Science Use Case 3

Data driven materials discovery for dye sensitized solar cells

ATPESC
Aug 9, 2019

Álvaro V Mayagoitia
Argonne CPS
ALCF Acknowledgement
This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.
My Summer project

“…make any problem a machine learning problem..”
4th Paradigm
Data-Intensive Scientific Discovery

History of Science, summarized by Jim Cray

Experiments

\[
\left(\frac{a}{a}\right)^2 = \frac{4\pi G \rho}{3} - K \frac{c^2}{a^2}
\]

1600 1950 2000

The 4th Paradigm
Data-Intensive Scientific Discovery

Chemical Compound Space

*Estimated Energetically Possible Organic Molecules

$>10^{60}$

*Nature, Insights, 2004
Chemical Compound Space

*Estimated Energetically Possible Organic Molecules >10^{60}

*Nature, Insights, 2004

Total number of water molecules in Earth: 10^{40}

Recent estimations say could exceed 10^{180}
Chemical Compound Space

*Estimated Energetically Possible Organic Molecules $>10^{60}$

Total number of water molecules in Earth: $10^{40}$

How much do we know of the chemical space?

Compiled from experiments since the early 1800s: $10^8$

Compilation:

- Computationally: eg. Harvard Clean Energy project
  - $10^7$ Molecules
  - $10^7$ CPU hrs
  - $10^9$ Calculations

https://cepdb.molecularspace.org/
Substances in CAS registry (ACS)

10M of new substances per year in average

155M in Aug 2019

Are they any useful?
**Materials Science Modeling**

DFT publications in the last 35 years

- All DFT
- Substance related DFT

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

*web of science queries*

Computational Science Division
Evolution of DNA sequencing

June 2019: trillion bases a day
Renewable sources of energy

Solar
10^5 TW at earth surface
10,000 TW tech. value

Wind
14 TW

Energy needed
2007: 15 TW
2017: 18 TW
2050: ~30 TW
2100: ~50 TW

Biomass
5-7 TW

Geothermal
1.9 TW

Tide/Ocean
0.7 TW

The Third Industrial Revolution by Jeremy Rifkin, 2011
IRENA report 2018
Cost per watt-hour of Solar energy

1977  $76.67  \uparrow
2019  $0.10  \downarrow
Share of total US energy consumption by end-use sectors 2018

- Comercial: 18%
- Residencial: 20%
- Industrial: 33%
- Transportation: 29%
Carbon footprint per sector 2018

- Residential, commercial, and industry: 26%
- Transportation: 34%
- Electric power: 40%
Artificial Satellites

SF bus shelters

Electric golf carts

Food courts
Solar windows

Harris Theatre – Chicago [north Millenium park]
Solar windows

SwissTech Convention Center EPFL
Solar windows
dye-sensitized SOlar cell

Operational Mechanism

A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO$_2$ films

Brian O'Regan* & Michael Grätzel†
Institute of Physical Chemistry, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland

1991 Conversion efficiency: 7.9%
DYE-SENSITIZED SOLAR CELL

- Emerging technology
- Cost effective (good price-to-performance ratio) 😊
- Less efficient than Si-base cells 😞

Organometallic cells
- Nazeeruddin et al JACS 1993 (ruthenium-based) N719: 10.4% eff.
- Yella et al Science 2011 (Zn-porphyrin-based): 12.3 % eff.

Organic cells
- Daeneke et al Nat. chem. 2011 (carbazole-base): 7.5% eff.
ENCODING STRUCTURE-FUNCTION

Screening with TDDFT is costly

Rules:

\[ \varepsilon_{LUMO} > E_{CT\text{TiO}_2} \quad \varepsilon_{HOMO} < E_{Electrolyte} \]
MAXIMIZING LIGHT HARVESTING EFFICIENCY

\[ AM = \frac{L}{L_0} \approx \frac{1}{\cos z} \]

- \( L_0 \) = zenith path length
- \( L \) = path length to the atmosphere
- \( z \) = zenith angle

https://en.wikipedia.org/wiki/Air_mass_(solar_energy)
MAXIMIZING LIGHT HARVESTING EFFICIENCY

Maximizing light harvesting efficiency
Work flow

Experimental Data

• 30 years of Academic Literature

Data extraction

• Composition
• Properties
• Spectra

Theoretical Studies

• Geometry
• Physical properties
Funnel approach
Screening the chemical space

Funnel / Filters

- Size, spectra, charges
- Optoelectronics rules
- Semi-empirical Methods
- Density Functional Methods
- “Gold” standard methods

From 100 kilo molecules, which ones could be good dyes?
<table>
<thead>
<tr>
<th>Compound</th>
<th>Lamda max</th>
<th>Ext. Coef.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triphenylmethane</td>
<td>480</td>
<td>4320</td>
</tr>
<tr>
<td>Indigo</td>
<td>613</td>
<td>320</td>
</tr>
<tr>
<td>anthraquinone</td>
<td>320</td>
<td>-----</td>
</tr>
</tbody>
</table>

How long would it take a person to get this information?
A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

http://chemdataextractor.org

Figure 2 shows the UV-vis absorption spectra of 3a (red) and 3b (blue) in acetonitrile. The peak at...

**NN** 2 shows the **VBZ** absorption **NN** spectra of 3a (red) and 3b (blue) in **CC** acetonitrile.

**NN** 2 shows the **DT** UV-vis absorption **NN** spectra of 3a (red) and 3b (blue) in **IN** acetonitrile.

**NN** 2 shows the **VBZ** absorption **NN** spectra of 3a (red) and 3b (blue) in **CM** acetonitrile.

- **Type**: UV-vis absorption
- **Of**: 3a red, 3b blue
- **In**: acetonitrile

**3a** → 2-[2-[4-(dimethylamino)phenyl]diazenyl]-benzoic acid

**3b** → 2-[2-[4-(dipropylamino)phenyl]diazenyl]-benzoic acid
The dye 2-[2-[4-(dimethylamino)phenyl]diazenyl]-benzoic acid (3a) was added...

UV-vis spectra were recorded using an Agilent8453 diode array spectrophotometer.

```json
{
    "name": "2-[2-[4-(dimethylamino)phenyl]diazenyl]-benzoic acid",
    "label": "3a",
    "uvvis": [ {
        "solvent": "acetonitrile",
        "apparatus": "Agilent8453 diode array spectrophotometer",
        "peaks": [ {
            "wavelength": "448",
            "extinction": "29,000"
        } ]
    } ]
}
```

Figure 2: UV-vis absorption spectra of 3a in acetonitrile.

<table>
<thead>
<tr>
<th>Dye</th>
<th>$\lambda_{\text{max}}$ (nm)</th>
<th>$\varepsilon$ (M$^{-1}$ cm$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td>448</td>
<td>29,000</td>
</tr>
<tr>
<td>3b</td>
<td>415</td>
<td>48,000</td>
</tr>
</tbody>
</table>
Informed high throughput computing

- **Global minima**
- **2D to 3D**

**Force Fields**
- Conformers
- Optimization
- Vibrations

**Semiempiricals**
- Geometry
- Excitations
- Multipoles

**DFT**
- Excitations
- Multipoles

**Coupled Cluster**
- Excitations
- Multipoles

Composite of codes:
- Babel
- Rdkit
- MOPAC
- ORCA
- NWChem
<table>
<thead>
<tr>
<th>Sample name</th>
<th>AFM parameters</th>
<th>XRR parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singly sensitized working electrodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1 only</td>
<td>5 ± 1</td>
<td>7 ± 2</td>
</tr>
<tr>
<td>8c only</td>
<td>5 ± 1</td>
<td>6 ± 2</td>
</tr>
<tr>
<td>XS6 only</td>
<td>4.9 ± 0.4</td>
<td>6.0 ± 0.7</td>
</tr>
<tr>
<td>H3 only</td>
<td>9 ± 1</td>
<td>15 ± 3</td>
</tr>
<tr>
<td>15 only</td>
<td>8 ± 2</td>
<td>15 ± 3</td>
</tr>
<tr>
<td>Co-sensitized working electrodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1 then 15</td>
<td>6 ± 2</td>
<td>10 ± 3</td>
</tr>
<tr>
<td>C1 and 15</td>
<td>7 ± 2</td>
<td>12 ± 4</td>
</tr>
<tr>
<td>H3 then C1</td>
<td>8 ± 2</td>
<td>16 ± 4</td>
</tr>
<tr>
<td>C1 and H3</td>
<td>5 ± 1</td>
<td>8 ± 3</td>
</tr>
<tr>
<td>8c then 15</td>
<td>6 ± 1</td>
<td>9 ± 2</td>
</tr>
<tr>
<td>8c and 15</td>
<td>4.6 ± 0.3</td>
<td>5.8 ± 0.4</td>
</tr>
<tr>
<td>H3 then 8c</td>
<td>5.5 ± 0.7</td>
<td>8 ± 1</td>
</tr>
<tr>
<td>8c and H3</td>
<td>5.2 ± 0.7</td>
<td>7 ± 2</td>
</tr>
<tr>
<td>XS6 then 15</td>
<td>6 ± 1</td>
<td>8 ± 2</td>
</tr>
<tr>
<td>XS6 and 15</td>
<td>7.8 ± 0.7</td>
<td>11 ± 1</td>
</tr>
<tr>
<td>XS6 then H3</td>
<td>5.5 ± 0.7</td>
<td>7.6 ± 0.8</td>
</tr>
<tr>
<td>XS6 and H3</td>
<td>5.3 ± 0.8</td>
<td>7 ± 1</td>
</tr>
</tbody>
</table>
Over fitting

Under fitting

Prediction + uncertainty
Gaussian Process

Quick introduction

Target function

\[ y_i = f(x_i) + \epsilon_i \]

Noise function

\[ \epsilon_i = \begin{bmatrix} y \\ y_\ast \end{bmatrix} \sim N \left( 0, \begin{bmatrix} K & K^T \\ K^T & K_{**} \end{bmatrix} \right) \]

Given \( y \), the probability of \( y_\ast \) is:

\[ y_\ast | y \sim N(K_\ast K^{-1} y, K_{**} - K_\ast K^{-1} K^T) \]

Prediction (or kriging)

\[ \bar{y}_\ast = K_\ast K^{-1} y \]

Variation

\[ \text{var}(y_\ast) = K_{**} - K_\ast K^{-1} K^T \]

Here* Means a point that we want to predict
Gaussian process

Covariance matrix

Covariance function

\[ k(x, x') = \sigma_f^2 \exp \left[ \frac{-(x - x')^2}{2l^2} \right] \]

Example: Square exponential

Covariance matrix

\[
K = \begin{bmatrix}
    k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\
    k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\
    \vdots & \vdots & \ddots & \vdots \\
    k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n)
\end{bmatrix}
\]
Molecular Fingerprint examples

Morgan Circular Fingerprint
Learn from data and feedback to experiments

Transition prediction

TDDFT gap prediction – We used Gaussian Process and Circular Morgan Fingerprints to predict the first transition of the a reduce scale TDDFT (sTDA//wB97X-D3/TZVP), we found that this value is predictable. Similar result found for HOMO-LUMO DFT gap.
Learn from data and feedback to experiments
Is this optically active?

Oscillator Strength prediction – Transitions could not be optically active. We can predict which of electronic transitions have an oscillator strength < 0.8 a.u. with an error 3%.

<table>
<thead>
<tr>
<th>True condition</th>
<th>Predicted condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>71.5</td>
</tr>
<tr>
<td>Non-Active</td>
<td>28.5</td>
</tr>
<tr>
<td>Active</td>
<td>2.7</td>
</tr>
<tr>
<td>Non-Active</td>
<td>97.3</td>
</tr>
</tbody>
</table>

Support:
- 428 (100%)
- 186 (0%)
Future work

• Extinction coefficients prediction could be improved adding extra information that could be slightly costly to get, such as orbital dipole moments or results from lower scale methods.
• Variational autoencoders (VAE) could help us to discover the most important molecular features of good dyes in the dataset. This is work in progress.
• Using Generative models to produce new molecules that optimize the physical chemical properties we want, and that are likely to exist (chemically stable) and can be synthetized in the lab.
• We will release a suit of code to simplify data driven materials research, with building blocks to tailor workflows for similar problems.