Combined Evolutionary Optimization & Machine Learning

Efficient discovery of Polymer Materials

Chemistry

Prof. Geoffrey Hutchison

Departments of Chemistry & Chemical Engineering
University of Pittsburgh

Artificial Intelligence for Materials Chemistry

March 23, 2020

<http://hutchison.chem.pitt.edu>

Funding

DOE DE-SC0019335

Students / Postdoc

Dr. Marcus Hanwell	Steve Owens
Aaron Crandall	Dr. Chris Marvin
Dr. Xialing Chen	Dr. Ilana Kanal
Dr. Xinfeng Quan	Dr. Nate Miller
Dr. Tamika Madison	Chris Petroff
Dr. Adam Gagorik	Anneliese Schmidt
Dr. Andrey Sharapov	Dakota Folmsbee
Dr. Paula Hoffmann	Danielle Heiner

Undergrads

Arthur Davis	Matt McCain
Casey Campbell	Jonathan Bechtel
Alexandra McSorley	Shannon Davies
Evan Richards	Leandro Gil-Silva
Leah Seebald	Eric LeCato
Bryan Siegfried	Benjamin Frank
Terry Paske	Isabel Foreman-Ortiz
Valerie Lentz	Luke Langkamp

How do you find the “right” molecule?

**10^{60} Possible
Molecules!**

Composition Space for π Co-Polymers

**Overall Search Space:
>20-100 billion compounds**

Monomers

**Polymerization
Sites**

X

+

**Composition
(2, 3, 4?)**

X

+

**Monomer Sequence
(ABAB vs. ABBA)**

**Sequence Length
(AABBABABA...)**

Organic Heterojunction Solar Cells

Two key parameters

- Optical absorption (band gap)
 - Energy level alignment for charge transfer
- Parameters do not tell us which molecule to synthesize ... it's an "inverse design" problem**

Statistics for Tetramers: Brute Force

Genetic Algorithm Approach

Efficient discrete optimization strategy

Represent search space as a set of genes

Balance of **exploration** and **exploitation**

Exploration: basin hopping / mutation / diversity

Exploitation: local optimization (crossover)

Genetic Algorithm Approach

1. Evaluate 64 candidates at once

2. Good monomers rise, bad monomers fall

3. Add weighting for "similar" monomers

4. Mutate: Generate candidates with new monomers

Repeat

GA Results: Hundreds of Efficient Targets

What Do We Learn?

How are current polymers designed?

GA "Discovers" New Strategies

Global Color Scales	Minimum	Mid-Point	Maximum
HOMO Scale (in eV)	-10.72	-8.54	-6.94
LUMO Scale (in eV)	-6.60	-2.03	-0.39

**Complex
Sequences!**

Efficiency	HOMO	LUMO	HOMO Barcode (energies in eV)					
11.02	-5.75	-4.31	-7.89	-7.98	-7.89	-7.98	-7.98	-7.89
10.59	-5.88	-4.29	-7.98	-8.35	-7.98	-8.35	-8.35	-7.98
10.58	-5.83	-4.28	-7.98	-8.24	-8.24	-7.98	-8.24	-7.98
10.18	-6.03	-4.29	-7.98	-8.32	-8.32	-7.98	-8.32	-7.98
10.10	-6.06	-4.30	-7.29	-7.60	-7.60	-7.29	-7.29	-7.60
10.10	-5.98	-4.27	-7.98	-7.34	-7.34	-7.98	-7.34	-7.98
10.09	-6.01	-4.28	-7.98	-8.27	-8.27	-7.98	-8.27	-7.98
10.02	-5.74	-4.23	-8.35	-7.98	-7.98	-8.35	-7.98	-8.35
10.01	-5.74	-4.28	-7.98	-7.48	-7.48	-7.98	-7.48	-7.98
10.01	-6.06	-4.29	-7.03	-7.54	-7.03	-7.54	-7.54	-7.03

**Donor-donor
pairs?!**

How Does GA Scale to Large Sets?

Up to 8000x faster
Up to 25M compounds

Across Multiple Sets

Other Properties of Interest

- **Photovoltaics**
Polymer Acceptors, Multi-Chromaphores
- Static Permittivity - **Polarizability**
- **Electromechanical Coupling**
Marcus Reorganization Energy
Piezoelectric Coefficients
- Low-lying triplet states (Singlet fission?)
- New ideas, collaborations welcome...

**Evolutionary optimization rapidly finds
new molecular targets...**

Where Does Machine Learning Help?

- **Property predictions**
Fast Surrogate for Electronic Structure
- **Clustering** (what trends/properties in top sets?)
 - Property space
 - Chemical structure space
- **Accelerated Convergence**
- And GA helps ML: generate targeted, diverse training, test sets, "hard" molecules

Clausius–Mossotti Relationship

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N}{V} \frac{\alpha}{3\epsilon_0}$$

**We can connect molecular properties (α , μ)
to bulk relative permittivity**

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N}{V} \frac{1}{3\epsilon_0} \left(\alpha + \frac{\mu^2}{3k_b T} \right)$$

ML Polarizability Predictions

Polarizability - already good accuracy on QM9 set
But look at the scale... $\sim 2\text{\AA}^3$.. really small!

New Approach: Wave Convolutions

New ML model
Harder test set

Collaboration with
S. Joshua Swamidass
(WUSTL Pathology & Immunology)

Train:

Test:

Evolving New Targets?

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N}{V} \frac{1}{3\epsilon_0} \left(\alpha + \frac{\mu^2}{3k_bT} \right)$$

Debye Eqn = Ideal high α , μ
Small volume

GA never "hits" both
high α and μ

Poster: Danielle Hiener

"Genetic algorithm method for rapid screening ..."

Evolving New Targets?

GA can be tuned to explore high μ or balanced or high α

favoring μ

"equal" weights

favoring α

Polarizability preprint coming soon...

Poster: Danielle Hiener

"Genetic algorithm method for rapid screening ..."

Take-Home Messages

Evolutionary / Genetic algorithms work well

- **Efficient discrete optimization of molecules**
 - Many properties to explore...
- **Can work with ML methods**
 - Train-on-the-fly as DFT surrogates
 - Finding diverse / hard training sets
(e.g., high polarizability, high dipole, etc.)

Any Questions?