Statistical Modeling In Material Sciences

Abraham Yosipof, Oren Nahum, Hanoch Senderowitz
Bar-Ilan University, Israel

The Second Kazan Summer School of Chemoinformatics, July 2015, Kazan
Acknowledgments

Avi Yosipof    Oren Nachum

Funding: INNI
Presentation Outline

- Statistical modeling in material Sciences
- Photovoltaic cells
- All oxides photovoltaic cells
  - Models
  - Experimental design
- Summary and conclusion
Big Data and Material Informatics

A set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

- Commerce: Analyze consumer purchases to increase profit
- Health: Analyze data to predict the outbreak of epidemics
- Biology: Analyze the genome to develop drugs
- Material Sciences: Analyze data to discover new materials
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.
Statistical Modeling

- Statistical modeling
- QSAR
- Machine learning

Time:
- Years
- Hours
- Minutes
- Seconds
- Microseconds
- Nanoseconds
- Picoseconds
- Femtoseconds

Distance:
- 1 Å
- 1 nm
- 1 μm
- 1 mm
- meters

- Engineering design
- Unit process design
- Finite element analysis
- Process simulation

- e.g., morphology, drug delivery
- e.g., permeation, crystallization
- e.g., electronics, catalysis

Statistical modeling, QSAR, QSAR, Machine learning
In 1863, A.F.A. Cros at the University of Strasbourg observed that toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased.

In the 1890's, Hans Horst Meyer of the University of Marburg and Charles Ernest Overton of the University of Zurich, working independently, noted that the toxicity of organic compounds depended on their lipophilicity.

Louis Hammett (1894-1987), correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity.

Hansch (1969) recognized the importance of the lipophilicity, expressed as the octanol-water partition coefficient, on biological activity (bioavailability).
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

Katritzki et al., ChemPlusChem 2012, 77, 507
Statistical Models

- Accurate experimental data
- 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
What is the Same? What is Different?

• Data
  - Size of data sets (?)
  - Complexity (?)

• Methods and Validation
  - Data reduction techniques (e.g., PCA)
  - Clustering
  - Classification models (e.g., Random Forests)
  - Quantitative models (e.g., MLR, SVM, kNN)
  - “OECD” principles and validation (not always)

• Descriptors
  - Something new something old something borrowed (something blue and a silver sixpence in her shoe)
  - And here lies the challenge
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

- 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)
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

Isayev et al., Chem. Mater. 2015, 27, 735–743
Spectra-Based Descriptors

Useful when composition / structure of the compound is unknown
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)
  - \[ PCE = \frac{FF \times J_{SC} \times V_{OC}}{P_{in}} \]
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 for PV Cells

• Goals
  ❖ Identify factors responsible for PV properties
  ❖ Experimental design
  ❖ Insight into factor affecting PV

• Assumptions
  ❖ A correlation exists between PV properties and cells characteristics:
    \[ PV = f(\text{Material Descriptors}) \]
    ○ Nature of correlation not necessarily known
  ❖ PV properties depend on measurable cell characteristics
Organic Photovoltaics

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

Olivares-Amaya et al., Energy Environ. Sci., 2011, 4, 4849–4861
Color coding refers to the top 10% molecules with highest predicted Voc (green), Jsc (blue), and Voc x Jsc (red). Best molecules are located in the upper left.
Dye-Sensitized Solar Cells (DSSC)

- Ruthenium sensitizers design
- Training set: 65 compounds
- Volsurf+ (MIFs-based), QM and “classical” descriptors
- PLS regression

Different dyes sensitized by incident radiation

Tortorella et al., RSC Adv., 2015, 5, 23865–23873
Dye-Sensitized Solar Cells (DSSC)

- Inversely correlated: presence of NO₂ and NH₂, PSA/HAS and PSA/SA ratios, H-bond
- Correlated: P n-oct, P c-Hex, log D5/log D10, flexibility

Tortorella et al., RSC Adv., 2015, 5, 23865–23873
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
~60 “useful” elements leading to
- ~30K inorganic compounds
- 3600 binary compounds (ABO$_x$); mostly known
- 216K ternary compounds (ABCO$_x$) almost all unknown

High Tc superconductors
Synthesis of Libraries of All Oxide PV Cells
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

Experimental Design → Experimental Data → Library Characterization → Model Building → Validated Model → Data Visualization

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

- Libraries
  - Window: TiO$_2$; Absorber: Cu-O
  - Window: TiO$_2$, Absorber: Cu$_2$O
  - Window: TiO$_2$, Absorber: Fe$_2$O$_3$
  - Window: TiO$_2$, Absorber: Co$_2$O$_3$

- Dependent variables
  - $V_{OC}$ (Open circuit voltage)
  - $J_{SC}$ (Short circuit current)
  - IQE (Internal quantum efficiency)
Material Descriptors

- Thickness of window layer \((T_w)\)
- Thickness of the absorber layer \((T_a)\)
- Thickness ratio \((T_a / (T_a + T_w))\)
- Distance from center of deposition plume \((D_{center})\)
- Maximum calculated photocurrent \((J_{max})\)
- Resistance \((R_a)\)
- Band gap \((BGP)\)

Descriptors pre-processing
- Remove constant descriptors
- Remove nearly constant descriptors
- Remove correlated descriptors
- Remove descriptors whose averaged value for “active” cells is similar to their averaged value for “inactive” cells
** Principle Component Analysis (PCA) **

** TiO$_2$/Cu-O **

** TiO$_2$/Fe$_2$O$_3$ **
(treated at different temperatures)

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{TiO2}$</td>
<td>TEMP</td>
</tr>
<tr>
<td>$d_{Fe2O3}$</td>
<td>BGP$_{Fe2O3}$</td>
</tr>
<tr>
<td>$[Fe2O3] / ([TiO2] + [Fe2O3])$</td>
<td>max$_J$</td>
</tr>
</tbody>
</table>
Focus on the TiO$_2$/Cu-O Library

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{Cu-O}}$</td>
<td>$T_{\text{TiO}_2}$</td>
</tr>
<tr>
<td>Ratio</td>
<td>$R_x$</td>
</tr>
<tr>
<td>Depth</td>
<td>$D_{\text{depth}}$</td>
</tr>
<tr>
<td>Int.</td>
<td>$I_{\text{int}}$</td>
</tr>
<tr>
<td>BGP</td>
<td></td>
</tr>
</tbody>
</table>

(A) PC1

(B) PC2

(C) $T_{\text{Cu-O}}$

(D) $T^{x}_{\text{TiO}_2}$
PCA of the TiO$_2$/Cu$_2$O Library
Property Space and $k$ Nearest Neighbors ($k$NN)

- Axis correspond to material descriptors
- Points correspond to solar cells
- Distance between points corresponds to similarity between solar cells

$k$NN predicts PV of a solar cell from the averaged PV of its $k$ nearest neighbors

The “trick” is to identify the relevant descriptors space
Identifying the Relevant Property Space

For $n$ descriptors, $2^n - 1$ options

Randomly select subset of descriptors

Calculate distances between compounds

Repeatedly leave out 1 compound

Predict activity for each compound from its $k$ nearest neighbors

Calculate model performance

$$q^2 = 1 - \frac{\sum (y_{exp} - y_{pred})}{\sum (y_{exp} - \bar{y})}$$

Select best QSAR model

Optimize $q^2$ by changing descriptors

Best model defined in terms of descriptors and $k$
Monte Carlo/Simulated Annealing (MC/SA)

Random Move → "Trial" → \( \Delta E \)

NO

YES

\[ \Delta E < 0 \quad \text{or} \quad \exp(-\Delta E/RT) > X_{[0,1]} \]

\( X_{[0,1]} \) is a random number in the range 0 to 1.

T\text{max} \quad \downarrow \quad \text{MC} \quad \downarrow \quad T_{\text{min}}

Temperature

MC Steps

Temperature

MC Steps

Temperature

MC Steps

Temperature

MC Steps
**kNN-Based Removal of Outliers**

1. **Data set**
2. Run kNN optimization to obtain the best model
3. For each solar cell, calculate the improvement in $Q^2_{LOO}$ upon its removal from the data set
4. Remove that solar cell which provides the largest increase in $Q^2_{LOO}$ upon its removal from the data set
5. **Stopping criteria**

---

$k$NN-Based Removal of Outliers and Activity Cliffs

![Graph and heat maps for TiO$_2$/CuO]
Splitting the Data

- **Models should be validate on an external test set**
- Data split into a modeling set and an independent test set using a new algorithm
### Approaches for Model Building

#### k Nearest Neighbors (kNN)

- **The idea:** Similar cells have similar photovoltaic properties
- **The method:** kNN 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 equation (chromosomes)
  - Rank equations according to performances (fitness function)
  - Optimize fitness function using genetic operators
- **Advantages:** Multiple models, variable importance
Proof of Concept: TiO$_2$/Cu-O Library

<table>
<thead>
<tr>
<th>End point</th>
<th>$Q_{100}^2$</th>
<th>No applicability domain</th>
<th>With applicability domain</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{SC}(Ag)$</td>
<td>0.87</td>
<td>0.86 (0.88)</td>
<td>0.01</td>
<td>0.86 (0.89)</td>
</tr>
<tr>
<td>$V_{OC}(Ag)$</td>
<td>0.86</td>
<td>0.73 (0.74)</td>
<td>0.02</td>
<td>0.75 (0.77)</td>
</tr>
<tr>
<td>$IQE(Ag)$</td>
<td>0.77</td>
<td>0.80 (0.84)</td>
<td>0.05</td>
<td>0.83 (0.86)</td>
</tr>
</tbody>
</table>

**Predicted vs Actual:**

### Jsc

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.25</td>
</tr>
<tr>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>0.25</td>
<td>0.45</td>
</tr>
<tr>
<td>0.35</td>
<td>0.55</td>
</tr>
<tr>
<td>0.45</td>
<td>0.65</td>
</tr>
</tbody>
</table>

### Voc

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.25</td>
</tr>
<tr>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>0.25</td>
<td>0.45</td>
</tr>
<tr>
<td>0.35</td>
<td>0.55</td>
</tr>
<tr>
<td>0.45</td>
<td>0.65</td>
</tr>
</tbody>
</table>

### IQE

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>1.0</td>
<td>1.2</td>
</tr>
</tbody>
</table>

**Model Equations:**

- $J_{SC} = 0.062 + 0.0004 \times T_{Cu-O} - 430384.1022/R_b$
- $V_{OC} = 0.011 \times J_{max} + 1.201 \times 10^{-5} \times T_{TiO_2} \times D_{center} - 0.04 - 6.62 \times 10^{-17} \times T_{Cu-O} \times R_b$
- $IQE = 1.784 \times Ratio + 0.072/Ratio - 2642279.244/(2356681.705 + R_b)$

<table>
<thead>
<tr>
<th>Model</th>
<th>$R_{CV}^2$</th>
<th>$Q_{ext}^2 (R^2)$</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{SC}$</td>
<td>0.88</td>
<td>0.86 (0.87)</td>
<td>0.01</td>
</tr>
<tr>
<td>$V_{OC}$</td>
<td>0.62</td>
<td>0.54 (0.55)</td>
<td>0.03</td>
</tr>
<tr>
<td>$IQE$</td>
<td>0.65</td>
<td>0.74 (0.74)</td>
<td>0.06</td>
</tr>
</tbody>
</table>
### Proof of Concept: TiO$_2$/Cu$_2$O Library

**Table 5.** Results obtained with the kNN algorithm for the two TiO$_2$|Cu$_2$O sub-libraries (back contacts are given in parenthesis).

<table>
<thead>
<tr>
<th>End point</th>
<th>$Q^2_{LOO}$</th>
<th>No applicability domain</th>
<th>With applicability domain</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q^2_{ext}$ ($R^2$)</td>
<td>MAE</td>
<td>$Q^2_{ext}$ ($R^2$)</td>
<td>MAE</td>
</tr>
<tr>
<td>$J_{SC}$ (Ag)</td>
<td>0.92</td>
<td>0.92 (0.92)</td>
<td>0.02</td>
<td>0.92 (0.92)</td>
</tr>
<tr>
<td>$V_{OC}$ (Ag)</td>
<td>0.78</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>0.89 (0.89)</td>
</tr>
<tr>
<td>$IQE$ (Ag)</td>
<td>0.91</td>
<td>0.87 (0.87)</td>
<td>0.18</td>
<td>0.87 (0.87)</td>
</tr>
<tr>
<td>$J_{SC}$ (Ag/Cu)</td>
<td>0.92</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>0.88 (0.89)</td>
</tr>
<tr>
<td>$V_{OC}$ (Ag/Cu)</td>
<td>0.92</td>
<td>0.88 (0.89)</td>
<td>0.02</td>
<td>0.89 (0.89)</td>
</tr>
<tr>
<td>$IQE$ (Ag/Cu)</td>
<td>0.90</td>
<td>0.91 (0.91)</td>
<td>0.16</td>
<td>0.89 (0.89)</td>
</tr>
</tbody>
</table>

**Table 6.** Results obtained with the GP algorithm for the two TiO$_2$|Cu$_2$O sub-libraries (back contacts are given in parenthesis).

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2_{CV}$</th>
<th>$Q^2_{ext}$ ($R^2$)</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{SC}$ (Ag) = 0.0009 $T_{CuO}$ - 0.22</td>
<td>0.74</td>
<td>0.76 (0.76)</td>
<td>0.04</td>
</tr>
<tr>
<td>$V_{OC}$ (Ag) = 0.00047 $T_{TiO_2}$ + 0.0004 $T_{CuO}$</td>
<td>0.65</td>
<td>0.78 (0.77)</td>
<td>0.02</td>
</tr>
<tr>
<td>$IQE$ (Ag) = 0.0058 $T_{CuO}$ - 1.26</td>
<td>0.70</td>
<td>0.72 (0.73)</td>
<td>0.28</td>
</tr>
<tr>
<td>$J_{SC}$ (Ag/Cu) = 0.0009 $T_{CuO}$ - 0.22</td>
<td>0.76</td>
<td>0.74 (0.76)</td>
<td>0.04</td>
</tr>
<tr>
<td>$V_{OC}$ (Ag/Cu) = 0.00048 $T_{TiO_2}$ + 0.0004 $T_{CuO}$</td>
<td>0.61</td>
<td>0.50 (0.50)</td>
<td>0.04</td>
</tr>
<tr>
<td>$IQE$ (Ag/Cu) = 0.0059 $T_{CuO}$ - 1.34</td>
<td>0.72</td>
<td>0.72 (0.73)</td>
<td>0.28</td>
</tr>
</tbody>
</table>
### The Effect of the Library’s Quality: \( kNN \)

#### Table 2. Results obtained with the \( kNN \) algorithm for the TiO\(_2\)–Cu–O library.

<table>
<thead>
<tr>
<th>End point</th>
<th>( Q_{100}^2 )</th>
<th>No applicability domain</th>
<th></th>
<th>With applicability domain</th>
<th></th>
<th>%coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_{SC}(Ag) )</td>
<td>0.87</td>
<td>0.86 (0.88)</td>
<td>0.01</td>
<td>0.86 (0.89)</td>
<td>0.01</td>
<td>83%</td>
</tr>
<tr>
<td>( V_{OC}(Ag) )</td>
<td>0.86</td>
<td>0.73 (0.74)</td>
<td>0.02</td>
<td>0.75 (0.77)</td>
<td>0.02</td>
<td>75%</td>
</tr>
<tr>
<td>( I_{QE}(Ag) )</td>
<td>0.77</td>
<td>0.80 (0.84)</td>
<td>0.05</td>
<td>0.83 (0.86)</td>
<td>0.04</td>
<td>87%</td>
</tr>
</tbody>
</table>

#### Table 5. Results obtained with the \( kNN \) algorithm for the two TiO\(_2\)–Cu\(_2\)O sub-libraries (back contacts are given in parenthesis).

<table>
<thead>
<tr>
<th>End point</th>
<th>( Q_{100}^2 )</th>
<th>No applicability domain</th>
<th></th>
<th>With applicability domain</th>
<th></th>
<th>%coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J_{SC} (Ag) )</td>
<td>0.92</td>
<td>0.92 (0.92)</td>
<td>0.02</td>
<td>0.92 (0.92)</td>
<td>0.02</td>
<td>91%</td>
</tr>
<tr>
<td>( V_{OC} (Ag) )</td>
<td>0.78</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>84%</td>
</tr>
<tr>
<td>( I_{QE}(Ag) )</td>
<td>0.91</td>
<td>0.87 (0.87)</td>
<td>0.18</td>
<td>0.87 (0.87)</td>
<td>0.19</td>
<td>91%</td>
</tr>
<tr>
<td>( J_{SC} (Ag/Cu) )</td>
<td>0.92</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>0.88 (0.89)</td>
<td>0.02</td>
<td>79%</td>
</tr>
<tr>
<td>( V_{OC} (Ag/Cu) )</td>
<td>0.92</td>
<td>0.88 (0.89)</td>
<td>0.02</td>
<td>0.89 (0.89)</td>
<td>0.02</td>
<td>82%</td>
</tr>
<tr>
<td>( I_{QE}(Ag/Cu) )</td>
<td>0.90</td>
<td>0.91 (0.91)</td>
<td>0.16</td>
<td>0.89 (0.89)</td>
<td>0.18</td>
<td>73%</td>
</tr>
</tbody>
</table>

**Descriptors**
- Ratio, BGP, \( D_{center} \)
- \( T_{TiO_2}, J_{max} \)
- \( T_{TiO_2}, Ratio, R_s \)

**Descriptors**
- \( T_{TiO_2}, T_{Cu_2O} \)
- \( T_{TiO_2}, T_{Cu_2O} \)
- \( T_{Cu_2O}, Ratio \)
- \( T_{Cu_2O}, Ratio \)
- \( T_{Cu_2O}, Ratio \)
The Effect of the Library’s Quality: GA

Table 3. Results obtained with the GP algorithm for the TiO\textsubscript{2} | Cu-O library.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2_{CV}$</th>
<th>$Q^2_{ext}$ (R\textsuperscript{2})</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{SC} = 0.062 + 0.0004 \times T_{Cu-O} - 430384.1022/R_0$</td>
<td>0.88</td>
<td>0.86 (0.87)</td>
<td>0.01</td>
</tr>
<tr>
<td>$V_{OC} = 0.011 \times J_{max} + 1.201 \times 10^{-3} \times T_{NO_2} \times D_{center} - 0.04 - 6.62 \times 10^{-13} \times T_{Cu-O} \times R_0$</td>
<td>0.62</td>
<td>0.54 (0.55)</td>
<td>0.03</td>
</tr>
<tr>
<td>$I_{QE} = 1.784 \times \text{Ratio} + 0.072/\text{Ratio} - 2642279.244/(2356681.705 + R_0)$</td>
<td>0.65</td>
<td>0.74 (0.74)</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 6. Results obtained with the GP algorithm for the two TiO\textsubscript{2} | Cu\textsubscript{2}O sub-libraries (back contacts are given in parenthesis).

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2_{CV}$</th>
<th>$Q^2_{ext}$ (R\textsuperscript{2})</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{SC} (Ag) = 0.00009 \times T_{Cu_2O} - 0.22$</td>
<td>0.74</td>
<td>0.76 (0.76)</td>
<td>0.04</td>
</tr>
<tr>
<td>$V_{OC} (Ag) = 0.00047 \times T_{NO_2} + 0.0004 \times T_{Cu_2O}$</td>
<td>0.65</td>
<td>0.78 (0.77)</td>
<td>0.02</td>
</tr>
<tr>
<td>$I_QE (Ag) = 0.0058 \times T_{Cu_2O} - 1.26$</td>
<td>0.70</td>
<td>0.72 (0.73)</td>
<td>0.28</td>
</tr>
<tr>
<td>$J_{SC} (Ag/Cu) = 0.00009 \times T_{Cu_2O} - 0.22$</td>
<td>0.76</td>
<td>0.74 (0.76)</td>
<td>0.04</td>
</tr>
<tr>
<td>$V_{OC} (Ag/Cu) = 0.00048 \times T_{NO_2} + 0.0004 \times T_{Cu_2O}$</td>
<td>0.61</td>
<td>0.50 (0.50)</td>
<td>0.04</td>
</tr>
<tr>
<td>$I_QE (Ag/Cu) = 0.0059 \times T_{Cu_2O} - 1.34$</td>
<td>0.72</td>
<td>0.72 (0.73)</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Experimental Design I: Virtual Library of PV Cells

- TiO$_2$/Co$_3$O$_4$ library with Ag/Cu back contacts
- GFA yielded $r^2_{test}(J_{sc}) = 0.79$; $r^2_{test}(IQE) = 0.91$; $r^2_{test}(V_{oc}) = 0.71$
- 1000 virtual cells generated and predicted
- Validation ongoing

![IQE (%) vs Virtual cell graph]
Experimental Design II: TiO$_2$Cu$_2$ONiO Library

- Gradients of TiO$_2$/Cu$_2$O
- NiO: 0, 5, 10 nm
- Search for correlation between thickness of NiO layer and PV parameters
Insight into Factors Affecting PV: The Role of the Absorber

In the absorber layer, photons of incident radiation are absorbed resulting in the creation of electron-hole pairs.

**Single Absorber**

**Binary Absorber**

**Ternary Absorber**

\[ \text{Cu}_x\text{O}_y \]  
\[ \text{NiO} \]  
\[ \text{In}_2\text{O}_3 \]

- \( \text{CuO-NiO-In}_2\text{O}_3 \)
Photoconductivity Ratio

A measure of how well does the absorber creates electron-hole pairs when irradiated

Hypothesis:
• Ni atoms improve photoconductivity
• Ni atoms serve as doping agents
QSAR Models for Photoconductivity Ratio

- Y: Photoconductivity ratio
- X: CuO-NiO-In$_2$O$_3$ thickness, %O, %Cu, %Ni, %In, Band gap

- $k$NN Model: %Ni, %In; $Q^2_{cv} = 0.51$
- GFA Model: $2.42 + 1.62 \times %Ni + 0.87 \times %In - 0.17 \times %Ni \times %In$; $Q^2_{cv} = 0.51$
- MLR Model:

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Ni</td>
<td>2.4</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>%In</td>
<td>1.4</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>%Ni X %In</td>
<td>-2.0</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
Insight Into the Interaction Term (%Ni x %In)

• Separate data into 2 classes based on %In: High %In (> 7%); Low %In (< 7%)

• Correlation
  - High %In: Photoconductivity ratio Vs %Ni (r = 0.156, p-value > 0.05)
  - Low %In: Photoconductivity ratio Vs %Ni (r = 0.883, p-value < 0.01)

• Linear regression
PV Parameters for the TiO$_2$Cu$_2$ONiO Library

\[ J_{sc} = 0.018 -0.001\%Cu + 0.000032\%In\%Cu + 0.000094\%Ni\%Cu \]
\[ R^2 = 0.67 \]

\[ V_{OC} = 0.786 – 0.032\%Ni – 0.017\%Cu – 0.023\%In \]
\[ + 0.002\%In\%Ni + 0.001\%In\%Cu + 0.001\%Ni\%Cu \]
\[ R^2 = 0.83 \]
Model Interpretation ($J_{SC}$)

<table>
<thead>
<tr>
<th></th>
<th>Cu-$J_{sc}$</th>
<th>Ni-$J_{sc}$</th>
<th>In-$J_{sc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Cu</td>
<td>-</td>
<td>0.232</td>
<td>-0.038</td>
</tr>
<tr>
<td>High Cu</td>
<td>-</td>
<td><strong>0.682</strong></td>
<td><strong>0.309</strong></td>
</tr>
<tr>
<td>Low Ni</td>
<td>-0.795</td>
<td>-</td>
<td><strong>0.767</strong></td>
</tr>
<tr>
<td>High Ni</td>
<td>-0.223</td>
<td>-</td>
<td><strong>0.092</strong></td>
</tr>
<tr>
<td>Low In</td>
<td>-0.743</td>
<td><strong>0.765</strong></td>
<td>-</td>
</tr>
<tr>
<td>High In</td>
<td>-0.323</td>
<td><strong>0.328</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Graph showing the relationship between %Cu and $J_{sc}$.
Experimental Design

$J_{sc}$  $V_{oc}$
Summary

Experimental Design: Ongoing

Experimental Data → Library Characterization → Data Visualization → Model Building → Validated Model
Conclusions

• Models with good prediction statistics are attainable for PV properties using both \( kNN \) and GP
• The similar properties principle holds for PV cells
• The thickness of the window and absorber layers are important determinant of PV properties of solar cells
• Data mining tools are useful in the field of PV
  - Insight
    - Experimental design
• Their weakness lies in their “ignorance”
• Their strength lies in their “ignorance”