra</td>
</tr>
<tr>
<td>1,2-anthra</td>
</tr>
<tr>
<td>1,4-anthra</td>
</tr>
<tr>
<td>1,10-anthra</td>
</tr>
<tr>
<td>2,9-anthra</td>
</tr>
<tr>
<td>1,5-anthra</td>
</tr>
<tr>
<td>1,7-anthra</td>
</tr>
<tr>
<td>2,3-anthra</td>
</tr>
<tr>
<td>2,6-anthra</td>
</tr>
<tr>
<td>1,10-anthra</td>
</tr>
<tr>
<td>1,5-anthra</td>
</tr>
<tr>
<td>1,7-anthra</td>
</tr>
<tr>
<td>2,3-anthra</td>
</tr>
<tr>
<td>2,6-anthra</td>
</tr>
<tr>
<td>1,10-anthra</td>
</tr>
<tr>
<td>1,5-anthra</td>
</tr>
<tr>
<td>1,7-anthra</td>
</tr>
<tr>
<td>2,3-anthra</td>
</tr>
<tr>
<td>2,6-anthra</td>
</tr>
</tbody>
</table>
Theoretical calibration of quinone redox potentials

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

R²: 0.989
RMSD: 0.046 V
> 300 new candidate quinones predicted

S. Er., C. Suh, M. P. Marshak, A. Aspuru-Guzik,
Chemical Science (2015)
Our metal-free aqueous flow battery

Computational screening of 10,000 quinone molecules

Intense design cycle

Select molecule

Synthesize molecules
Test in flow battery
Theory-experiment collaboration

Michael Aziz
Engineering

Roy Gordon
Chemistry

Alán Aspuru-Guzik
Chemistry

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
Moelcular Space Shuttle: advanced molecular discovery platform

High-throughput materials discovery process and tools

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)
Complex quinone redox pathways

Highly reduced

Highly oxidized
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 ($\Delta G^0_{\text{solv}} < -0.75 \text{ eV}$): 2,052 couples

- X-axis: Quinone redox potential ($E^{0}_{1}$) / Y-axis: Stability ($K_{\text{hyd}}$)
- Warmer colors represent higher density of molecules
Additional current-theory work 2: Second-oxidation quinones

Screening Procedures with consideration of Michael addition

1\textsuperscript{st} screening with Michael addition
- 762 couples

2\textsuperscript{nd} screening:
- Fewer than two R-groups
- 98 couples

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

Molecules most likely to be of interest

8,252 couples
Molecular baseball cards including stability
Beyond quinones

Sulfolobus archaebacteria

Long-lasting blue organic LED

The Harvard Clean Energy Project
Generating renewable energy

Organic flow batteries
Storing renewable energy

Blue Organic LED
For your next gadget or TV

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: Sample random molecules in the DB
2. Calculate: Run a few thousand calculations on those molecules
3. Train: Train ML system on calculation results
4. Prioritize: Use predictions to rank candidates and prioritize the calculations

Enhanced discovery rate
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
Selecting molecules is like dating.
Organic LED Screening

Synthetic accessibility voting tool
Neural Net Training Workflow
Data mining 500,000 quantum calculations

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.
Sample Baseball card

Not the true structure

High performance devices

Device 16.7% EQE!!

Synthesis of compounds

Molecule

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

22.5%
Key breakthroughs in efficiency: Strength

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

- 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

1. Sample
   Sample random molecules in the DB

2. Calculate
   Run a few thousand calculations on those molecules

3. Train
   Train ML system on calculation results

4. Prioritize
   Use predictions to rank candidates and prioritize the calculations

Enhanced discovery rate
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.