Materials Informatics: The Marriage of Materials and Data Sciences

Omer Kaspi,[a] Hadar Binyamin,[a] Olga Girshevitz,[b] Abraham Yosipof,[c] Hanoch Senderowitz[a]

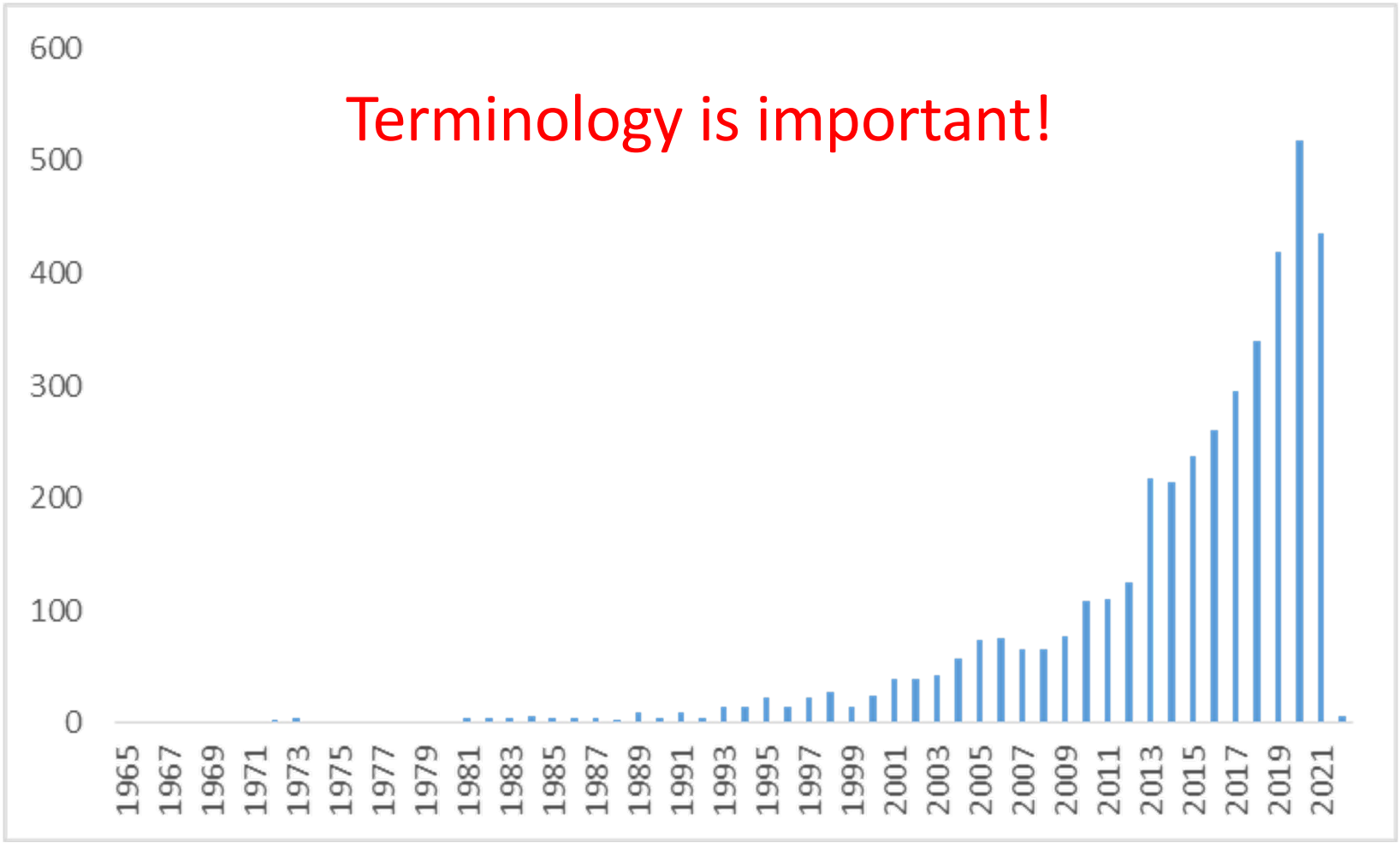
[a]Department of Chemistry, Bar Ilan University, Israel
[b]Bar Ilan Institute of Nanotechnology and Advanced Materials, Bar-Ilan University, Israel
[c]Department of Information Systems, College of Law & Business, Israel

AIDD School, October 21st, Leuven, Belgium

What Can You Do With Simple Tools?

- No neural networks
- No complex ML algorithms
- But a few cool applications

Materials Informatics is Rapidly Growing



Senderowitz and Tropsha, JCIM 2018, 58, 1313-1314

The Rise and Rise of Material Informatics

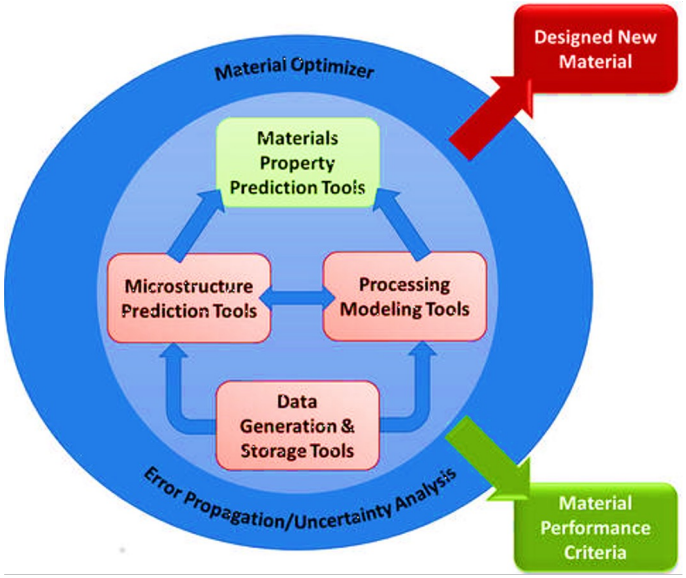
ICSD

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

> 260,000 crystal structures

> 209,000 experimental inorg. structures	> 30,000 metal-organic structures	> 23,000 theoretical structures	~ 26,000 derived structures
>9,800 structure types	> 124,000 authors	> 1,600 journals	> 96,000 articles

NIST



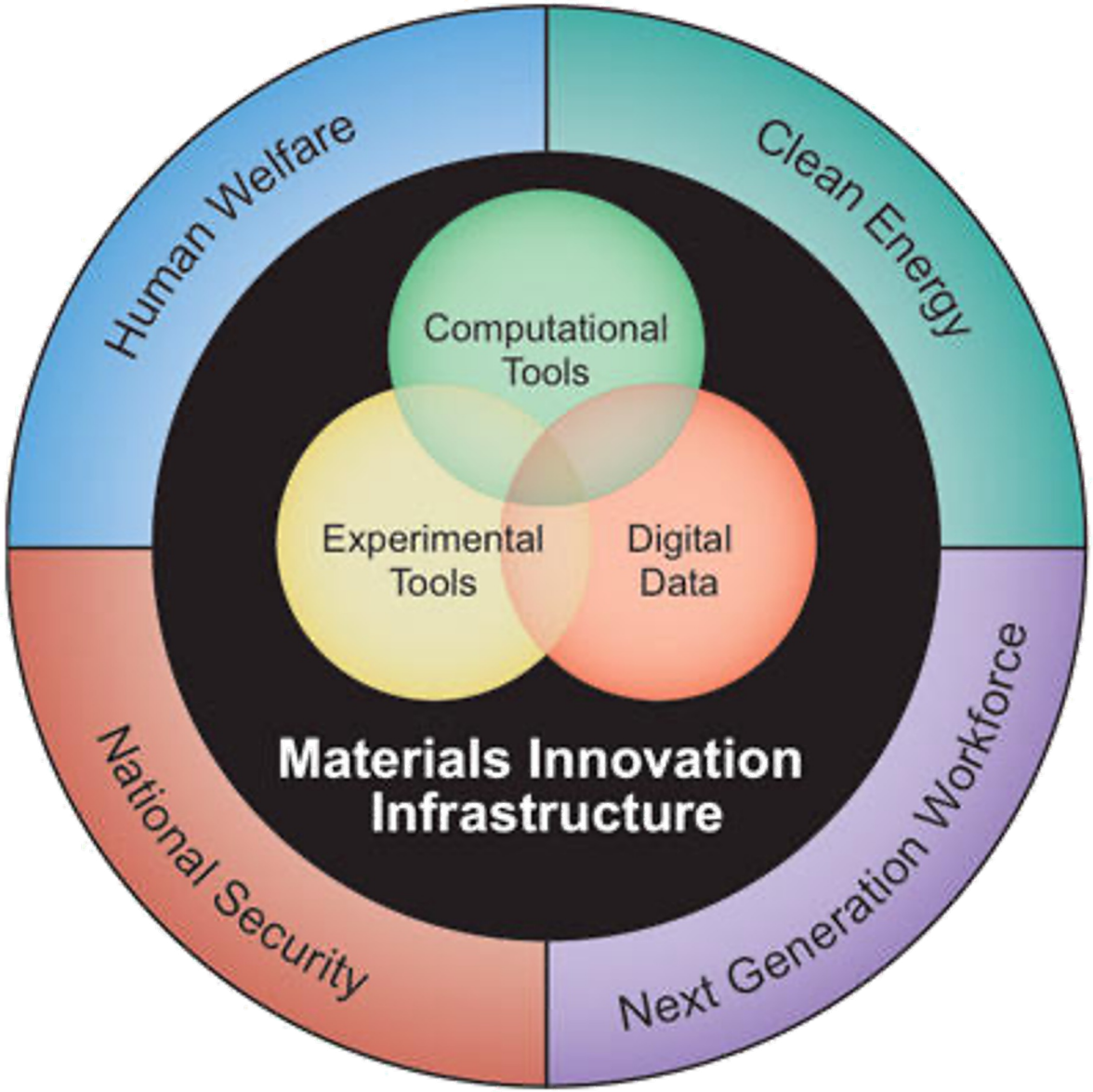
> 40000 compounds

MatWeb
MATERIAL PROPERTY DATA

> 160000 compounds

Material Genome Initiative (NIST)

The Materials Genome Initiative is a federal multi-agency initiative for discovering, manufacturing, and deploying advanced materials twice as fast and at a fraction of the cost compared to traditional methods. The initiative creates policy, resources, and infrastructure to support U.S. institutions in the adoption of methods for accelerating materials development.



AFLOW (https://aflowlib.org/)

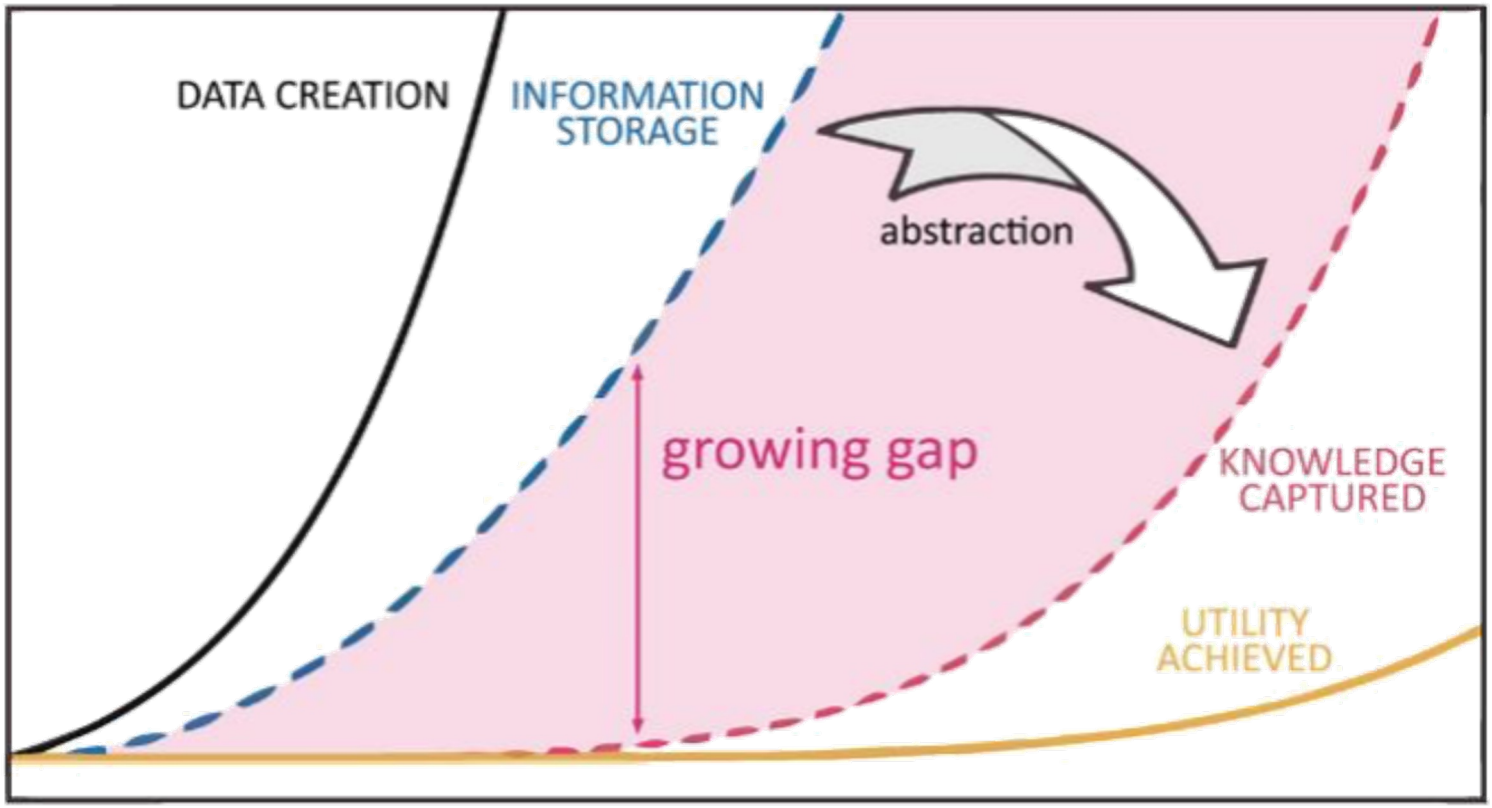
Welcome to AFLOW, a globally available database of **3,528,653** material compounds with over **733,959,824** calculated properties, and growing.

3,477,380 **form. enthalpies**	366,978 **band structures**	172,478 **Bader charges**	5,650 **elastic properties**
5,664 **thermal properties**	1,738 **binary systems**	30,282 **ternary systems**	150,659 **quaternary systems**

Too Much Information or Too Little Information?

Materials space is HUGE ~$10^{30\text{-}50}$ candidates

Drowning in Information but Starving for Knowledge (John Naisbitt)

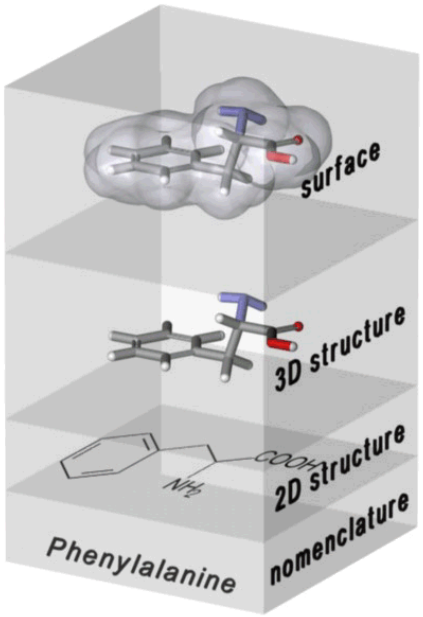


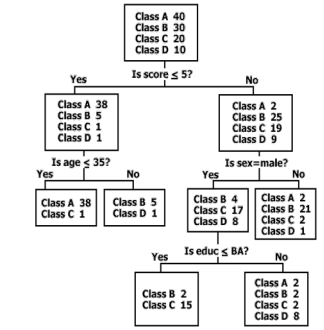
Tremendous opportunities for materials-informatics

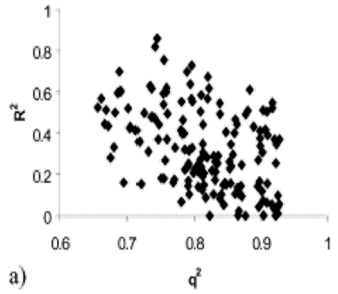
Important lessons to be taught from chemoinformatics

Structure-Property Predictions

- Accurate experimental data (e.g., activities, dependent variables)
- Descriptors (independent variables)
  - Structure-derived , process-derived (measured; calculated)
- A mathematical model
  - *e.g.*, quantitative, qualitative, linear, non-linear
- Model validation
  - Models developed on a training set and tested on an independent test set
  - Models should be simple and interpretable

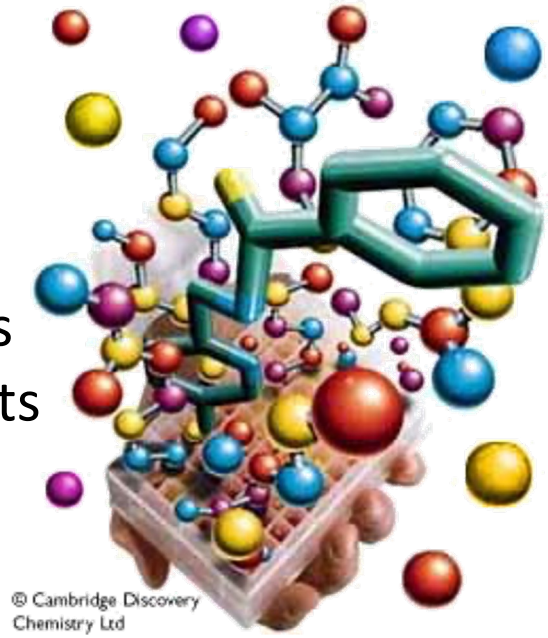




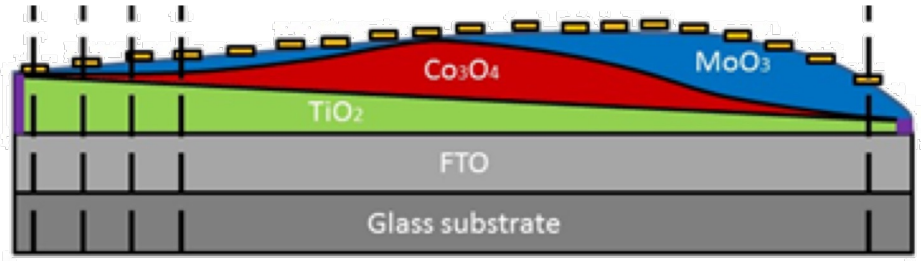


The Compounds

- Chemoinformatics
  - Structures typically well-defined
  - Potential exceptions: Polymers and mixtures (but constituting building blocks/components are known)
  - True even for combinatorial chemistry



- Materials informatics
  - Structures sometimes well-defined
  - Not true for combinatorial material synthesis



Yosipof et al., Mol Inform. 2016, 35, 568-579

The Data

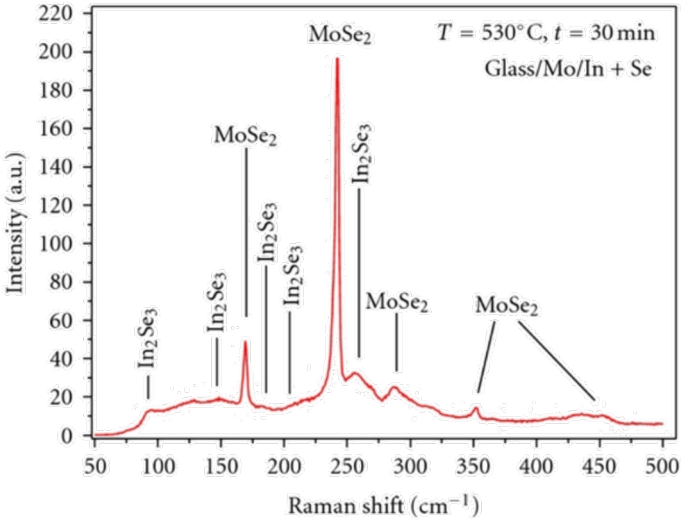
- Chemoinformatics
  - Primarily concerned with pharmacokinetics / pharmacodynamic related activities
  - Diversity comes from the targets / ligands
  - Medium / large / very large data sets

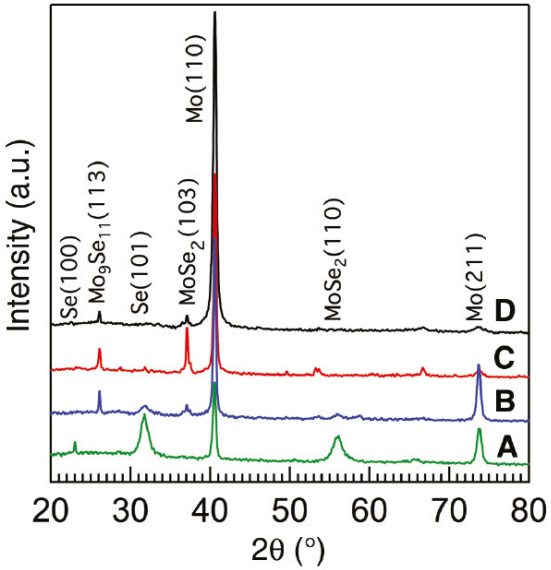
- Materials informatics
  - Diversity comes from activities and the nature of the materials
  - Solubility of materials
  - Biological activities
  - Young's modulus
  - Thermal conductivity
  - Atomization energies
  - Glass transition temperatures
  - Half decomposition temperature
  - Melting point of ionic liquids
  - Viscosity
  - Photovoltaic properties
  - Tiny / small / medium / large / very large data sets

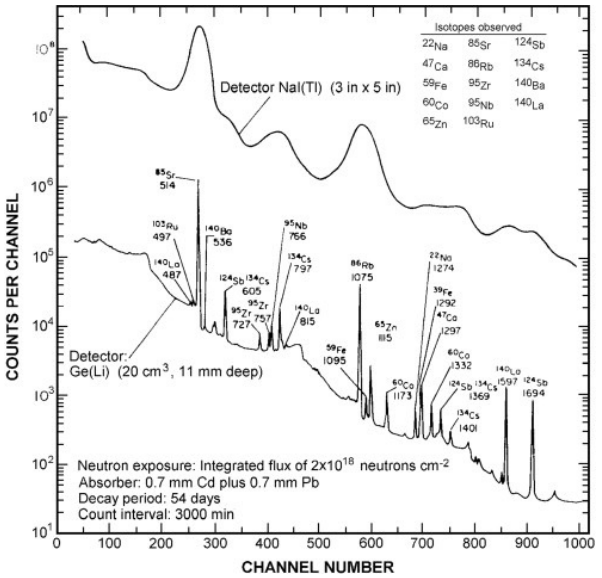
The Descriptors

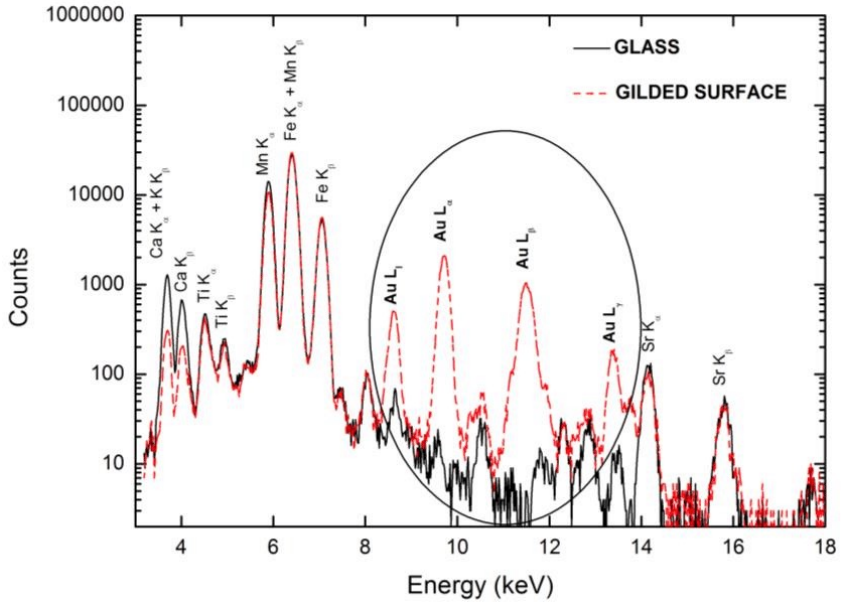
- Chemoinformatics
  - Typically nD (n = 1,5) “classical” descriptors
  - Limited usage of QM-derived descriptors
- Materials informatics
  - Typically nD (n = 1,5) “classical” descriptors
  - Heavy reliance on QM descriptors
  - Usage of experimental conditions as descriptors
  - Heavy reliance on measured descriptors (for undefined structures)

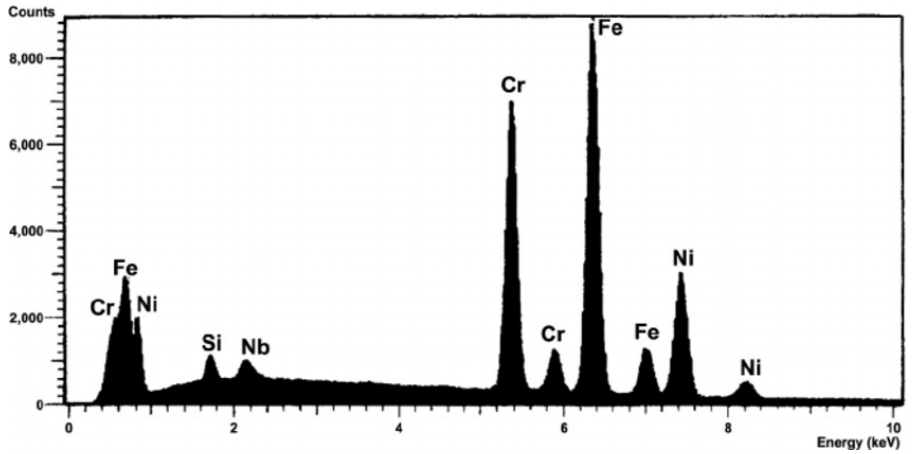
Focus on Spectra (Raw data, Images)











The Algorithms

- Chemoinformatics and materials informatics
  - Unsupervised methods
    - Data reduction techniques (e.g., PCA)
    - Clustering
  - Supervised methods
    - Classification models (e.g., Random Forests)
    - Quantitative models (e.g., MLR, SVM, *kNN*, neural networks, DL)

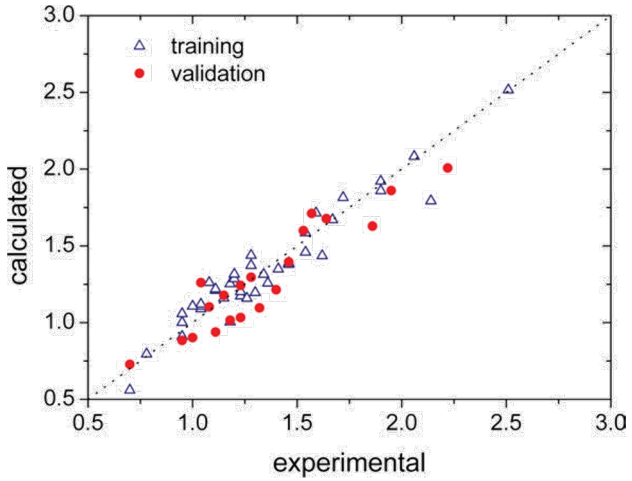
Validation

- Chemoinformatics
  - "OECD" principles available and frequently followed
- Materials informatics
  - Insufficient external validation
  - Inappropriate control for chance correlation

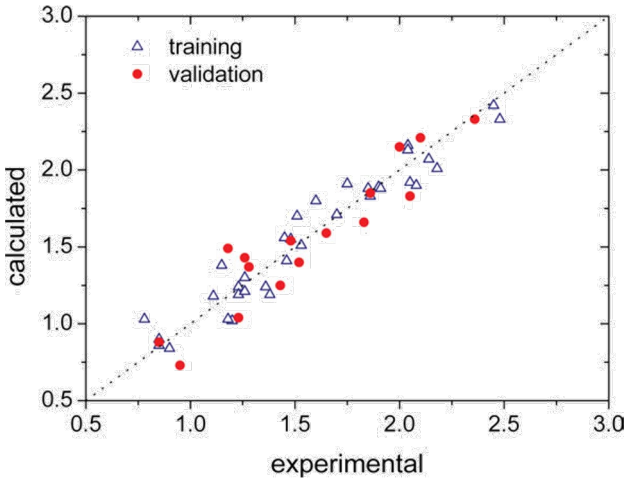
Prediction of Impact Sensitivity for Energetic Materials

- Prediction of impact sensitivity of nitro compounds
- 161 compounds, specific and global models MLR, “OECD” validation

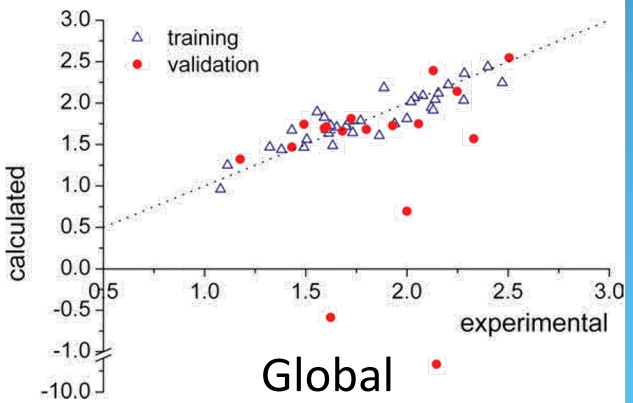
Nitramines



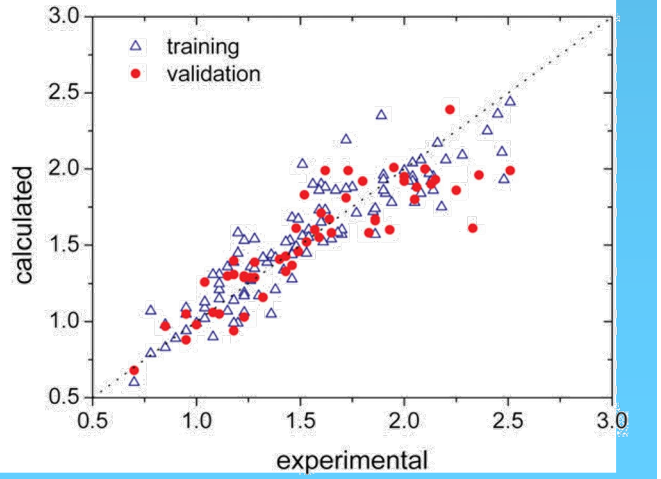
Nitroaliphatic



Nitroaromatic



Global

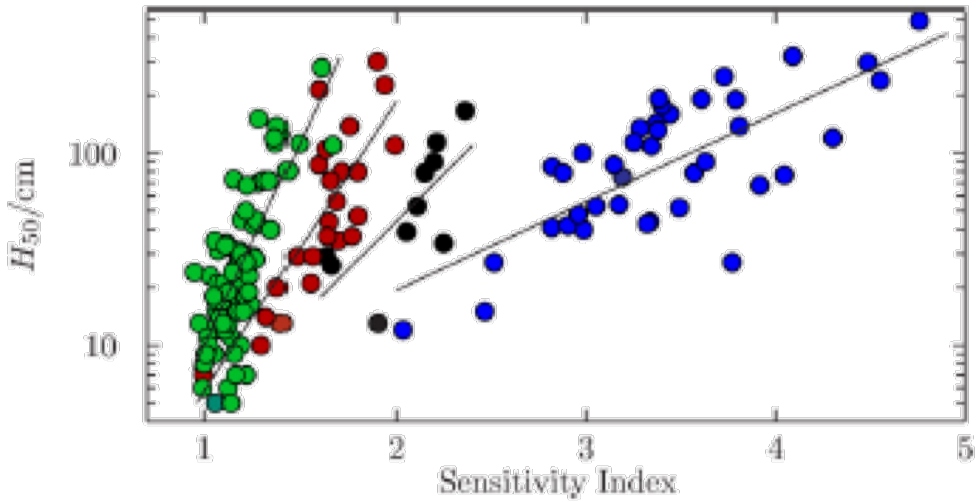


- Good models for nitramine and nitroaliphatic but not for nitroaromatic compounds

Fayet et al., Process Safety Progress (Vol.31, No.3)

Prediction of Impact Sensitivity for Energetic Materials

- Prediction of impact sensitivity of nitro compounds from “physical principles”
- Sensitivity Index (SI)
  - Number of atoms
  - Dissociation energy of the weakest X-NO_2 bond
  - Energy released upon the decomposition of 1 mole of compound

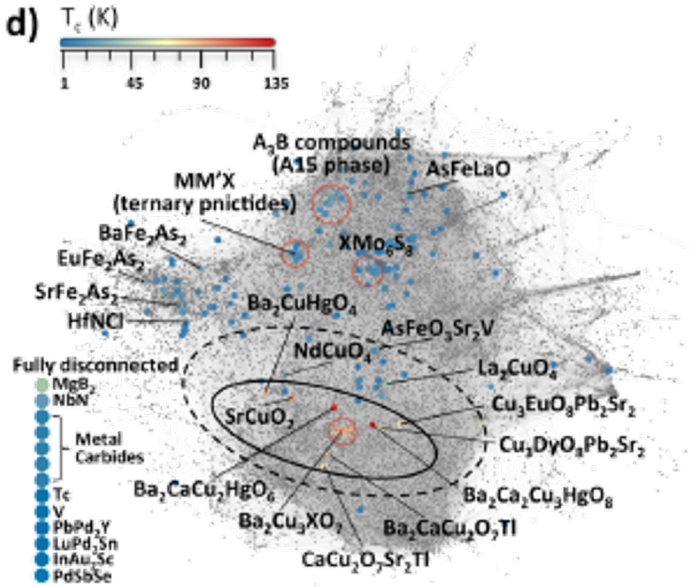


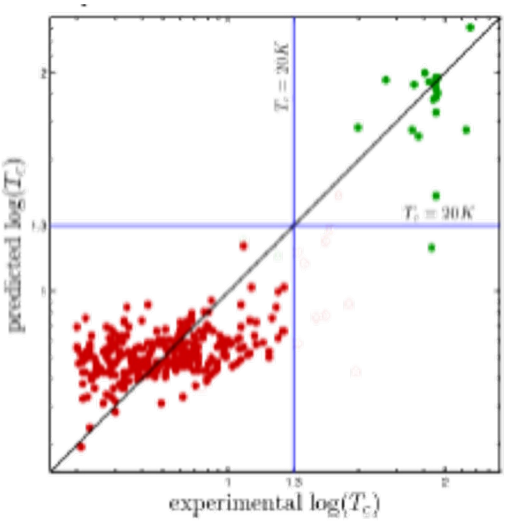
Mathieu, J. Phys. Chem. A 2013, 117, 2253−2259

Material Cartography

- Purpose
  - Displaying material space (AFLOWLIB)
  - Similarity-based Identification of specific materials
  - QMSPR models
- Descriptors
  - Band structure fingerprints
  - SiRMS (fragment-like)
  - QM
- Methods
  - Clustering, RF, PLS

Super-Conductivity Critical Temperature





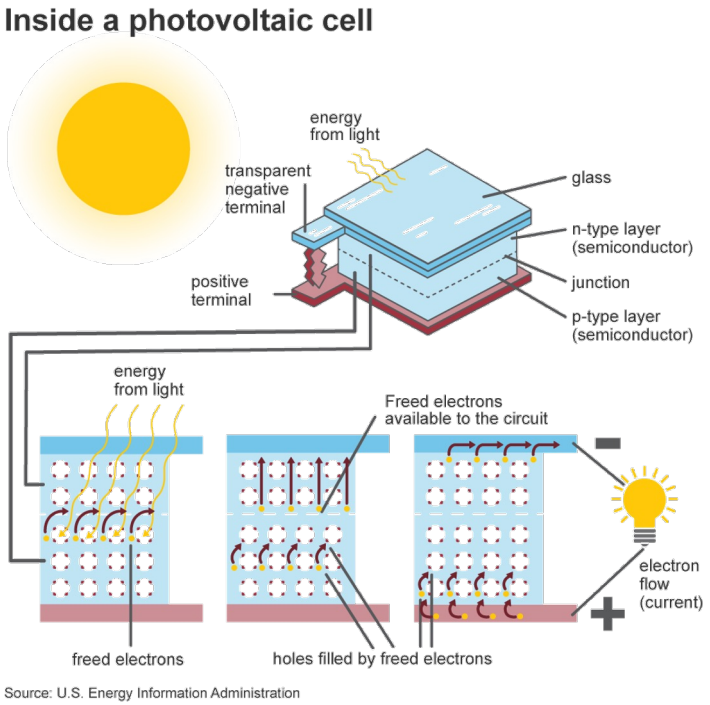
Isayev et al., Chem. Mater. 2015, 27, 735−743

60 Seconds on Solar Cells

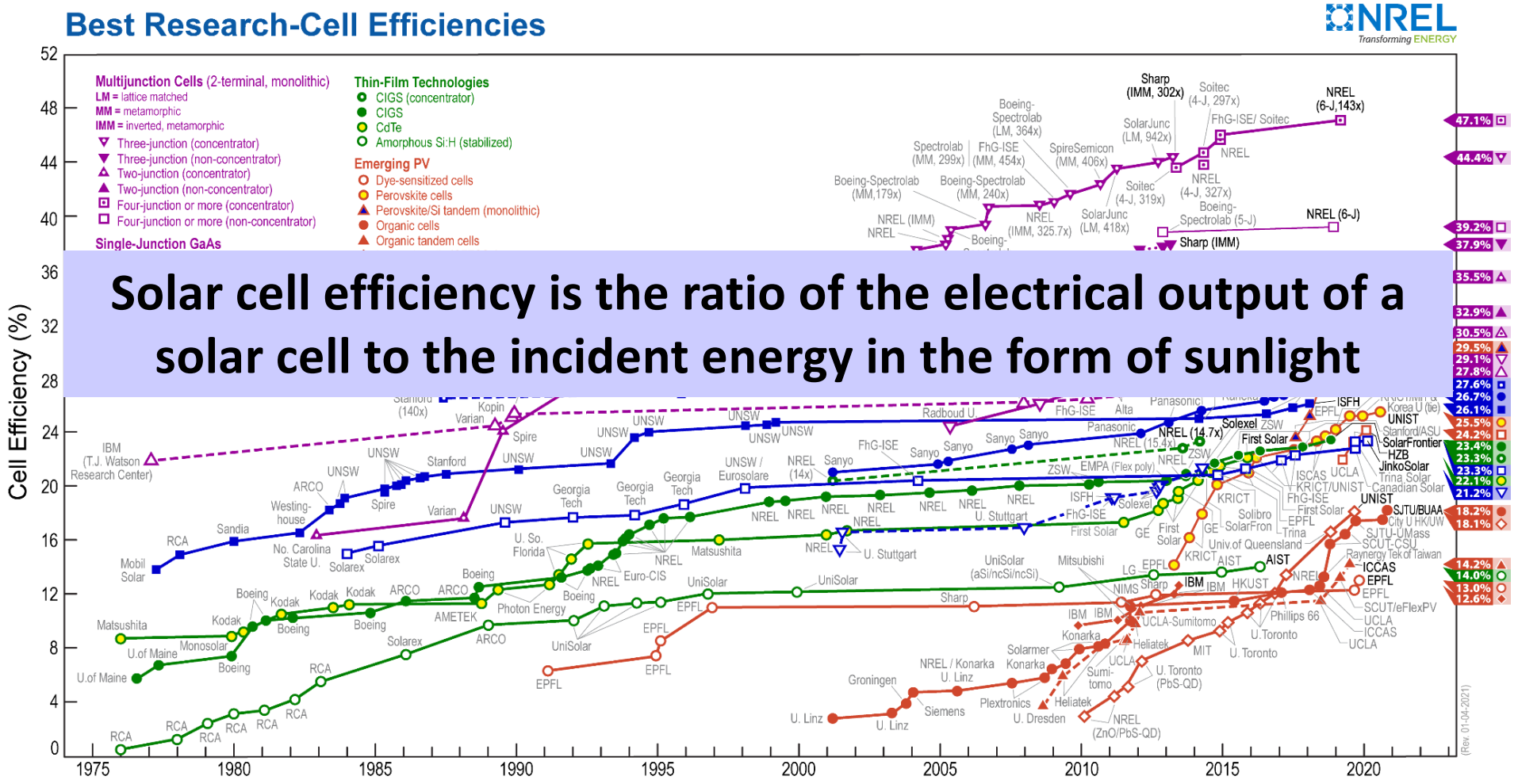
1. Generation of the charge carriers (electrons and holes) due to the absorption of photons
2. Separation of the photo-generated charge carriers in a junction via *n*-type (high electron conductivity) and *p*-type (high hole conductivity) semi-conductors
3. collection of the photo-generated charge carriers at the termini of the junction

- Key Parameters
  - Open circuit voltage (V_{OC})
  - Short circuit current (J_{SC})
  - Internal quantum efficiency (IQE)
  - Fill factor (FF)
  - Power Conversion Efficiency (PCE)

$$= \frac{FF \times J_{SC} \times V_{OC}}{P_{in}}$$



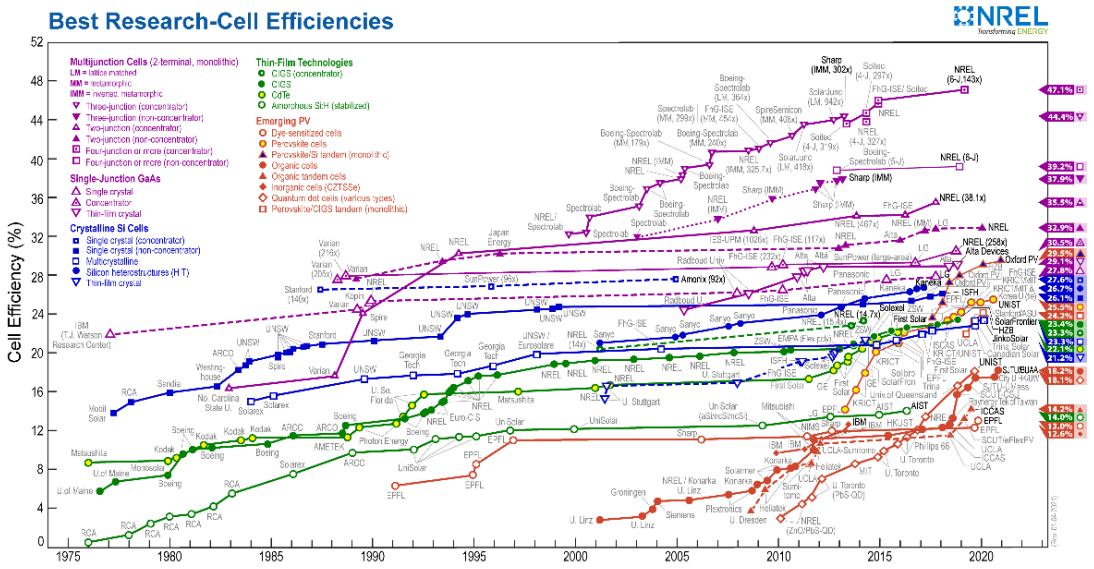
Efficiencies of Solar Cells



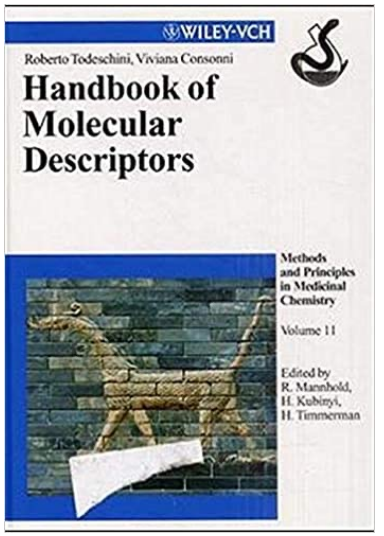
Solar cell efficiency is the ratio of the electrical output of a solar cell to the incident energy in the form of sunlight

Statistical Modeling of Solar Cells

- Goals
  - Identify factors responsible for PV properties
  - Build predictive model for PV properties
  - Experimental design

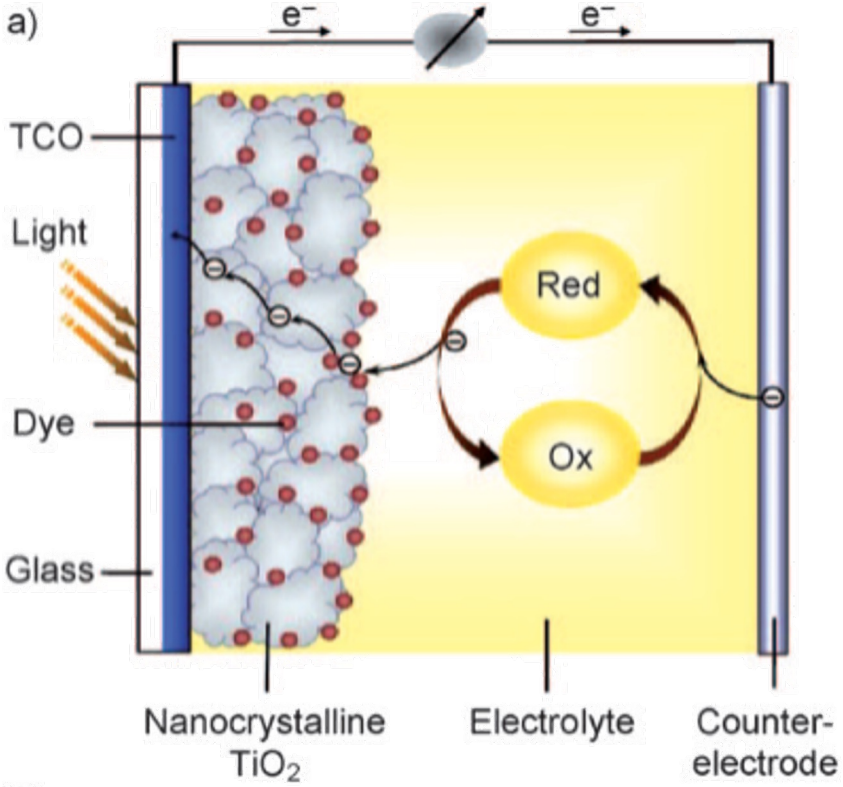


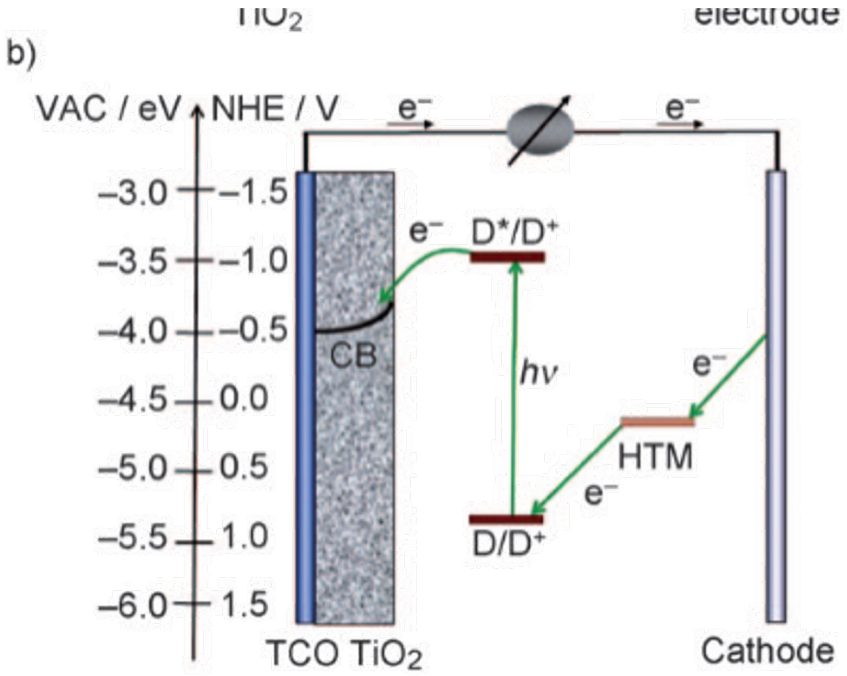
$$= f$$



Dye Sensitized Solar Cells (Grätzel Cells)

- Illumination → photon absorbed by dye
- Photo-excited electrons from sensitized dye transferred to TiO_2
- Dye regenerated by electrolyte
- Electrolyte reduced by counter electrode

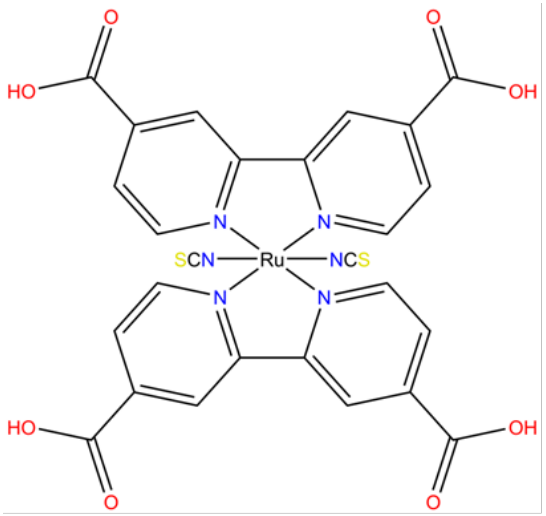




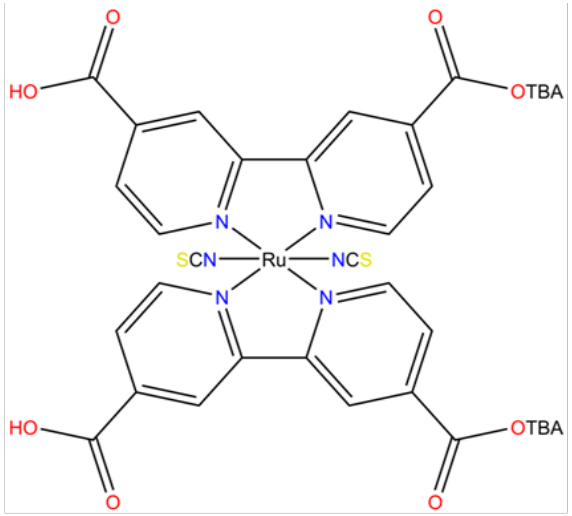
The Importance of the Dye

- Photon Harvesting
- Electron Injection
- Overall cell performances

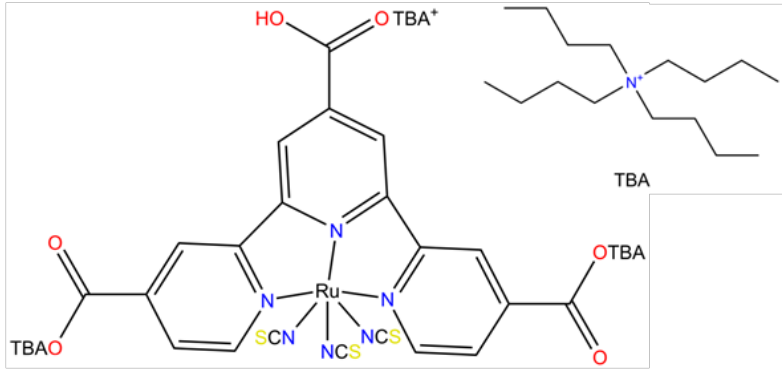
- Metal-based dyes (Ruthenium)
  - Rare and expensive metals
  - Complex synthesis
  - Low molar extinction coefficients
    (How strongly a compound absorbs light at a particular wave-length)



N3



N719



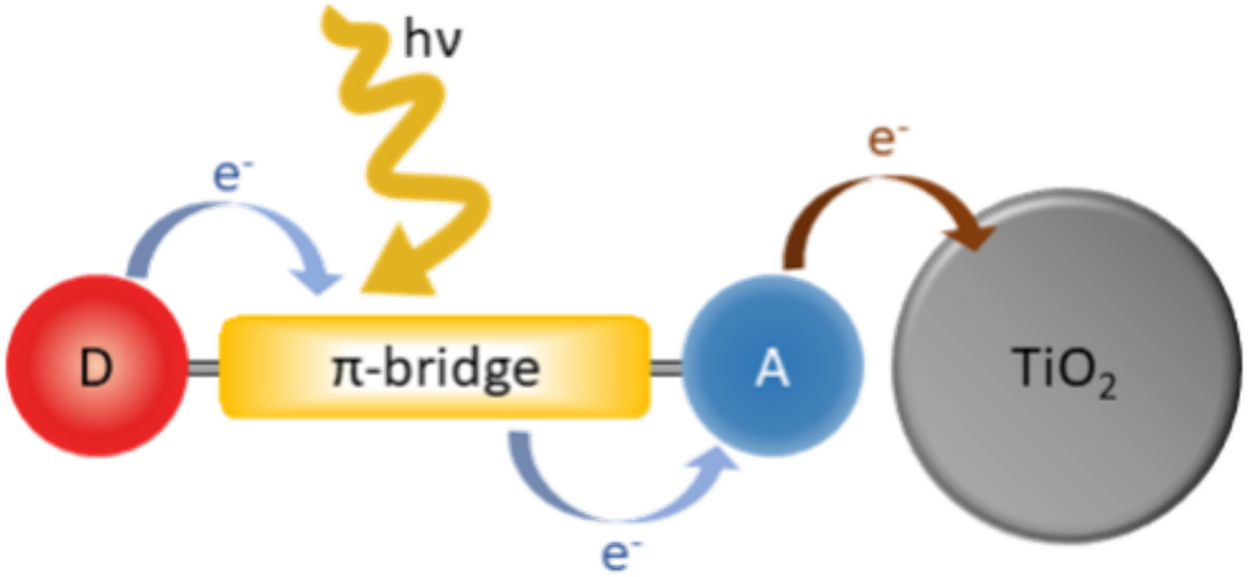
Black dye

Binyamin and Senderowitz, NPJ computational Materials, 8, 142 (2022)

The Importance of the Dye

- Metal-free dyes
  - Inexpensive
  - Easy synthesis
  - High molar extinction coefficients
  - Tunable photovoltaic and electrochemical properties

  - The main disadvantage: lower PCE (power conversion efficiency)



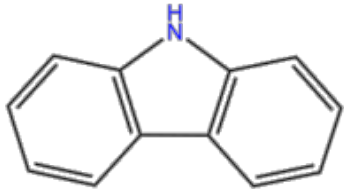
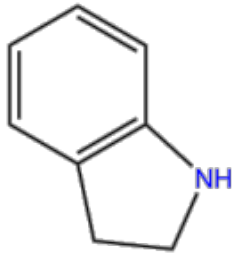
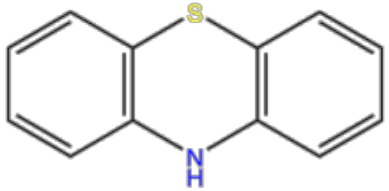
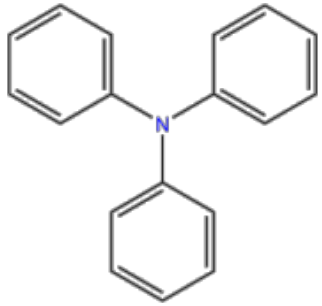
The Dye Sensitized Solar Cells Database

- Over 4000 entries of experimentally tested DSSCs
- Preprocessing:
  - Only metal-free dyes
  - No duplicates
  - Exclude cases with co-sensitizers & co-adsorbents
  - Uniform testing conditions
- Resulting database: ~1400 entries

Scaffold	Number of entries
Triphenylamine	621 (42.44%)
Phenothiazine	270 (18.46%)
Carbazole	182 (12.44%)
Indoline	115 (7.86%)
Coumarin	53 (3.62%)
Diphenylamine	33 (2.26%)
Bodipy	16 (1.09%)
Imidazole	14 (0.96%)
Cyanine	14 (0.96%)

Venkatraman et al., J Cheminform. 2018, 10(1):18

The Dye Sensitized Solar Cells Database

Scaffold	Structure	'Actives' set	'Decoys' Set
Carbzole		Carbazole dyes subset (165 compounds)	ZINC carbzole library (5916 compounds)
Indoline		Indoline dyes subset (115 compounds)	ZINC indoline library (19620 compounds)
Phenothiazine		Phenothiazine dyes subset (254 compounds)	ZINC phenothiazine library (392 compounds)
Triphenylamine		Triphenylamine dyes subset (520 compounds)	ZINC triphenyamine library (488 compounds)

Photovoltaphores: Pharmacophores for Dyes

A pharmacophore can be considered as the highest common denominator of a group of molecules exhibiting a specific “activity”

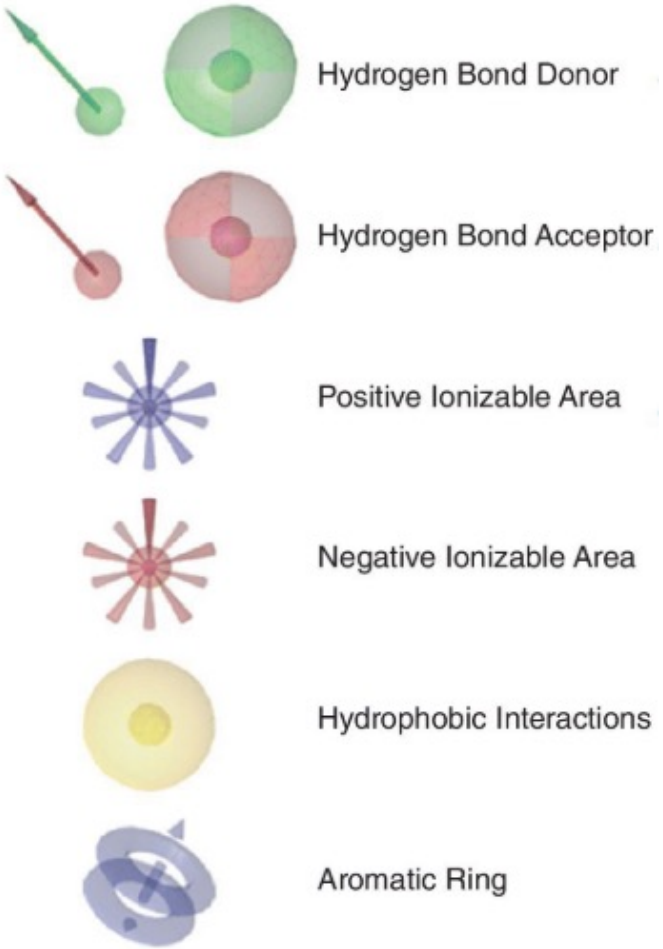
Conformational Search
- Up to 200 conformers

↓

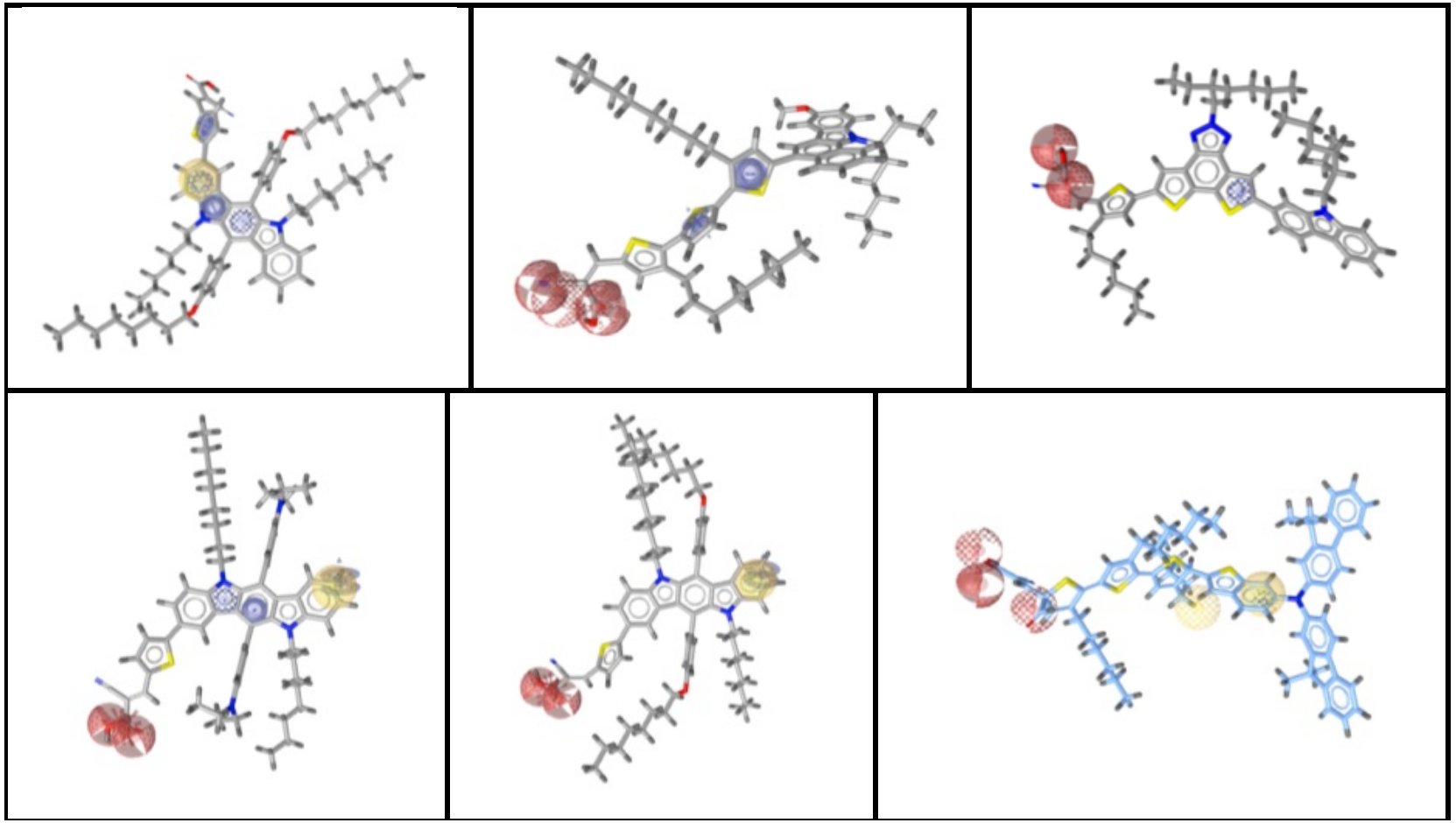
Clustering
- RDF similarity
- Pharmacophore alignment score

↓

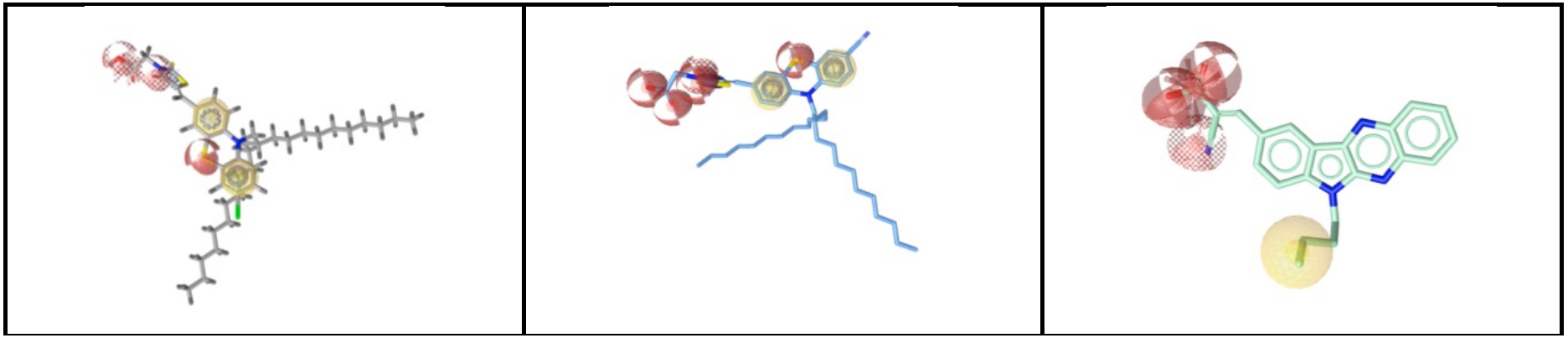
Model generation
- Scaffold specific
- Scaffold agnostic (most active/inactive dyes)



Photovoltaphores from Active Dyes

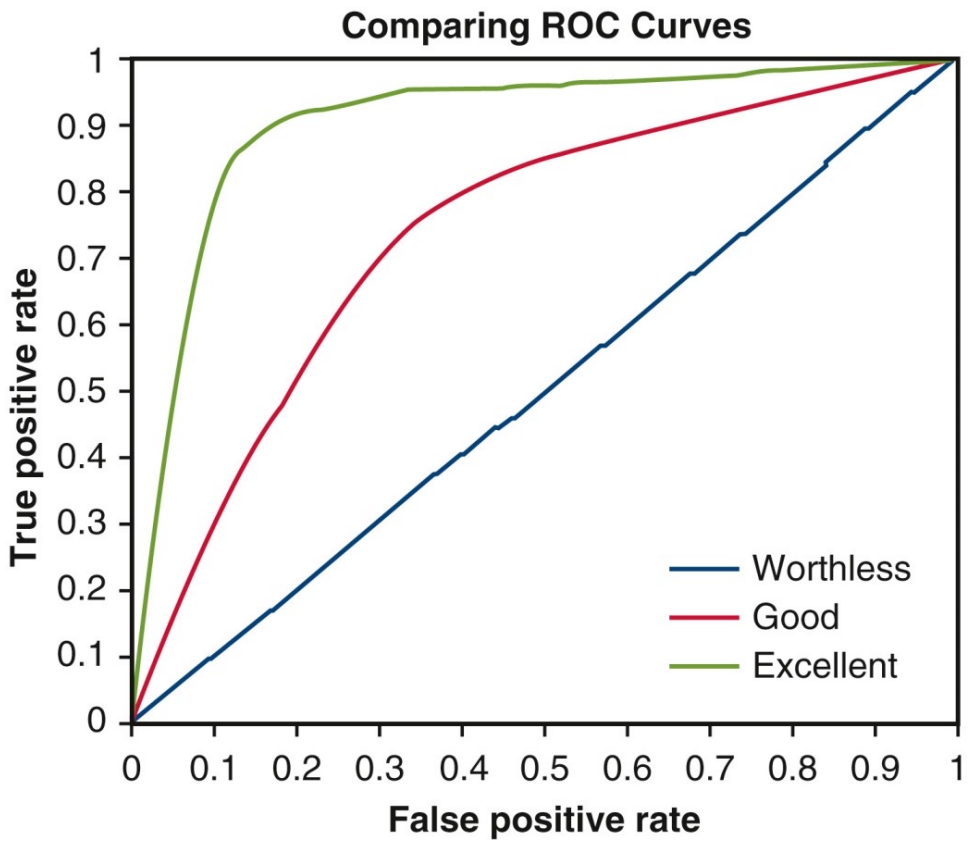


Photovoltaphores from Inactive Dyes

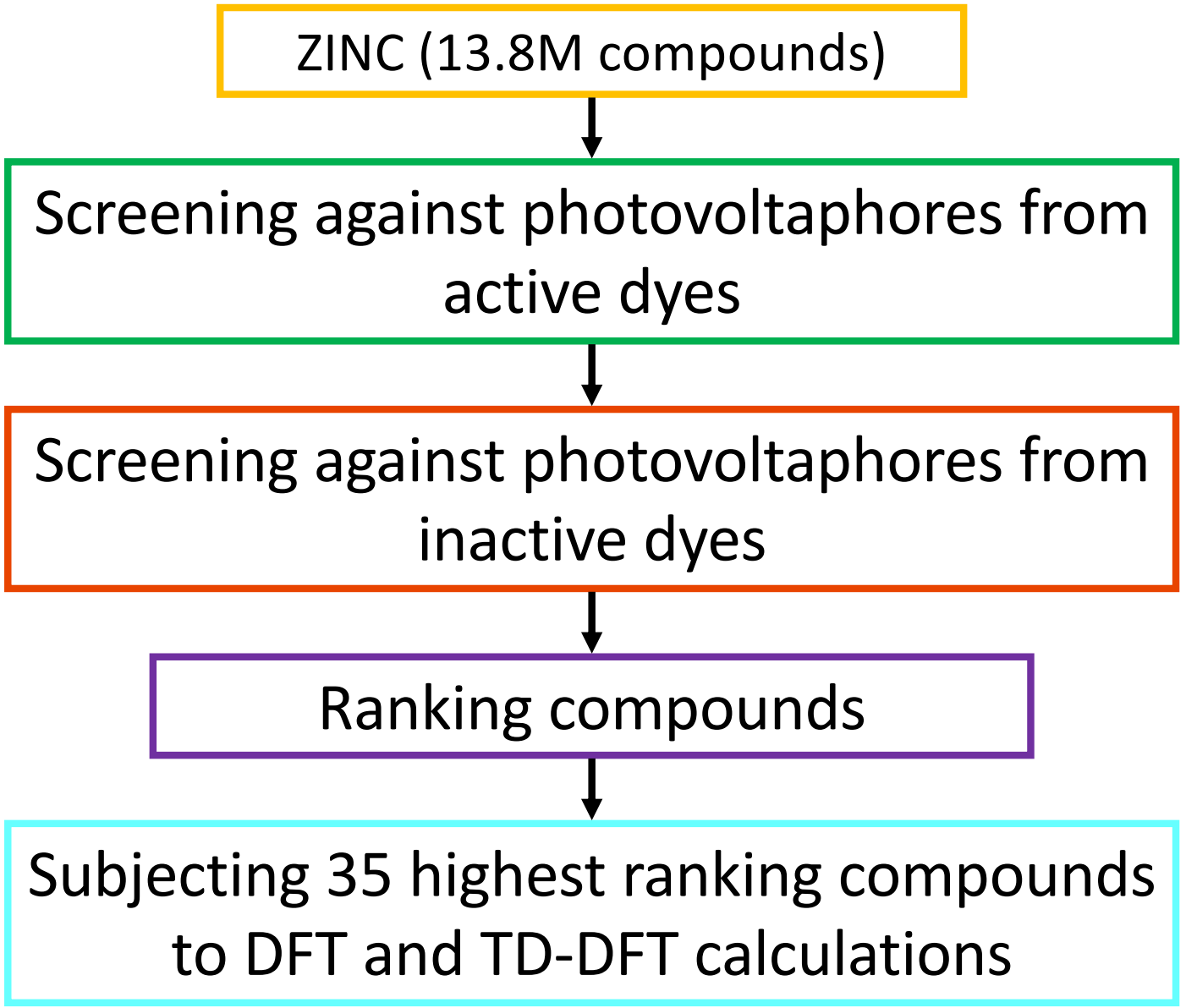


Model Validation

- Model tested using small-scale VS campaign
  - Active compounds from DSSCDB
  - Inactive compounds from ZINC (similar scaffolds)
- Models validated using ROC curve and numerical metrics

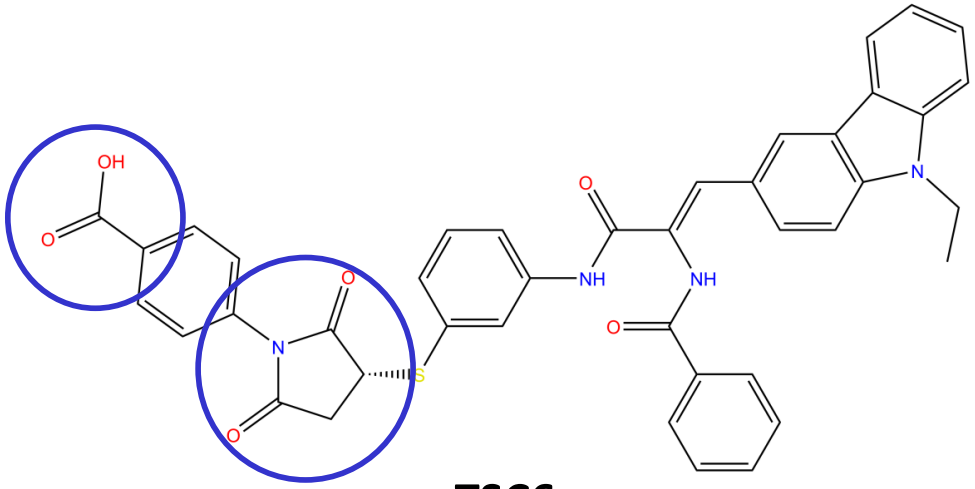


Virtual Screening

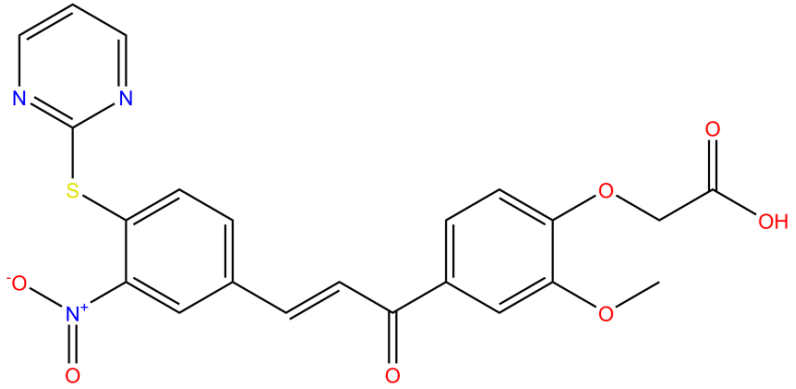


Virtual Hits

- HOMO/LUMO
  - DFT
  - B3LYP
  - 6-31G(d,p)
- UV-vis absorption spectrum
  - TD-DFT
  - CAM-B3LYP
  - 6-31G(d,p)



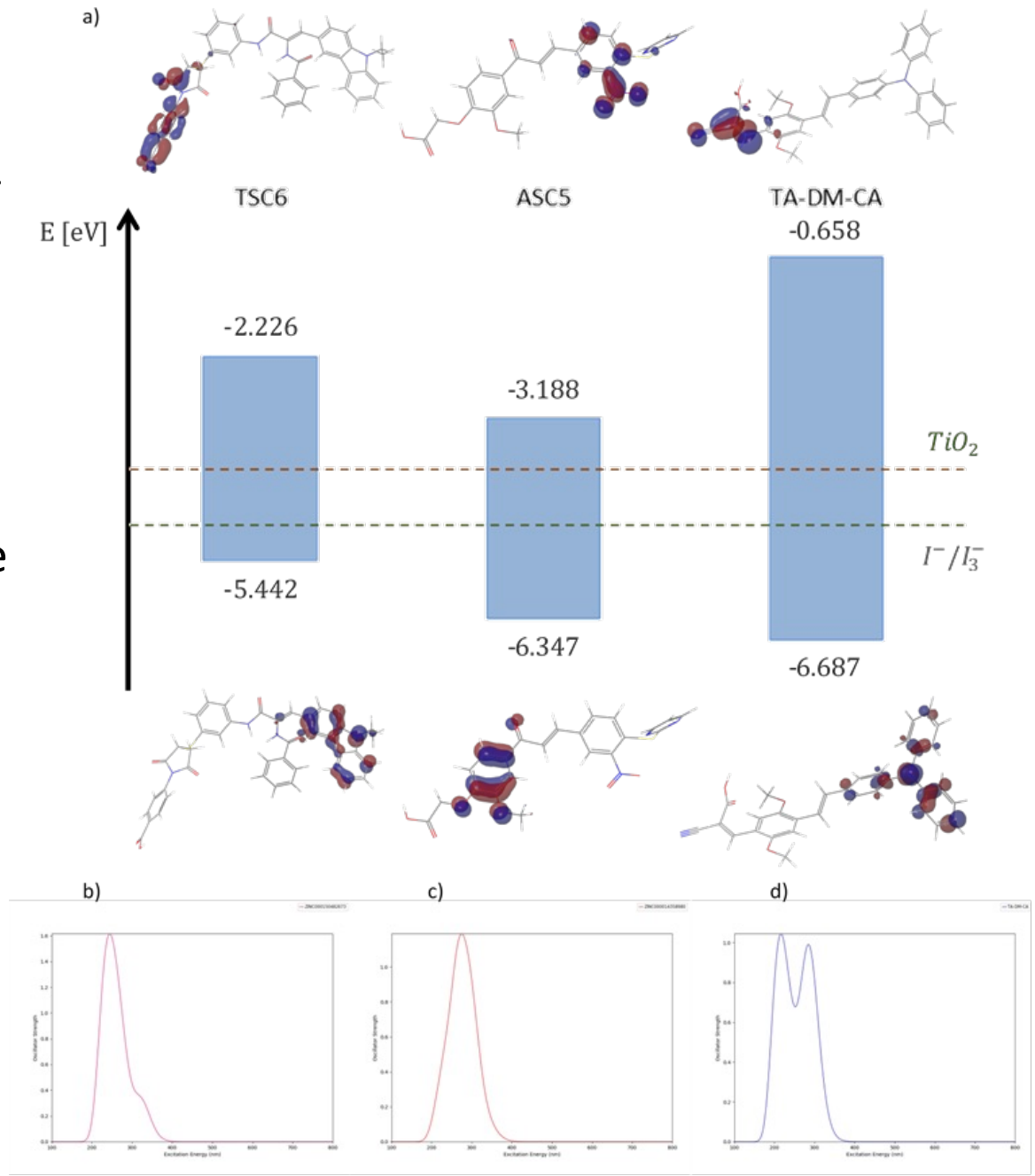
TSC6
(ZINC000150482673)



ASC5
(ZINC000014358980)

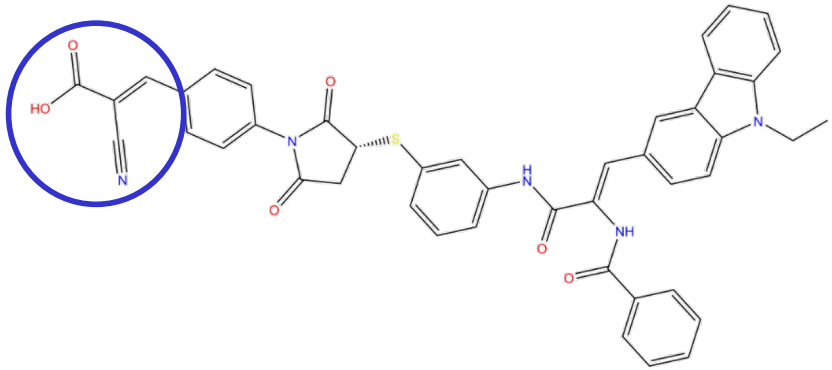
Virtual Hits

- Luminescent, absorbs in UV-vis and NIR regions, high molar extinction coefficient
- HOMO remote from semi-conductor's conduction band and below the energy level of the redox electrolyte
- LUMO close to semiconductor's surface, and higher than the semiconductor conduction band potential
- Hydrophobic periphery in order to enhance the cell's stability

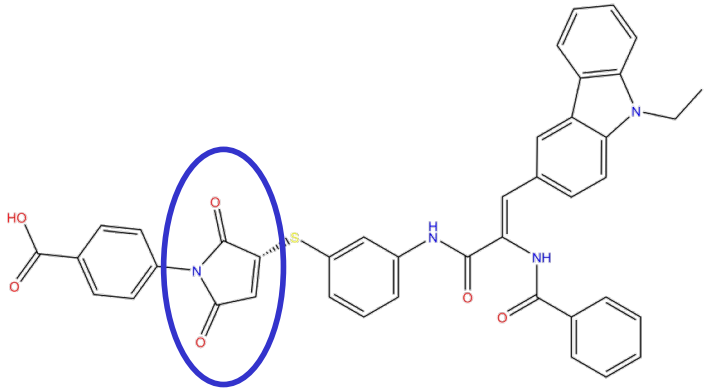


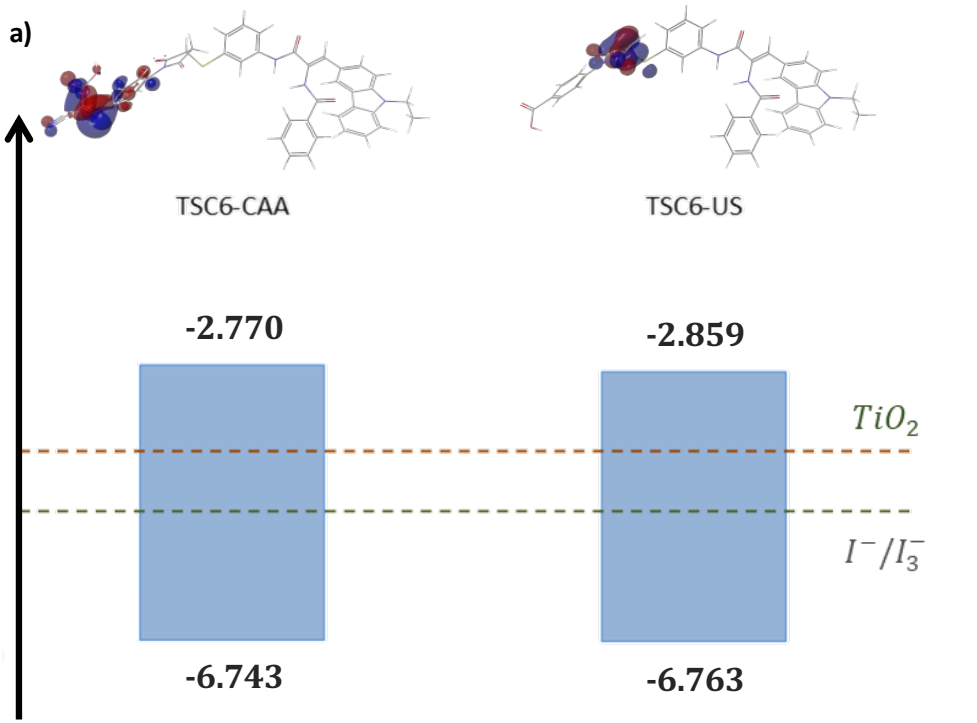
Virtual Hits

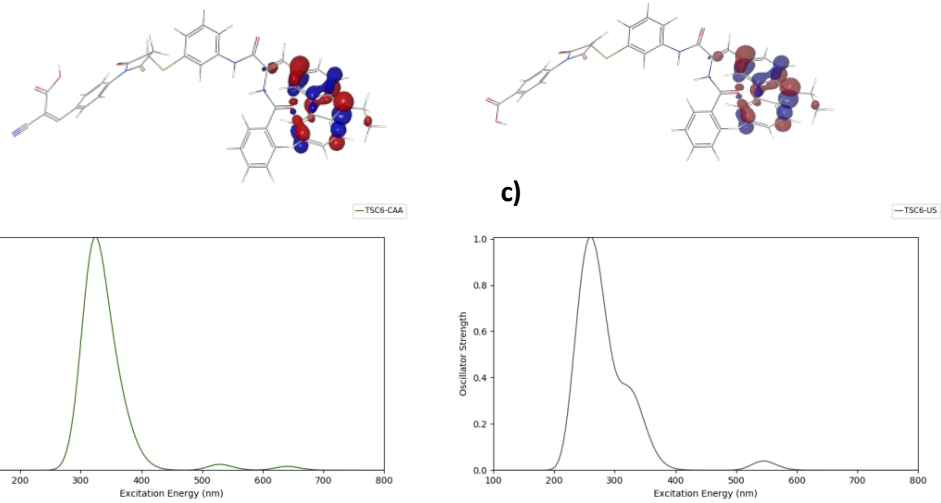
- Carboxyl group on phenyl → cyanoacrylic acid



- Succinimide unit → pyrroledione moiety

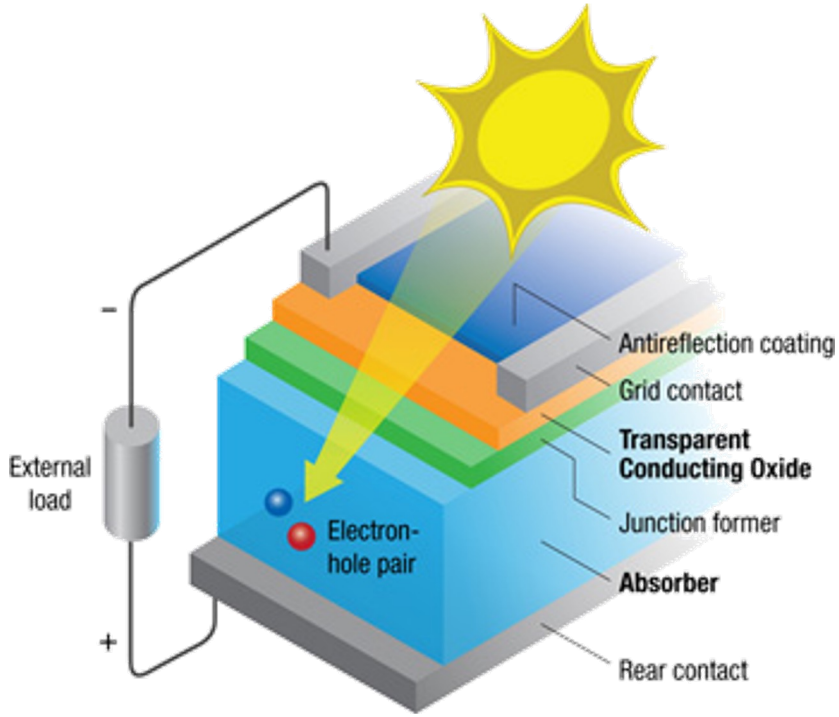






Solar Cells Based on Metal Oxides (MO)

- Material
  - Abundant
  - Environmentally safe
  - Optimizeable
  - Low cost
- Fabrication
  - Cheap fabrication methods
- Operation
  - Long term operation (stability)

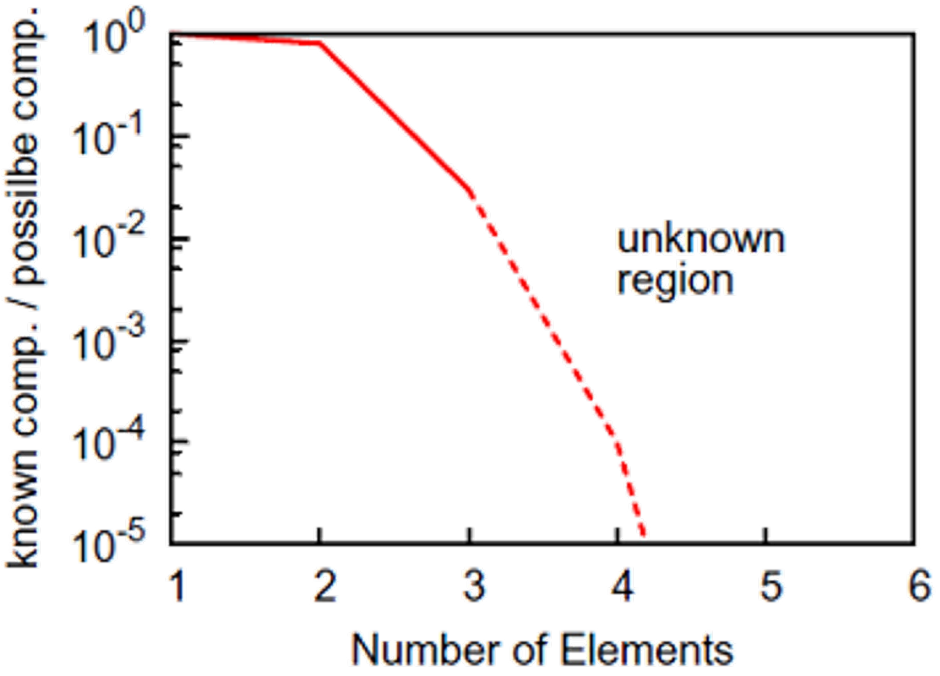


But Cells Not Efficient Enough

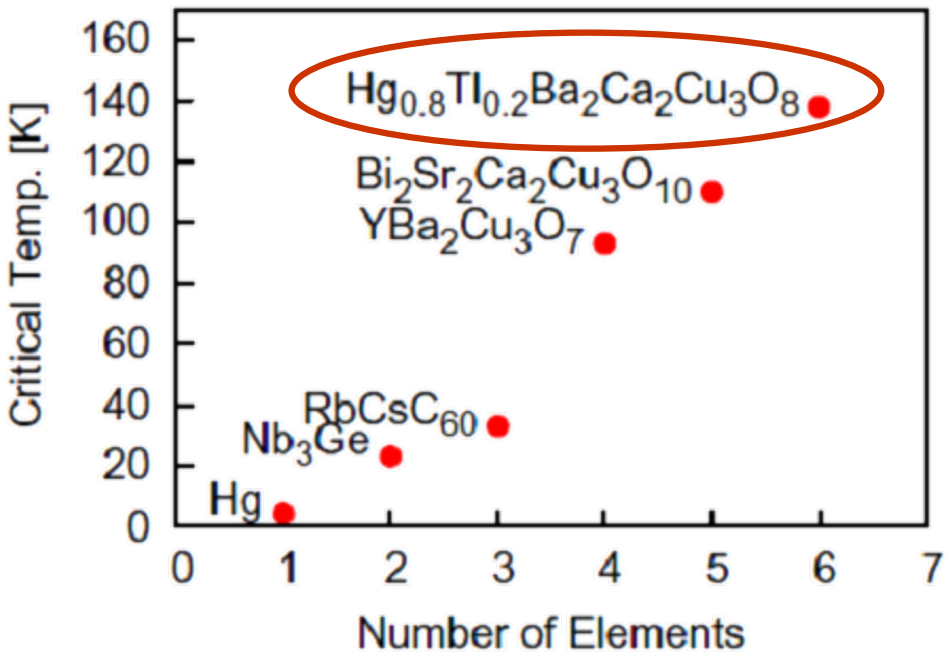
New Metal Oxides (MO) Required

Combinatorial Material Science

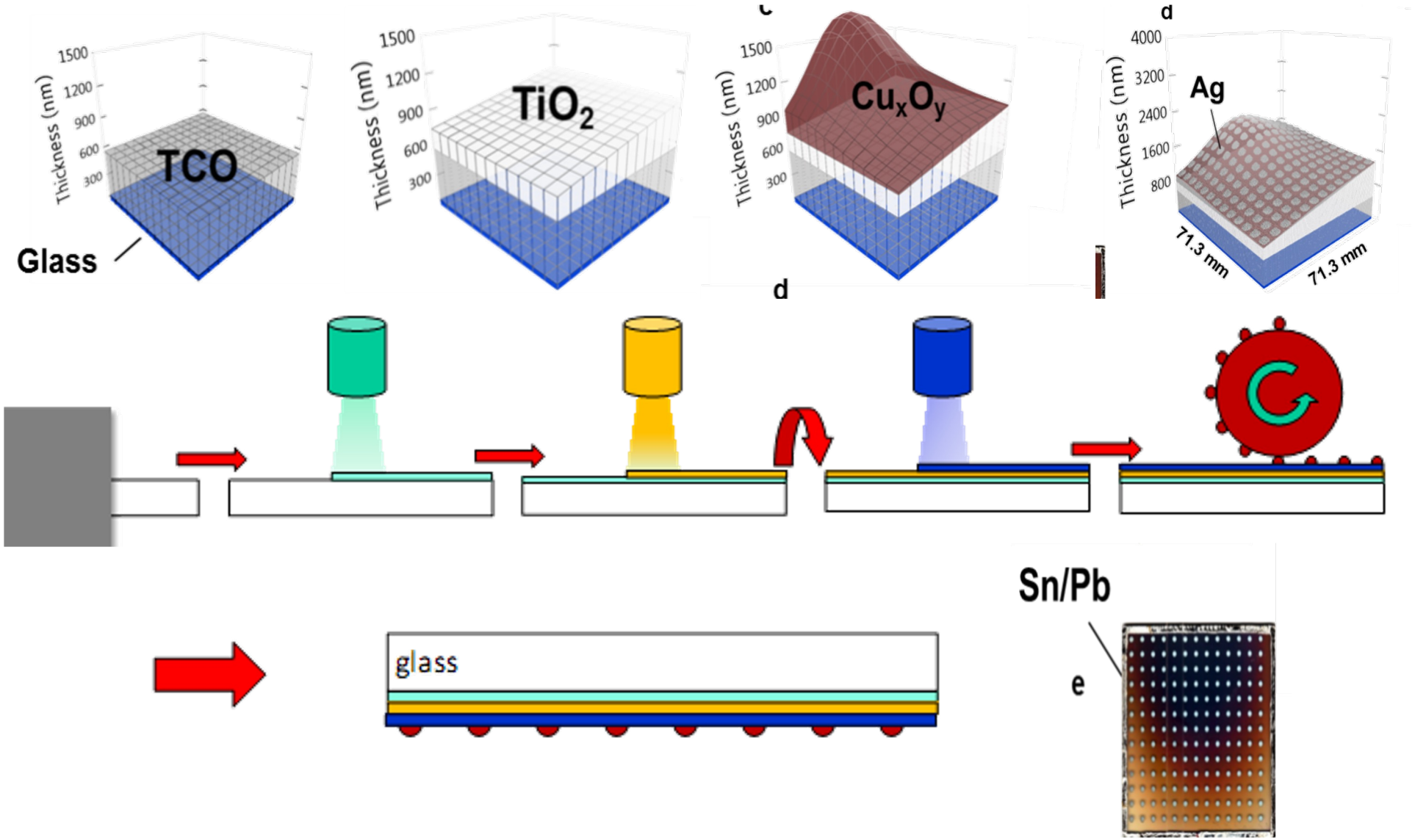
- ~60 “useful” elements leading to
  - ~30K inorganic compounds
  - 3600 binary compounds (ABO_X); mostly known
  - 216K ternary compounds ($ABCO_X$) many of which unknown



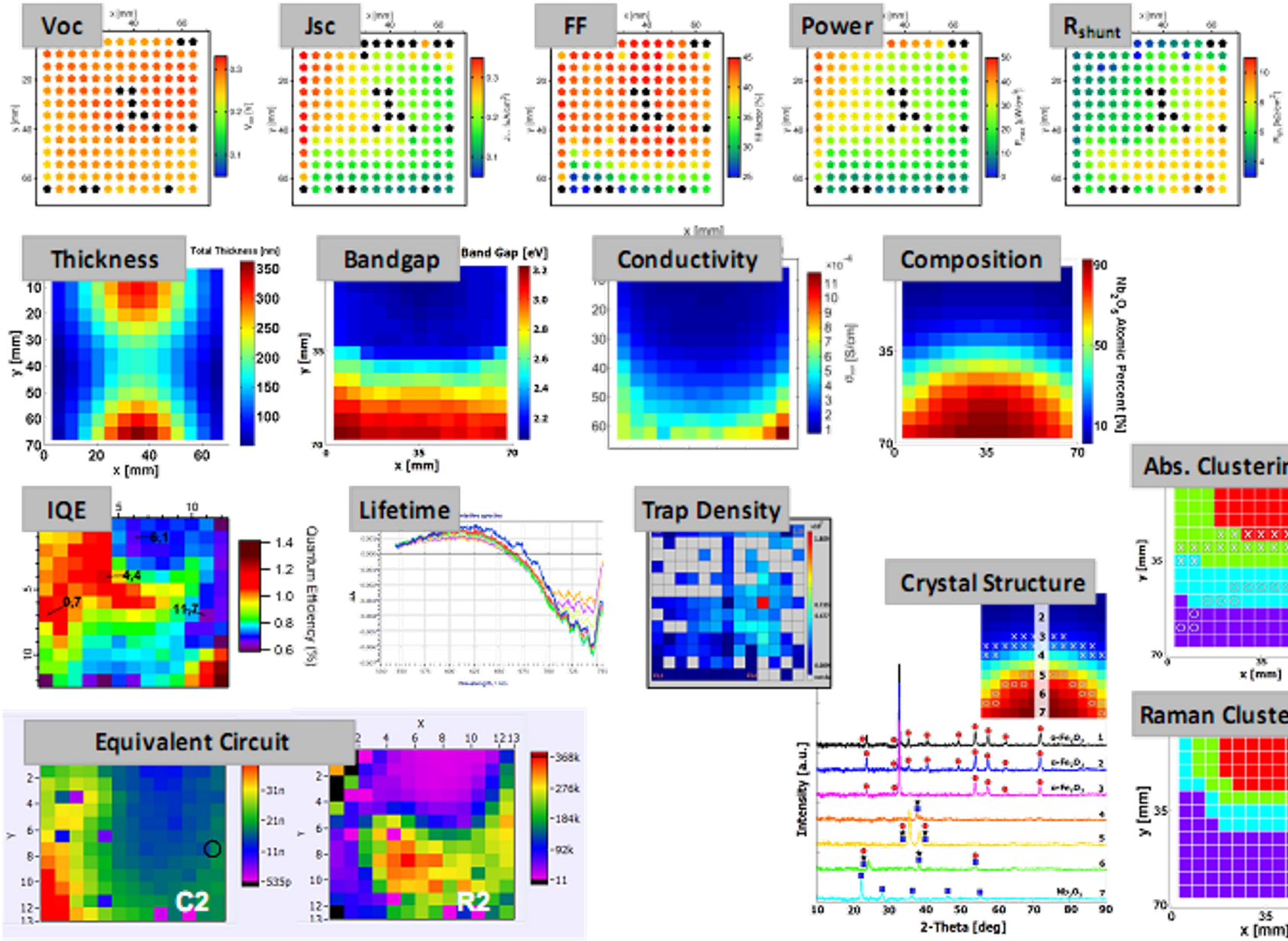
High Tc superconductors



Combinatorial Material Synthesis



Analysis


Voc
Jsc
FF
Power
R_{shunt}
Thickness
Total Thickness [nm]
y [mm]
x [mm]
Bandgap
Band Gap [eV]
Conductivity
Composition
Nb_2O_5 Atomic Percent [%]
IQE
Quantum Efficiency [%]
Lifetime
Trap Density
Abs. Clustering
Crystal Structure
Intensity [a.u.]
2-Theta [deg]
Raman Clustering
Equivalent Circuit
C2
R2

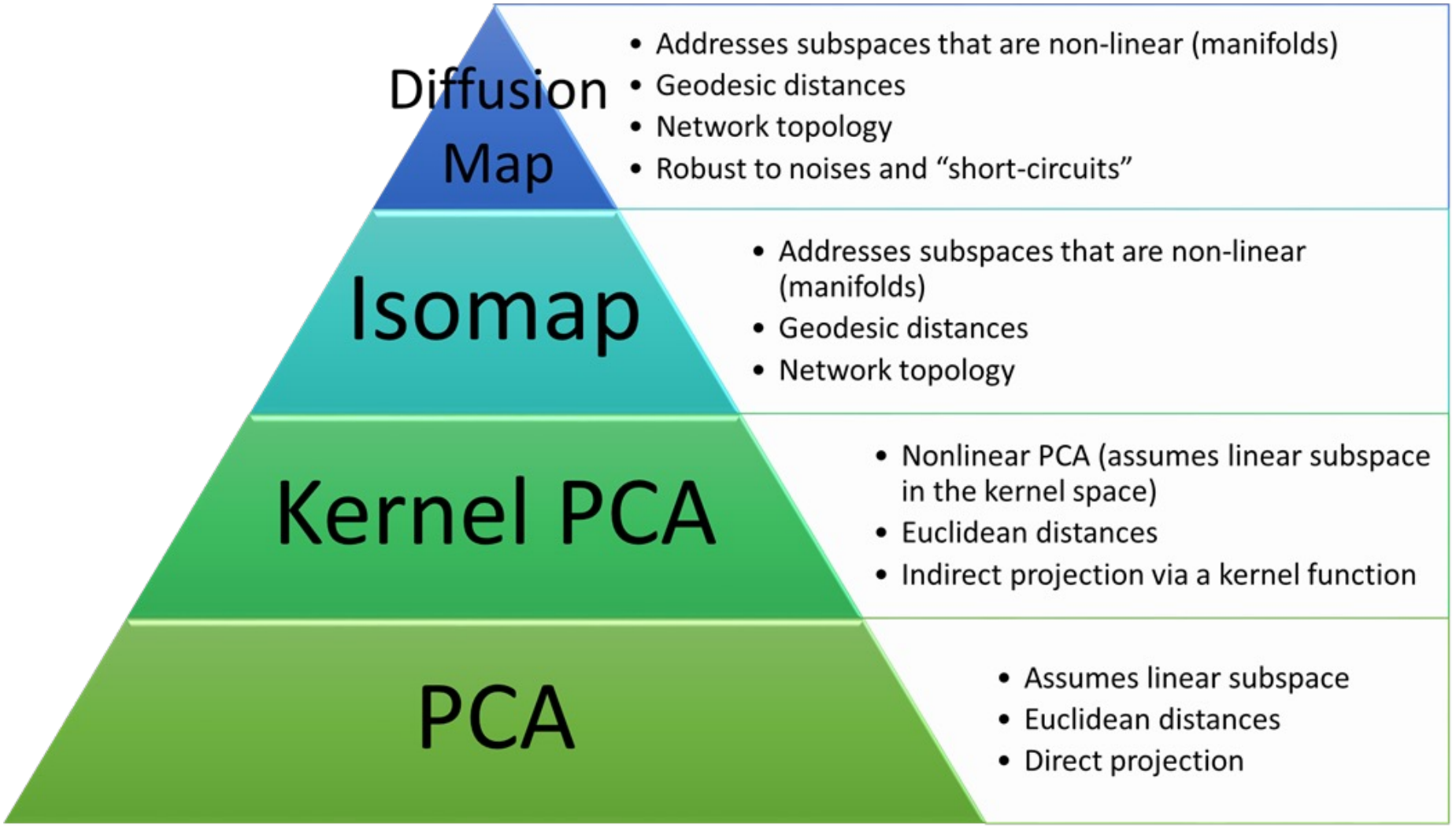
Visualizing the Solar Space

~1350 cells from 5 libraries

	TiO_2||Cu-O	TiO_2||Cu_2O	TiO_2||Co_3O_4||MoO_3	TiO_2||Co_3O_4	TiO_2||CuO-NiO-In_2O_3
# Cells	169	338	169	338	338
V_{OC} *[mV]*	31-380	6.6-354	24-620	172-443	111-509
J_{SC} [μA cm^{-2}]	73-290	13.9-406	25-Oct	5.5-11	Jul-54
P_{max} [μW cm^{-2}]	0.02-1.02	0-1.26	0.0018-0.11	0.016-0.042	0.015-0.11
FF [%]	23-63	0.16-40	23-41	32-53	26-41
R_s [Ohm cm^2]	18-23x10^3	9-0.5x10^6	5x10^3-0.6x10^6	8x10^3-0.5x10^6	5x10^3-0.6x10^6
R_{sh} [Ohm cm^2]	10^3-1.9x10^6	5x10^3-0.6x10^6	0.7x10^6-3.3x10^6	2.8x10^6-12x10^6	0.15x10^6-2.6x10^6
IQE [%]	0.57-1.16	0.1-2.67	0.05-0.3	0.06-0.32	0.02-0.47

Kaspi et al., JCIM 2018 , 58, 2428-2439

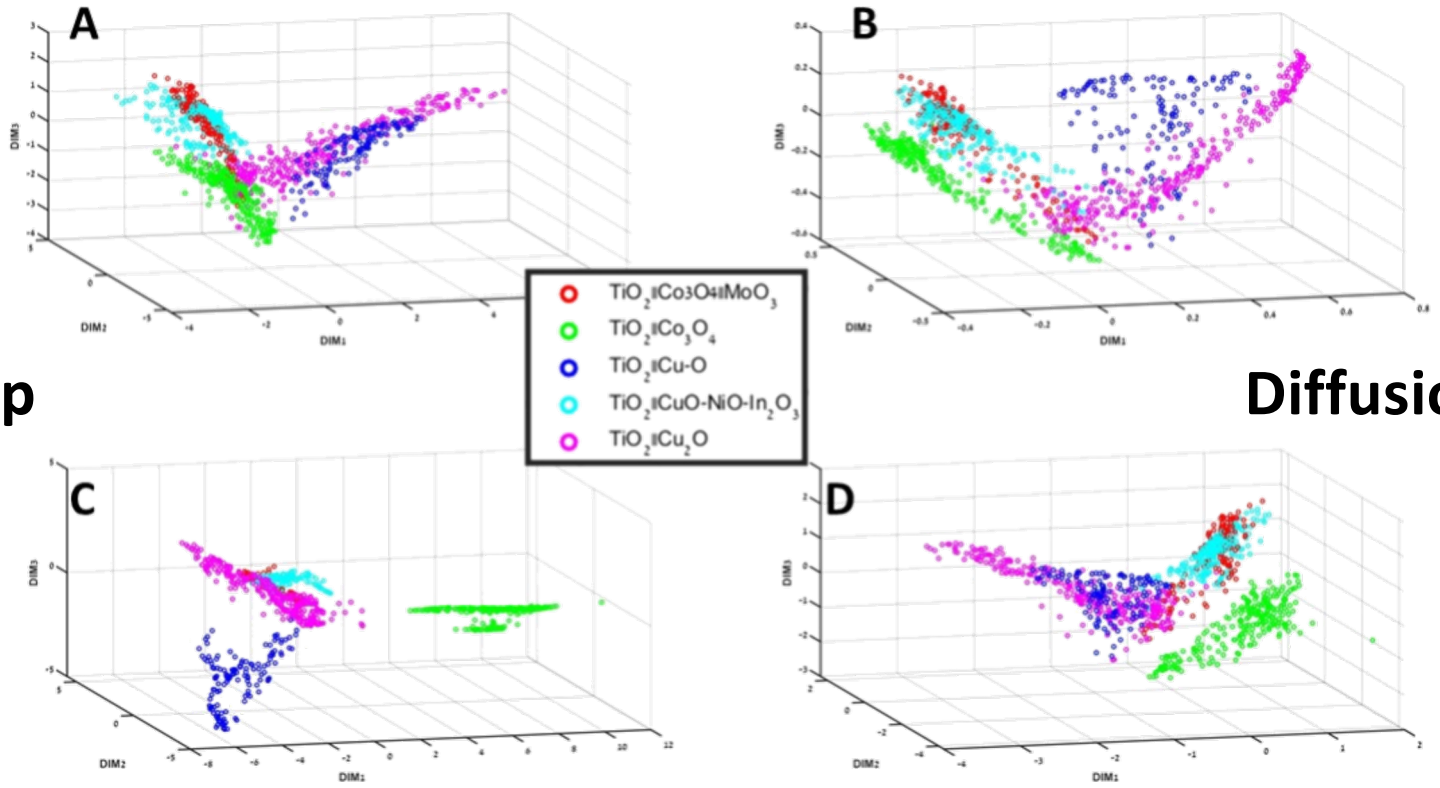
Visualizing the Solar Space



Visualizing the Solar Space

PCA

Kernel PCA



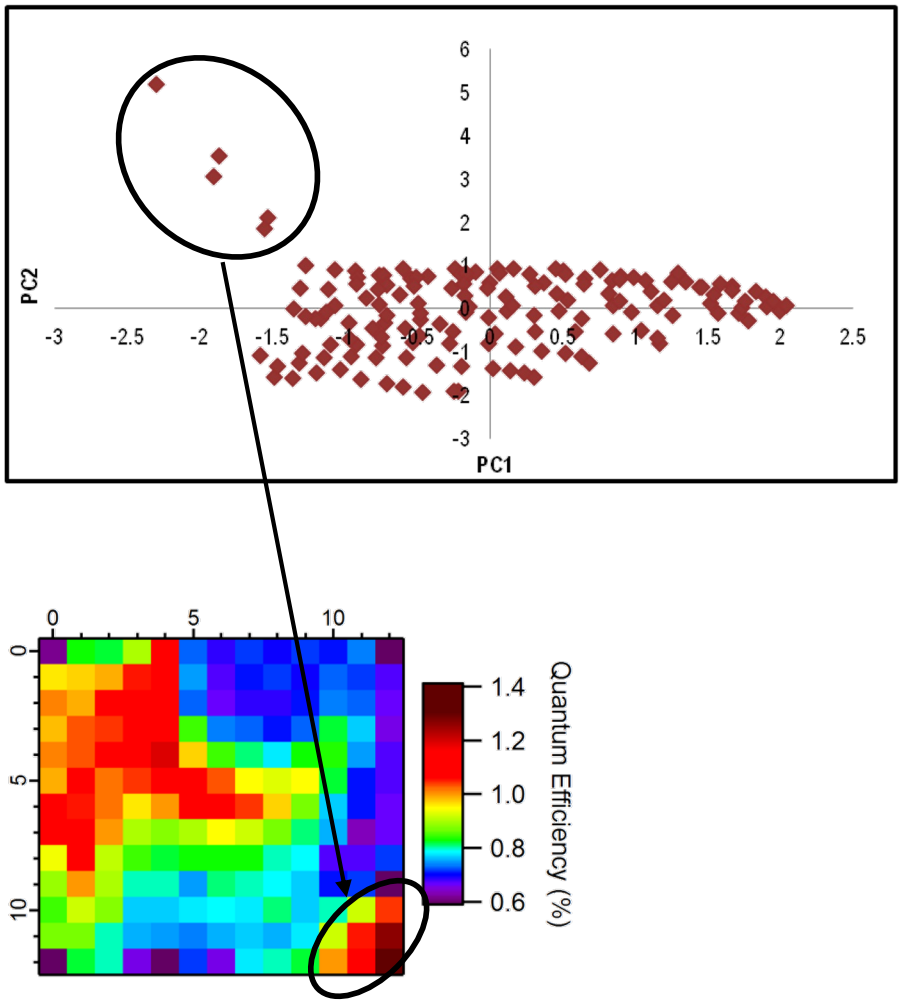
Isomap

Diffusion Map

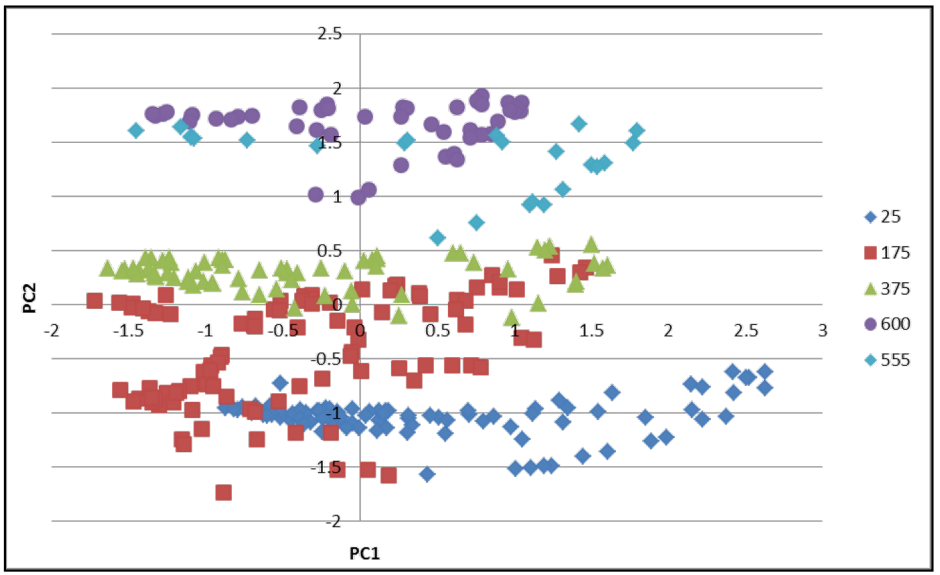
	PCA	Kernel PCA	Isomap	Diffusion map	Original space
Trust	0.82	0.78	0.70	0.78	-----
1NN Classification	0.92	0.91	0.92	0.92	0.95
3NN Classification	0.92	0.90	0.93	0.91	0.95
5NN Classification	0.90	0.90	0.93	0.90	0.94

Principle Component Analysis (PCA)

TiO_2/Cu-O



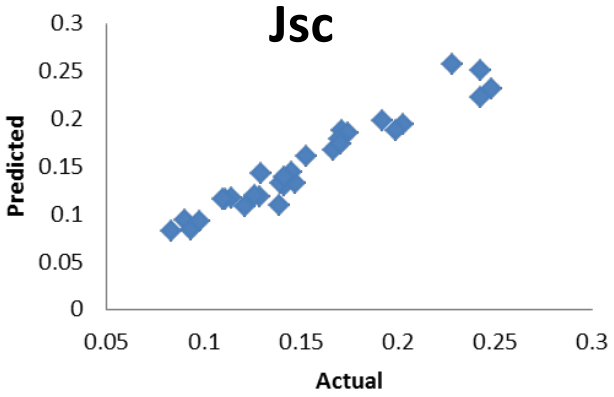
TiO_2/Fe_2O_3
(treated at different temperatures)

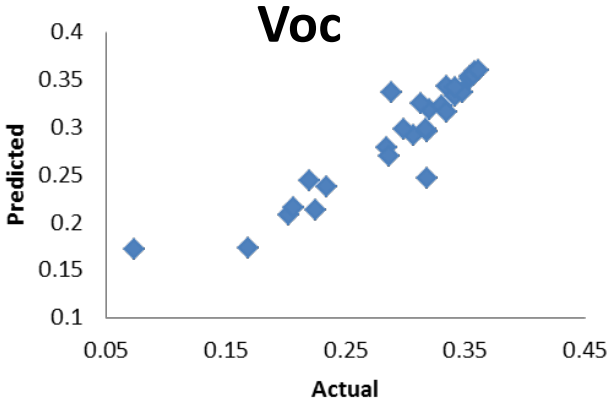


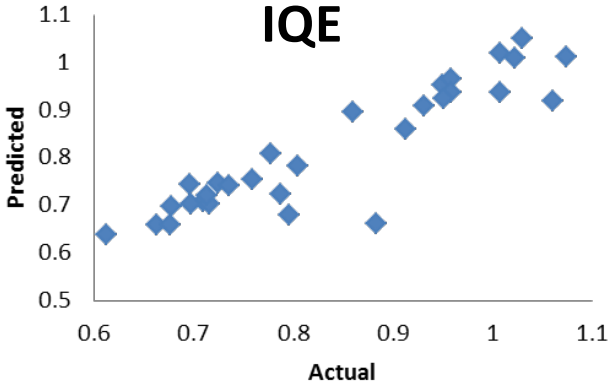
PC1	PC2
d_TiO2	TEMP
d_Fe2O3	BGP_Fe2O3
[Fe2O3] / ([TiO2] + [Fe2O3])	
max_J	

Machine Learning: *k*NN and GFA

End point	Q^2_{LOO}	No applicability domain		With applicability domain			Descriptors
		Q^2_{ext} (R^2)	*MAE*	Q^2_{ext} (R^2)	*MAE*	%coverage	
J_{SC}(Ag)	0.87	0.86 (0.88)	0.01	0.86 (0.89)	0.01	83 %	*Ratio, BGP,* D_{center}
V_{OC}(Ag)	0.86	0.73 (0.74)	0.02	0.75 (0.77)	0.02	75 %	T_{TiO_2}, J_{max}
IQE(Ag)	0.77	0.80 (0.84)	0.05	0.83 (0.86)	0.04	87 %	T_{TiO_2}, *Ratio,* R_a



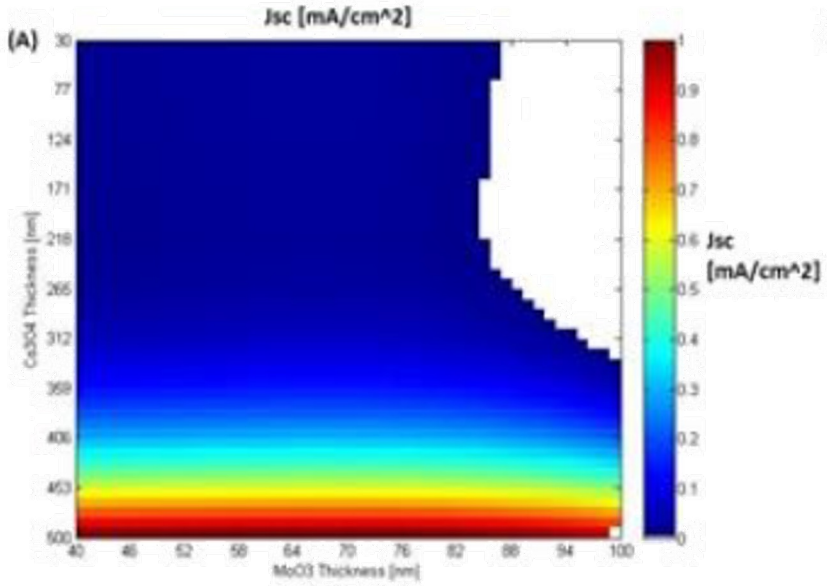


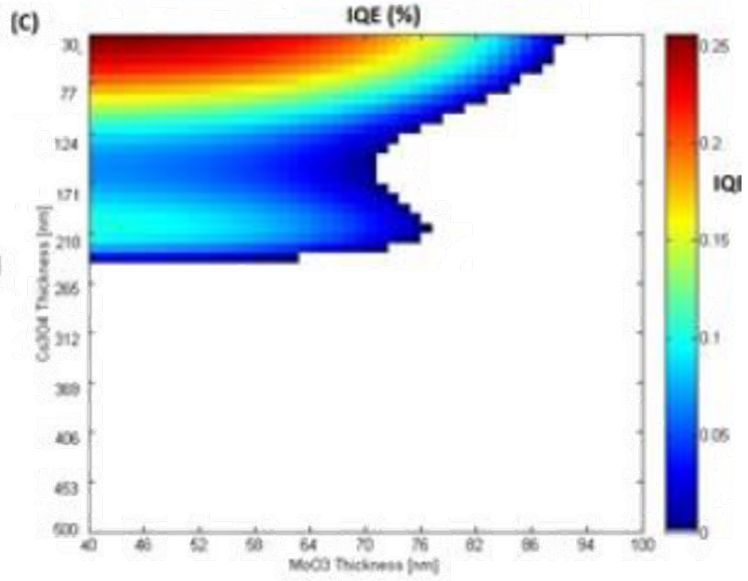


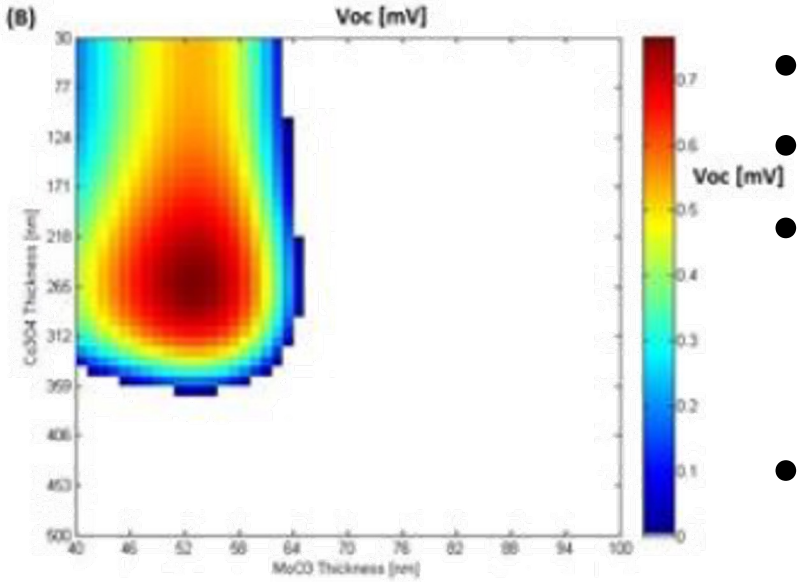
Model	R^2_{CV}	Q^2_{ext} (R^2)	*MAE*
$J_{SC} = 0.062 + 0.0004 \times T_{Cu-O} - 430384.1022/R_a$	0.88	0.86 (0.87)	0.01
$V_{OC} = 0.011 \times J_{max} + 1.201 \times 10^{-5} \times T_{TiO_2} \times D_{center} - 0.04 - 6.62 \times 10^{-13} \times T_{Cu-O} \times R_a$	0.62	0.54 (0.55)	0.03
$IQE = 1.784 \times Ratio + 0.072/Ratio - 2642279.244/(2356681.705 + R_a)$	0.65	0.74 (0.74)	0.06

Yosipof et al., Molecular Informatics, 2015, 34, 367--79

Virtual Cell



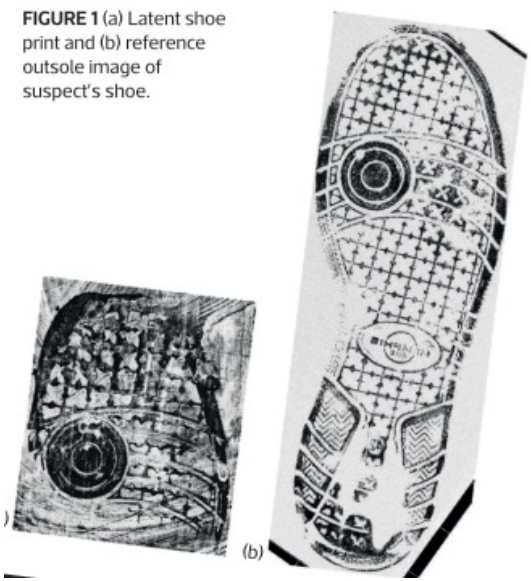




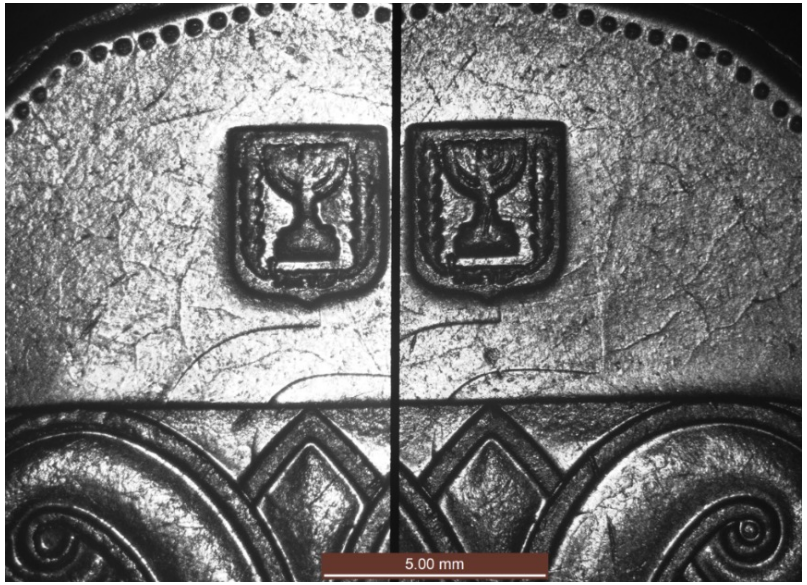
- Expanded range of layers thickness
- Predict and plot activities
- Large parts of the “thickness space” are outside the applicability domain of the models
- Difficulties in producing cells with good V_{OC}, J_{SC} and *IQE* values

Forensic Informatics

shoeprints



faked coins

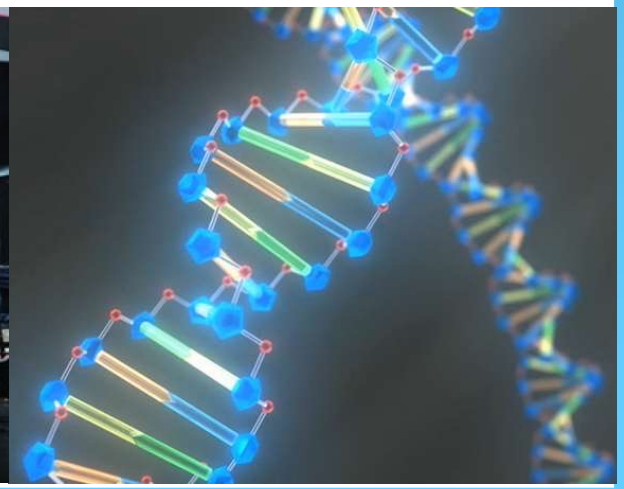


ammunition

gunshot residues

glass fragments

DNA

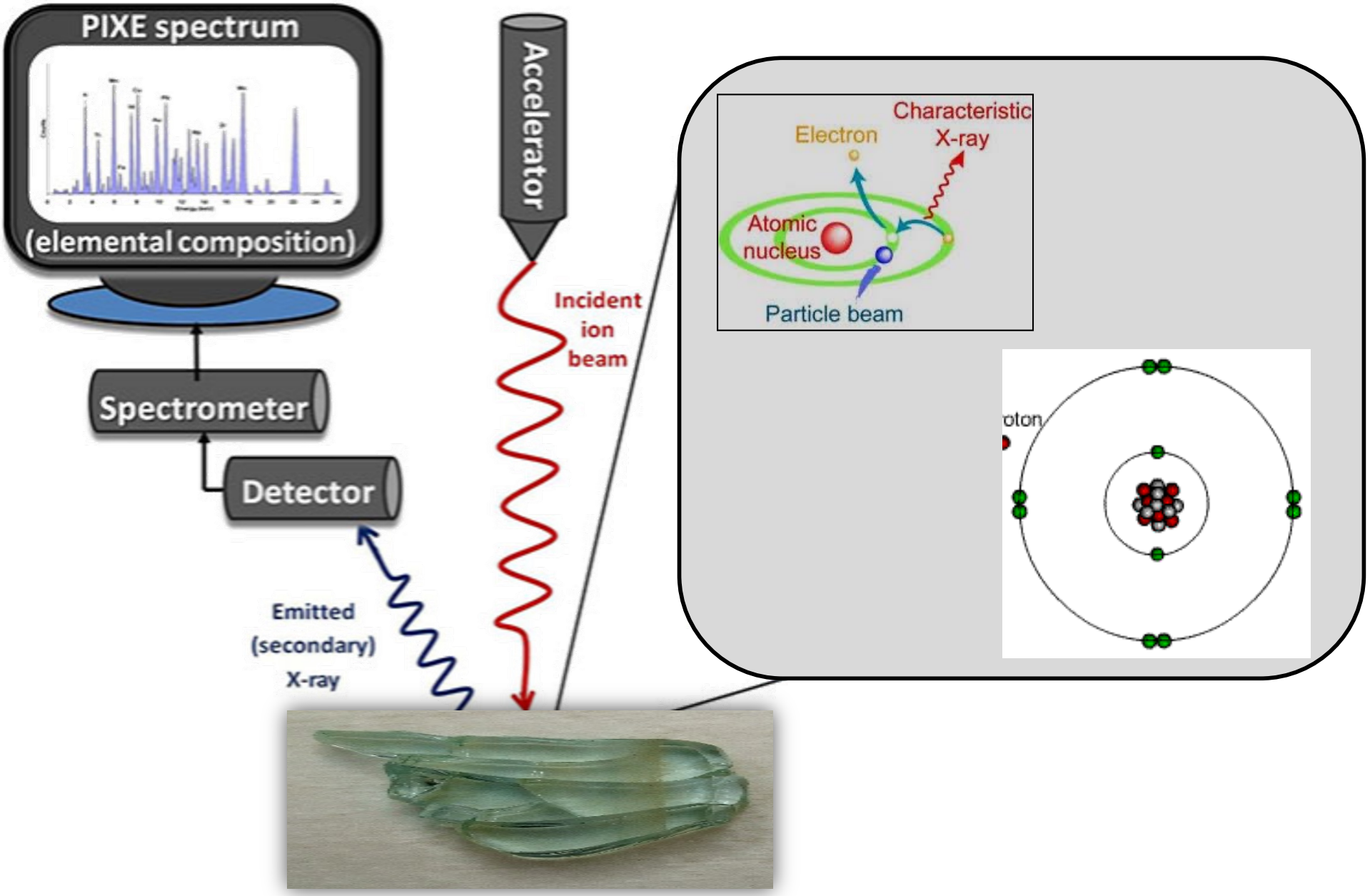


Glass Fragments

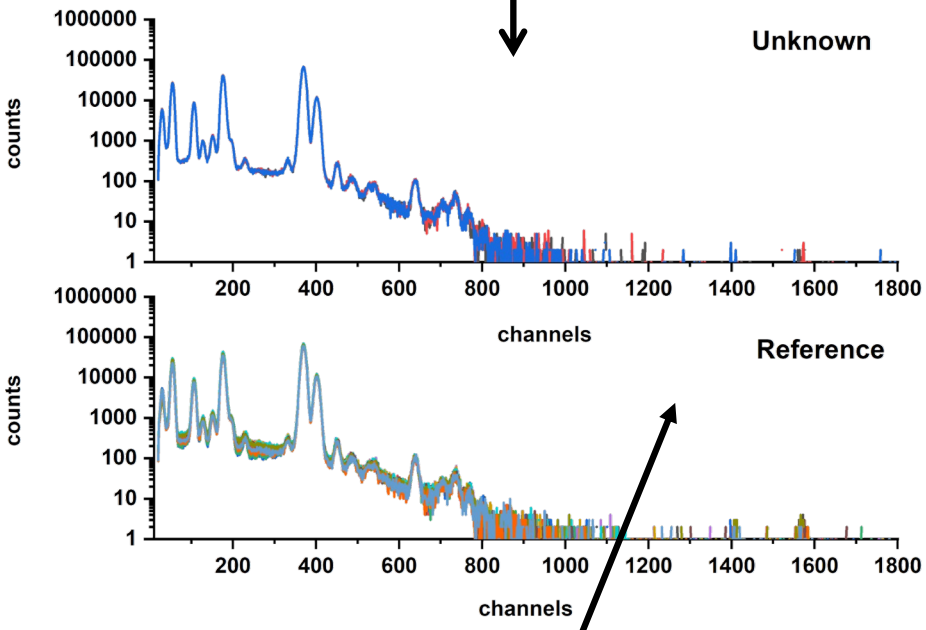
Every year, the Israeli Police Force reports on dozens of crimes that involved glass fragment as evidence which did not realize their forensic potential

Case from Israeli Police

PIXE as a Tool for Evidence Characterization



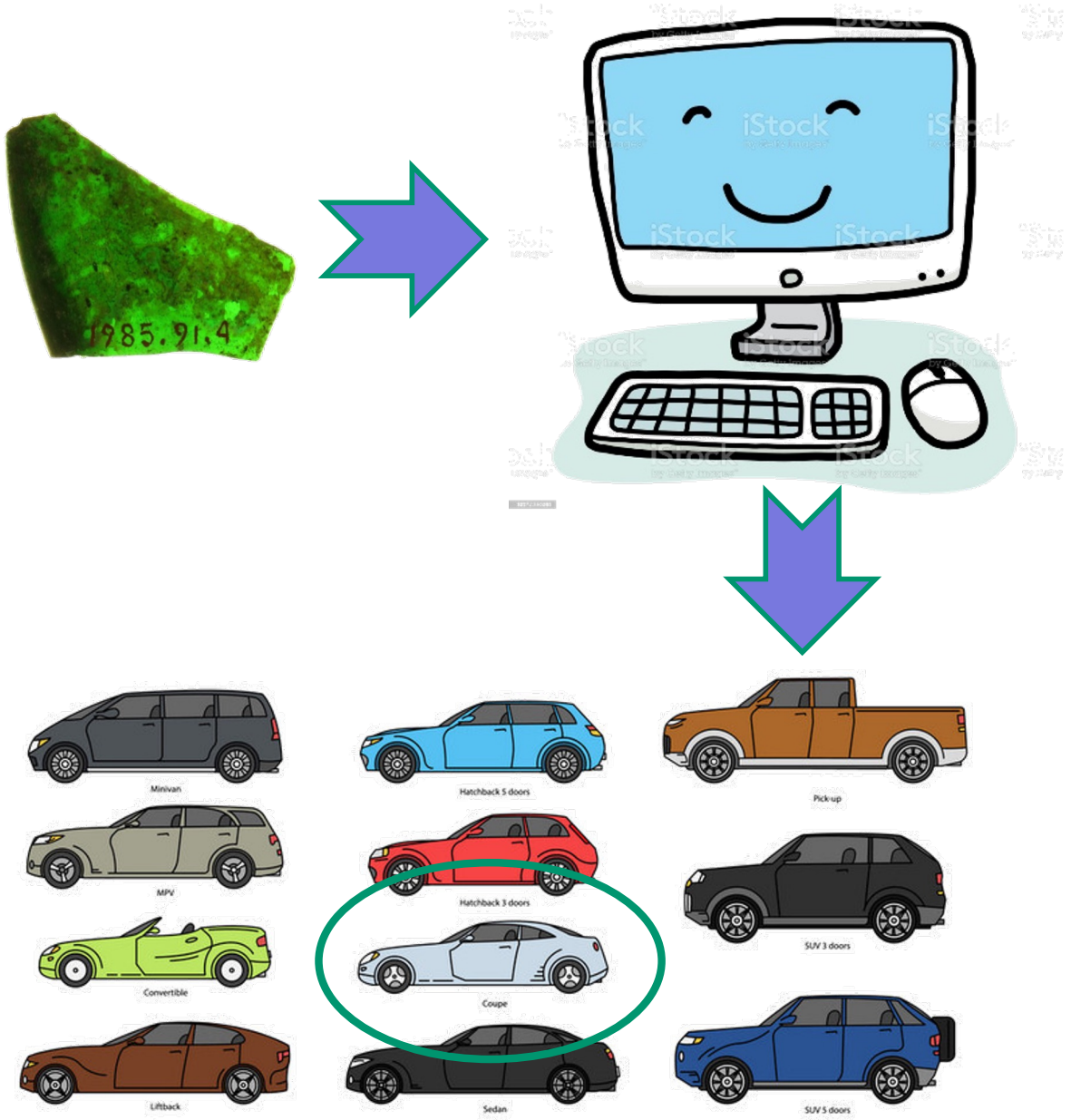
Case from the Israeli Police: Assocoation



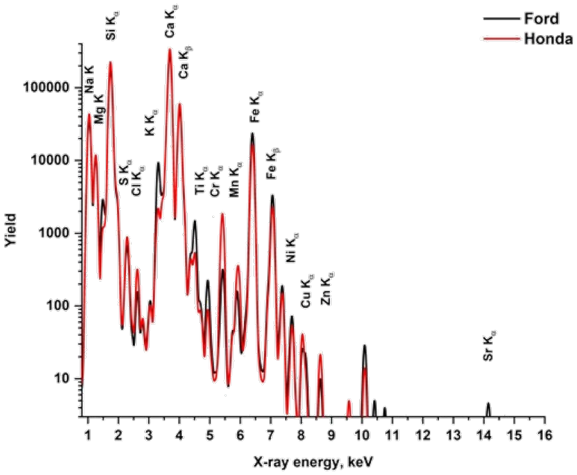
Perfect Match!!!!

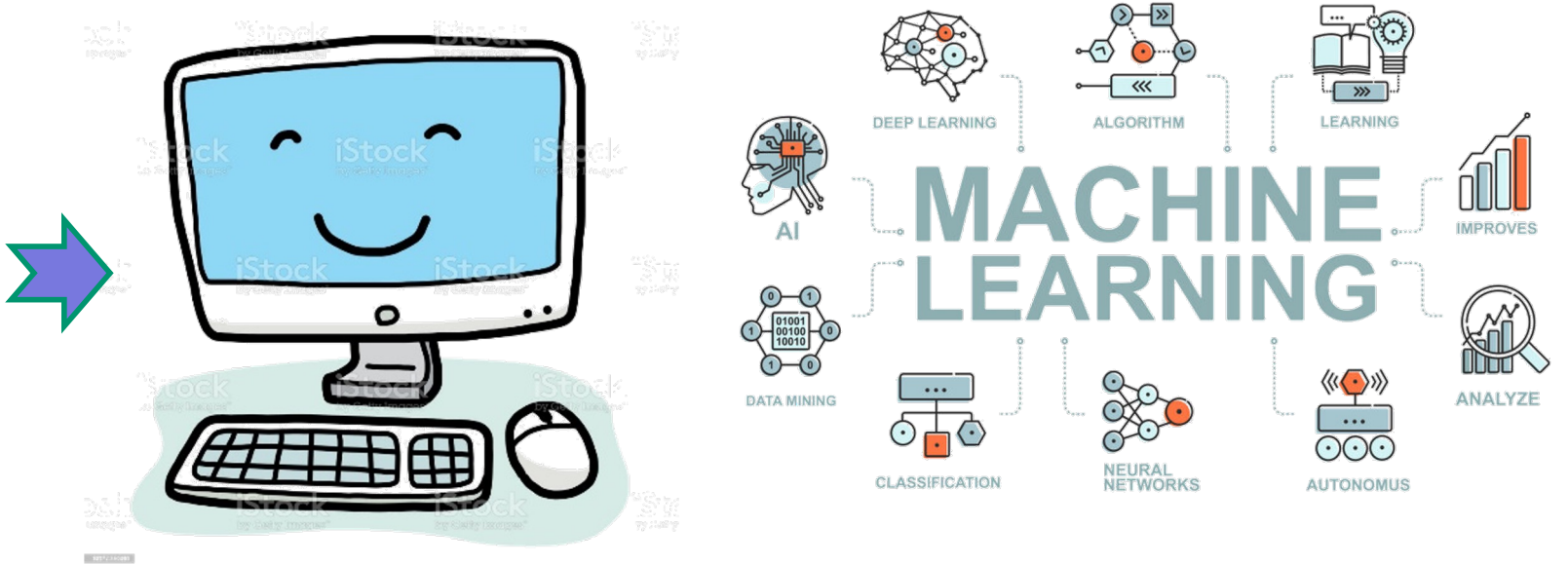
However, association requires a set of reference structures

Increasing Complexity: Classification



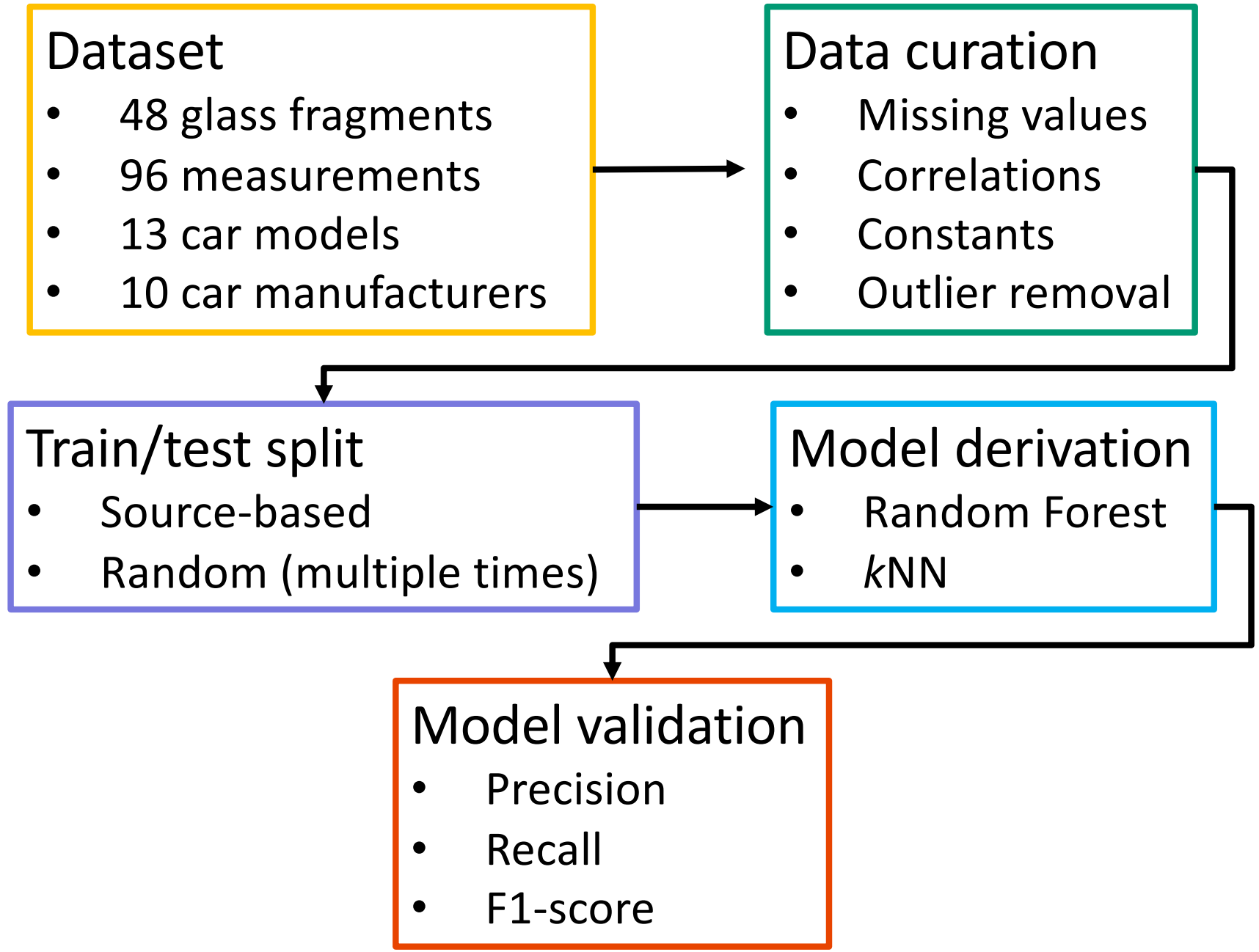
Increasing Complexity



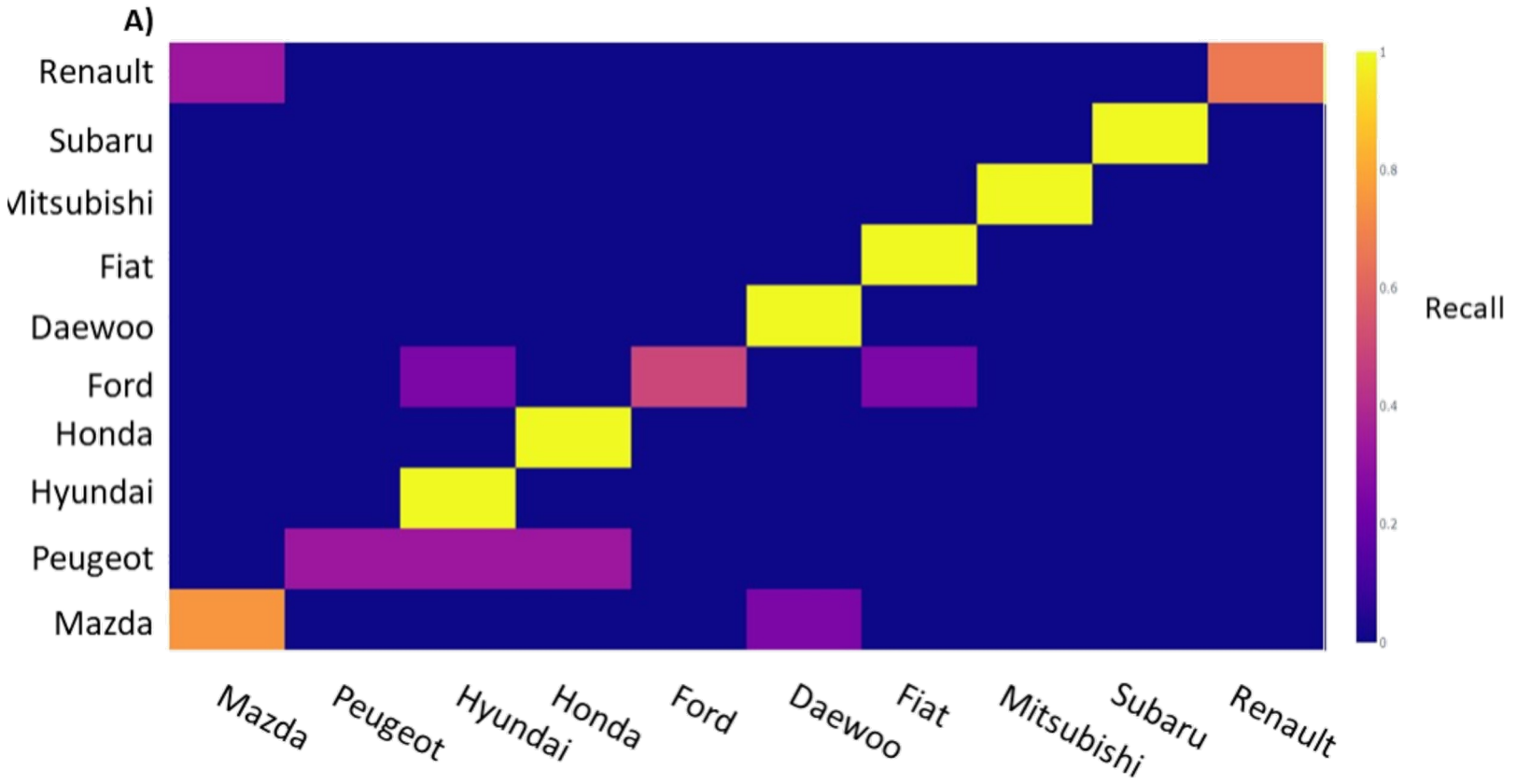


Kaspi et al., Talanta 2021, 234, 122608

Machine Learning



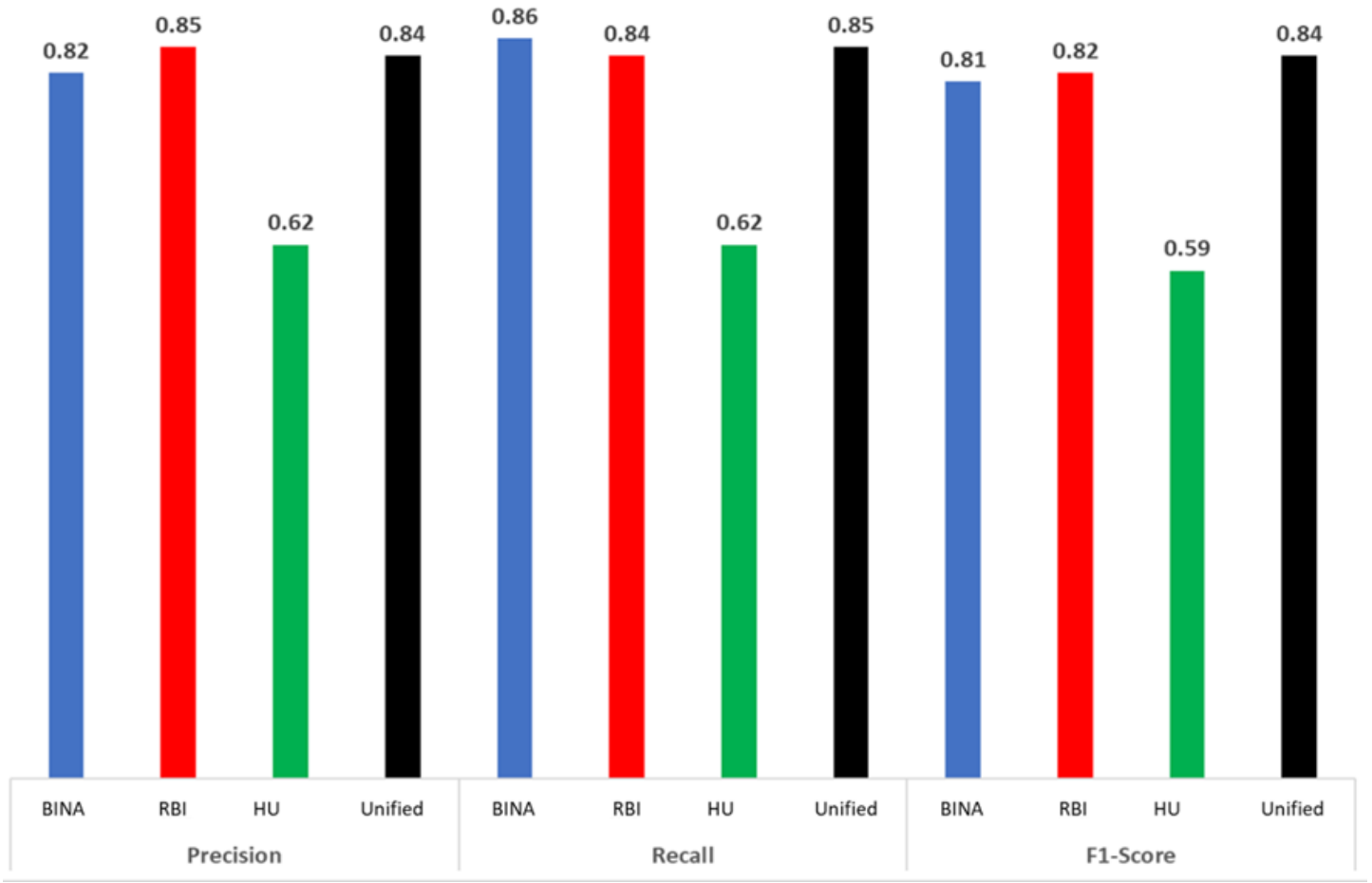
Results



	Random Train | Test (%)		Train – Surface Test – Bulk (%)		Train – Bulk Test – Surface (%)	
	RF	*k*NN	RF	*k*NN	RF	*k*NN
Recall	0.76 ± 0.07	0.74 ± 0.07	0.86	0.83	0.87	0.77
Precision	0.80 ± 0.09	0.77 ± 0.10	0.88	0.86	0.82	0.75
F1 - Score	0.73 ± 0.07	0.71 ± 0.08	0.84	0.82	0.84	0.75

Increasing Diversity

Uniting results from three different laboratories



Kaspi et al., Forensic Science International 333(1–3):111216

Further Increasing Diversity

Uniting results obtained with different analytical techniques from different laboratories

	PIXE (RBI+BINA)	PIGE +INAA	LA-ICP-MS	EDS
PIXE (RBI+BINA)	P: 0.87 ± 0.04 R: 0.90 ± 0.03 F1: 0.87 ± 0.05	P: 0.78 ± 0.05 R: 0.81 ± 0.04 F1: 0.77 ± 0.05	P: 0.92 ± 0.02 R: 0.93 ± 0.02 F1: 0.92 ± 0.03	P: 0.81 ± 0.04 R: 0.84 ± 0.03 F1: 0.81 ± 0.05
PIGE + INAA		P: 0.92 ± 0.06 R: 0.87 ± 0.10 F1: 0.89 ± 0.08	P: 0.86 ± 0.06 R: 0.83 ± 0.09 F1: 0.83 ± 0.07	P: 0.49 ± 0.06 R: 0.50 ± 0.09 F1: 0.46 ± 0.07
LA-ICP-MS			P: 0.79 ± 0.09 R: 0.76 ± 0.12 F1: 0.76 ± 0.10	P: 0.88 ± 0.06 R: 0.89 ± 0.06 F1: 0.86 ± 0.07
EDS				P: 0.51 ± 0.09 R: 0.52 ± 0.11 F1: 0.48 ± 0.10

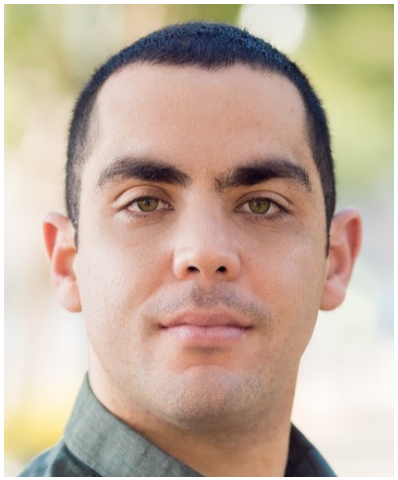
Kaspi et al., JCIM submitted for publications

Conclusions

- Statistical modeling is useful in the field of materials sciences
  - Insight
  - Experimental design
- Examples discussed
  - Solar cells
  - Forensics
  - Many others
- Challenges
  - Well curated large datasets
  - New descriptors for non well-defined compositions

Collaborating with experimentalists is indispensable

Acknowledgments



Omer Kaspi

Avi Yosipof

Hadar Binyamin

Olga Girshevitz

- Arie Zaban (Bar-Ilan University)
- Iva Bogdanović Radović (Ruđer Bošković Institute)
- Jyrki Räisänen (Helsinki University)
- Israel Police Force (Division of Identification and Forensic)

INNI Israel National Nanotechnology Initiative

IAEA International Atomic Energy Agency