Science Use Case 3

Data driven materials discovery for dye sensitized solar cells

ATPESC
Aug 9, 2019

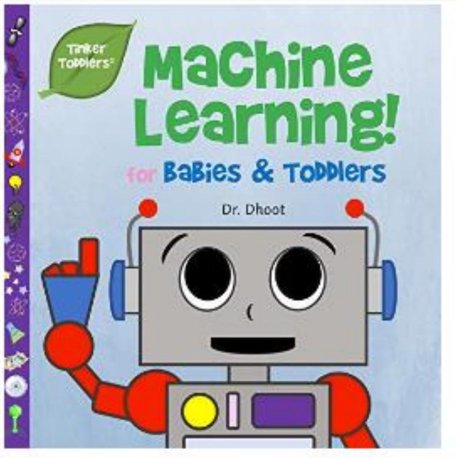
Álvaro V Mayagoitia
Argonne CPS

Acknowledgement

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..”

History of Science, summarized by Jim Cray

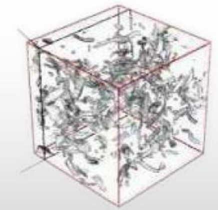
4^{th} Paradigm Data-Intensive Scientific Discovery

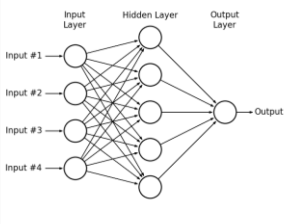
Experiments

Theory

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

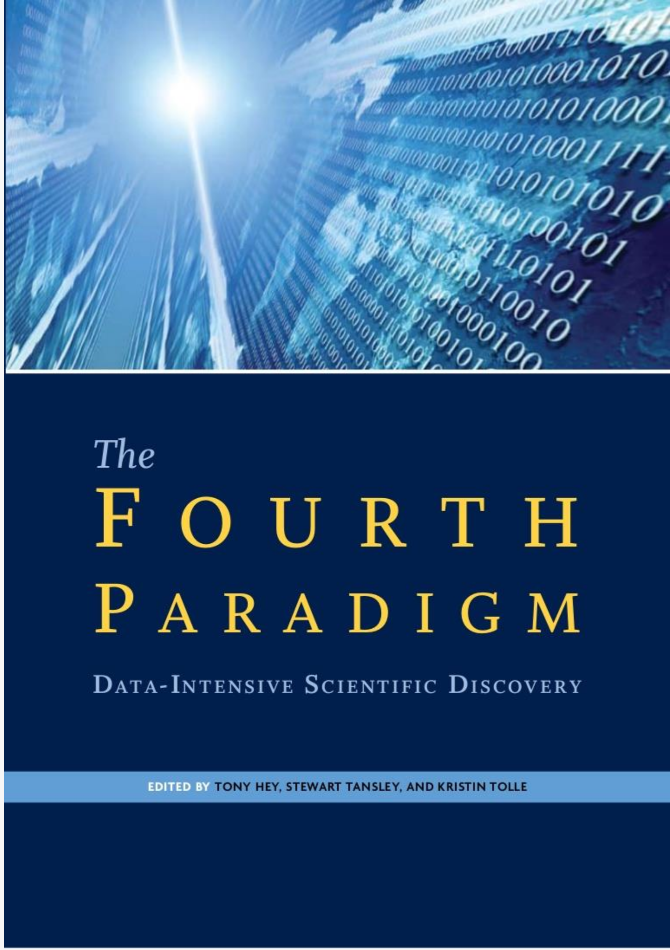
Computation





Data science

1600 1950 2000



https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Chemical Compound Space

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

*Nature, Insights, 2004

Chemical Compound Space

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

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

*Nature, Insights, 2004

*Nature, Insights, 2004

Recent estimations say could exceed 10^{180}

Chemical Compound Space

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

*Nature, Insights, 2004

* Nature, Insights,

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

How much do we know of the chemical space?

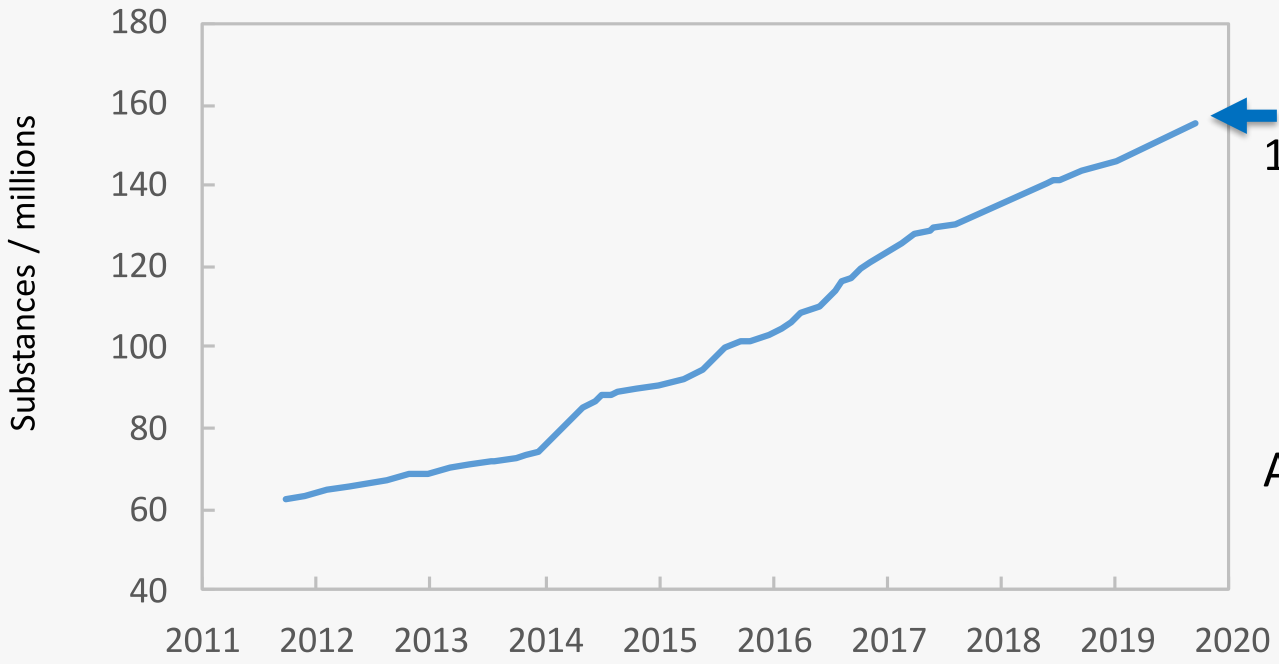
Compiled from experiments since the early 1800s: 10^{8}

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)

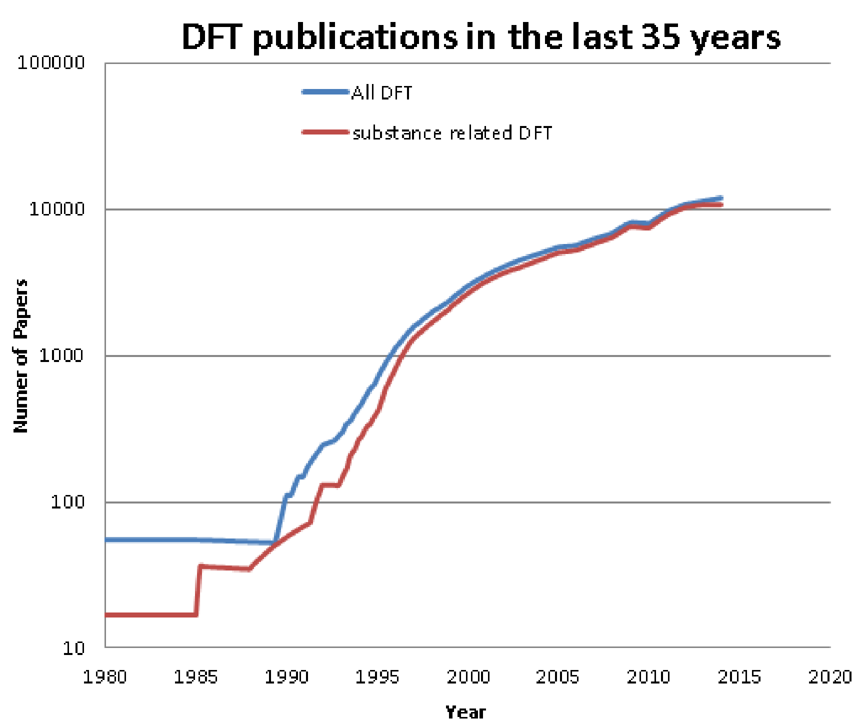


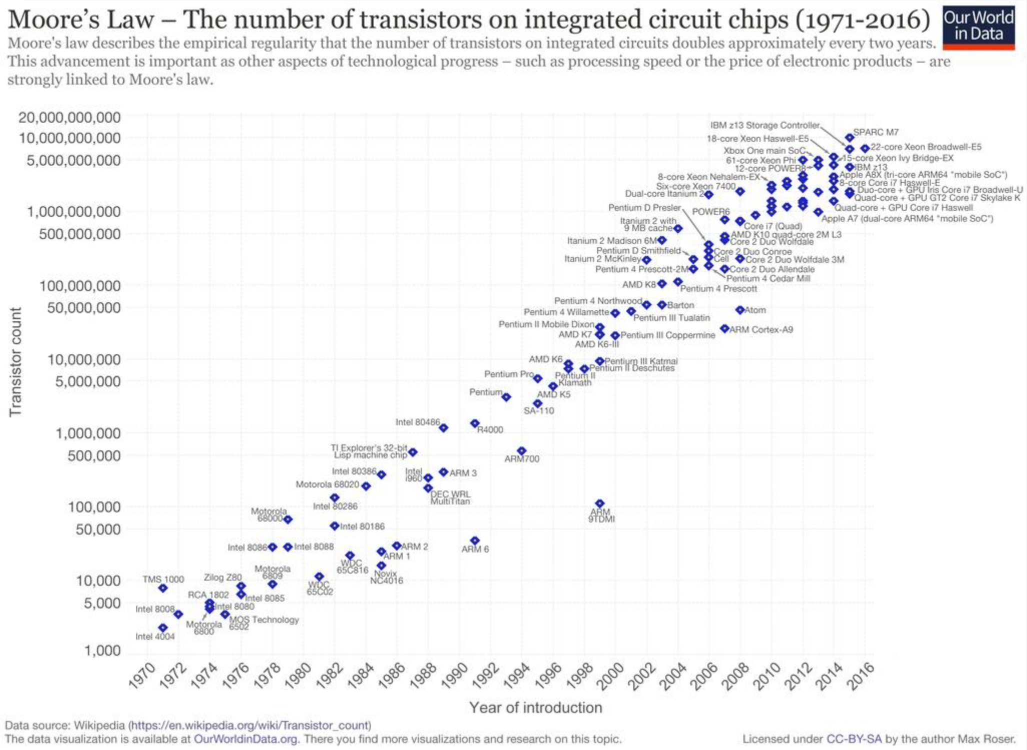
155M in Aug 2019

Are they any useful?

<u>10M of new substances per year in average</u>

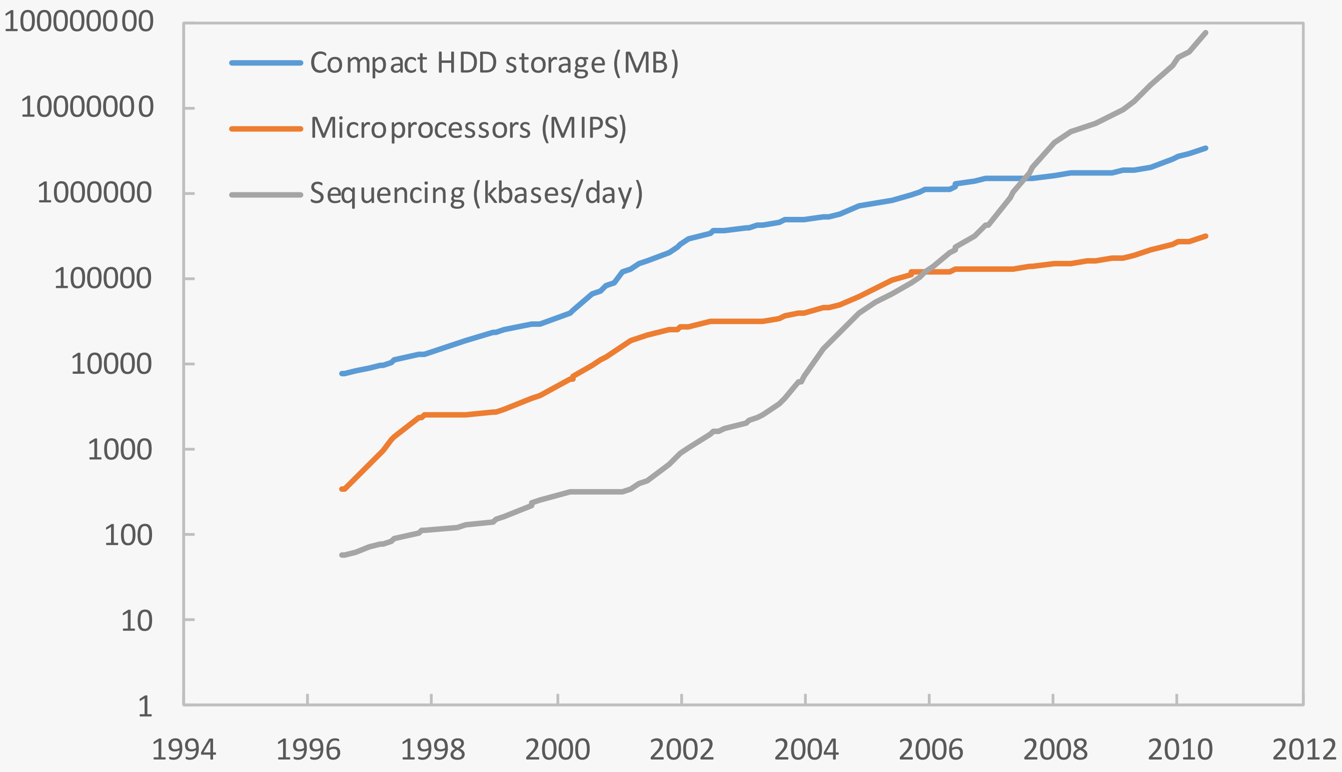
MATERIALS SCIENCE MODELING





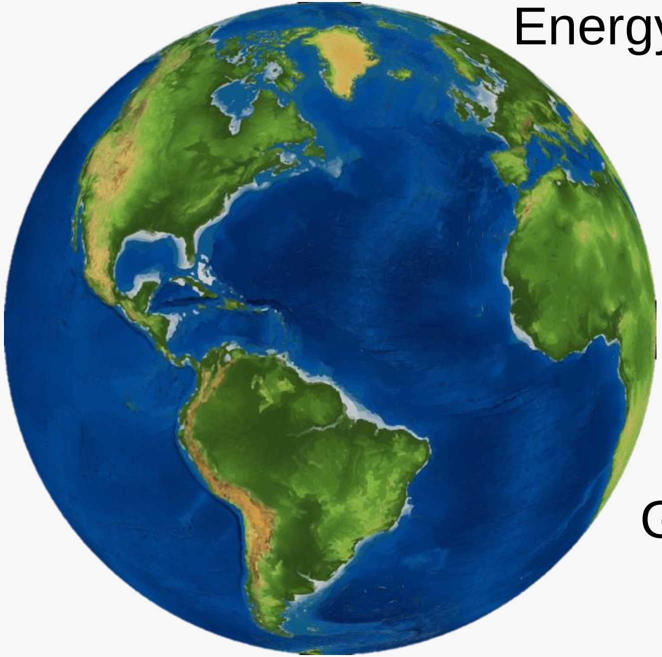
*Web of science queries

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

Energy needed

2007: **15 TW**

2017: **18 TW**

2050: ~**30 TW**

2100: ~**50 TW**

Biomass 5-7 **TW**

Wind 14 **TW**

Geothermal 1.9 **TW**

Tide/Ocean 0.7 **TW**

The Third Industrial Revolution by Jeremy Rifkin, 2011

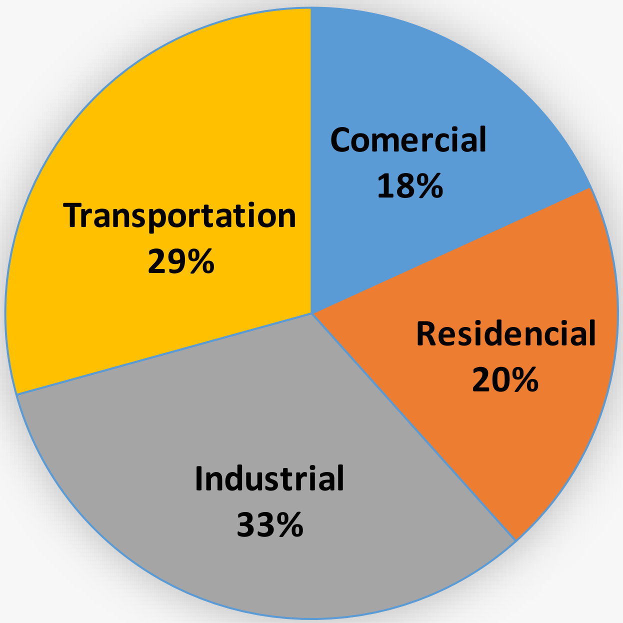
IRENA report 2018

Cost per watt-hour of Solar energy

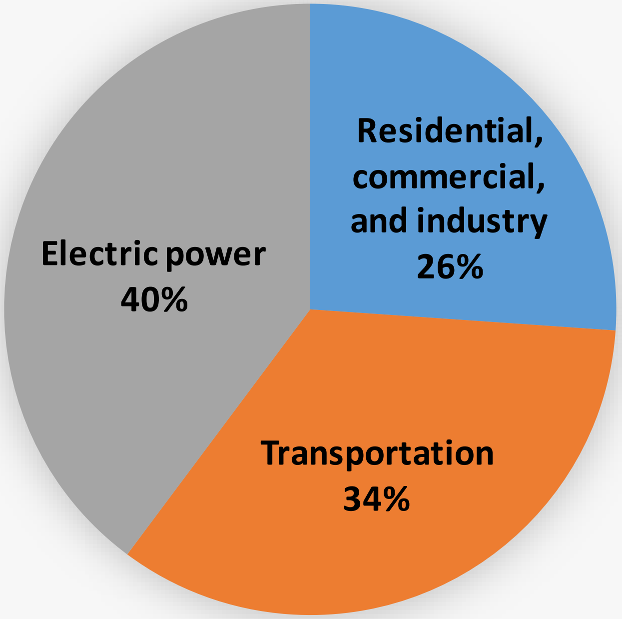
1977	2019
$76.67	$0.10

pvinsights.com

Share of total US energy consumption by end-use sectors 2018



Carbon footprint per sector 2018



Artificial Satellites

SF bus shelters

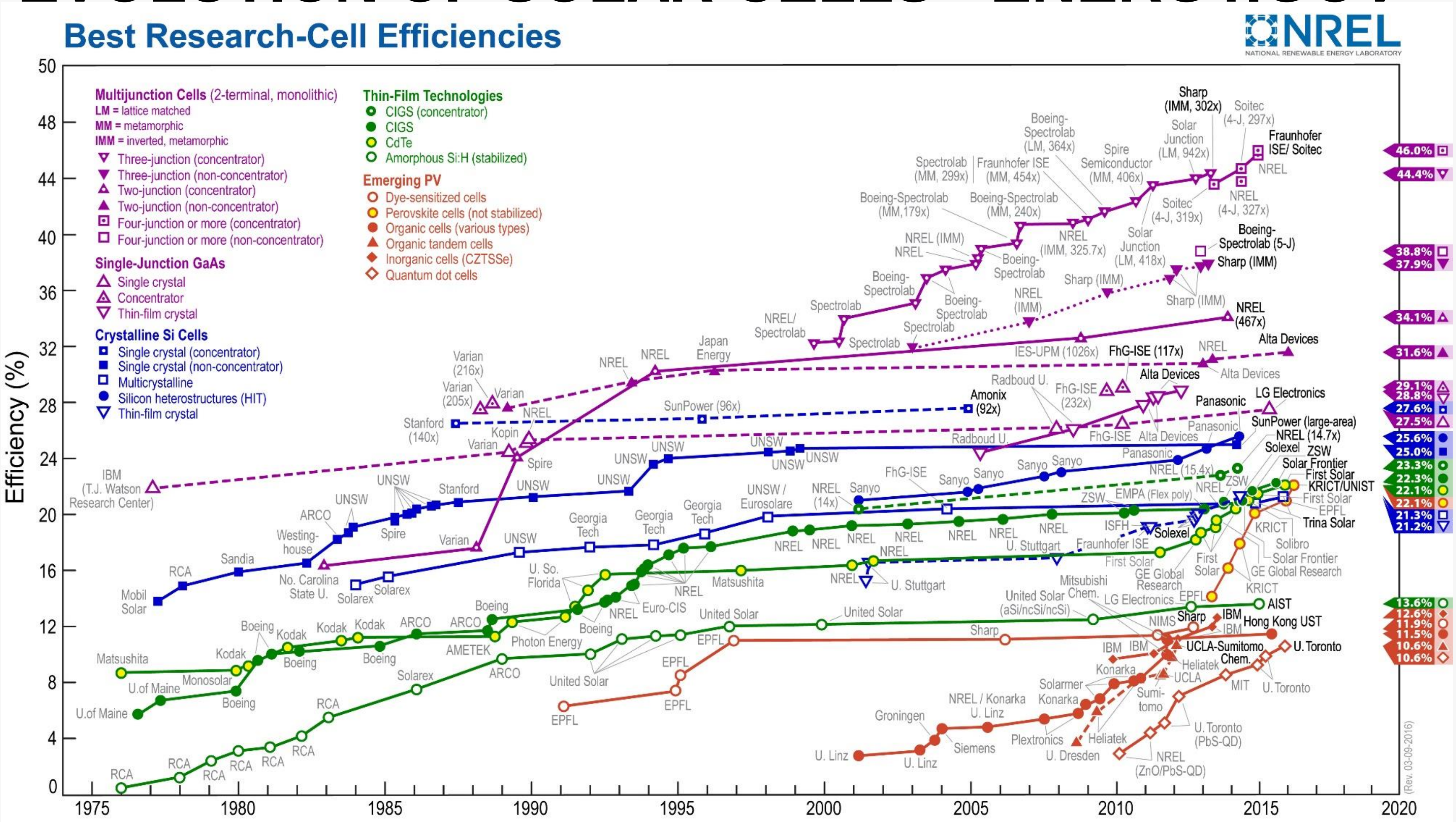
Food courts

Electric golf carts

Solar windows

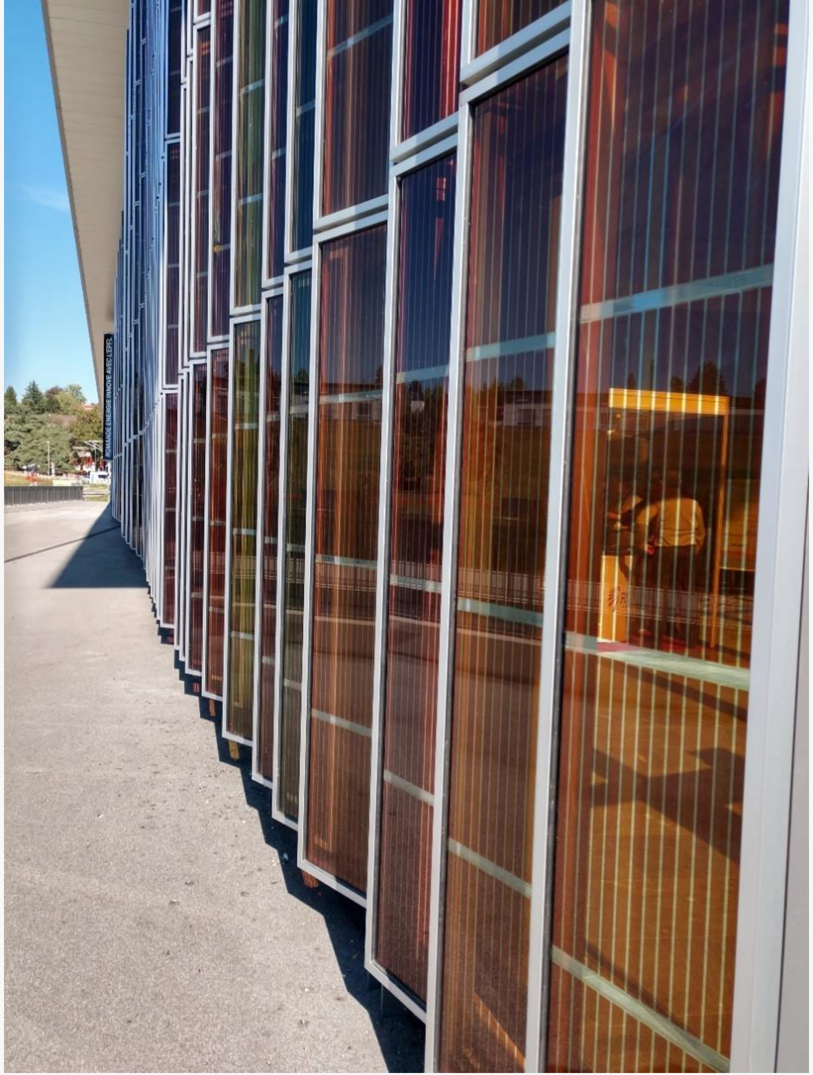
Harris Theatre – Chicago
[north Millenium park]

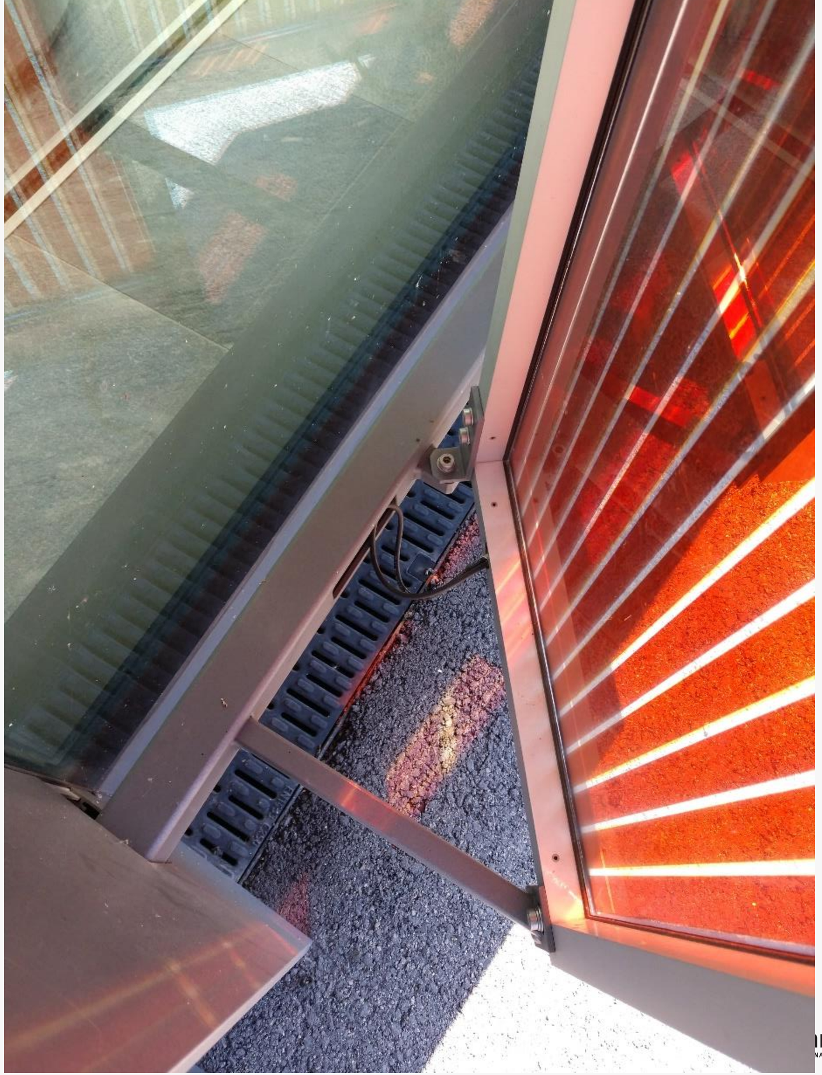
EVOLUTION OF SOLAR CELLS - ENERGY.GOV



Solar windows

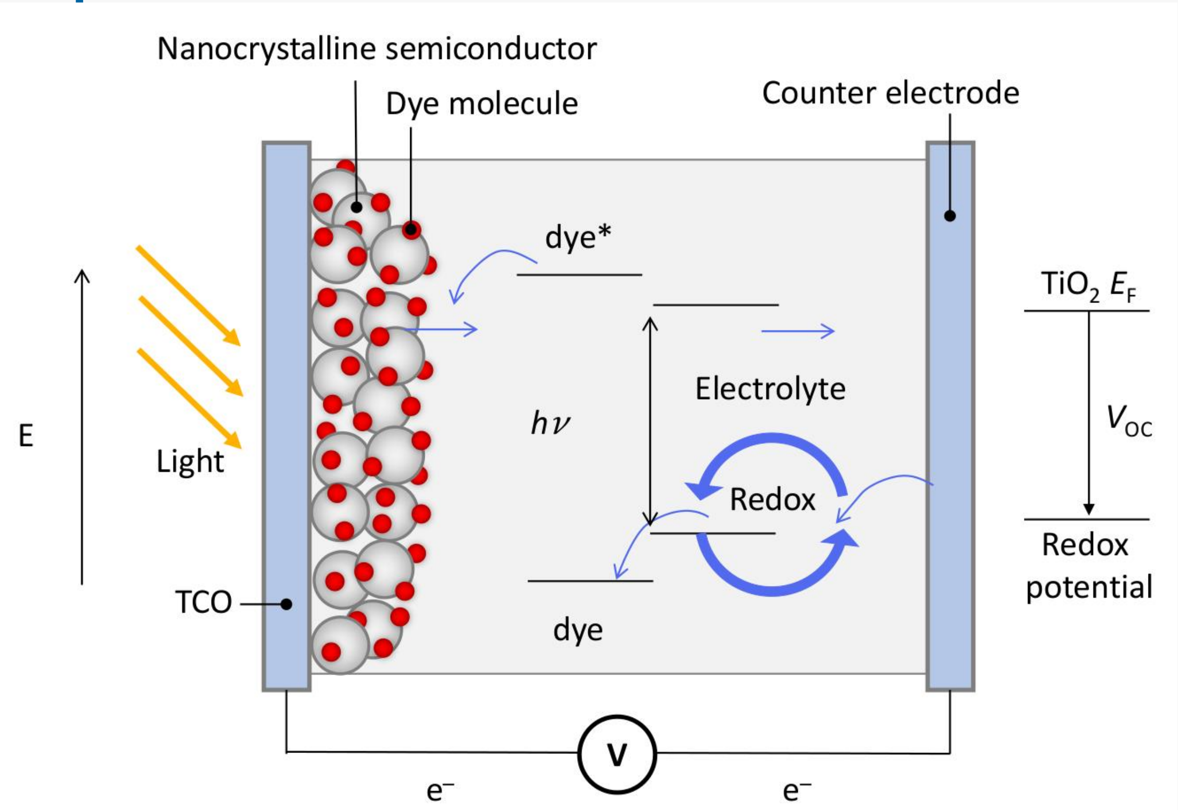
SwissTech Convention Center EPFL


ROMANDE ENERGIE INNOVE AVEC L'EPFL



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

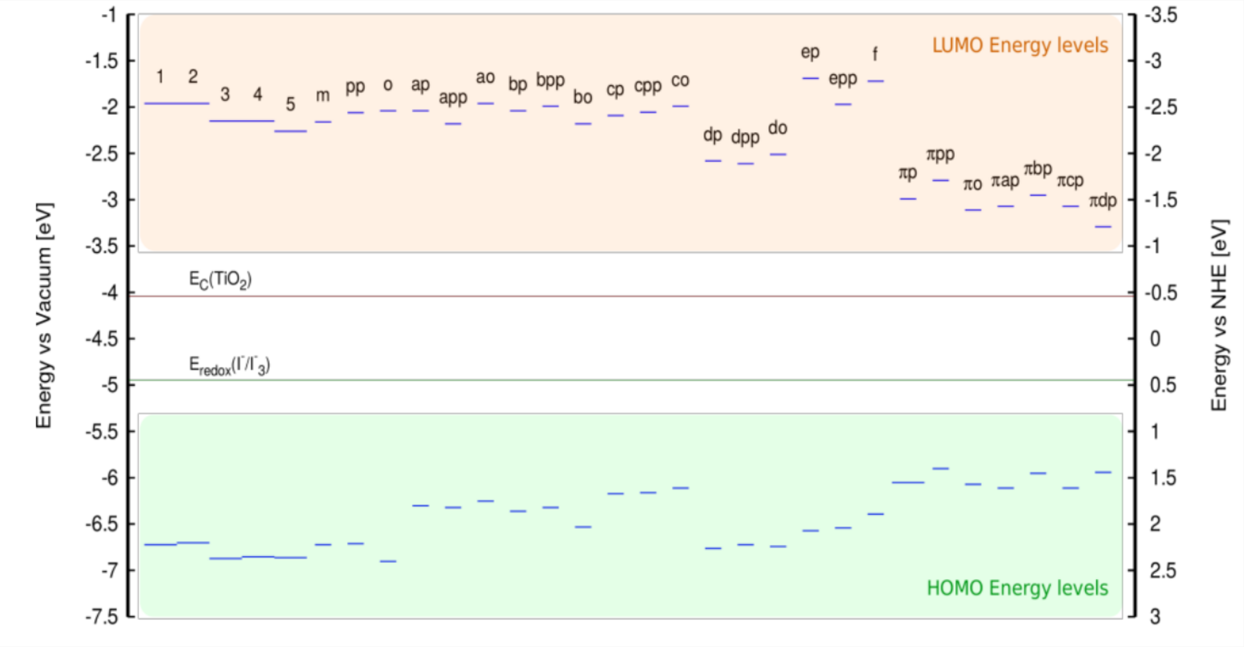
• Nazeeruddin et al JACS 1993 (ruthenium-based) N719:	10.4% eff.
• Yella et al Science 2011 (Zn-porphyrin-based) :	12.3 % eff.
• Burschka et al Nature 2013 (lead-iodide-based):	15% eff.

Organic cells

• Daeneke et al Nat. chem. 2011 (carbazole-base):	7.5% eff.
• Zeng et al Chem. Mater. 2010 (thiophene-base):	10.3% eff.

ENCODING STRUCTURE-FUNCTION

Screening with TDDFT is costly



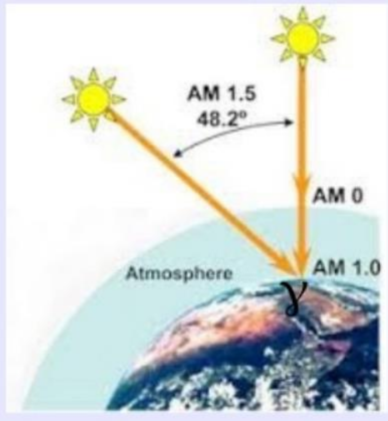
Rules:

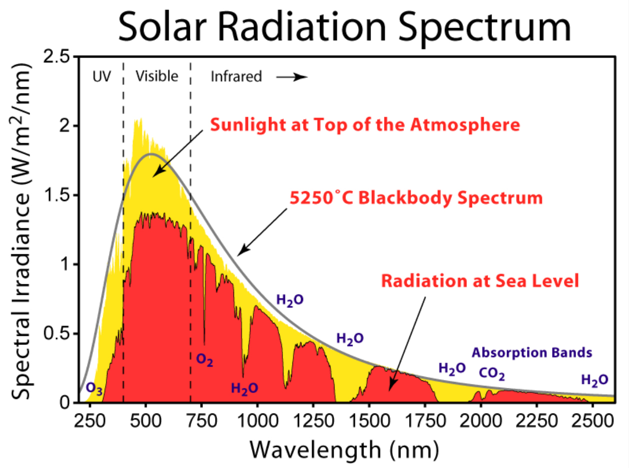
$$\varepsilon_{LUMO} > Ec_{TiO_2} \qquad \varepsilon_{HOMO} < E_{Electrolyte}$$

MAXIMIZING LIGHT HARVESTING EFFICIENCY

$$AM = \frac{L}{L_o} \approx \frac{1}{\cos z}$$

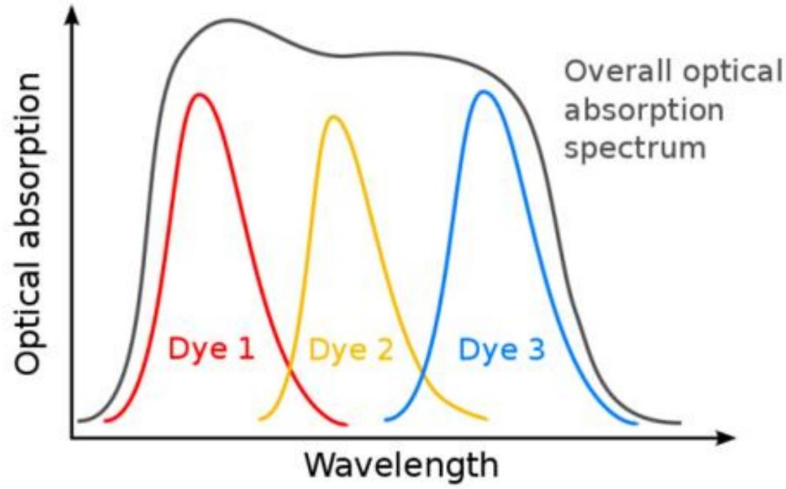
L_o = zenith path length
L= path length to the atmosphere
z = zenith angle





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

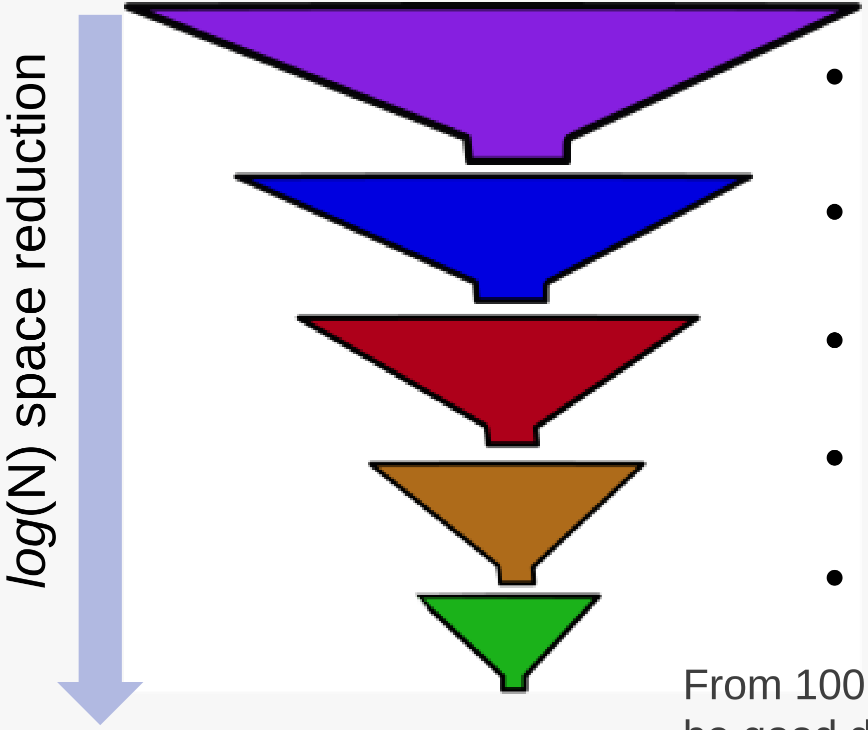
Theoretical Studies

- Geometry
- Physical properties

Funnel approach

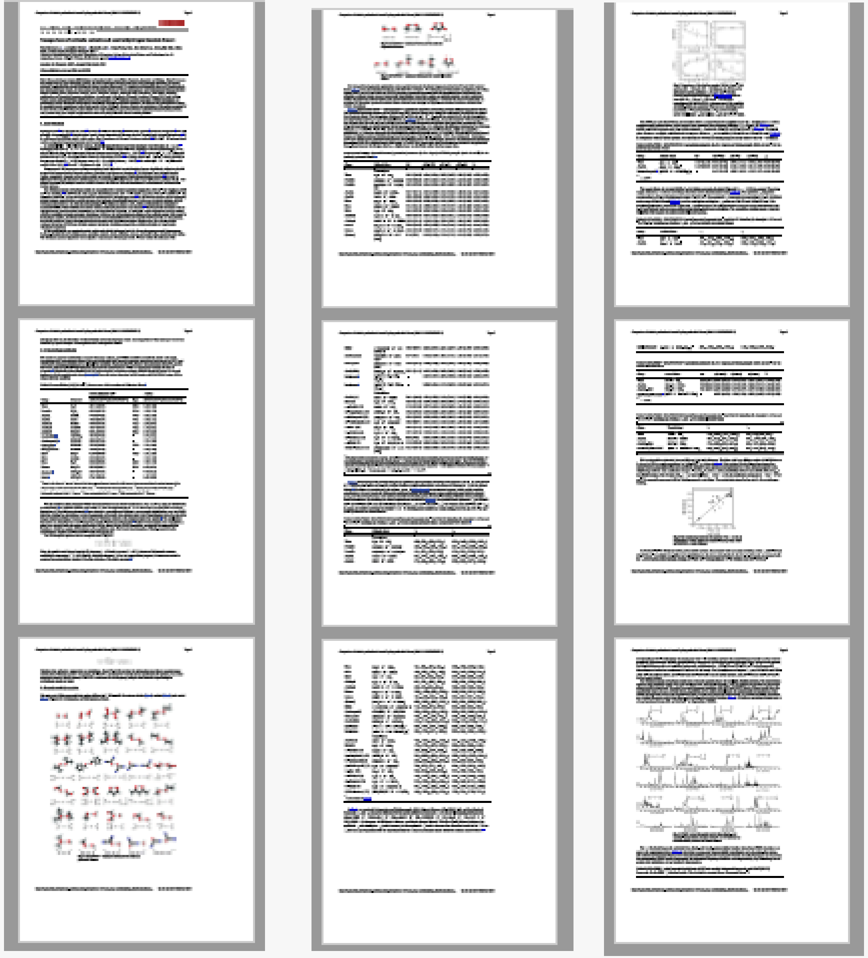
Screening the chemical space

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?

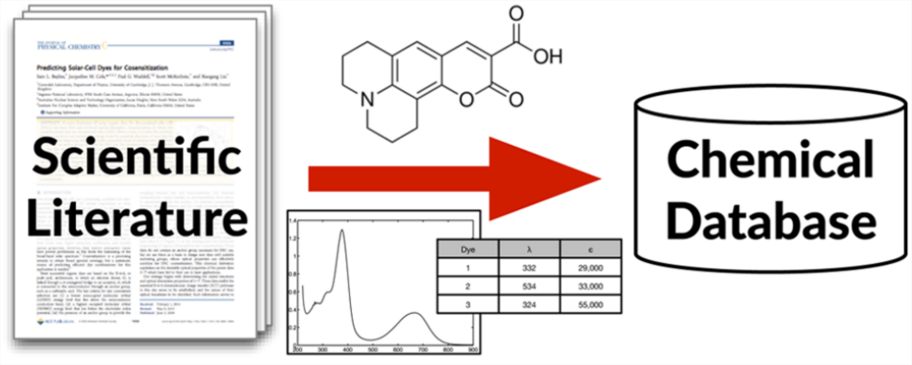


Compound	Lamda max	Ext. Coef.
Triphenylmethane	480	4320
Indigo	613	320
anthraquinone	320	-----

How long would it take a person to get this information?

ChemDataExtractor

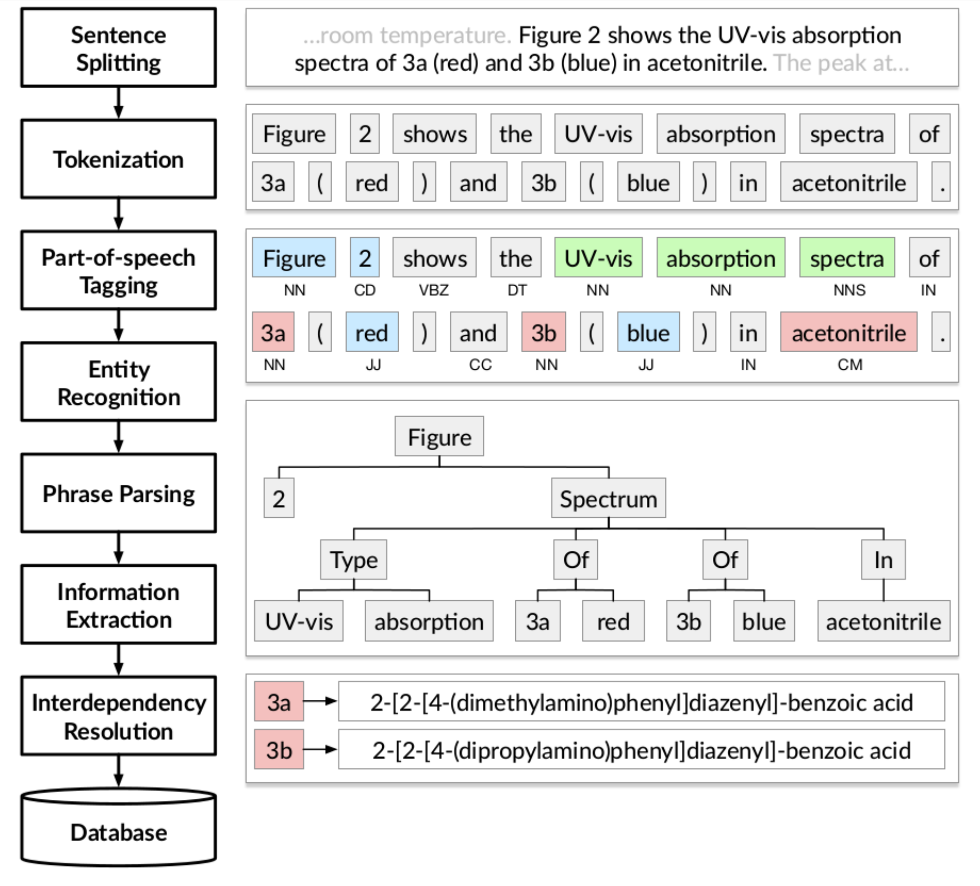
A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature



http://chemdataextractor.org

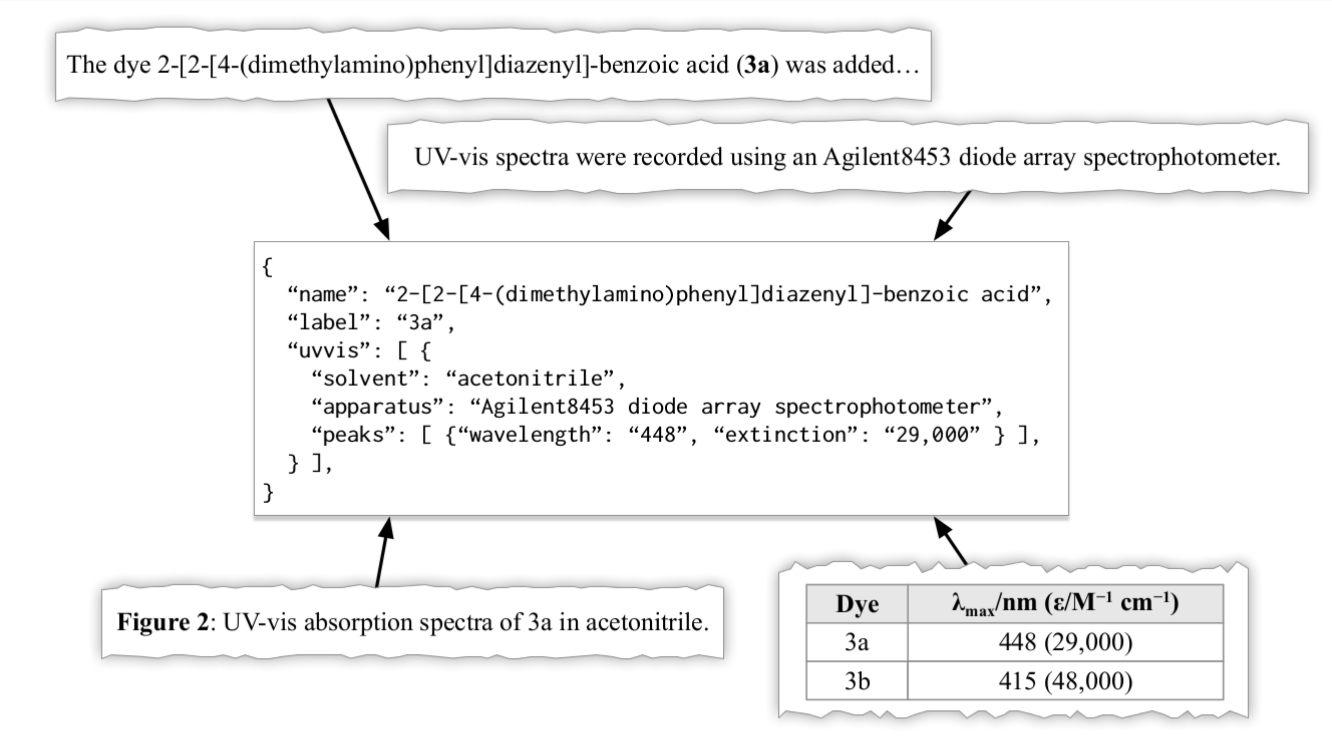
Swain, M. C., & Cole, J. M. J. Chem. Inf. Model. 2016

NATURAL LANGUAGE PROCESSING PIPELINE



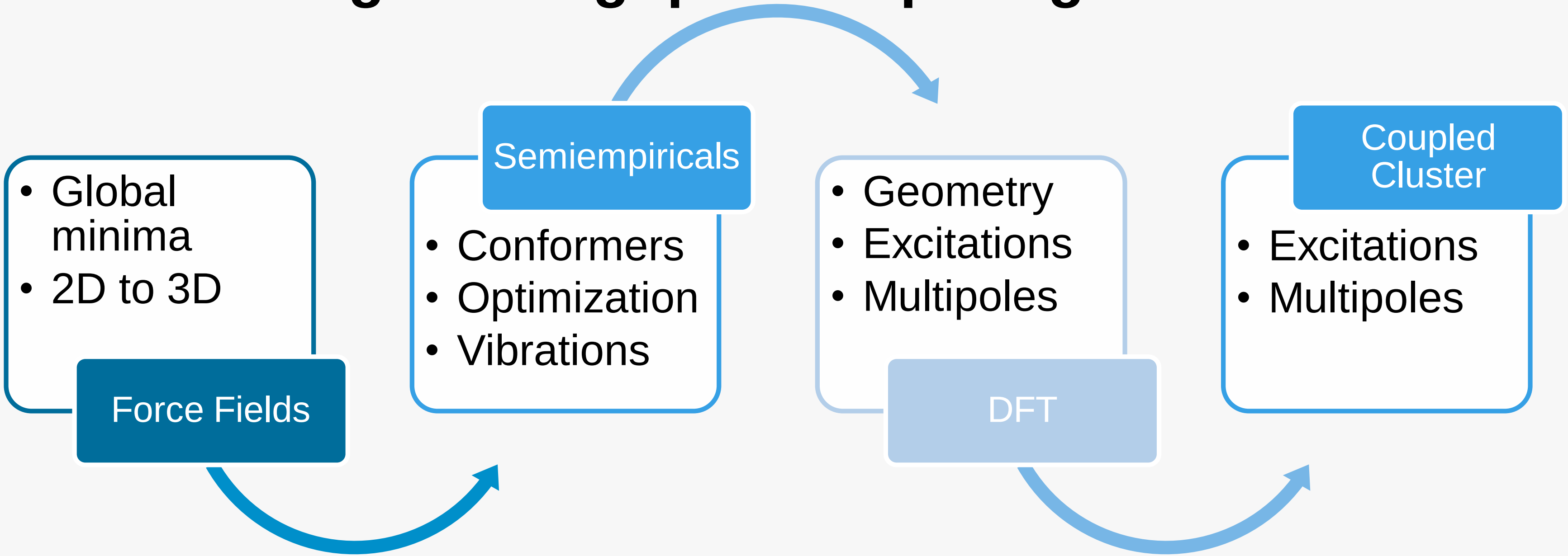
Swain, M. C., & Cole, J. M. J. Chem. Inf. Model. 2016

DATABASE



Swain, M. C., & Cole, J. M. J. Chem. Inf. Model. 2016

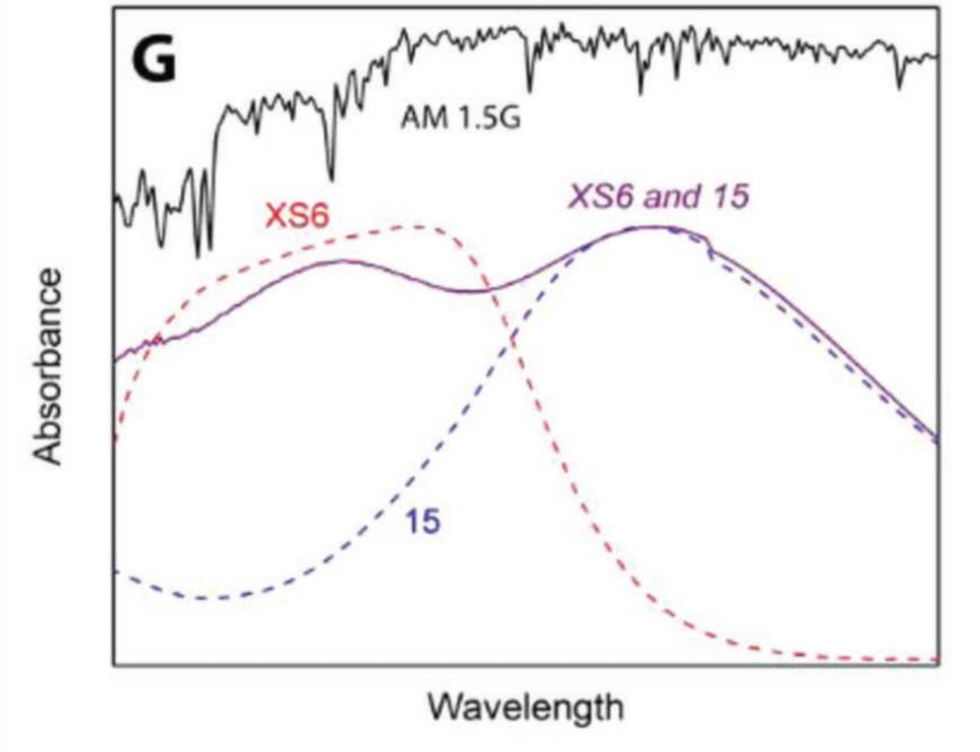
Informed high throughput computing



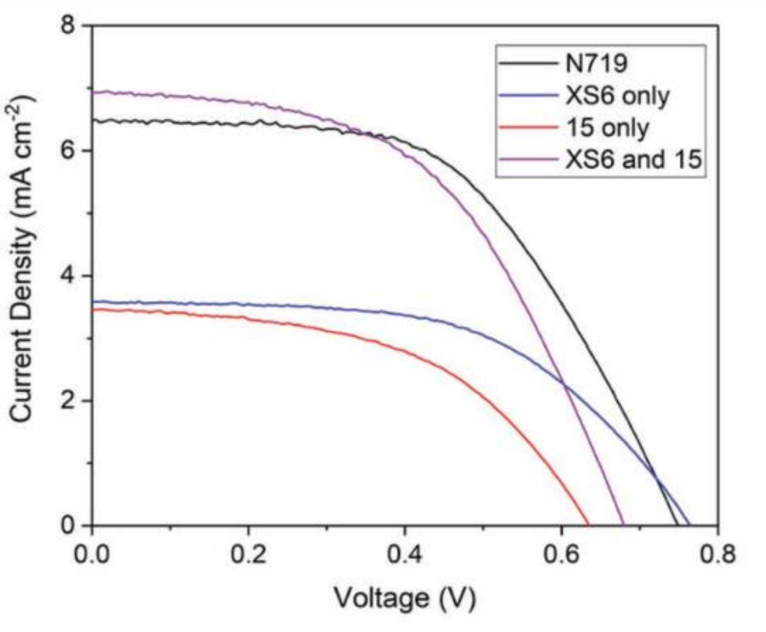
Composite of codes:

- Babel
- Rdkit
- MOPAC
- ORCA
- NWChem

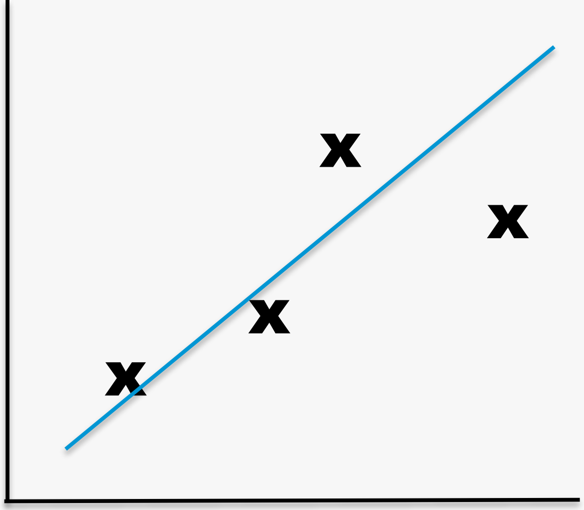
Sample name	AFM parameters				XRR parameters			
	Mean height [nm]	Max height [nm]	Aggregate coverage [%]	Number of aggregates [μm^{-2}]	Dye layer thickness [Å]	SLD_{dye} [$\times 10^{-6}$ Å^{-2}]	Surface roughness [Å]	Surface coverage [%]
Singly sensitized working electrodes								
C1 only	5 ± 1	7 ± 2	3 ± 6	2 ± 3	43.5 ± 0.9	6.6 ± 0.5	5.6 ± 0.7	55 ± 4
8c only	5 ± 1	6 ± 2	3 ± 2	3 ± 2	26.6 ± 0.9	5.1 ± 0.9	3.3 ± 0.8	39 ± 7
XS6 only	4.9 ± 0.4	6.0 ± 0.7	1.0 ± 0.1	0.3 ± 0.2	23.6 ± 0.5	8.7 ± 0.4	3.7 ± 0.5	73 ± 3
H3 only	9 ± 1	15 ± 3	0.3 ± 0.1	0.18 ± 0.05	27 ± 1	6.7 ± 0.5	3.7 ± 0.5	55 ± 4
15 only	8 ± 2	15 ± 3	7 ± 2	1.1 ± 0.4	24.3 ± 0.3	7.8 ± 0.4	2.7 ± 0.3	62 ± 3
Co-sensitized working electrodes								
C1 then 15	6 ± 2	10 ± 3	1.3 ± 0.5	0.7 ± 0.2	33.7 ± 0.5	5.9 ± 0.7	3.1 ± 0.6	49 ± 6
C1 and 15	7 ± 2	12 ± 4	2.0 ± 0.5	0.9 ± 0.6	21.5 ± 0.8	6.3 ± 0.9	3.8 ± 0.7	52 ± 7
H3 then C1	8 ± 2	16 ± 4	3 ± 3	0.4 ± 0.2	42 ± 1	6.0 ± 0.6	5.2 ± 0.7	49 ± 5
C1 and H3	5 ± 1	8 ± 3	2 ± 1	2 ± 2	25.4 ± 0.4	8.5 ± 0.4	3.0 ± 0.5	69 ± 3
8c then 15	6 ± 1	9 ± 2	1.1 ± 0.2	0.7 ± 0.6	30.9 ± 0.4	6.9 ± 0.4	3.9 ± 0.6	54 ± 3
8c and 15	4.6 ± 0.3	5.8 ± 0.4	12 ± 9	16 ± 5	31 ± 2	5.7 ± 0.5	7 ± 2	45 ± 4
H3 then 8c	5.5 ± 0.7	8 ± 1	3 ± 2	1 ± 1	37.2 ± 0.2	9.0 ± 0.7	2.9 ± 0.4	70 ± 5
8c and H3	5.2 ± 0.7	7 ± 2	2 ± 2	1 ± 1	27.5 ± 0.4	8.0 ± 0.4	3.3 ± 0.6	63 ± 3
XS6 then 15	**6 ± 1**	**8 ± 2**	**0.7 ± 0.3**	**0.8 ± 0.3**	**18.8 ± 0.3**	**8.7 ± 0.5**	**3.6 ± 0.4**	**72 ± 4**
XS6 and 15	**7.8 ± 0.7**	**11 ± 1**	**0.3 ± 0.1**	**0.24 ± 0.09**	**18.6 ± 0.3**	**8.8 ± 0.5**	**3.4 ± 0.4**	**73 ± 4**
XS6 then H3	5.5 ± 0.7	7.6 ± 0.8	0.3 ± 0.1	0.25 ± 0.04	21.0 ± 0.3	9.6 ± 0.5	4.1 ± 0.4	79 ± 4
XS6 and H3	5.3 ± 0.8	7 ± 1	0.3 ± 0.1	0.2 ± 0.1	21.6 ± 0.6	8.7 ± 0.6	4.0 ± 0.5	71 ± 5

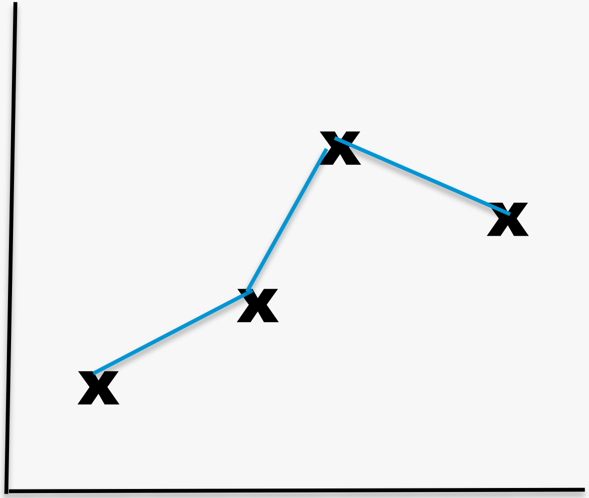


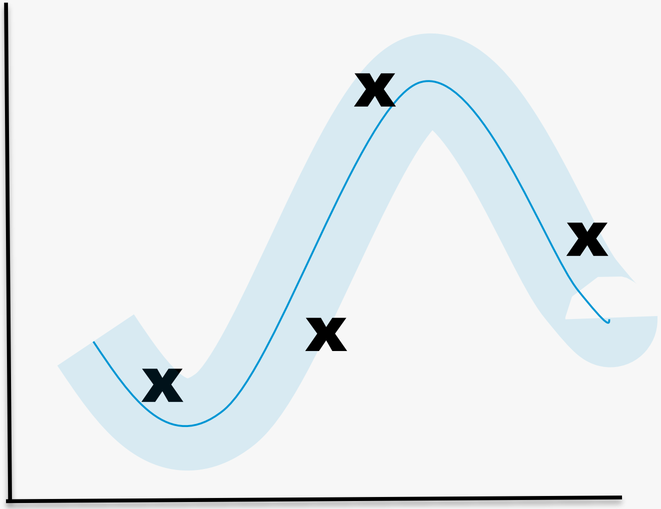
pvinsights.com

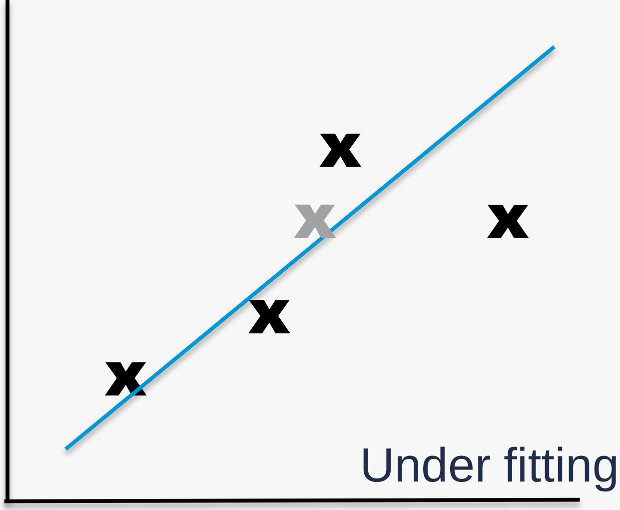


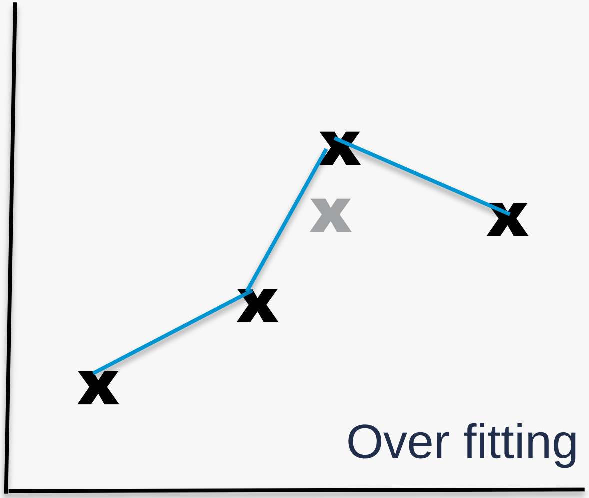
pvinsights.com

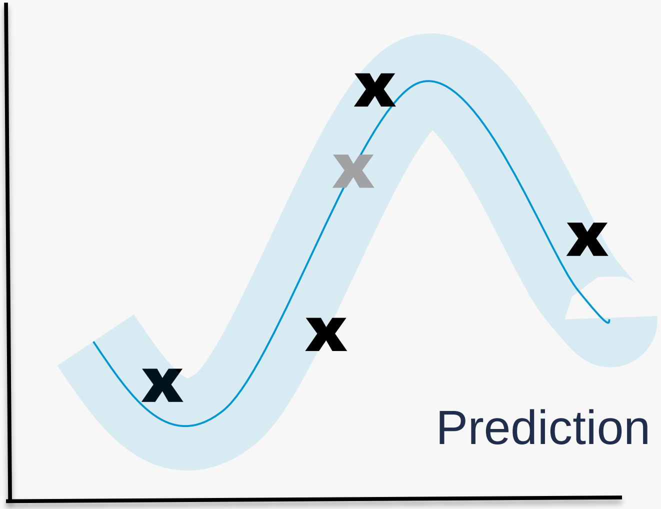







Under fitting


Over fitting


Prediction + uncertainty

Gaussian Process

Quick introduction

Target function

$$y_i = f(x_i) + \epsilon_i$$

Noise function

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

Here* Means a point that we want to predict

Given **y** , the probability of y_* is:

$$y_* | \mathbf{y} \sim \mathcal{N}(K_* K^{-1} \mathbf{y},\ K_{**} - K_* K^{-1} K_*^T)$$

Prediction (or kriging)

$$\overline{y}_* = K_* K^{-1} \mathbf{y}$$

Variation

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

Gaussian process

Covariance matrix

Covariance function

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

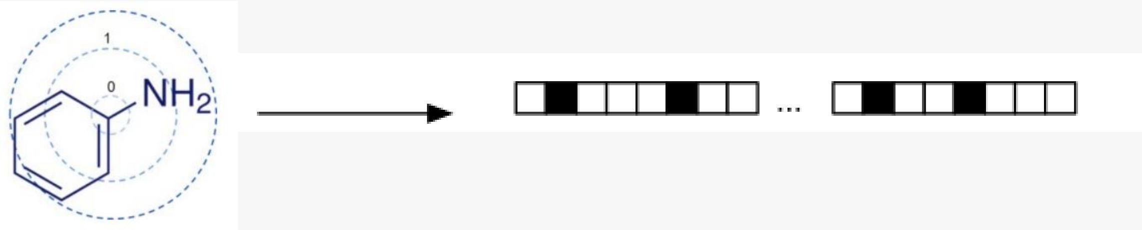
Example: Square exponential

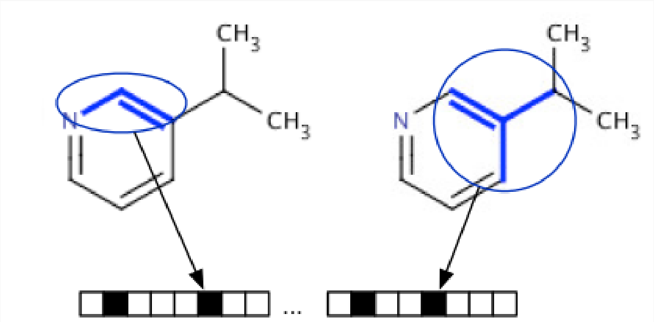
Covariance matrix

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

Molecular Fingerprint examples

Morgan Circular Fingerprint

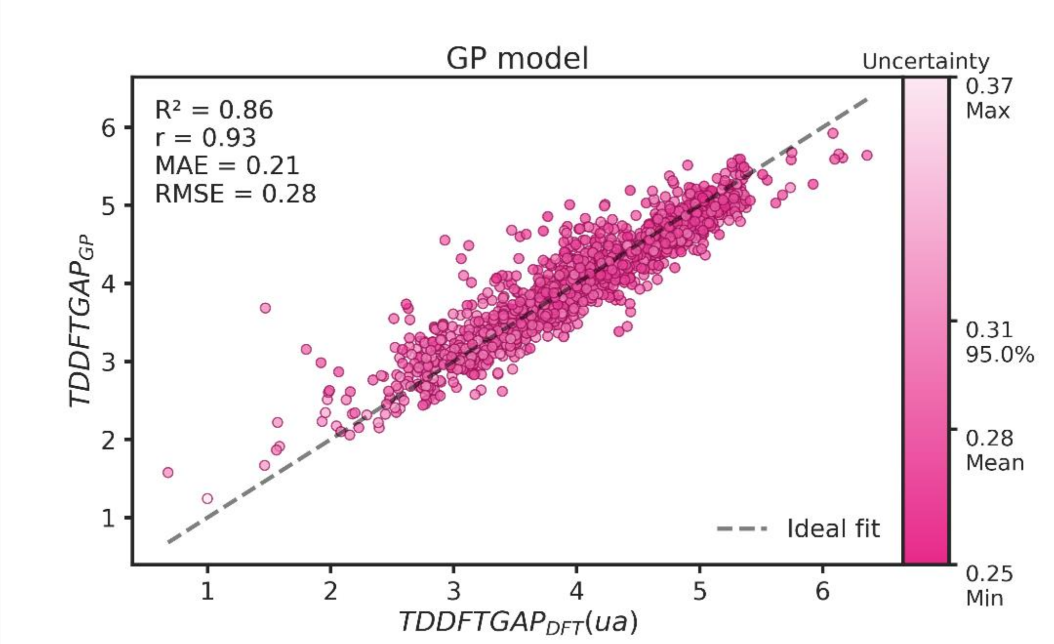




Learn from data and feedback to experiments

Transition prediction

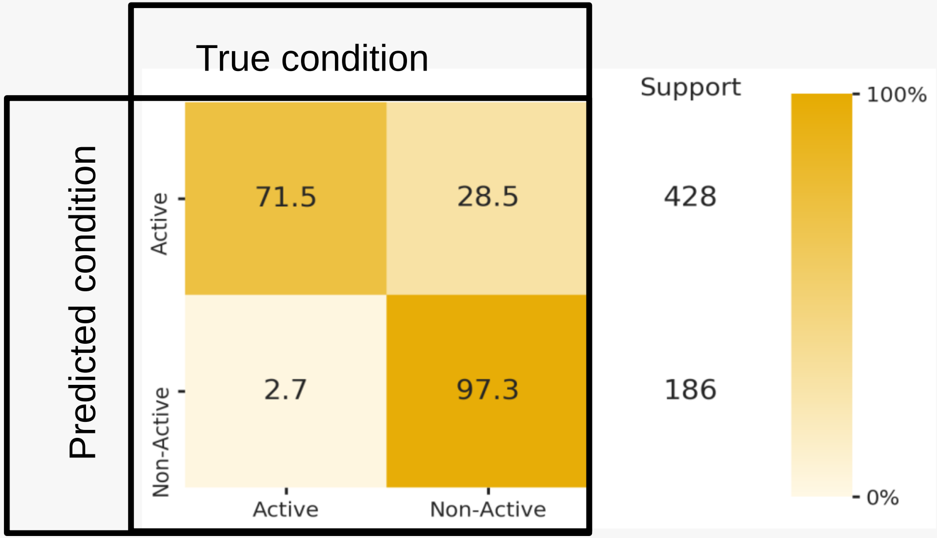
TDDFT gap prediction – We used Gaussian Process and Circular Morgan Fingerprints to predict the first transition of the a reduce scale TDDFT (sTDA//wB97X-D3/TZVP), we found that this value is predictable. Similar result found for HOMO-LUMO DFT gap.



Learn from data and feedback to experiments

Is this optically active?

Confusion matrix



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%.

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.**

Q&A

pvinsights.com