

Materials Informatics: Statistical Modeling in Material Science

Hanoch Senderowitz
Bar-Ilan University, Israel

Presentation Goals

- Present material informatics and compare “classic” with material-related statistical models
- Provide a flavor of the diversity of material-related “activities”
- Examples from solar cells (going beyond the models)



Avi Yosipof



Oren Nachum



Omer Kaspi



Funding

Experiments



Arie Zaban

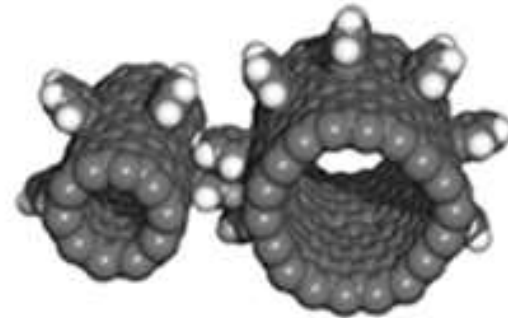
First Scientific Publication on Nanotechnology in Russian

Компьютерное моделирование в молекулярной нанотехнологии

Опубликовано в журнале *"Компьютерра"* №41 от 13 октября 1997 года
Автор: ИГОРЬ БАСКИН | Раздел: ТЕМА НОМЕРА

Представьте себе такую картину: на конвейере сборочного цеха днем и ночью управляемые компьютером роботы при помощи механических рук-манипуляторов собирают из поставляемых на предприятие либо производимых в других цехах деталей самые разнообразные машины. Картина, казалось бы, знакомая и не должна вызывать в конце XX века ни у кого особого удивления, за исключением только одного момента: все эти машины, роботы, компьютеры, конвейер и даже сам завод по размеру не превосходят вируса. Все детали механизмов таких машин представляют собой индивидуальные молекулы, либо супрамолекулярные (то есть состоящие из нескольких молекул) комплексы. Хотя все это выглядит как фантастическая картина из очень далекого будущего, на самом же деле это вполне конкретные научно-технические разработки последних нескольких лет, которые и составляют предмет молекулярной нанотехнологии.

Молекулярная нанотехнология занимается дизайном, моделированием и производством молекулярных машин и молекулярных устройств. Пионером этого направления можно по праву считать Эрика Дрекслера, опубликовавшего пять лет назад книгу Eric Drexler, *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, John Wiley & Sons, 1992 г. По своим



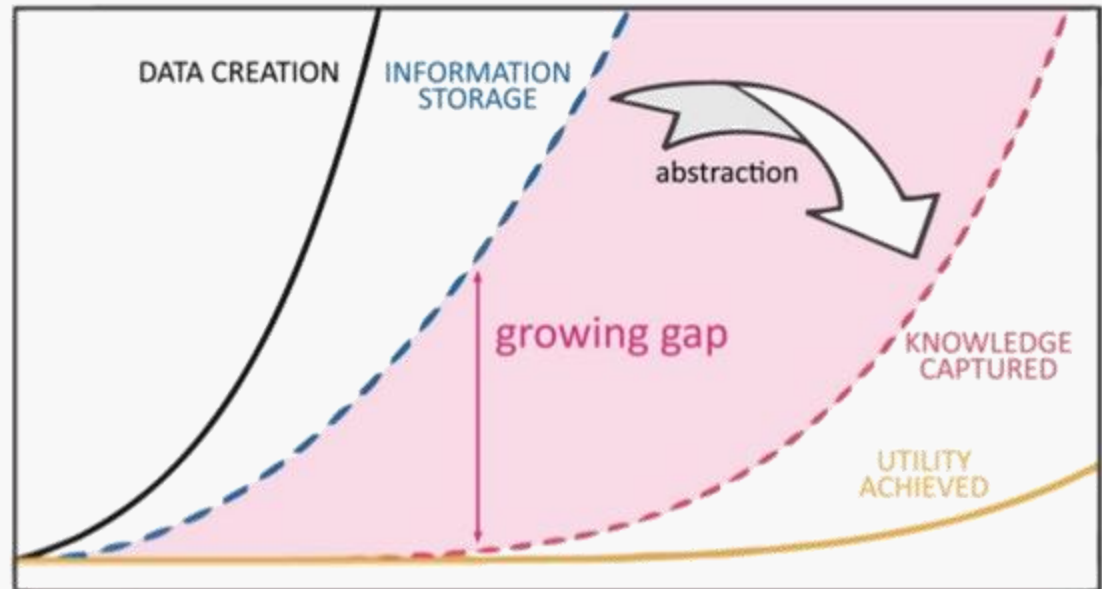
Material Informatics: Turning Data into Knowledge

Materials informatics is a field of study that applies the principles of informatics to materials science and engineering to better understand the use, selection, development, and discovery of materials. This is an emerging field, with a goal to achieve high-speed and robust acquisition, management, analysis, and dissemination of diverse materials data.



WIKIPEDIA
The Free Encyclopedia

- Related to big data
- Makes use of machine learning



Areas of Applications of QSAR/QSPR in Material Sciences

- Medicinal chemistry, drug design, pharmaceuticals
- Personal care products and cosmetics
- Food industry
- Catalysts design
- Anticorrosive material design
- Optical devices design
- Nanotechnology
- Explosives
- Solar cells

Statistical Models (aka QSAR/QSPR)

- Accurate experimental data



$$\Delta G = -2.3RT \log K = -2.3RT \log \frac{[ES]}{[E][S]} \propto \log \frac{1}{[S]}$$

- Descriptors

- ❖ Structure-derived (measured; calculated)

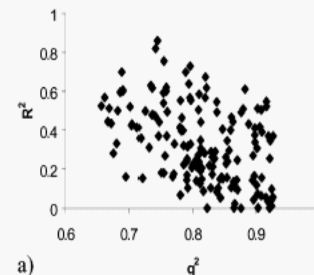
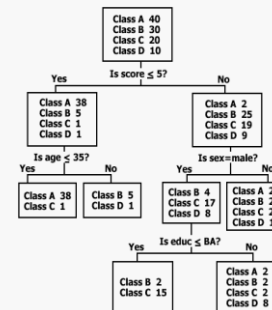
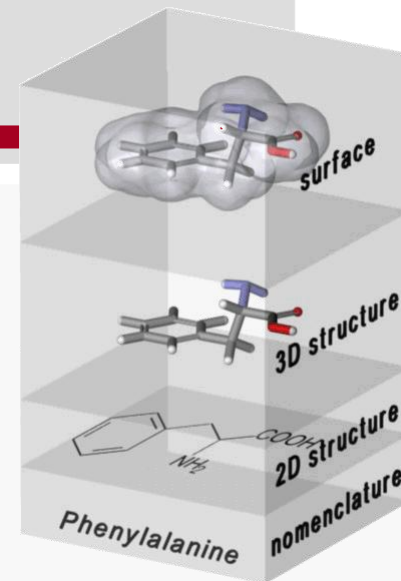
- A mathematical model

- ❖ *e.g.*, quantitative, qualitative, linear, non-linear

- Model validation

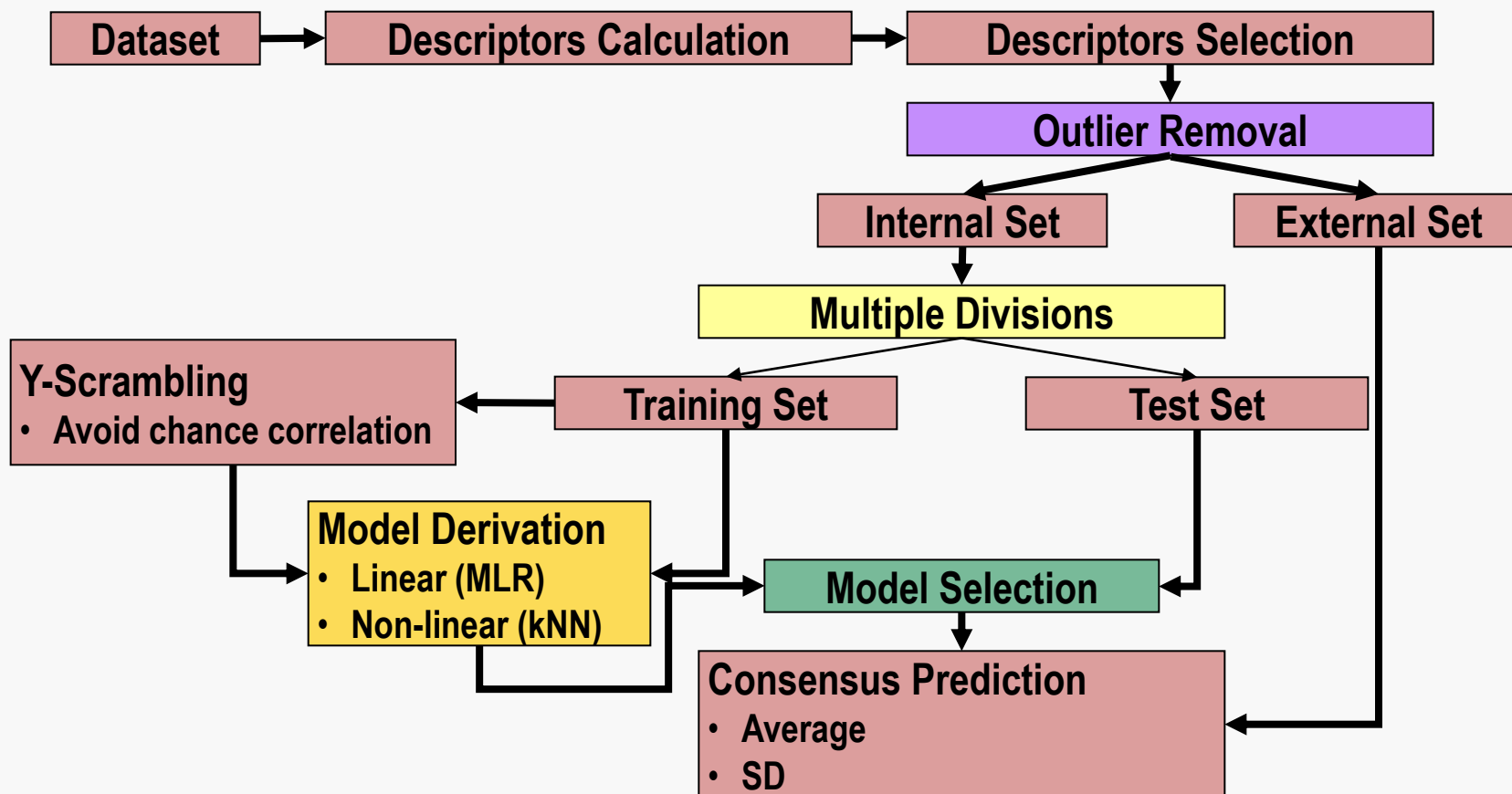
- ❖ Models developed on a training set and tested on an independent test set

- ❖ **Models should be simple and interpretable**



QSAR Engine

1. Descriptors selection
2. Outliers removal
3. Generation of multiple models
4. Model(s) validation and selection
5. Consensus model
6. Validation
7. Predictions



The Compounds

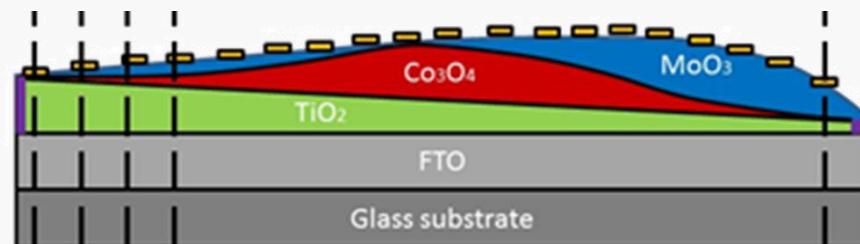
- Classical QSAR

- ❖ Structures typically well-defined
- ❖ Potential exception: Polymers and mixtures (but consisting elements are known)
- ❖ True even for combinatorial chemistry



- Material QSAR

- ❖ Structures sometimes well-defined
- ❖ Not true for combinatorial material synthesis



The Data

- Classical QSAR

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

- Material QSAR

- ❖ 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

- Classical QSAR
 - ❖ Typically nD (n = 1,5) “classical” descriptors
 - ❖ Limited usage of QM-derived descriptors
- Material QSAR
 - ❖ 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)

Raman Spectroscopy

Raman spectroscopy is a spectroscopic technique used to observe vibrational, rotational, and other low-frequency modes in a system. Raman spectroscopy is commonly used in chemistry **to provide a fingerprint by which molecules can be identified.**

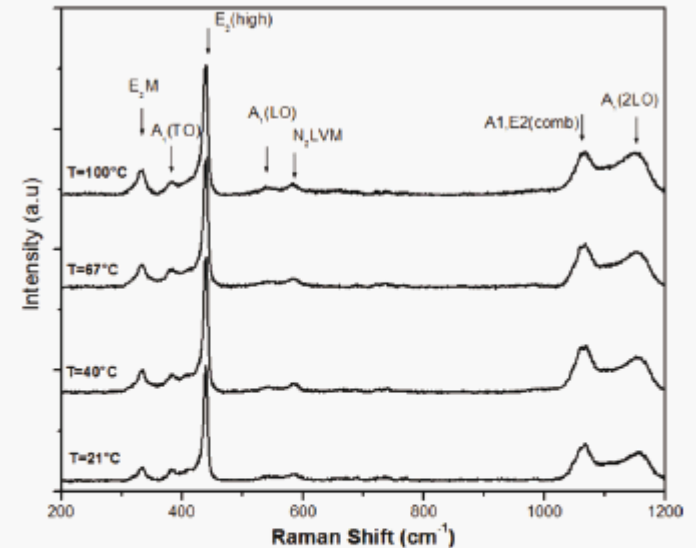
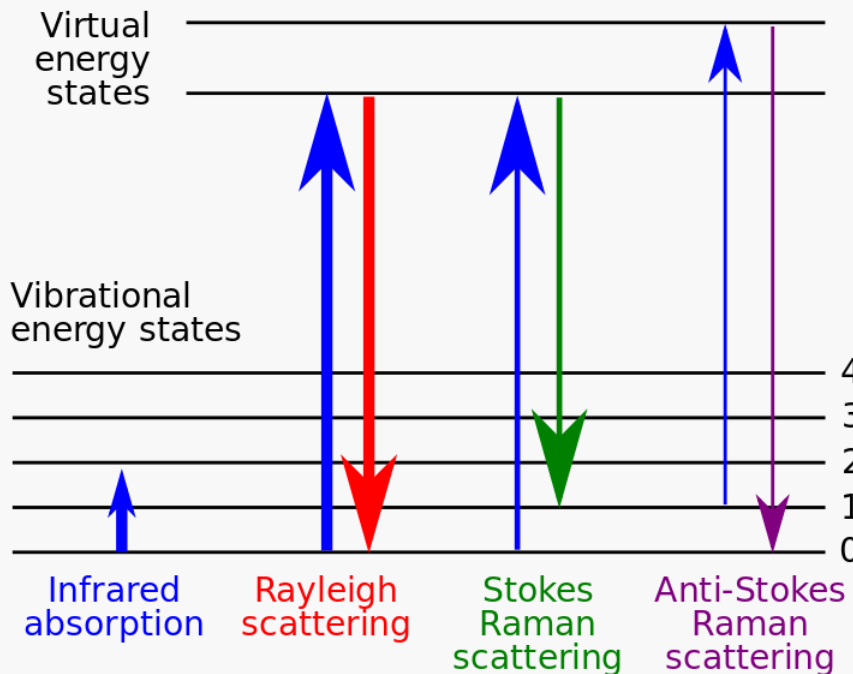
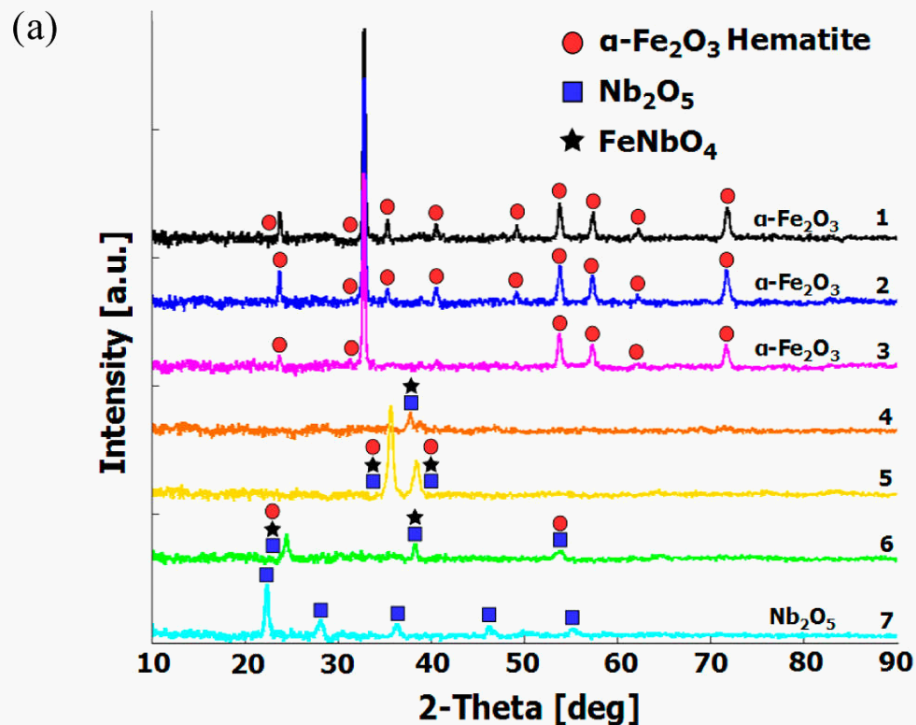
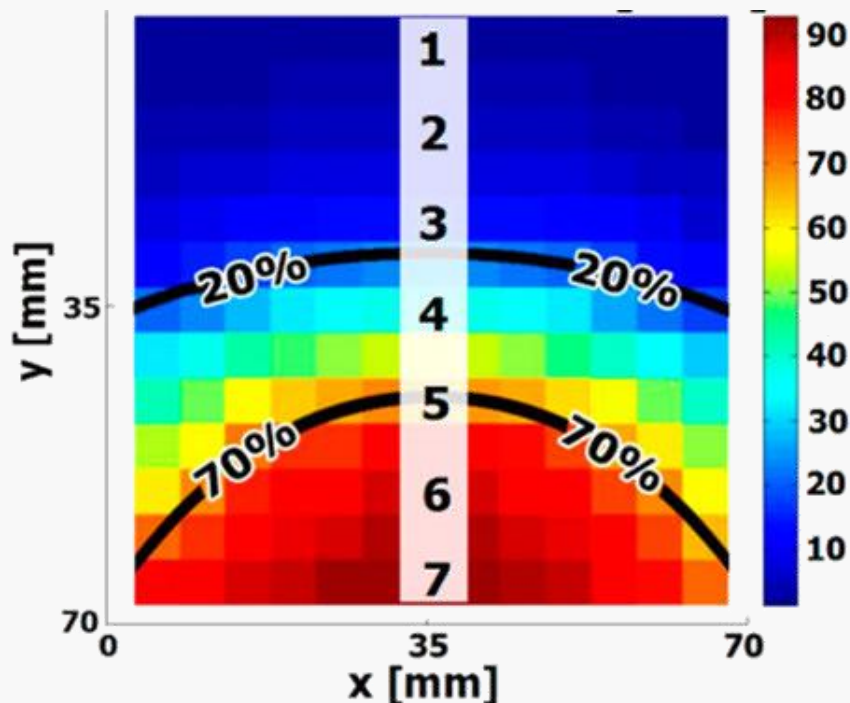


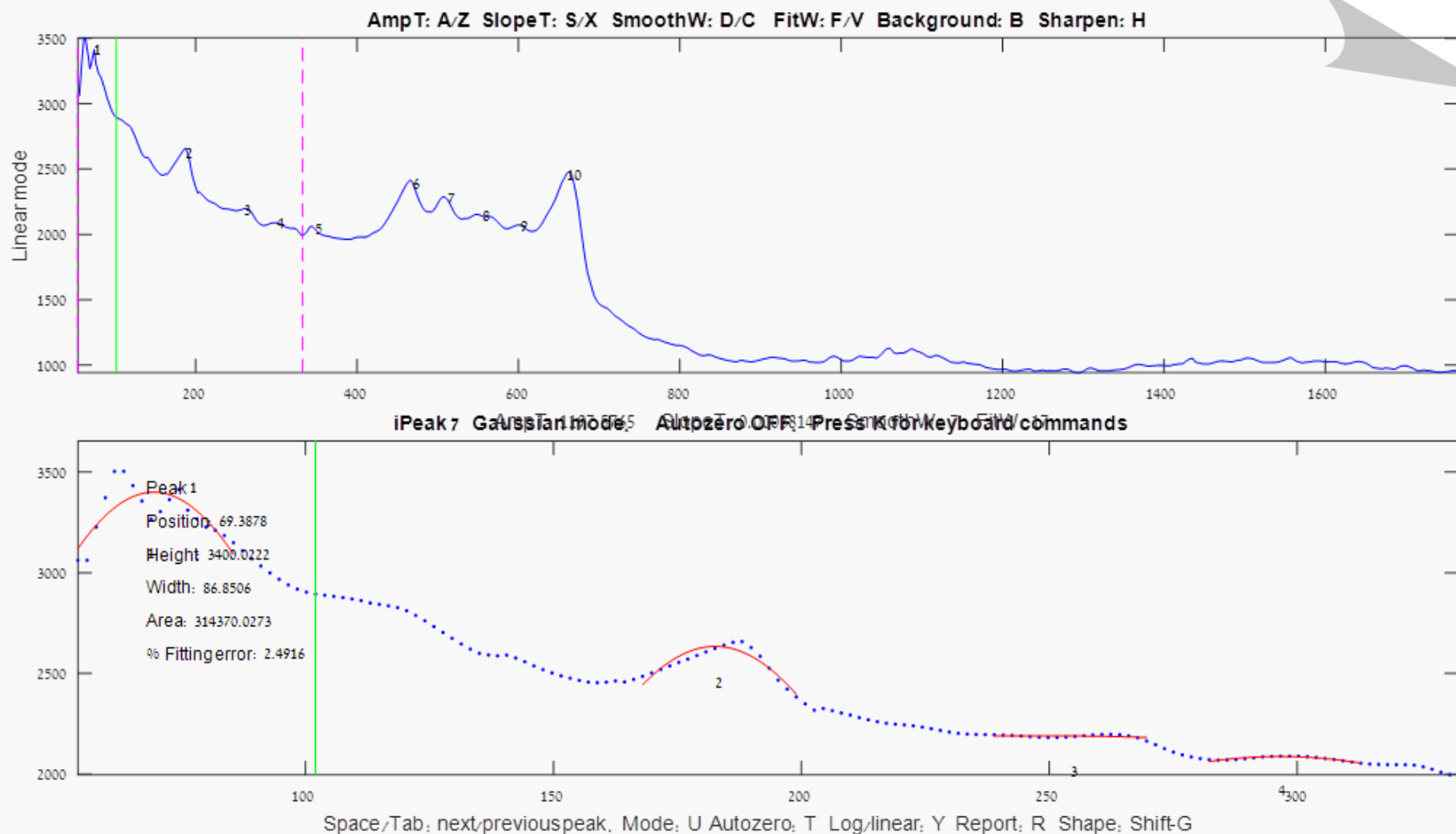
Figure 4. Variation with temperature of the Raman spectra of ZnO. The most intensive Raman line E₂(high) was assigned as a control peak to calculate phase relationship in the composites prepared.

X-Ray Diffraction (XRD)

X-ray diffraction has been in use in two main areas, for the **fingerprint** characterization of crystalline materials and the determination of their structure. Each crystalline solid has its unique characteristic X-ray powder pattern which may be used as a "fingerprint" for its identification.



Using Spectra as Descriptors



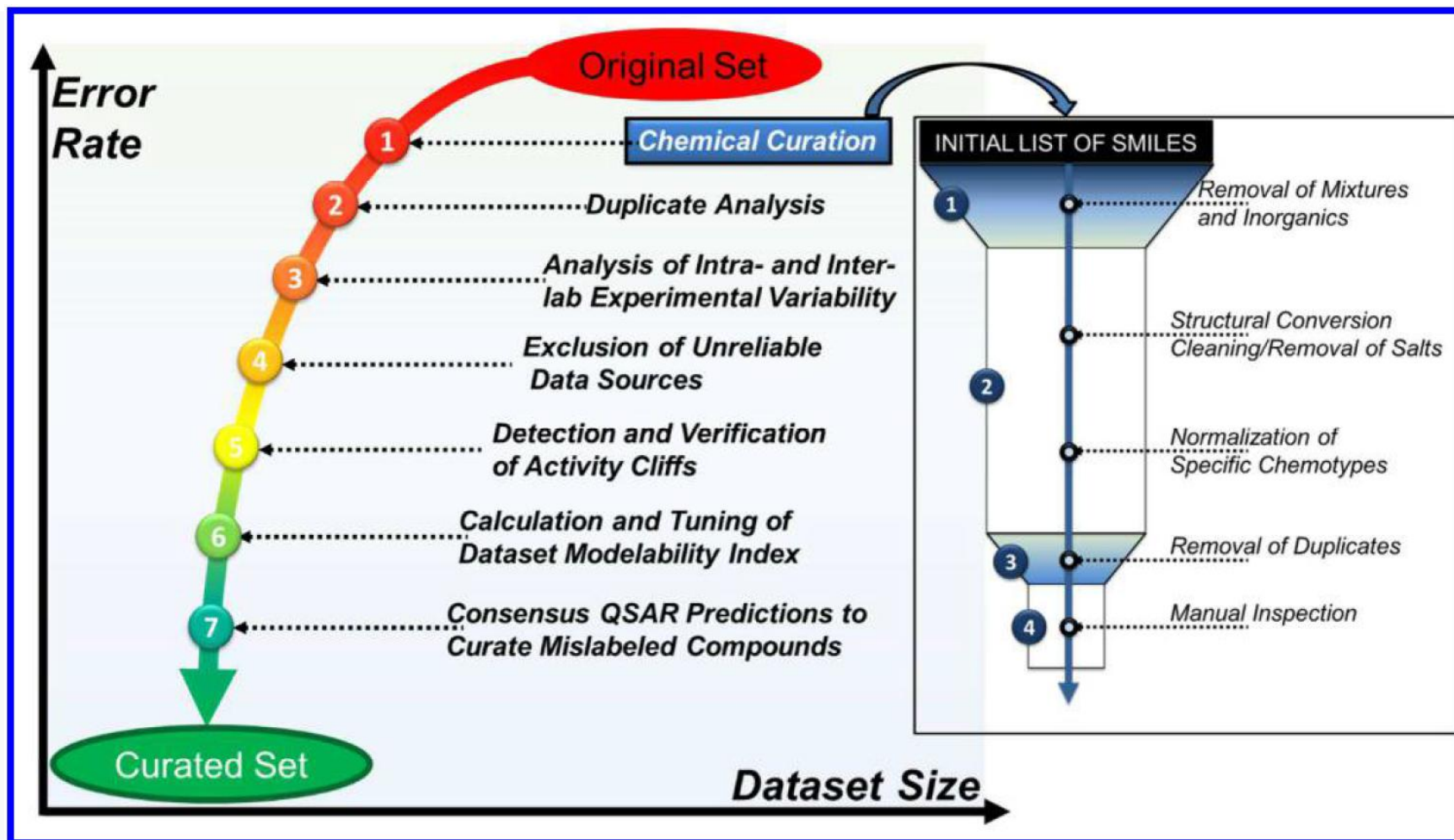
Methods

- Classical and material QSAR
 - ❖ Data reduction techniques (e.g., PCA)
 - ❖ Clustering
 - ❖ Classification models (e.g., Random Forests)
 - ❖ Quantitative models (e.g., MLR, SVM, *kNN*)

Validation

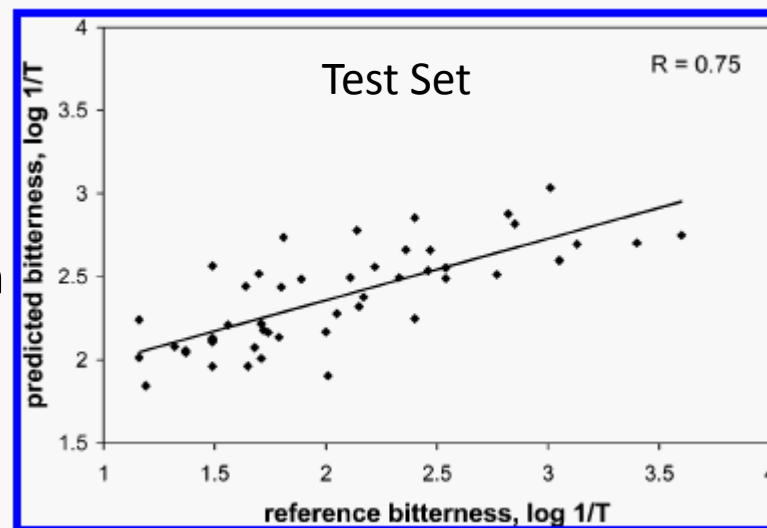
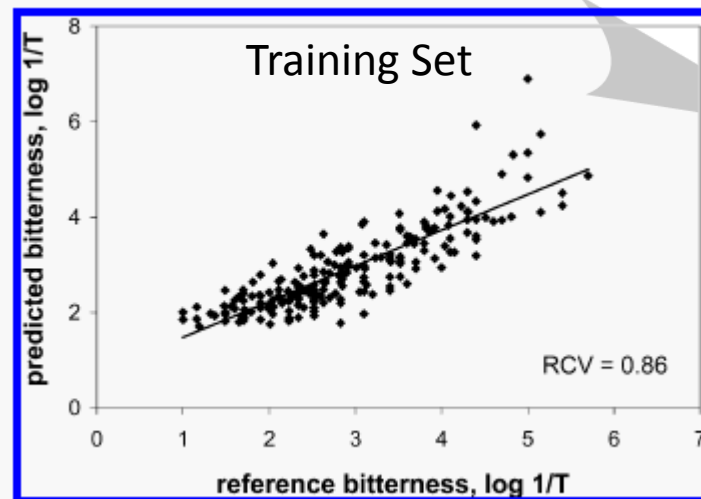
- Classical QSAR
 - ❖ “OECD” principles available and frequently followed
- Material QSAR
 - ❖ Insufficient external validation
 - ❖ Inappropriate control for chance correlation

Data Curation



Bitter Taste Predictions

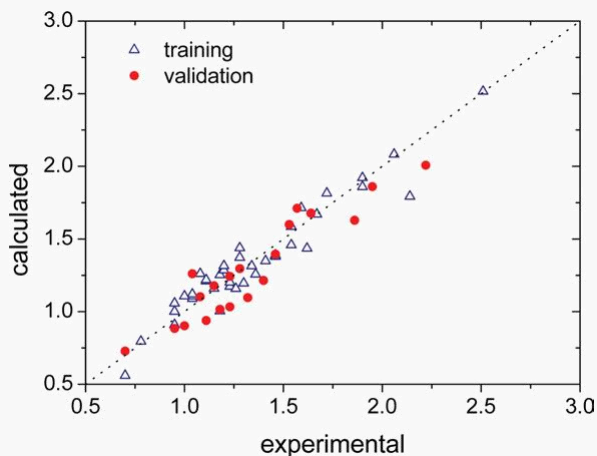
- Prediction of peptide bitterness
 - Training set: 176 short peptides
 - Test set: 48 short peptides
 - Residue-based and global descriptors
 - PLS regression
-
- Global descriptors are more important than residue-based descriptors



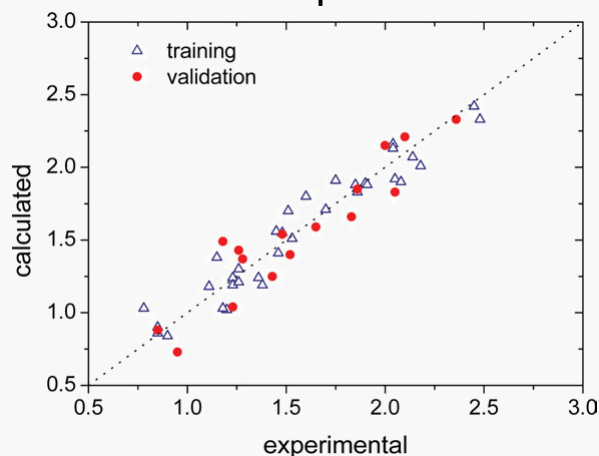
Explosives Prediction I

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

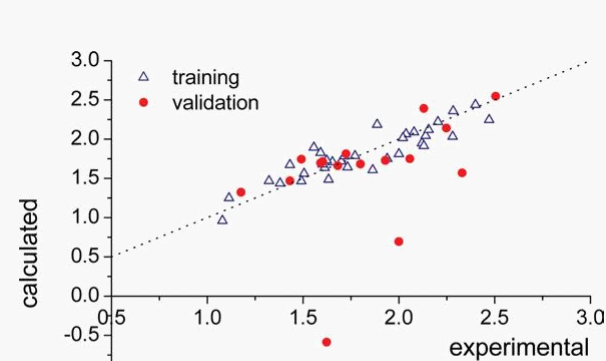
Nitramines



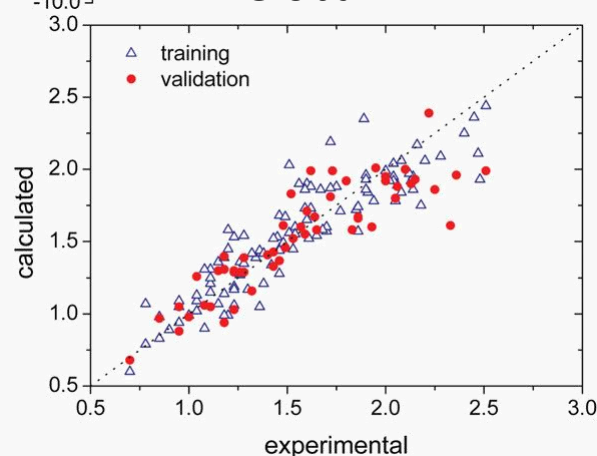
Nitroaliphatic



Nitroaromatic



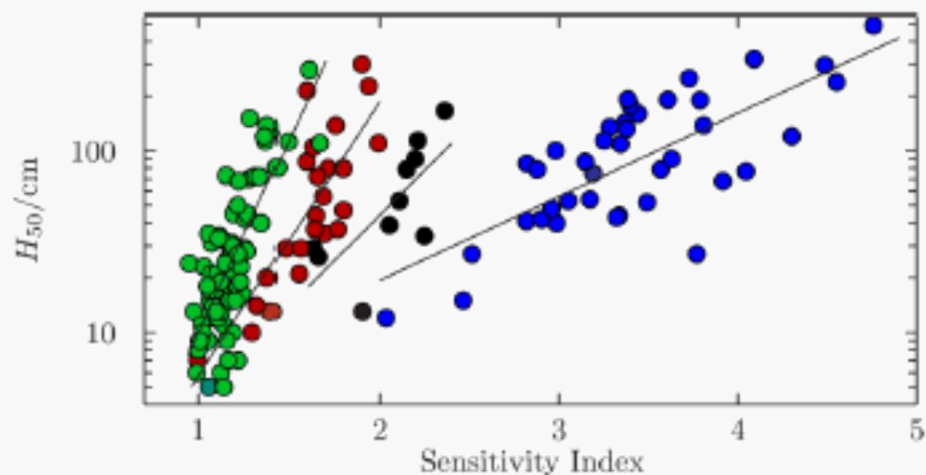
Global



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

Explosives Prediction II

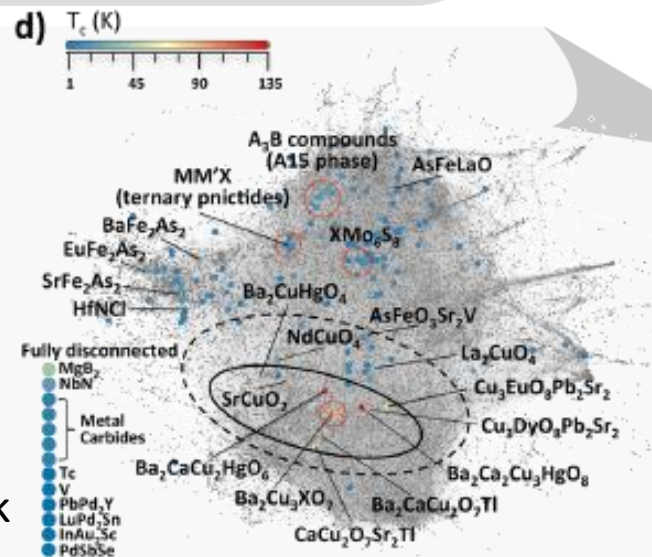
- Prediction of impact sensitivity of nitro compounds from “physical principles”
- Sensitivity Index (SI)
 - ❖ Number of atoms
 - ❖ Dissociation energy of the weakest X-NO₂ bond
 - ❖ Energy released upon the decomposition of 1 mole of compound



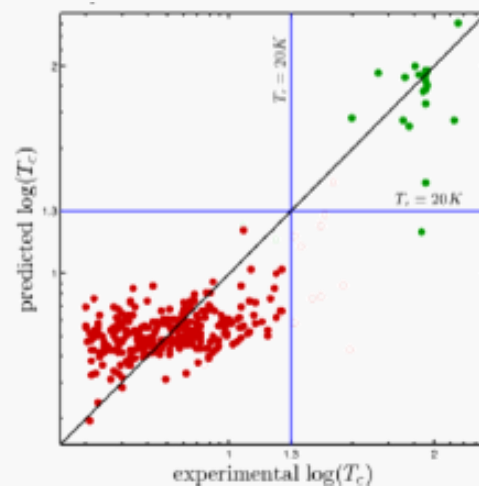
Material Cartography

Super-Conductivity Critical Temperature

- 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



Network



Model

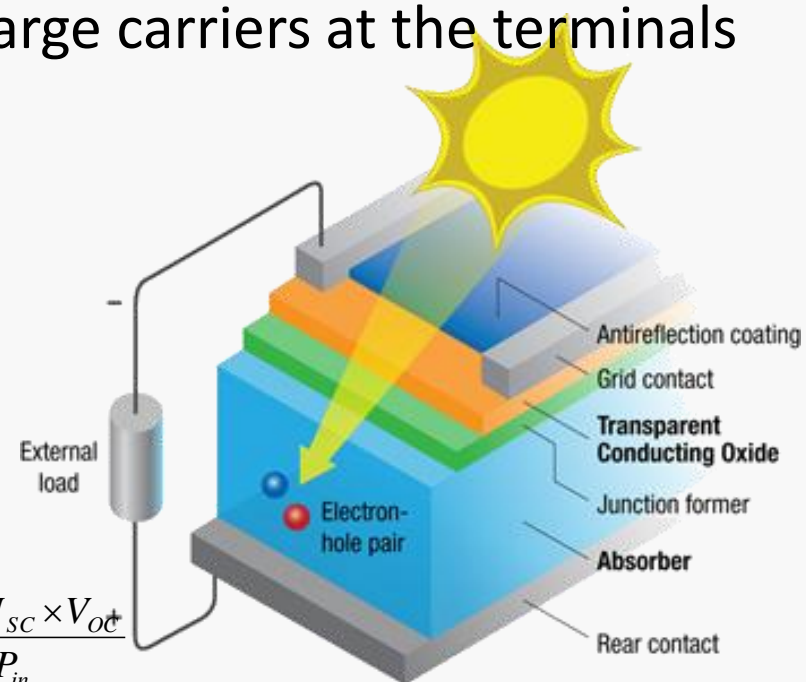
60 Seconds on Photovoltaic (PV) Cells

1. Generation of the charge carriers (electrons and holes) due to the absorption of photons
2. Separation of the photo-generated charge carriers in the 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 terminals of the junction

- Key Parameters

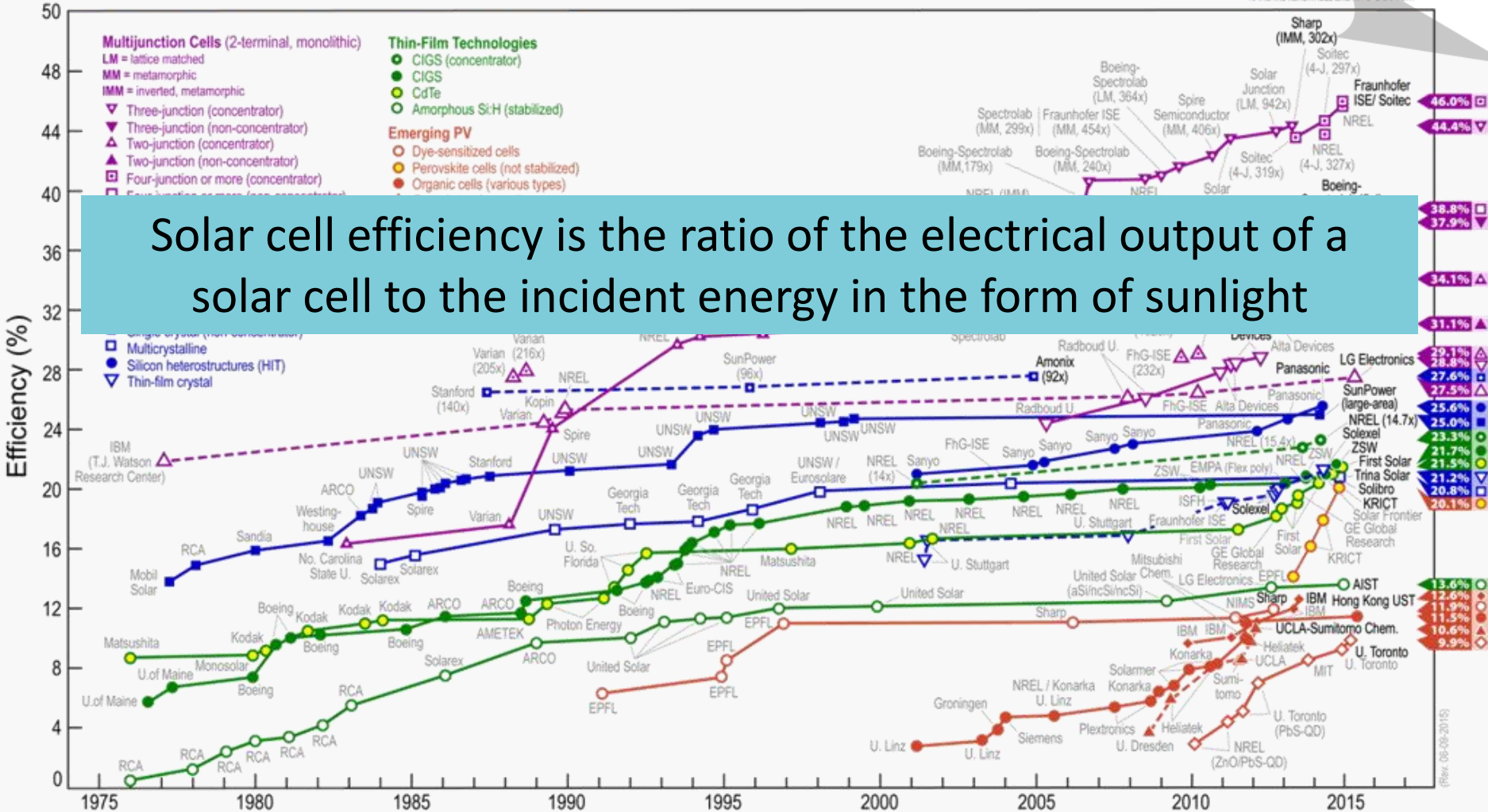
- ❖ Open circuit voltage (V_{OC})
- ❖ Short circuit current (J_{SC})
- ❖ Internal quantum efficiency (IQE)
- ❖ Fill factor (FF)

- ❖ Power Conversion Efficiency (PCE) =
$$\frac{FF \times J_{SC} \times V_{OC}}{P_{in}}$$



Efficiencies of Solar Cells

Best Research-Cell Efficiencies



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 for PV Cells

- Goals

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

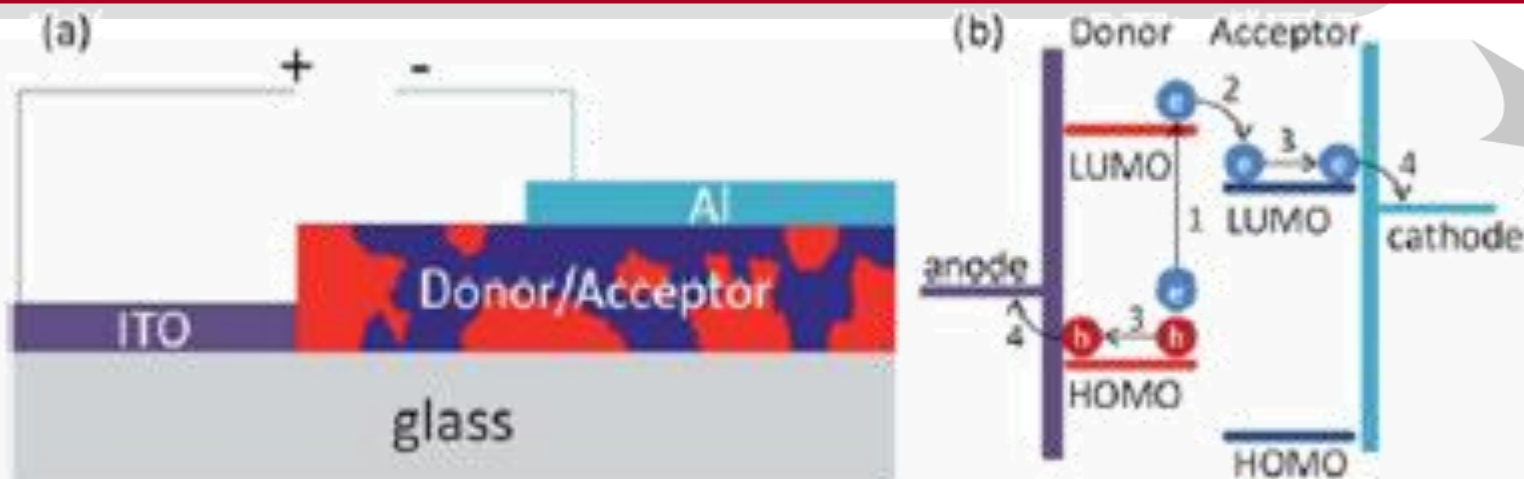
- Assumptions

- ❖ A correlation exists between PV properties and cells characteristics:

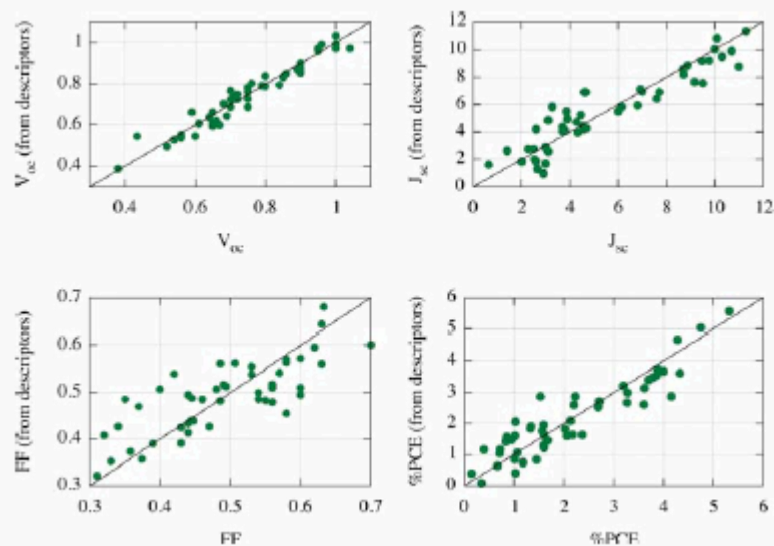
$$PV = f(\textit{Material Descriptors})$$

- Nature of correlation not necessarily known

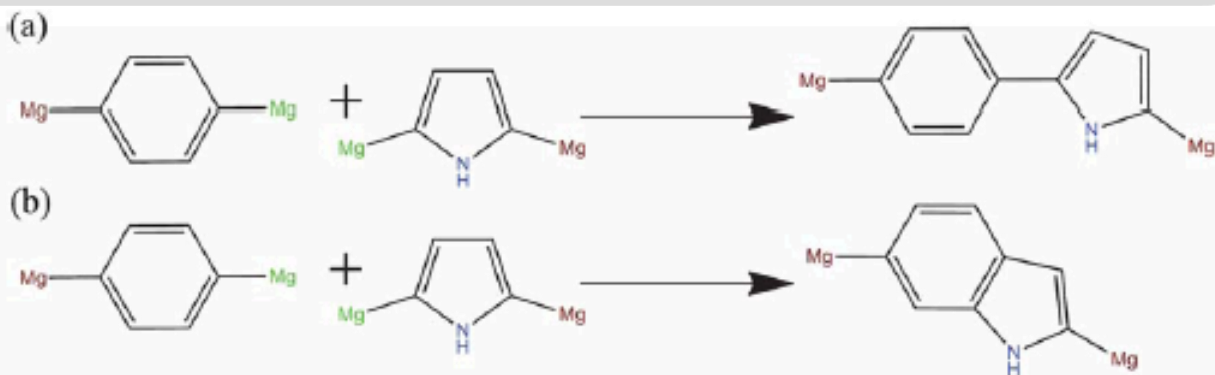
Organic Photovoltaics



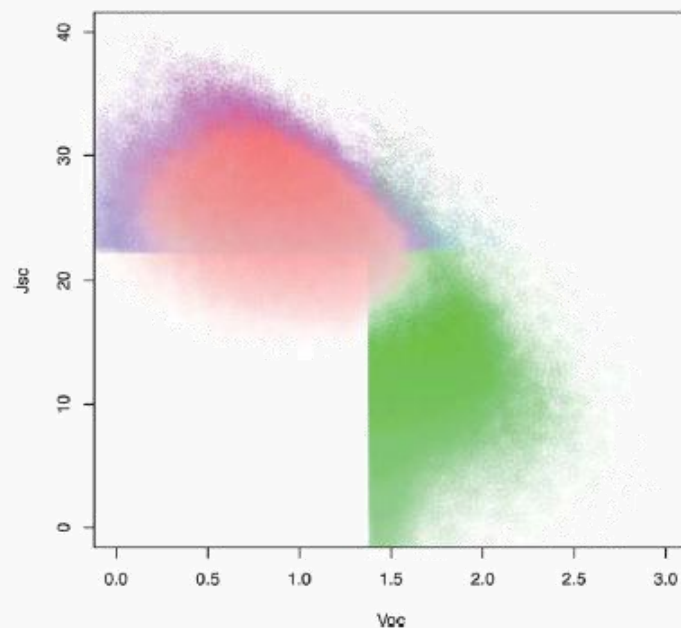
- Acceptor design
- Training set: 50 compounds
- ChemAxon descriptors
- Linear regression



Organic Photovoltaics

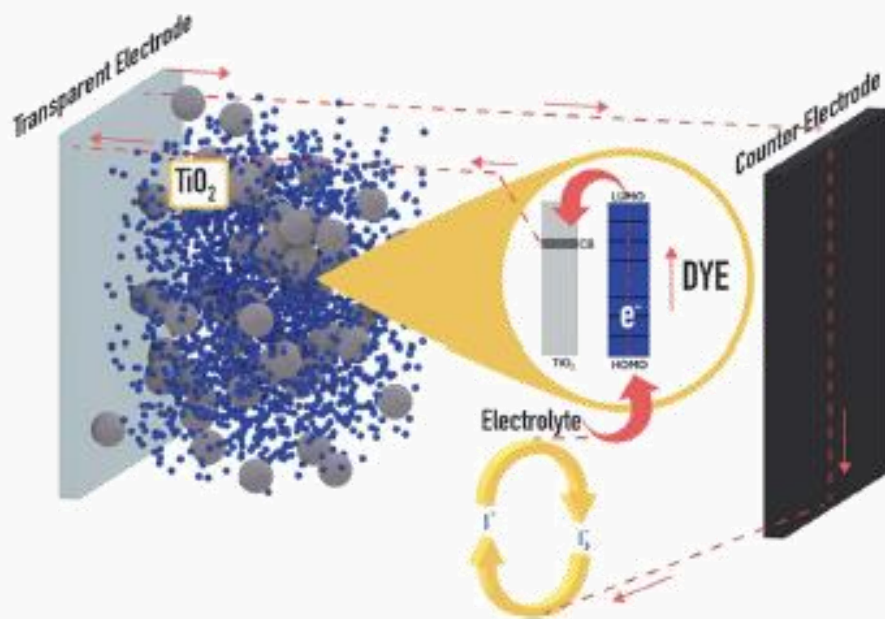


Color coding refers top 10% molecules with highest predicted Voc (green), Jsc (blue), and Voc x Jsc (red). Best molecules located upper left



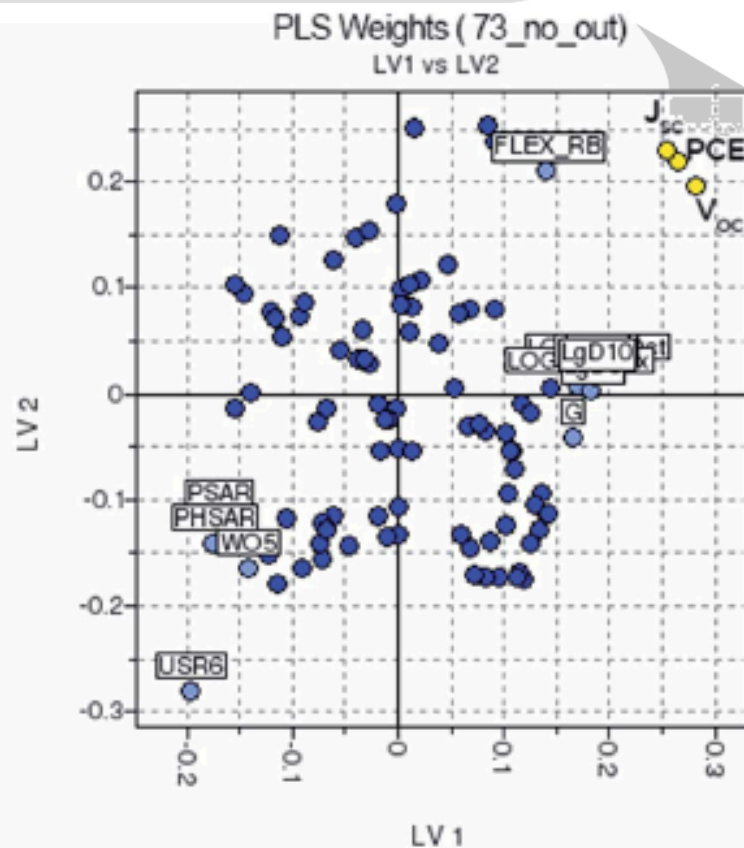
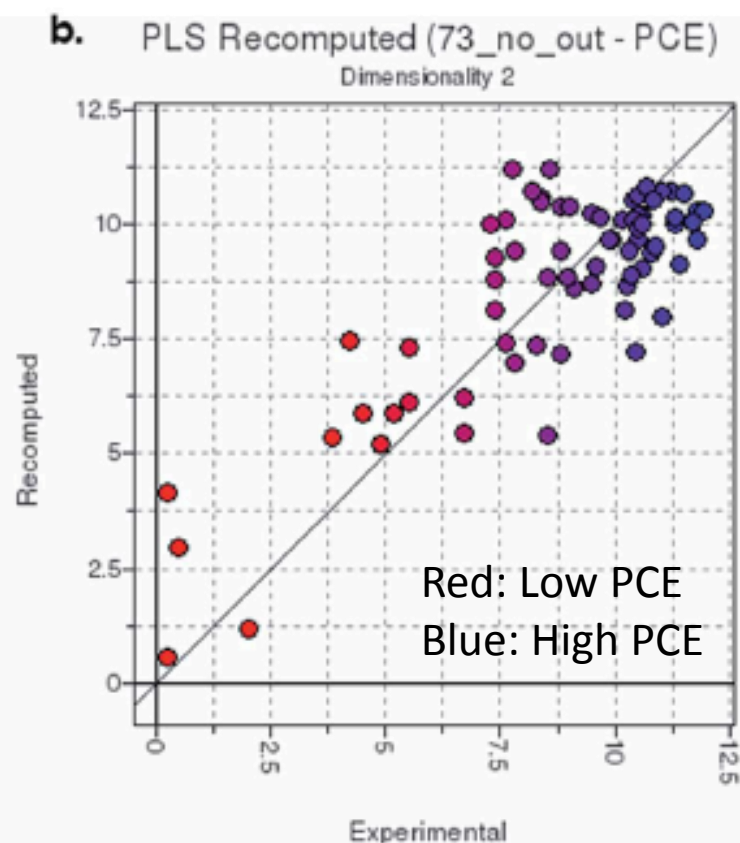
Dye-Sensitized Solar Cells (DSSC)

- Different dyes sensitized by incident radiation
- Photo-excited electrons transferred to TiO_2
- Holes transferred to the electrolyte



- Ruthenium sensitizers design
- Training set: 73 compounds
- No validation set
- Volsurf+ (MIFs-based), QM and “classical” descriptors
- PLS regression

Dye-Sensitized Solar Cells (DSSC)



- **Inversely correlated:** presence of NO_2 and NH_2 , PSA/HAS and PSA/SA ratios, H-bond
- **Correlated :** P n-oct, P c-Hex, log D5/log D10, flexibility

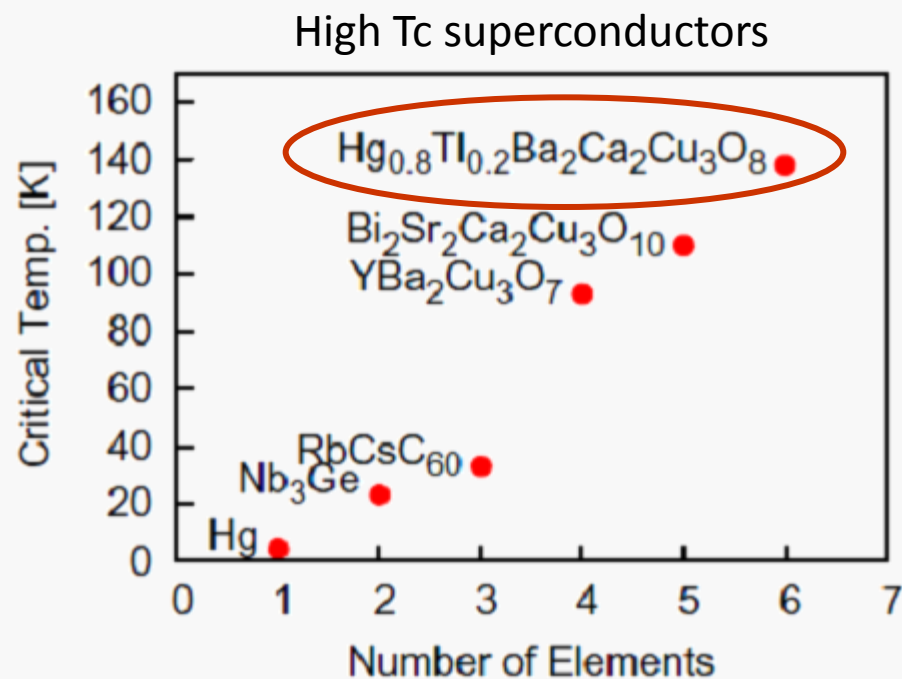
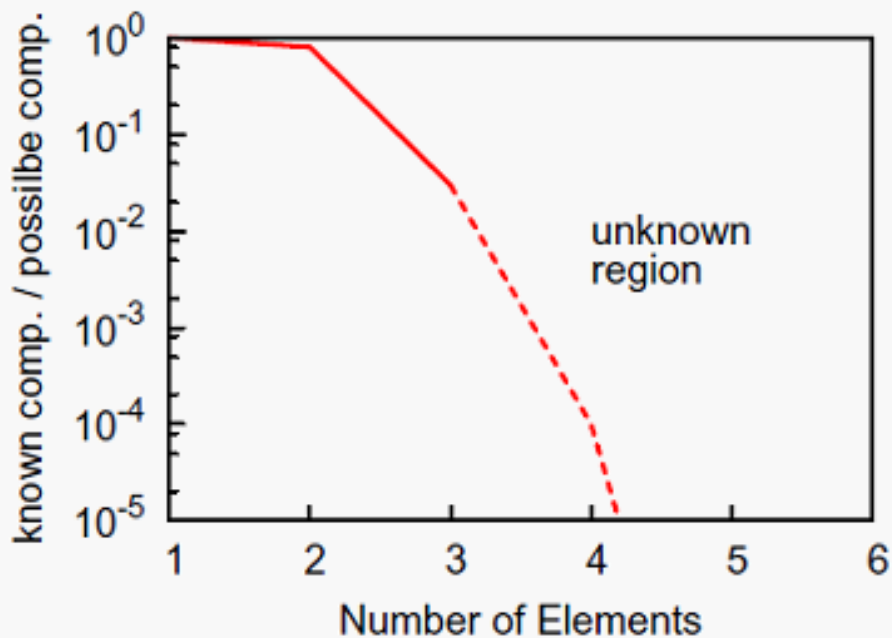
All Metal Oxide PV Cells

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

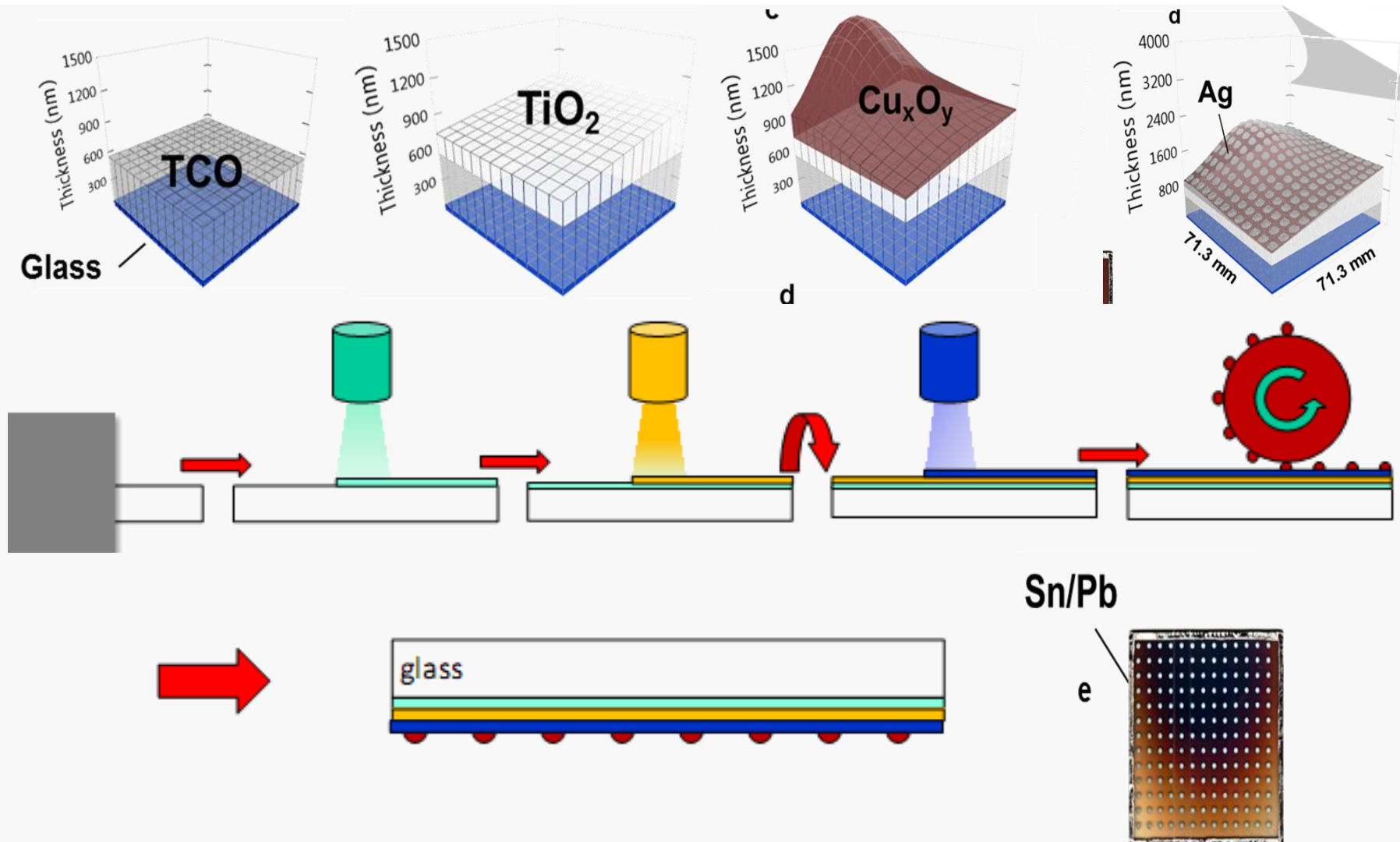
But Cell Not Efficient Enough
New Metal Oxides (MO) Required

Combinatorial Material Science

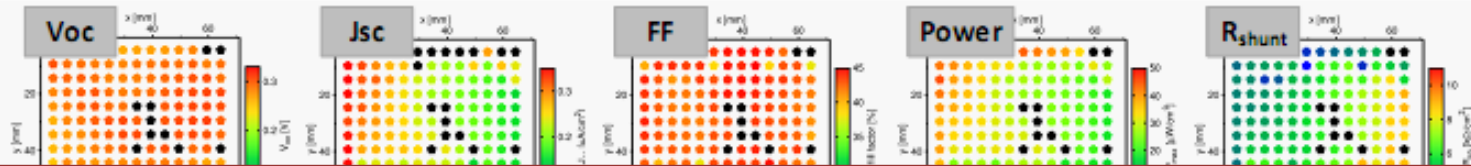
- ~60 “useful” elements leading to
 - ❖ ~30K inorganic compounds
 - ❖ 3600 binary compounds (ABO_x); mostly known
 - ❖ 216K ternary compounds ($ABCO_x$) almost all unknown



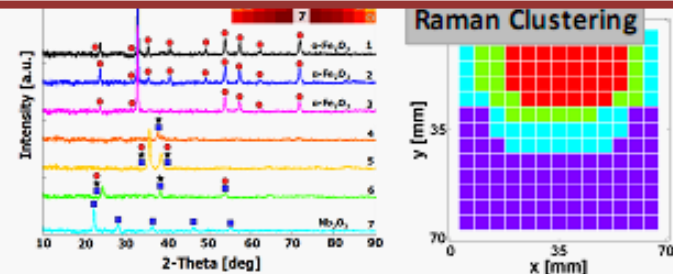
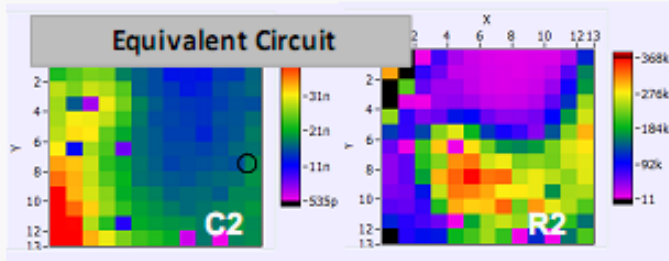
Synthesis of Libraries of All Oxide PV Cells via Combinatorial Material Synthesis



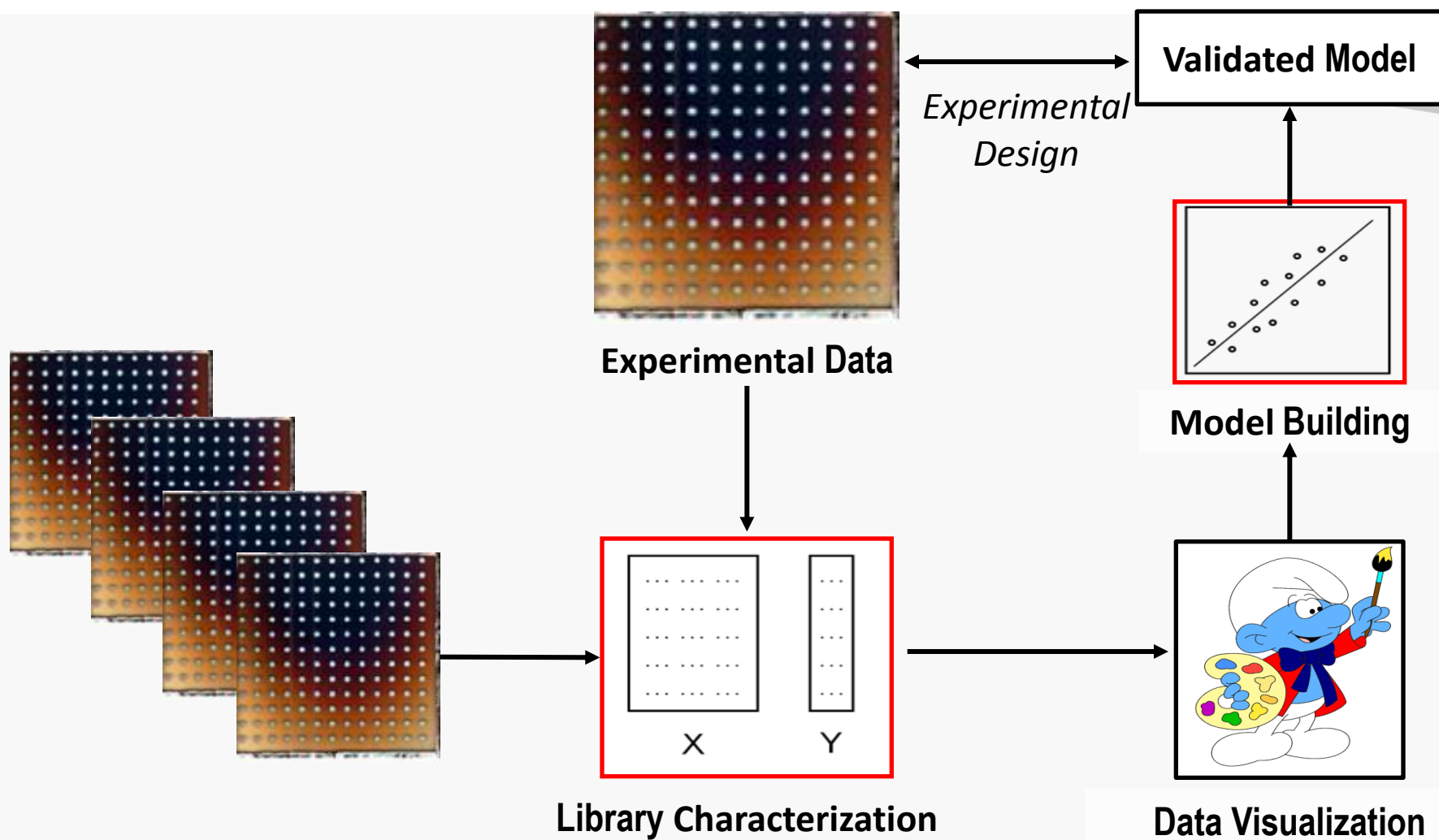
Analysis of Libraries of PV Cells



- Band gap: The energy difference (in electron volts) between the top of the valence band and the bottom of the conduction band
- IQE reflects the charge separation and collection efficiencies of a device
- Fill Factor is the ratio of the maximal theoretical power of the cell per unit volume divided by $V_{OC} \times J_{SC}$

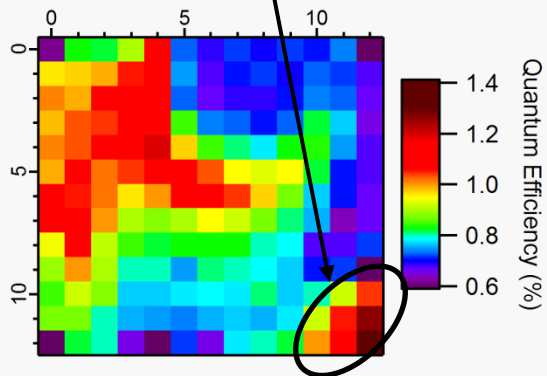
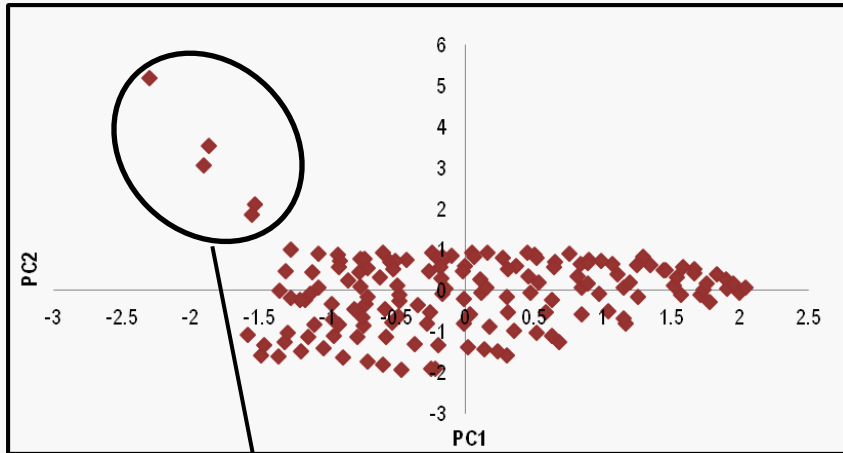


Material Informatics Workflow



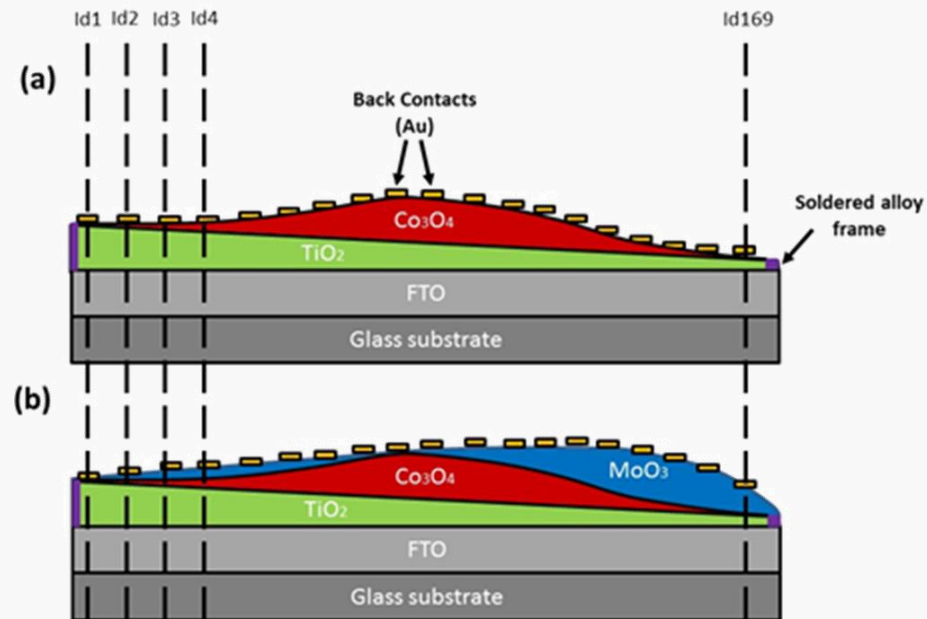
Principle Component Analysis (PCA)

TiO₂/Cu-O



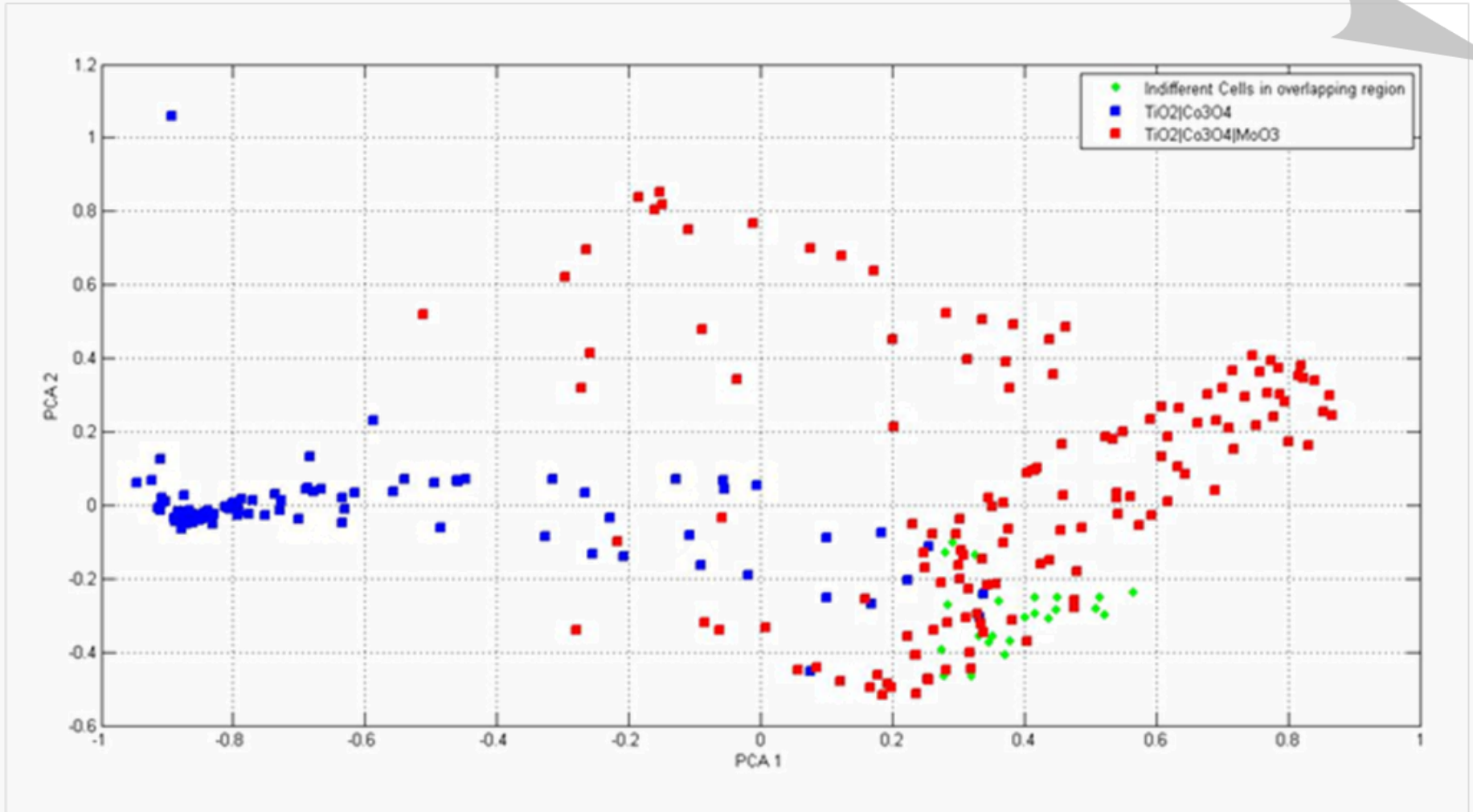
PCA of $\text{TiO}_2|\text{Co}_3\text{O}_4$ and $\text{TiO}_2|\text{Co}_3\text{O}_4|\text{MoO}_3$ Libraries

- V_{OC}
- J_{SC}
- I_{QE}
- FF
- P_{max}

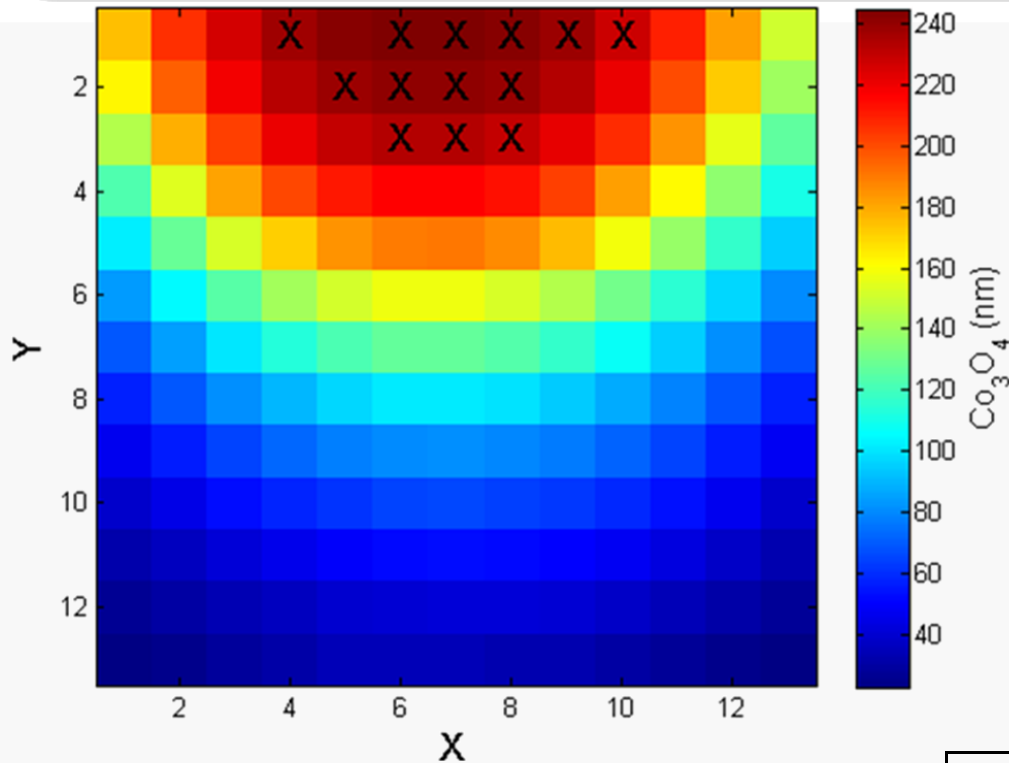


Activity	$\text{TiO}_2 \text{Co}_3\text{O}_4$		$\text{TiO}_2 \text{Co}_3\text{O}_4 \text{MoO}_3$		t-test	df	p-value
	Mean	SD	Mean	SD			
V_{oc} (mV)	173	149	446	146	-14.79	258	<0.001
J_{sc} ($\mu\text{A}/\text{cm}^2$)	8.76	5.48	17.41	3.82	-14.97	258	<0.001
I_{QE} (%)	0.04	0.01	0.13	0.07	-12.39	258	<0.001
FF (%)	25.76	4.21	29.8	3.6	-8.30	258	<0.001
P_{max} ($\mu\text{W}/\text{cm}^2$)	0.17	0.22	0.6	0.23	-14.84	258	<0.001

PCA of $\text{TiO}_2|\text{Co}_3\text{O}_4$ and $\text{TiO}_2|\text{Co}_3\text{O}_4|\text{MoO}_3$ Libraries

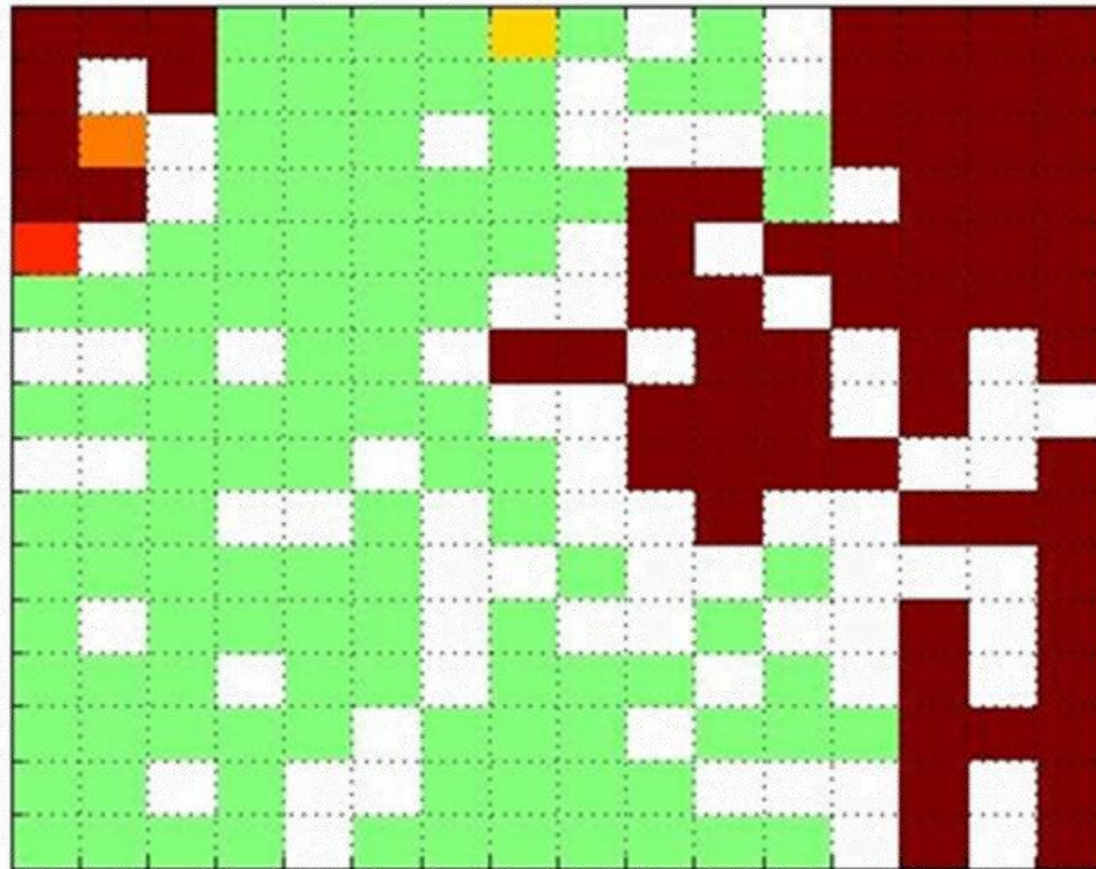







PCA of $\text{TiO}_2|\text{Co}_3\text{O}_4$ and $\text{TiO}_2|\text{Co}_3\text{O}_4|\text{MoO}_3$ Libraries



Activity	$\text{TiO}_2 \text{Co}_3\text{O}_4$		$\text{TiO}_2 \text{Co}_3\text{O}_4 \text{MoO}_3$		t-test	df	p-value
	Mean	SD	Mean	SD			
V_{oc} (mV)	431	18.9	484	36.8	-7.08	12	<0.001
J_{sc} ($\mu\text{A}/\text{cm}^2$)	17.4	0.7	15.3	1.8	6.26	12	<0.001
IQE (%)	0.07	0.002	0.06	0.006	6.27	12	<0.001
FF (%)	32.38	2.08	32.21	2.34	0.51	12	>0.05
P_{max} ($\mu\text{W}/\text{cm}^2$)	0.63	0.05	0.62	0.06	0.98	12	>0.05

SOM of $\text{TiO}_2 | \text{Co}_3\text{O}_4$ and $\text{TiO}_2 | \text{Co}_3\text{O}_4 | \text{MoO}_3$ Libraries

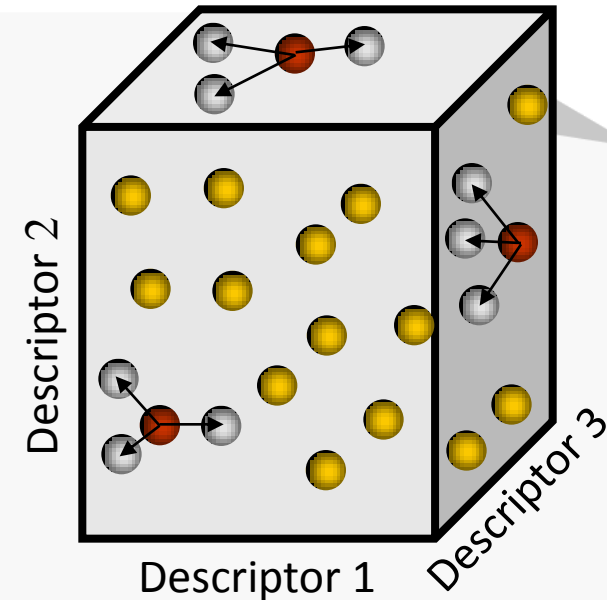


-  $\text{TiO}_2 | \text{Co}_3\text{O}_4$
-  $\text{TiO}_2 | \text{Co}_3\text{O}_4 | \text{MoO}_3$
-  A node with one cell from $\text{TiO}_2 | \text{Co}_3\text{O}_4$ and another cell from $\text{TiO}_2 | \text{Co}_3\text{O}_4 | \text{MoO}_3$
-  A node with one cell from $\text{TiO}_2 | \text{Co}_3\text{O}_4 | \text{MoO}_3$ and two cells from $\text{TiO}_2 | \text{Co}_3\text{O}_4$
-  Empty Node

Approaches for Model Building

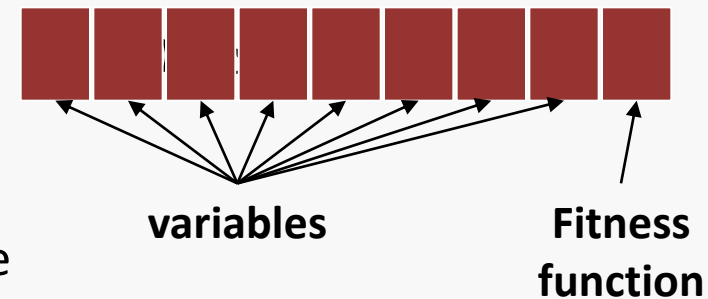
k Nearest Neighbors (*k*NN)

- The idea: Similar cells have similar photovoltaic properties
- The method: *k*NN predicts the property of a cell from the averaged properties of its *k* nearest neighbors
- The challenge: Identify the relevant descriptors space
- Advantages: Non-linear



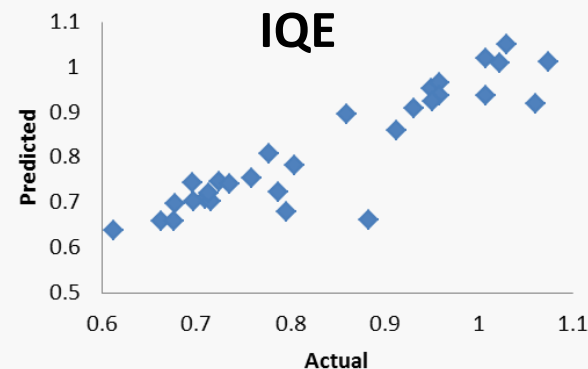
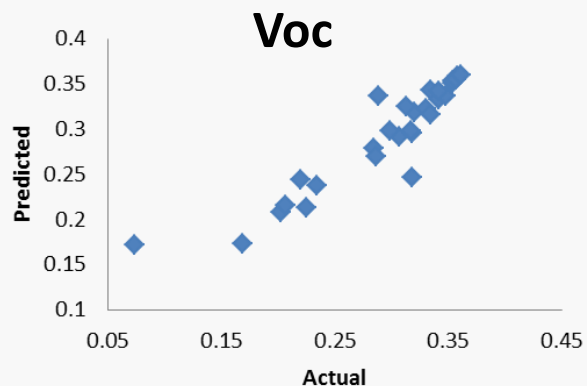
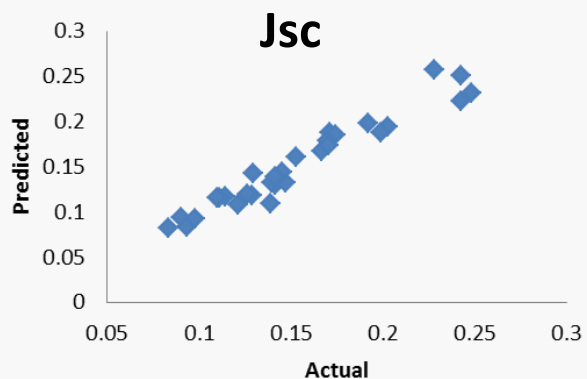
Genetic Function Approximation (GFA)

- The idea:
 - ❖ Create a population of equations (chromosomes)
 - ❖ Rank equations according to performances (fitness function)
 - ❖ Optimize fitness function using genetic operators
- Advantages: Multiple models, variable importance



Proof of Concept: TiO₂/Cu-O Library

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 _{TiO₂} , J _{max}
IQE(Ag)	0.77	0.80 (0.84)	0.05	0.83 (0.86)	0.04	87 %	T _{TiO₂} , 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

Proof of Concept: TiO₂/Cu₂O Library

Table 5. Results obtained with the kNN algorithm for the two TiO₂|Cu₂O sub-libraries (back contacts are given in parenthesis).

End point	Q_{Loo}^2	No applicability domain		With applicability domain			Descriptors
		$Q_{\text{ext}}^2 (R^2)$	MAE	$Q_{\text{ext}}^2 (R^2)$	MAE	%coverage	
J_{SC} (Ag)	0.92	0.92 (0.92)	0.02	0.92 (0.92)	0.02	91%	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
V_{OC} (Ag)	0.78	0.89 (0.89)	0.02	0.89 (0.89)	0.02	84%	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
IQE (Ag)	0.91	0.87 (0.87)	0.18	0.87 (0.87)	0.19	91%	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
J_{SC} (Ag/Cu)	0.92	0.89 (0.89)	0.02	0.88 (0.89)	0.02	79%	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$
V_{OC} (Ag/Cu)	0.92	0.88 (0.89)	0.02	0.89 (0.89)	0.02	82%	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$
IQE (Ag/Cu)	0.90	0.91 (0.91)	0.16	0.89 (0.89)	0.18	73%	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$

Table 6. Results obtained with the GP algorithm for the two TiO₂|Cu₂O sub-libraries (back contacts are given in parenthesis).

Model	R_{CV}^2	$Q_{\text{ext}}^2 (R^2)$	MAE
J_{SC} (Ag) = $0.0009 \times T_{\text{Cu}_2\text{O}} - 0.22$	0.74	0.76 (0.76)	0.04
V_{OC} (Ag) = $0.00047 \times T_{\text{TiO}_2} + 0.0004 \times T_{\text{Cu}_2\text{O}}$	0.65	0.78 (0.77)	0.02
IQE (Ag) = $0.0058 \times T_{\text{Cu}_2\text{O}} - 1.26$	0.70	0.72 (0.73)	0.28
J_{SC} (Ag/Cu) = $0.0009 \times T_{\text{Cu}_2\text{O}} - 0.22$	0.76	0.74 (0.76)	0.04
V_{OC} (Ag/Cu) = $0.00048 \times T_{\text{TiO}_2} + 0.0004 \times T_{\text{Cu}_2\text{O}}$	0.61	0.50 (0.50)	0.04
IQE (Ag/Cu) = $0.0059 \times T_{\text{Cu}_2\text{O}} - 1.34$	0.72	0.72 (0.73)	0.28

The Effect of the Library's Quality: *k*NN

Less uniform library

Table 2. Results obtained with the *k*NN algorithm for the TiO₂|Cu-O library.

End point	Q_{LOO}^2	No applicability domain		With applicability domain			Descriptors
		Q_{ext}^2 (R^2)	MAE	Q_{ext}^2 (R^2)	MAE	%coverage	
$J_{\text{SC}}(\text{Ag})$	0.87	0.86 (0.88)	0.01	0.86 (0.89)	0.01	83 %	$\text{Ratio}, \text{BGP}, D_{\text{center}}$
$V_{\text{OC}}(\text{Ag})$	0.86	0.73 (0.74)	0.02	0.75 (0.77)	0.02	75 %	$T_{\text{TiO}_2}, J_{\text{max}}$
$\text{IQE}(\text{Ag})$	0.77	0.80 (0.84)	0.05	0.83 (0.86)	0.04	87 %	$T_{\text{TiO}_2}, \text{Ratio}, R_s$

More uniform library

Table 5. Results obtained with the *k*NN algorithm for the two TiO₂|Cu₂O sub-libraries (back contacts are given in parenthesis).

End point	Q_{LOO}^2	No applicability domain		With applicability domain			Descriptors
		Q_{ext}^2 (R^2)	MAE	Q_{ext}^2 (R^2)	MAE	%coverage	
$J_{\text{SC}}(\text{Ag})$	0.92	0.92 (0.92)	0.02	0.92 (0.92)	0.02	91 %	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
$V_{\text{OC}}(\text{Ag})$	0.78	0.89 (0.89)	0.02	0.89 (0.89)	0.02	84 %	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
$\text{IQE}(\text{Ag})$	0.91	0.87 (0.87)	0.18	0.87 (0.87)	0.19	91 %	$T_{\text{TiO}_2}, T_{\text{Cu}_2\text{O}}$
$J_{\text{SC}}(\text{Ag}/\text{Cu})$	0.92	0.89 (0.89)	0.02	0.88 (0.89)	0.02	79 %	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$
$V_{\text{OC}}(\text{Ag}/\text{Cu})$	0.92	0.88 (0.89)	0.02	0.89 (0.89)	0.02	82 %	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$
$\text{IQE}(\text{Ag}/\text{Cu})$	0.90	0.91 (0.91)	0.16	0.89 (0.89)	0.18	73 %	$T_{\text{Cu}_2\text{O}}, \text{Ratio}$

The Effect of the Library's Quality: GA

Less uniform library

Table 3. Results obtained with the GP algorithm for the TiO₂|Cu-O library.

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

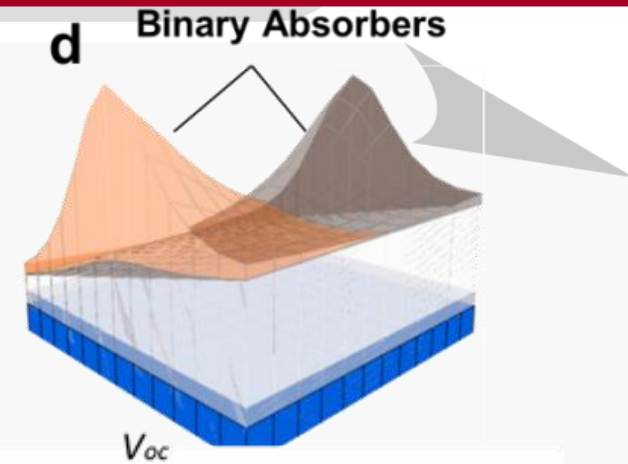
More uniform library

Table 6. Results obtained with the GP algorithm for the two TiO₂|Cu₂O sub-libraries (back contacts are given in parenthesis).

Model	R_{CV}^2	$Q_{ext}^2 (R^2)$	MAE
$J_{SC} (Ag) = 0.0009 \times T_{Cu_2O} - 0.22$	0.74	0.76 (0.76)	0.04
$V_{OC} (Ag) = 0.00047 \times T_{TiO_2} + 0.0004 \times T_{Cu_2O}$	0.65	0.78 (0.77)	0.02
$IQE (Ag) = 0.0058 \times T_{Cu_2O} - 1.26$	0.70	0.72 (0.73)	0.28
$J_{SC} (Ag/Cu) = 0.0009 \times T_{Cu_2O} - 0.22$	0.76	0.74 (0.76)	0.04
$V_{OC} (Ag/Cu) = 0.00048 \times T_{TiO_2} + 0.0004 \times T_{Cu_2O}$	0.61	0.50 (0.50)	0.04
$IQE (Ag/Cu) = 0.0059 \times T_{Cu_2O} - 1.34$	0.72	0.72 (0.73)	0.28

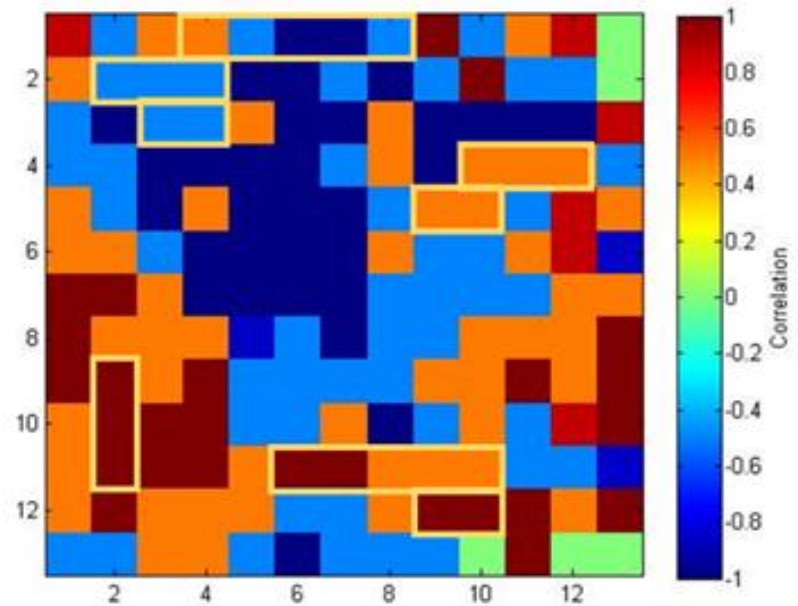
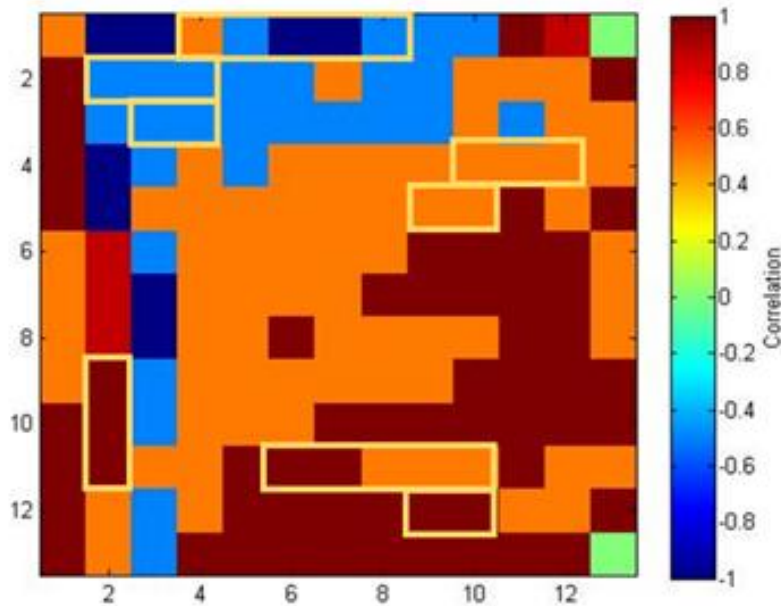
Experimental Design I: $\text{TiO}_2\text{Cu}_2\text{ONiO}$ Library

- Gradients of $\text{TiO}_2/\text{Cu}_2\text{O}$
- NiO: 0, 5, 10 nm
- Search for correlation between thickness of NiO layer and PV parameters

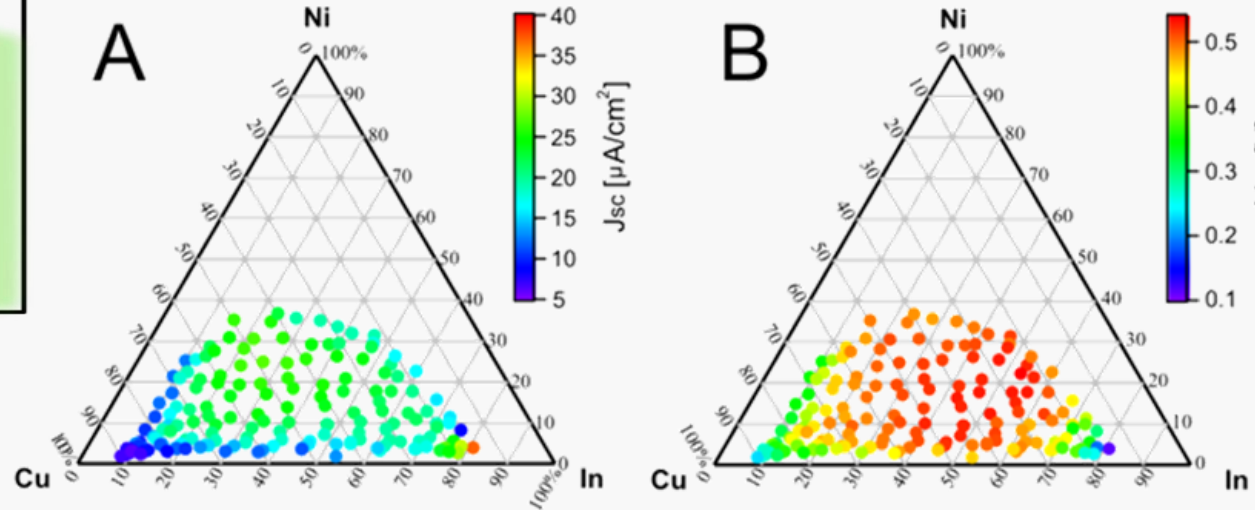
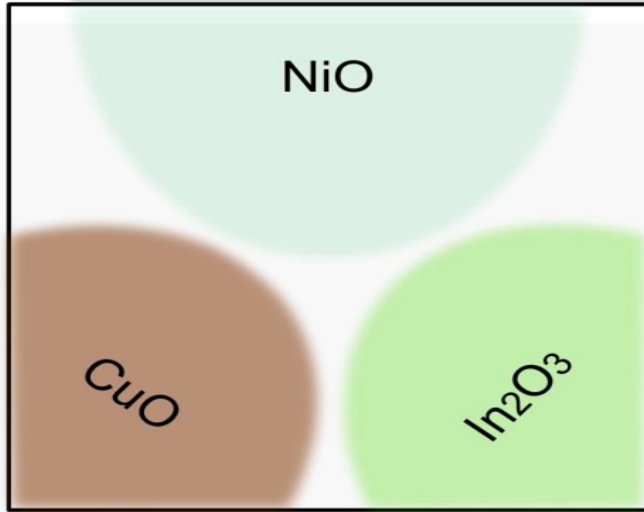


J_{sc}

V_{oc}



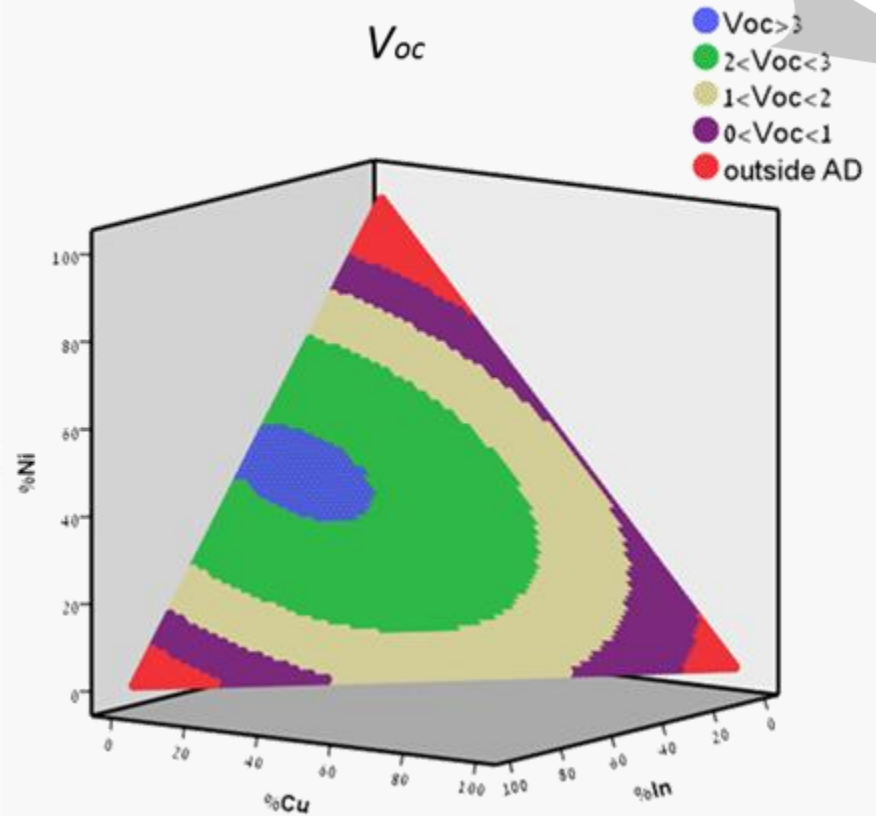
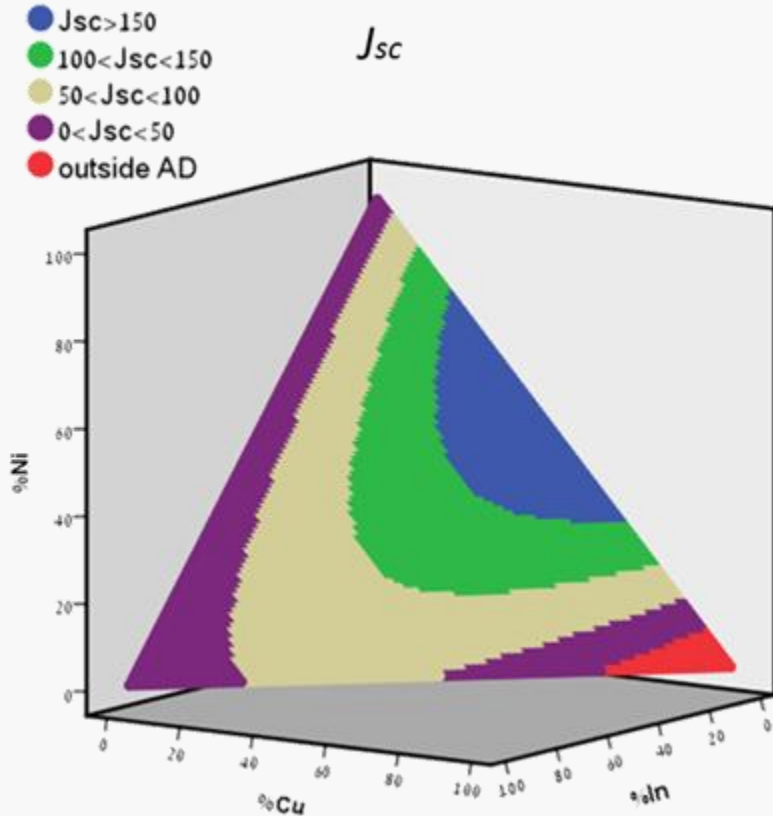
PV Parameters for a TiO_2 | $\text{CuO-NiO-In}_2\text{O}_3$ Library



$$J_{SC} = a_1 - a_2 \times \%Cu + a_3 \times \%In \times \%Cu + a_4 \times \%Ni \times \%Cu \quad R^2 = 0.67$$

$$V_{OC} = a_1 - a_2 \times \%Ni - a_3 \times \%Cu - a_4 \times \%In + a_5 \times \%In \times \%Ni + a_6 \times \%In \times \%Cu + a_7 \times \%Ni \times \%Cu \quad R^2 = 0.83$$

Experimental Design II



Simultaneous improvement of both V_{OC} and J_{SC} is complicated by conflicting requirements in terms of cell compositions

Conclusions

- Statistical modeling is useful in material science
 - ❖ Insight
 - ❖ Experimental design
- Challenges
 - ❖ Data curation
 - ❖ Problem specific descriptors
 - ❖ Model validation
- The similar properties principle holds for PV cells
- Both the strengths and the weaknesses of statistical modeling approaches lies in their “ignorance”