***M.Sc. of System Engineering***

**Decision Support System for Optimizing Pharma Research**

***M.Sc. Final Research Project***

***5/19/17***

***Shy Alon***

***038505665***

**Advisors:** Dr. Abraham Yosipof

Table of Contents

[Abstract 7](#_Toc480192465)

[1. Introduction 8](#_Toc480192466)

[1.1. The need for Decision Support System for Financial Stability 8](#_Toc480192467)

[2. Methods 11](#_Toc480192468)

[2.1. System Overview 11](#_Toc480192469)

[2.2. Input Data 13](#_Toc480192470)

[2.3. Feature Selection 13](#_Toc480192471)

[2.4. Run t-SNE 14](#_Toc480192472)

[2.5. Fitness Calculation 15](#_Toc480192473)

[2.6. Model Selection 15](#_Toc480192474)

[3. Results and Discussion 16](#_Toc480192475)

[3.1. Pre-Processing Implementation 16](#_Toc480192476)

[3.1.1. Data bases 16](#_Toc480192477)

[3.1.2. Library Characterization 17](#_Toc480192478)

[3.1.3. Basic statistical analysis 18](#_Toc480192479)

[3.1.4. Principal Component Analysis 18](#_Toc480192480)

[3.1.5. Heat Map 21](#_Toc480192481)

[3.1.6. Self-Organizing Map (SOM) 23](#_Toc480192482)

[3.2. RANSAC Implementation 24](#_Toc480192483)

[3.2.1. Databases 24](#_Toc480192484)

[3.3. kNN Implementation 30](#_Toc480192485)

[3.4. Virtual Cell Implementation 31](#_Toc480192486)

[3.4.1. (Ag) and (Ag|Cu) virtual libraries 31](#_Toc480192487)

[3.4.2. virtual library 33](#_Toc480192488)

[4. Conclusion 35](#_Toc480192489)

[5. List of Publications 36](#_Toc480192490)

[6. References 37](#_Toc480192491)

[Appendix A – GUI Screens 39](#_Toc480192492)

List of Figures

[Figure 1 Commonly Used Financial Indicators 9](#_Toc483041605)

[Figure 2 t-SNE Optimization Algorithm 12](#_Toc483041606)

[Figure 3 Bankruptcy Map Based on Financial Indicators 14](#_Toc483041607)

[Figure 4 Local as Opposed to Global Distances 15](#_Toc483041608)

[**Figure 5:** A schematic representation of the PV solar cells library. (a) TiO2|Co3O4 library (b) TiO2|Co3O4|MoO3 library. 18](#_Toc483041609)

[Figure 6: A PC plot for photovoltaic cells from the two solar cell libraries. Blue and red dots represent cells from the TiO2|Co3O4 and TiO2|Co3O4|MoO3 libraries, respectively. 20](#_Toc483041610)

[Figure 7: A PC plot of photovoltaic cells from the two solar cell libraries. Blue and red dots represent cells from the TiO2|Co3O4 and TiO2|Co3O4|MoO3 libraries, respectively. Green dots indicate 13 cell pairs from the two libraries with identical IDs and distances from one another in the lowest 10th percentile. 22](#_Toc483041611)

[zFigure 8: Heat map of the PV libraries, where each square represents a cell in the library. The heat map is colored according to the thickness of the Co3O4 layer. 'X' represents cells with similar PV behavior in both libraries according to the PCA. The X and Y axes represent the location of the cell on the 13 x 13 grid. 22](#_Toc483041612)

[Figure 9: A SOM composed of 16 x 16 = 256 neurons (i.e nodes) of the two solar cell libraries, where each colored square represents a node color coded according to its library membership (see legend). White squares represent empty neurons. 24](#_Toc483041613)

[Figure 10: A SOM composed of 10 x 10 = 100 neurons (i.e nodes) of the two solar cell libraries, where each colored square represents a node color coded according to its library membership (see legend). White squares represent empty neurons. 24](#_Toc483041614)

[**Figure 11:** A schematic representation of the PV solar cells TiO2|Co3O4 library 25](#_Toc483041615)

[Figure 12: Virtual cells based on the with Ag back contacts (A: *JSC* *(μA/cm2*); B: *VOC* *(V);* C: IQE (%)) and With Ag|Cu Back Contacts (D: *JSC* *(μA/cm2*); E: *VOC* *(V);* F: IQE (%)) solar cells libraries. The white regions are outside of the model’s applicability domain. 33](#_Toc483041616)

[Figure 13: Virtual cells based on the library (A: *JSC* *(μA/cm2*); B: *VOC* *(V);* C: IQE (%)). The white regions are outside of the model’s applicability domain. 34](#_Toc483041617)

[Figure 14: Verification Configuration screen: (A) The main screen for the Data Input stage composed from two main sections. Section (1) allows the user to separate the input data into properties and descriptors and further classify the descriptors as relevant / irrelevant for the current research. Section (2) allows the user to configure the type of data verification (seen in B) and then visually verify the data chosen. 38](#_Toc483041618)

[Figure 15: SOM Screen divided into three components. Component (1) allows the user to configure the designated SOM, Component (2) presents the raw data, and Component (3) presents the SOM resulting maps. On the left hand map, neurons were colors based on the true library classification. On the right hand map, the neurons were colored based on the classification made by SOM. 39](#_Toc483041619)

[Figure 16: Test – Train screen which allows the user to configure the manner in which the sample space is divided into test and train groups. In this screen the user can choose various parameters of the simulated annealing process and accept as an output the acceptance ratio and the groups themselves. 40](#_Toc483041620)

[Figure 17: RANSAC Configuration screen 41](#_Toc483041621)

[Figure 18: Virtual Cell Configuration Screen 41](#_Toc483041622)

List of Tables

[Table 1**:** Results of the independent sample t-tests. 19](#_Toc475380985)

[Table 2: Correlation matrix between the five PV activities. 20](#_Toc475380986)

[Table 3: Results of the paired sample *t*-test. 23](#_Toc475380987)

[Table 4: Number of model-compatible samples for the three datasets based on the RANSAC models. 28](#_Toc475380988)

[Table 5: RANSC model performance for the three datasets. 28](#_Toc475380989)

[Table 6: A comparison of model coverage between RANSAC and *k*NN models 29](#_Toc475380990)

[Table 7: RANSAC derived models for different PV properties. 29](#_Toc475380991)

[Table 8: Featured selected for the libraries by the various methods. 31](#_Toc475380992)

# Abstract

This paper presents the development of a pharmaceutical research decision support system for winnowing thousands of candidate molecular compounds. This work presents the problem analysis, the method definition and the development of an innovative clustering algorithm focused on grouping molecular compounds with similar underlying nature. The results are validated using multiple drug analysis datasets and other datasets with similar characteristics.

# 1. Needs Analysis

## 1.1. Operations Analysis

The process of bringing a new drug to market (as described in figure 1) is long and expensive one by all accounts with costs estimated by non-pharma market members in hundreds of millions of US dollars and reported by pharma market members as high as one billion US dollars [‎1].

Figure Pharma Compound Funnel

According to the U.S. food and drug administration (FDA) before a new drug hits the market there are 4 required steps:

1. Discovery and development: discovery of new drugs through new insights into a disease process that allow researchers to design a product to stop or reverse the effects of the disease, broad range of tests of molecular compounds to find possible beneficial effects for multiple diseases or existing treatments that have unanticipated effects or new technologies.

At this stage in the process, many thousands of compounds are potentially candidates for development as a medical treatment. Early testing filters the candidates to a small number of compounds. Once a promising compound is identified experiments are conducted to gather information on how it is absorbed, distributed, metabolized, and excreted; Its potential benefits; the best dosage; the best way to give the drug (such as by mouth or injection); side effects; effects on different groups of people (such as by gender, race, or ethnicity); interaction with other drugs and treatments and its effectiveness as compared with similar drugs.

1. Preclinical Research: Before human trials the compound’s toxicity must be ascertained using two types of preclinical research: in vitro (using controlled environment outside of a living organism) and in vivo (using living organisms such as cells and animals). Preclinical studies are limited in scope but must provide detailed information on dosing and toxicity levels, leading to the decision whether the drug should be tested in people.
2. Clinical Research: The studies, or trials, that are conducted on people. Clinical trials vary greatly on scales of risk and process and follow a typical series from early, small-scale, Phase 1 studies to late-stage, large scale, Phase 3 studies.
3. FDA Review: when a drug has indicated from its early tests and preclinical and clinical research that it is safe and effective for its intended use, the developer can file an application to market it.

According to PhRMA (a U.S. based biopharmaceutical research company’s consortium) the process for researching and developing new medicines keeps growing in difficulty and length. On average, it takes at least ten years for a new medicine to complete the journey from initial discovery to the marketplace, with clinical trials alone taking six to seven years on average. The average cost to research and develop each successful drug is estimated to be $2.6 billion. This number incorporates the cost (incurred by academic and governmental agencies) of failures of the thousands and sometimes millions of compounds that may be screened and assessed early in the R&D process, only a few of which will ultimately receive approval. The overall probability of clinical success (the likelihood that a drug entering clinical testing will eventually be approved) is estimated to be less than 12%.

A worrying trend for pharmaceutical industry is the continued rise in costs of research and development (as seen in figure 2, based on [‎2]). In addition to regulation, competition in the global market and reduced government funding the size of the field of potential compounds.

Figure Average PhRMA Member Company R&D Expenditures, 1995-2015

The optimal compound to proceed with to the pre-clinical testing phase, given it even actually exists for a specific effect, is the proverbial needle in a haystack and finding it incurs a significant cost. It is so difficult (and therefore unviable economically) that it is manly funded by academic institutions (universities and public research foundations), governments and philanthropic organizations [‎4].

According to the Tufts Center for the Study of Drug Development [‎2] roughly 30% of the total cost of an approved new compound is attributed to the pre-human phase of the research which includes the drug discovery and the pre-clinical phase. That is a huge potential for savings which can be materialized using the tools and processes described in this paper.

A Nature’s Review article [5] presents a model which defines the distinct phases of drug discovery and development starting at the initial stage of “target-to-hit” to the final stage of releasing the drug to the market. The model describes (among other identifiers) the probability of a successful transition from one stage to the next and the phase cost for each project. The model estimates the total cost to achieve one drug launch per year at $1,778 million. per NME launch. It is important to note that this model does not include investments for exploratory discovery research (which as mentioned before is rarely performed in the industry), post-launch expenses or overheads (that is, salaries for employees not engaged in R&D activities but necessary to support the organization).

Figure 3 Candidates Required for a Single Release

## 1.2. Operational Requirements

The system, however complex in its implementation, has a very simple operational requirement to fulfill which is:

***The system shall analyze large corpus of compound data and generate recommendations with regard to the applicability of analyzed compounds as candidates for preclinical trials.***

## 1.3. Concept of Operations

The system operates within the context of a pharmaceutical company with a pre-existing corpus of analyzed compounds and a pool of potential compounds from which to select candidates for preclinical trials.

The system is to be used as a non-real-time support system. This means it will perform its function in parallel to the function of an active research and development team and augment its capabilities with decision support capabilities.

The system is to be used by specifically trained personnel so it has no strict user experience limitations on the learning curve it incurs.

The system, being software based, is agnostic to the hardware it runs on and can be executed in various environments such as cloud environment or on-premise environment.

system es

## 1.4. Functional Analysis

The designed decision support system is currently demonstrated using a combination of MATLAB (preforming pre-processing and the enveloping optimization process) and C++ code performing the dimensionality reduction process.

Input Database

Selection of Features

Run t-SNE

Optimization Engine (GA)

Calculate Fitness Function

Select Best Model

Figure t-SNE Optimization Algorithm

As depicted in figure 2 the system is comprised of 5 functional parts:

1. The input database which supplies the raw data to the system.
2. The feature selection mechanism which (at the beginning of the process) generates a rich enough population of feature selection combination and in subsequent steps uses Genetic Algorithm to create the next generation.
3. The dimensionality reduction process which reduces the multi-dimensional feature space into a two-dimensional space. The process is one which attributes more weight to close distances than to long distances and therefore prioritizes keeping neighbors together over preserving ratios of long distances.
4. A fitness function calculation using the tagged data as a measure of success. Even though a few metrics are being used for measurement the fitness function relies on the fraction of known data points with their closest known data point sharing the same tag.
5. Model selection and generation using genetic algorithms to create ever improving generations until a quality condition is satisfied or the maximal number of iterations is reached.

The system’s goal is to cluster all potential compounds with the selected features in the way which will allow us to assume in the highest probability that the neighbor of a compound with desirable attributes is highly likely to share those attributes. As the critical factors for identifying an effective compound are numerous and alternate as the chemical and biological environments shifts the algorithm identifies the critical factors in each dataset separately.

The underlying assumption is that inside the data there are features which are pertinent to the classification of the compounds into classes of desired effect capability and the algorithm attempts to isolate those features into a model which allows for clustering of compounds according to such classes. Such assumption relies on the fact that in different situations different features carry the most relevant data.

The purpose behind the tool is to identify potentially effective compounds using previously known information about a small number of compounds. For example, if we can tag a few of the compounds we can – by association – estimate with a high probability of success whether the compounds near them in the resulting map (such as in figure 4) have high effectiveness potential.

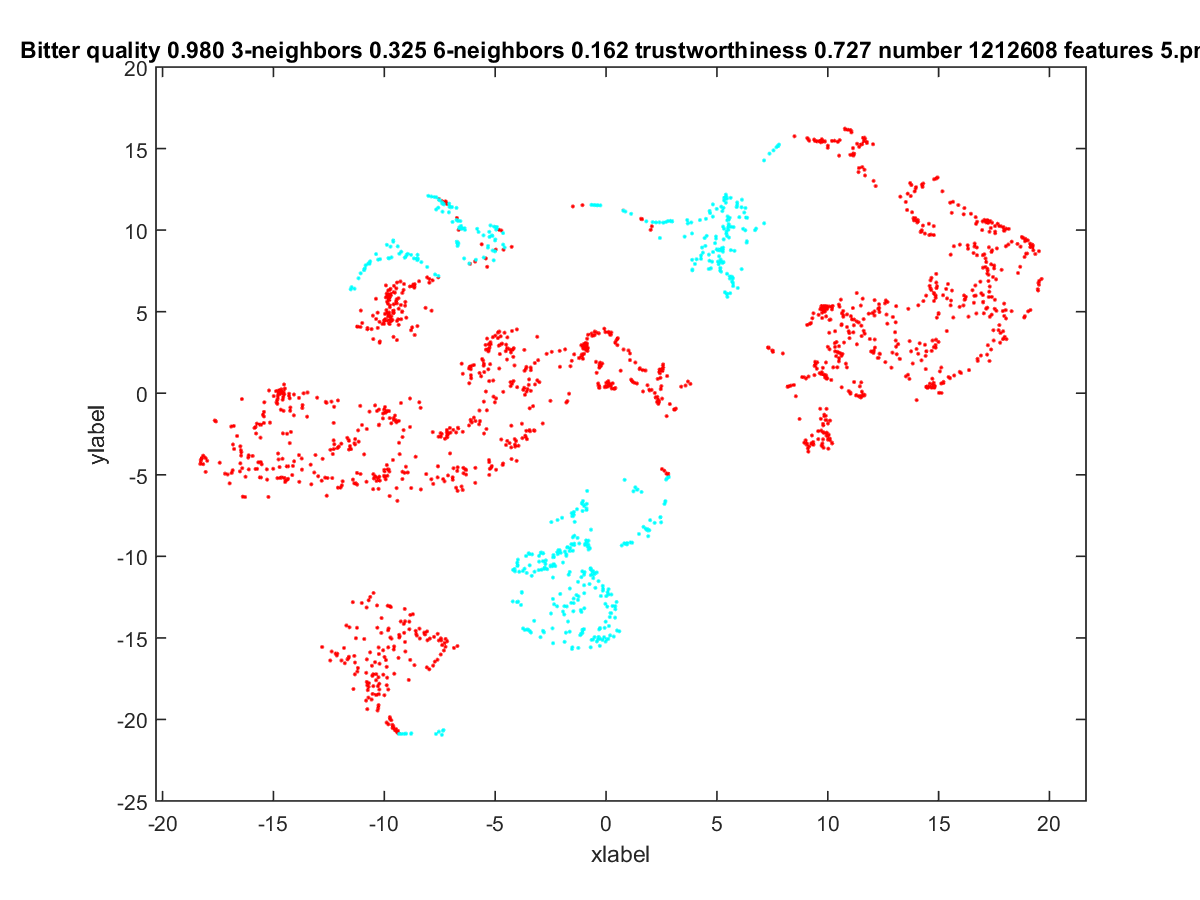


Figure Compounds Grouped by Bitterness

## 1.6 Capabilities and Operational Effectiveness

Figure 6 describes the top level operational needs:

Figure 6 Capabilities

The system needs to be able to measure the distance effectively and for that purpose needs to intelligently reduce the dimensions of the data space (as opposed to simply assuming equal weight on each dimension) and measure the quality of the dimensionality reduction by using clustering algorithm on a-priori tagged data.

The system needs to select the tested features of the data out of a large field of diverse features. For that purpose, the feature selection mechanism needs to be adaptive and to be able to progressively improve while allowing for and fixing mistakes.

# 2. Concept Exploration

## 2.1. Operations Analysis

### Dimensionality Reduction

High-dimensional data often lies on or near a much lower dimensional space. A good way to represent data points is by their low-dimensional coordinates. The low-dimensional representation of the data should capture information about high dimensional pairwise distances.

The system assumes that using a non-linear dimensionality reduction algorithm can provide a powerful classification tool.

### Quality Factor

Utilizing a dimensionality reduction derived clustering algorithm for the sake of classification requires a quality factor which improves with the quality of the resulting classification algorithm. The quality factor has to be simple enough to be used for all resulting data sets and positively defined.

### Adaptive Filtering

Since the system relies on the constant improvement of the feature set for the purpose of classification the adaptive filtering part calls for an evolutionary algorithm which will guide the progress of resulting clusters to constantly improving quality scores.

## 2.2. Concept Exploration

For the sake of this research paper where only a single concept could be explored and implemented no comparative concept exploration was performed – the only concept with realistic implementation chance was a MATLAB driven, PC platformed application.

# 3. Concept Definition

## 2.1. Operations Analysis

### Dimensionality Reduction

High-dimensional data often lies on or

## 2.2. Input Data

The first stage consists of data input. The system accepts digital comma separated value files created using publicly available financial indicators and, for some, classifications into their bankruptcy status. Then, the user separates the data into two groups, namely, Descriptors and Activities.

## 2.3. Feature Selection

Under the assumption that some of the features carry more information than others for the snapshot of the world the data represents the goal of the algorithm is to select the optimal feature set. As the basis for a genetic algorithm the initial population is 20 sets of randomly selected features. From each generation, the next generation is created as follows:

1. The fittest 30% are carried as they are to the next generation
2. Mutations of the fittest 20% are generated into the next generation
3. Crosses of the fittest 50% are generated into the next generation.

After the process stops improving (given at least 3 iterations without improvement) it’s assumed that the optimal feature set has been found.

## 2.4. Run t-SNE

t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction algorithm which has been chosen because as opposed to more commonly used dimensionality reduction algorithms (such as PCA) which are mainly concerned with preserving large pairwise distances t-SNE maintains structure by putting more weight on local distances.

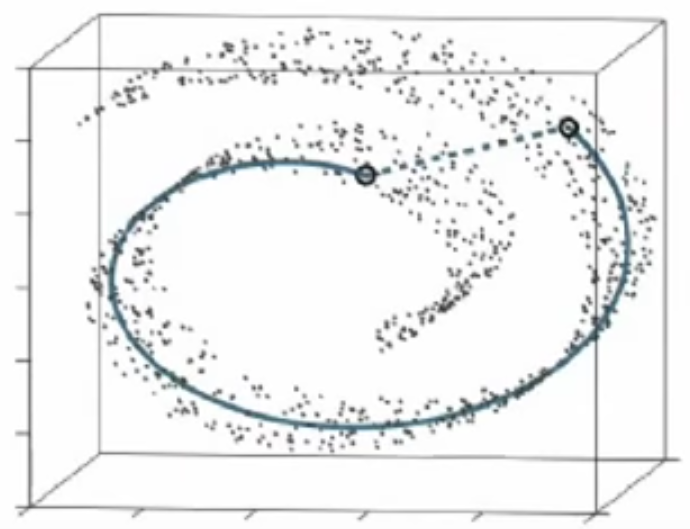


Figure Local as Opposed to Global Distances

In the domain of bankruptcy detection, it is critical that the process of preserving large distances (which carry little data) will not affect preserving local (short) distances since the shortest distance estimation along the selected features is what determines the nearest neighbor and for our purpose of classifying untagged samples this is the most important quality.

## 2.5. Fitness Calculation

The fitness of a feature set (representing the quality of the clustering algorithm) has been measured using multiple metrics including the nearest neighbor being of the same class, two out of three nearest neighbors and three out of five nearest neighbors. All metrics behave in a similar manner so finally the fitness score is determined by the percentage of tagged samples having a nearest tagged neighbor with the same tag. This quality metric indicates that the untagged neighbors are likely to share the same tag (thus indicating that the clustering succeeded to cluster data points with the same tag together).

## 2.6. Model Selection

The model with the best fitness score is selected for a dataset and the mapped data points (in 2 dimensions) are used for identifying corporations running the risk of bankruptcy – the untagged corporations adjacent to the tagged as bankrupt corporations constitute our target set.

# 3. Results and Discussion

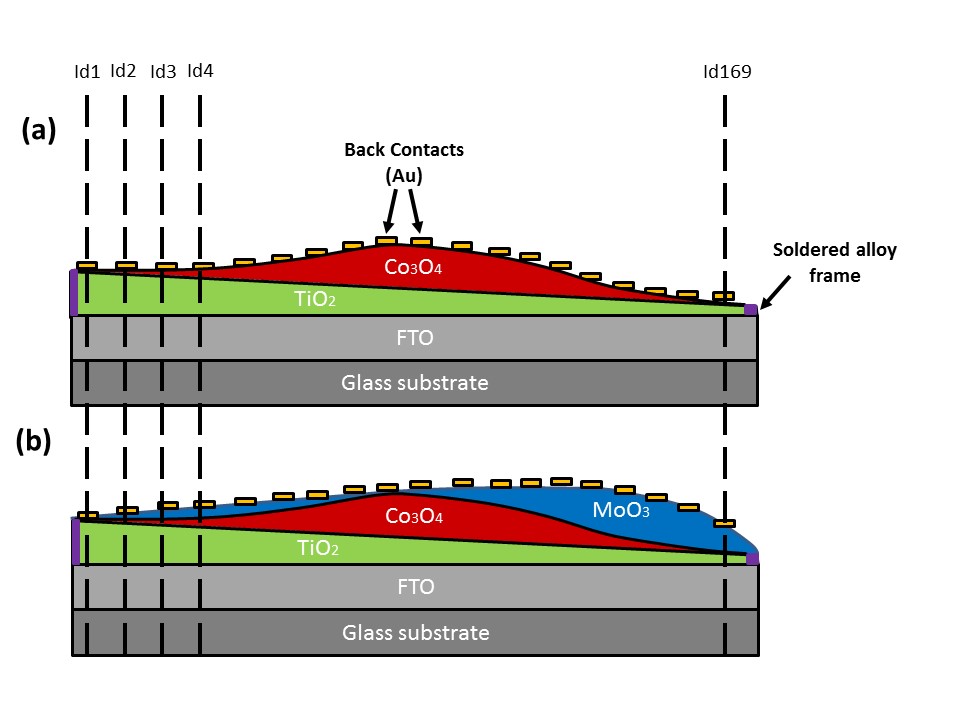
Below we describe the application of the above-describe workflow to “real world” problems.

## 3.1. Pre-Processing Implementation

The main objective of this stage is to demonstrate the usefulness of data reduction and visualization techniques for the analysis of solar cell libraries. For the purpose of validating the proposed implementation, two Photovoltaic libraries were chosen from the work of Majhi et al.*(*[*5*](#_ENREF_5)*)* The stages described in the Pre-Processing section are allow to comparatively analyze the libraries using simple statistics and data visualization methods (PCA and SOM).

### 3.1.1. Data bases

The implementation of this phase was tested using a TiO2|Co3O4 and a TiO2|Co3O4|MoO3 solar cells libraries The two libraries share the same preparation method and the same thickness of the TiO2 and Co3O4 layers. An additional layer of MoO3 was added to the second library. Both libraries were generated on precut glass coated with Fluorine Doped tin Oxide (FTO) substrates onto which a TiO2 window layer with a linear gradient was deposited, followed by Pulsed Laser Deposition (PLD)*(*[*6*](#_ENREF_6)*)* of a Co3O4 layer with the deposition profile presented in Figure 5a and 5b. For the second library (Figure 5b), an additional layer of MoO3 was PLD-deposited on top of the Co3O4 layer. A grid of 13 x 13 = 169 Au back contacts was used for both libraries. In the following analysis, each of the resulting 169 solar cells is identified by its location on the back contact grid and cells with the same TiO2|Co3O4 composition have identical IDs (Figure 5).

**Figure 8:** A schematic representation of the PV solar cells library. (a) TiO2|Co3O4 library (b) TiO2|Co3O4|MoO3 library.

### 3.1.2. Library Characterization

Each of the individual solar cells is characterized by its material descriptors (see below) and experimentally measured photovoltaic activities. The descriptors for these libraries are thickness of the window layer (), the thickness of the absorber layers (), the thickness ratio between the absorber layer and the total (absorber + window) layers (*ratio*), and the band gap of absorber layer (*BGP*) - The band gap is the energy difference (in electron volts) between the top of the valence band and the bottom of the conduction band. In this research we focus on three experimental measured PV activities: (1) The short circuit photocurrent density (*Jsc*). The short circuit photocurrent density is the current density through the solar cell when the voltage difference across the cell is zero. (2) The open circuit voltage (*Voc*). The open circuit voltage is the maximum voltage available from a solar cell - occurs at an open circuit. (3) The internal quantum efficiency (*IQE*), which reflects the charge separation and collection efficiencies of a device and is calculated by equation (1) where is the maximum theoretical calculated photocurrent.

(1)

### 3.1.3. Basic statistical analysis

Majhi et al.*(*[*5*](#_ENREF_5)*)* introduced the MoO3 layer as a hole selective material to allow for better extraction of holes under the assumption that this would lead to improved PV parameters. In accord with this hypothesis, an increase in all PV parameters for the TiO2|Co3O4|MoO3 library was indeed noted. To quantify this observation, we have utilized the independent sample *t*-test. The results are shown in Table 1 and suggest that: (i) means for all PV parameters obtained for the TiO2|Co3O4|MoO3 library are higher than those for the TiO2|Co3O4 library and (ii) these differences are statistically significant.

**Table 2:** Results of the independent sample t-test comparing the activities of the TiO2|Co3O4|MoO3 and TiO2|Co3O4 libraries.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TiO2|Co3O4** | | **TiO2|Co3O4|MoO3** | |  |  |  |
| **Activity** | **Mean** | **SD** | **Mean** | **SD** | **t-test** | **df** | **p-value** |
| *Voc* (mV) | 173 | 149 | 446 | 146 | -14.79 | 258 | <0.001 |
| *Jsc* (µA/cm2) | 8.76 | 5.48 | 17.41 | 3.82 | -14.97 | 258 | <0.001 |
| *IQE* (%) | 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 |
| *Pmax* ( µW/cm2) | 0.17 | 0.22 | 0.6 | 0.23 | -14.84 | 258 | <0.001 |

### 3.1.4. Principal Component Analysis

To compare between the libraries, we characterized each cell by using the five PV activities (*JSC, VOC, IQE, FF and Pmax*) that form the feature (i.e. activity) space. In this space each dimension corresponds to one activity, cells are represented by points and the distance between any two points represents the degree of similarity between the corresponding solar cells. Next, photovoltaic cells from the two libraries were subjected, together, to PCA. In this analysis, the original 5D activity space was reduced into a 2D space with the two principal components (PCs) covering 72.0% and 18.8% of the original variance, respectively, for a total of 90.8%. The ability to capture >90% of the variance using only two principal components is suggestive of high correlation between the original PV parameters. Table 2 presents the correlation between the five PV activities, where three of the ten values are indeed higher than 0.8 (*IQE- JSC*; *Pmax- JSC*; and *Pmax- VOC*). This is perhaps not surprising since some of the PV parameters are derived from one another (although not in a linear manner). For example, *IQE* is derived from *JSC* resulting in a high correlation (r = 0.83) between the two parameters. The resulting PC plot for the two libraries is presented in Figure 6.

Table : Correlation matrix between the five PV activities across both libraries, for a total of 260 cells (110 from TiO2|Co3O4 library and 150 cells from the TiO2|Co3O4|MoO3 library).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Activity** | *Voc* (mV) | *Jsc* (µA/cm2) | *IQE* (%) | *FF* (%) | *Pmax* ( µW/cm2) |
| *Voc* (mV) | 1 |  |  |  |  |
| *Jsc* (µA/cm2) | 0.73 | 1 |  |  |  |
| *IQE* (%) | 0.35 | 0.83 | 1 |  |  |
| *FF* (%) | 0.55 | 0.35 | 0.08 | 1 |  |
| *Pmax* ( µW/cm2) | 0.96 | 0.82 | 0.51 | 0.52 | 1 |

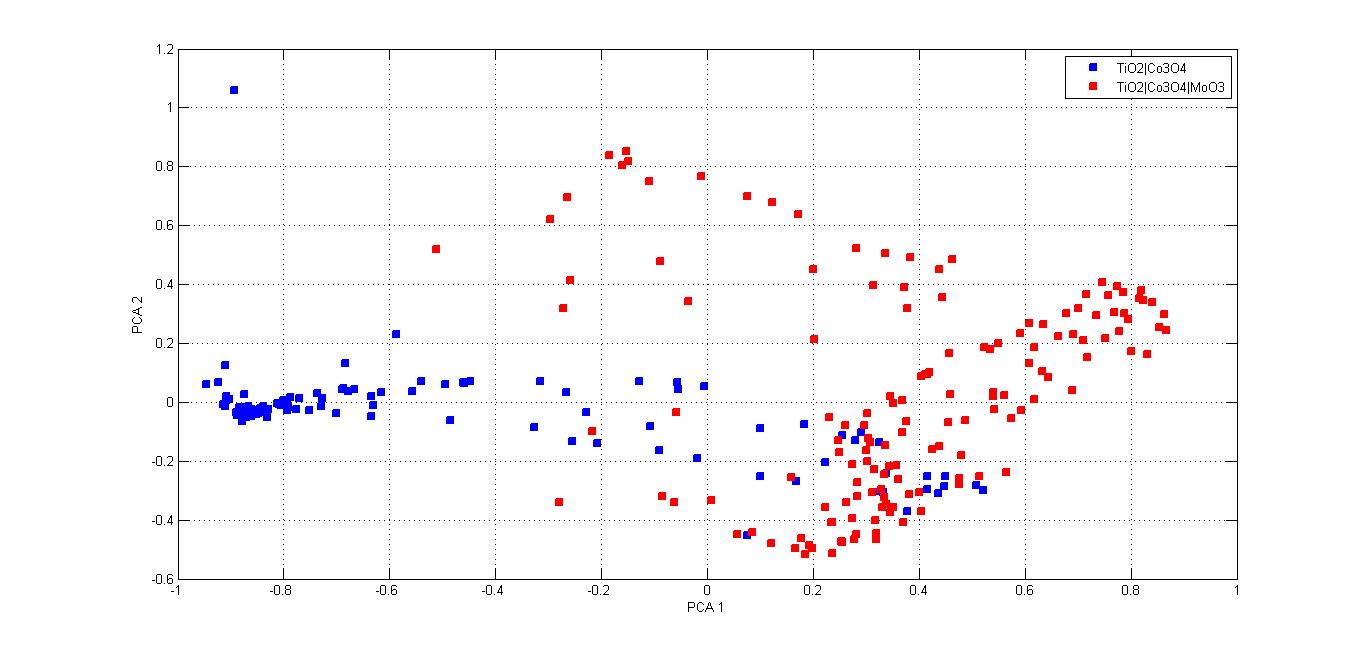


Figure 9: A PC plot for photovoltaic cells from the two solar cell libraries. Blue and red dots represent cells from the TiO2|Co3O4 and TiO2|Co3O4|MoO3 libraries, respectively.

Figure 6 clearly demonstrates that cells from the TiO2|Co3O4|MoO3 library are more spread out in the PC space than cells from the TiO2|Co3O4 library, in particular along PC2. This suggests a non-uniform effect of the MoO3 layer on the overall PV properties of cells characterized by different TiO2|Co3O4 profiles. Thus while certain TiO2|Co3O4 combinations benefit from the addition of MoO3, others are either indifferent or even suffer from reduced performances due to this change. These observations complement the one made by Majhi et al.,*(*[*5*](#_ENREF_5)*)* that the addition of the MoO3 layer increases PV activity by providing a more detailed analysis. Furthermore additional information is obtained regarding the region that is largely unaffected by the additional layer.

Interestingly, a region of overlap between the two libraries is clearly observable in the PC plot. Since the addition of MoO3 may change the PV parameters of any cell, it was necessary to test whether neighboring cells (i.e., cells with similar activity profiles) from the two libraries have identical IDs (i.e., they differ only by the addition of the MoO3 layer).

This test was performed by measuring the Euclidean distance between cells with identical IDs in both libraries in the PC space. The 13 pairs with the smallest distances (smaller than 10% of the maximum distance between two identical IDs) are presented as green dots in Figure 7 and as expected, reside within the overlap region.

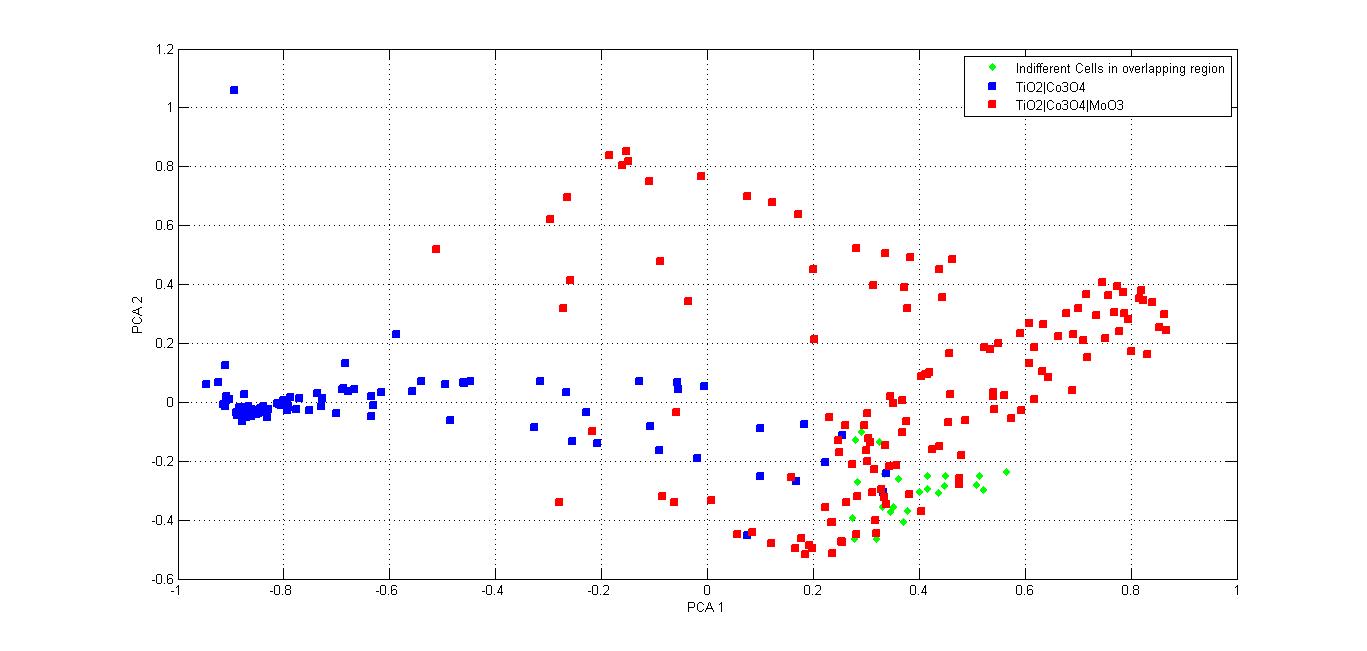


Figure 10: A PC plot of photovoltaic cells from the two solar cell libraries. Blue and red dots represent cells from the TiO2|Co3O4 and TiO2|Co3O4|MoO3 libraries, respectively. Green dots indicate 13 cell pairs from the two libraries with identical IDs and distances from one another in the lowest 10th percentile.

### 3.1.5. Heat Map

Based on the above, for certain cells, the addition of MoO3 causes only insignificant changes in the overall PV profile (however see below for an analysis of specific parameters). To further characterize these cells, we have mapped them onto the thickness profile of the Co3O4 layer (see Figure 8 ). All cells which were largely unaffected by the addition of the MoO3 layer are found to reside in library regions with thickness of the Co3O4 layer above 227 nm which corresponds to the top 6th percentile of overall layer's thickness. These cells also have a relatively thin MoO3 layer (ranging from 38 nm to 45 nm corresponding to the lower 30th percentile of the overall thickness). In addition, Figure 8 demonstrates that all these cells are physically close together and are thus constructed in a similar fashion and have similar material descriptors. These results are in accord with those presented by Majhi et al.*(*[*5*](#_ENREF_5)*)* (Figure 4 in reference *(*[*5*](#_ENREF_5)*)*) which show that above a Co3O4 thickness of 150-250 nm, both libraries have similar profiles for specific PV parameters. However, the analysis in reference [3] was performed at the level of the entire library and consequently provided only limited information on the effect of the MoO3 layer on the complete PV profile of individual cells. In contrast, the present analysis allows for the evaluation of the effect of MoO3 on the PV profile of cells with specific and identical TiO2|Co3O4 combinations.

zFigure 11: Heat map of the PV libraries, where each square represents a cell in the library. The heat map is colored according to the thickness of the Co3O4 layer. 'X' represents cells with similar PV behavior in both libraries according to the PCA. The X and Y axes represent the location of the cell on the 13 x 13 grid.

To further quantity the effect of the MoO3 layer on these cells we have used the paired sample *t*-test. The results are shown in Table 3 and indicate that in contrast with most of the cells, for two parameters, *JSC*, and *IQE*, the addition of MoO3 led to a statistically significant decrease in performances. Two other parameters, *FF* and *Pmax* demonstrated insignificant differences between the libraries. *VOC* was the only parameter that maintains a significantly higher value for the library containing MoO3. In reference (3) the overall decrease in *JSC* and *VOC* for the TiO2|Co3O4|MoO3 library at large Co3O4 thicknesses was attributed to a poor extraction of electrons at the TiO2|Co3O4 interface. While this decrease was observed at the level of the complete libraries, here we demonstrate that for a subset of cells characterized by a high Co3O4 thickness, the addition of MoO3 led to a statistically significant increase in *VOC*.

Table : Results of the paired sample *t*-test comparing the activities of the 13 “unaffected” cells from the TiO2|Co3O4|MoO3 and TiO2|Co3O4 libraries.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TiO2|Co3O4** | | **TiO2|Co3O4|MoO3** | |  | | |  |
| **Activity** | **Mean** | **SD** | **Mean** | **SD** | **t-test** | **df** | **p-value** | |
| *Voc* (mV) | 431 | 18.9 | 484 | 36.8 | -7.08 | 12 | <0.001 | |
| *Jsc* (µA/cm2) | 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 | |
| *Pmax* ( µW/cm2) | 0.63 | 0.05 | 0.62 | 0.06 | 0.98 | 12 | >0.05 | |

### 3.1.6. Self-Organizing Map (SOM)

In addition to the PCA, a self-organizing map (SOM) was constructed for the two libraries. As discussed above SOM uses a neural-network to cluster objects based on their similarity. In this work, the objects are the solar cells and similarity is defined by the five PV parameters. The resulting map is presented in Figure 9 with nodes (i.e. neurons) colored according to library membership.

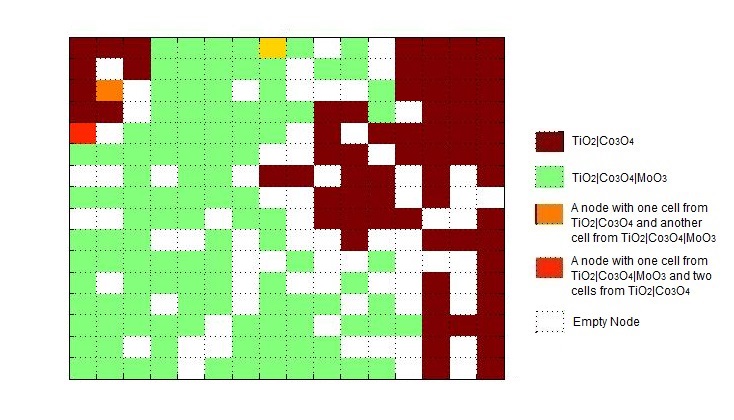


Figure 12: A SOM composed of 16 x 16 = 256 neurons (i.e nodes) of the two solar cell libraries, where each colored square represents a node color coded according to its library membership (see legend). White squares represent empty neurons.

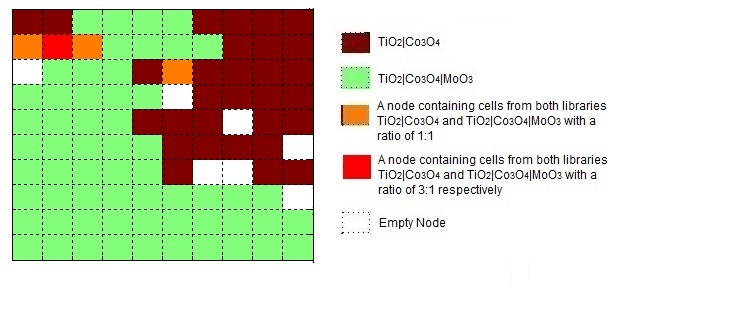


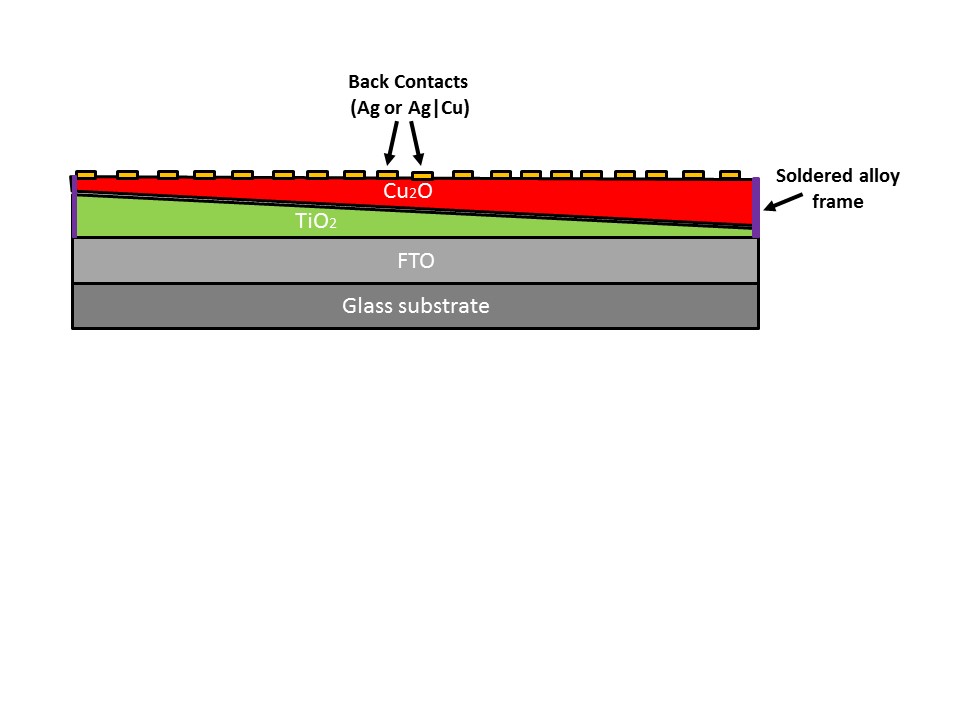
Figure 13: A SOM composed of 10 x 10 = 100 neurons (i.e nodes) of the two solar cell libraries, where each colored square represents a node color coded according to its library membership (see legend). White squares represent empty neurons.

Overall, a clear separation is observed between the two libraries with only minor overlap as seen previously by using the PCA subcomponent (thus providing addition validation towards the reliability of both implementations). In further agreement with the PCA results, members of the TiO2|Co3O4|MoO3 library occupy more nodes than members of the TiO2|Co3O4 library thus suggesting the relatively to TiO2|Co3O4, TiO2|Co3O4|MoO3 cells vary from one another in such a way that they can not be clustered into the same node, while many TiO2|Co3O4 cells are clustered in the same node in groups more than two in a node as Figure 9 and Figure 10 suggests (Figure 10 being a low resolution image of Figure 9).

## 3.2. RANSAC Implementation

### 3.2.1. Databases

In a similar manner to the Pre-Processing stage, the RANSAC module was validated via its application on the previously presented library as well as on additional library obtained from Pavanet al.*(*[*22*](#_ENREF_22)*)* This library was generated on precut glass coated with Fluorine Doped tin Oxide (FTO) substrates onto which a TiO2 window layer with a linear gradient was deposited, followed by an absorber layer of Cu2O (see Figure 11). Two different back-contacts were used, namely, silver only (Ag) and silver and copper (Ag|Cu) deposited one after the other, leading to two sub-libraries (datasets) each consisting of grid of 13 x 13 = 169 cells. In this work we omitted the non-photovoltaic cells leaving a total of 166 and 162 cells for the Ag and Ag|Cu back contact data base respectively. Characterization of the libraries was performed in the same manner as described in the Pre-Processing implementation.



**Figure 14:** A schematic representation of the PV solar cells TiO2|Co3O4 library

The RANSC algorithm was applied to the three datasets described above. For each dataset three models were derived to describe their Photovoltaic (PV) properties (*JSC, VOC* and *IQE*). Table 4 presents the number of training set and test set samples found to reside within a distance of one standard deviation of the train subsample from the derived model (i.e., model-compatible samples). Model-incompatible samples in the training and test sets are referred to as outliers and outside of the model’s AD, respectively. As can clearly be seen, the vast majority (≥ 85%) of the samples are included within the strip for both the training and test sets. This suggests that (i) Predictive models could likely be derived for this dataset and (ii) The model described by the strip forming curve approximates most of the training set and test set samples to within one standard deviation (the pre-defined strip width; see Methods section) from their experimental values. One could therefore propose that the majority of future samples will be similarly predicted. However, in two cases the number of acceptable cells was below the 85% threshold (the *VOC* models for the library with 83% and 73% of acceptable cells for the training and test sets, respectively), indicating higher variance for this property in this dataset in comparison with the other datasets. This may potentially lead to poor model performance (see below). Indeed the performances of the *VOC* model from the library are exceptionally poor in accord with the insufficient number of model-compatible samples (Table 4). This model was therefore excluded from the analysis reported below.

The RANSAC algorithm led to models with good statistical parameters (Table 5) for training set samples for *JSC* ( between 0.74-0.77), *Voc* ( between 0.41-0.57 excluding the library; see above) and IQE ( between 0.72-0.85). Upon the removal of outliers, the statistical parameters for all models improved with the largest improvement being obtained for *VOC* ( between 0.78-0.82, 0.73-0.74 (excluding the ) and 0.78-0.85 for *JSC*, *VOC* and IQE , respectively.

The performances of the RANSAC models on the test set samples followed a trend similar to that observed for the training set. Thus, for all test sets, was between 0.69-0.82, 0.62-0.80, and 0.68-0.79 for *JSC*, *VOC*(excluding the and IQE, respectively. As expected for datasets devoid of significant activity cliffs, when considering only samples within the models’ applicability domains, these numbers improved to 0.82-0.87 and 0.79-0.83 for *JSC*, and IQE, respectively. For *VOC* of the (Ag) library, no test set samples were filtered by the applicability domain leading to no change in model performances ( = 0.80). However for this property a significant increase in the (Ag|Cu) library upon the removal of only two samples was observed ( = 0.62 and 0.73 without and with the model’s AD, respectively).

Two of the above described datasets ( (Ag) and (Ag|Cu)) were previously modeled by Yosipof et al. using *k*NN and a Genetic Algorithm (GA) approach*(*[*23*](#_ENREF_23)*)* thereby allowing for a direct comparison between the performances of the resulting models. GA produced models with values between 0.74-0.76, 0.50-0.78 and 0.72 for *JSC*, *VOC* and IQE respectively. The corresponding numbers obtained by RANSAC are = 0.69-0.76, 0.62-0.80, and 0.68-0.78 for *JSC*, *VOC* and IQE, respectively, with no AD and = 0.84-0.87, 0.73-0.80 and 0.82-0.83 for *JSC*, *VOC* and IQE, respectively, with AD. These results suggest that the performances of the RANSAC models are similar to those of the GA with no consideration of the AD and provide significant improvement upon the application of AD. Of note, there is no inherent definition of AD in the GA method. For *k*NN was reported to be 0.89-0.92, 0.56-0.89, and 0.87-0.91 for *JSC*, *VOC* and IQE, respectively, with no AD and 0.88-0.92, 0.55-0.89 and 0.87-0.89 for *JSC*, *VOC* and IQE, respectively, with AD. Thus, *k*NN provides models with higher prediction statistics than RANSAC in particular when the AD is not considered. However, the performances of RANSAC approach those of *k*NN upon the introduction of the AD. Moreover, the test set coverage provided by RANSAC is generally higher than that provided by *k*NN (Table 6). Finally, in contrast with *k*NN, RANSAC provides a model in the form of a QSAR equation which enhances model interpretability.

Table : Number of model-compatible samples for the three datasets based on the RANSAC models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *JSC* | *VOC* | *IQE* |
| (Ag) | | | |
| # Model - Compatible Training Samples | 129/130 (99%) | 123/130 (95%) | 129/130 (99%) |
| # Model - Compatible Test Samples | 28/32 (88%) | 32/32 (100%) | 28/32 (88%) |
| (Ag|Cu) | | | |
| # Model - Compatible Training Samples | 131/134 (98%) | 127/134 (95%) | 129/134 (96%) |
| # Model - Compatible Test Samples | 30/32 (94%) | 30/32 (94%) | 31/32 (97%) |
|  | | | |
| # Model - Compatible Training Samples | 118/120 (98%) | 100/120 (83%) | 120/120 (100%) |
| # Model - Compatible Test Samples | 30/30 (100%) | 22/30 (73%) | 30/30 (100%) |

Table : RANSC model performance for the three datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Library | Activity |  | (no outliers) |  | (AD) |
| (Ag) | *JSC* | 0.76 | 0.82 | 0.69 | 0.87 |
| *VOC* | 0.41 | 0.74 | 0.80 | 0.80 |
| *IQE* | 0.72 | 0.79 | 0.68 | 0.83 |
| (Ag|Cu) | *JSC* | 0.74 | 0.78 | 0.76 | 0.84 |
| *VOC* | 0.57 | 0.73 | 0.62 | 0.73 |
| *IQE* | 0.72 | 0.78 | 0.78 | 0.82 |
|  | *JSC* | 0.77 | 0.78 | 0.82 | 0.82 |
| *VOC* | 0.14 | 0.59 | 0.41 | 0.26\* |
| *IQE* | 0.85 | 0.85 | 0.79 | 0.79 |

\*The significant decrease in model performance for *VOC* in the library upon the application of the AD ( = 0.41 and 0.26 without and with the model’s AD, respectively) is indicative of the likely presence of an activity cliffs for this property.

Table : A comparison of model coverage between RANSAC and *k*NN models

|  |  |  |  |
| --- | --- | --- | --- |
| Library | Activity | RANSAC coverage | *k*NN coverage\* |
| (Ag) | *JSC* | 88% | 91% |
| *VOC* | 100% | 84% |
| *IQE* | 88% | 91% |
| (Ag|Cu) | *JSC* | 94% | 79% |
| *VOC* | 94% | 82% |
| *IQE* | 97% | 73% |

\*The data for *k*NN were taken from Table 5 in reference*(*[*23*](#_ENREF_23)*)*.

*3.2 RANSAC as a feature selection tool*

Table 7 presents the model equations produced by RANSAC for the different PV properties of the three datasets.

Table : RANSAC derived models for different PV properties.

|  |  |
| --- | --- |
| PV Property | Model |
| (Ag) | *JSC* |
| *VOC* |
|  |
| (Ag|Cu) | *JSC* |
| *VOC* |
|  |
|  | *JSC* |
| *VOC* |
|  |
|  |  |

For both datasets it is evident that while four descriptors were evaluated by RANSAC, only two were picked by the algorithm as correlated with the photovoltaic activity. The dataset was characterized by three descriptors and all were selected by RANSAC.

The features selected by the RANSAC algorithm could be compared with those selected by the *k*NN and GA models reported by Yosipof et al.*(*[*23*](#_ENREF_23)*)* As can be deduced from Table 8, all methods selected the same descriptors for the (Ag) library while *k*NN replaced by the *ratio* descriptor for the (Ag|Cu) library. While GA sometimes selected a smaller number of “base descriptors”, it compensated for this smaller number by incorporating these descriptors in more complex mathematical equations. In contrast, the RANSAC algorithm is limited to simple polynomial equation (to the 5th power in this study).

Table : Featured selected for the libraries by the various methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Library | Activity | RANSAC | Genetic | *k*NN |
| (Ag) | *JSC* |  |  |  |
| *VOC* |  |  |  |
| IQE |  |  |  |
| (Ag|Cu) | *JSC* |  |  | , *Ratio* |
| *VOC* |  |  | , *Ratio* |
| IQE |  |  | , *Ratio* |

## 3.3. kNN Implementation

The technical challenge in the implementation of *k*NN was the unification of algorithms written in two different coding environments. The main infrastructure of the DSS is based on Matlab’s 2014b environment while the *k*NN algorithm including Outlier Removal and *k*NN modeling was written in a C++ environment. The process of integrating these two subcomponents involved transforming the C++ code into an executable that the Matlab code can call via the Windows operating system. While not ideal for either algorithms, this solution provides the middle ground that allows the full integration. The implementation was verified by processing the same data as presented by Yosipof et al.*(*[*24*](#_ENREF_24)*)* and obtaining the same features as in the original paper.

## 3.4. Virtual Cell Implementation

RANSAC derived models could be used to predict PV properties of virtual solar cell libraries. These predictions could serve two purposes: (i) Identify trends related to the dependence of PV properties on descriptors values, trends that are not easily discernible from the resulting equations. (ii) Provide a theoretical basis for and guide future experiments.

### 3.4.1. (Ag) and (Ag|Cu) virtual libraries

The original  *(Ag)* and *(Ag|Cu)* libraries were of identical compositions with between 70 nm and 311.5 nm and between 249 nm and 596 nm. The virtual cell should cover these ranges and expand upon them to allow RANSAC-based extrapolations. With this in mind, thickness values for the different layers were selected to be between 200 nm and 700 nm and between 40 nm and 400 nm for the Cu2O and TiO2 layers, respectively, where each range was divided into 100 bins (a total of 10,000 cells per virtual library). These specific ranges were selected following several iterations designed to find the model’s limits, beyond which the results would not be physically meaningful (i.e., have negative PV values). Next, the PV properties (*JSC*, *VOC*, IQE) of each cell were predicted using the RANSAC models presented in Table 7. The results of these predictions are presented in Figure 12 and demonstrate a few trends: (i) All PV activities primarily depend on the thickness of the Cu2O layer rather than on the thickness of the TiO2 layer. This trend was noted by Pavan et al.*(*[*3*](#_ENREF_3)*)* but only for *JSC*. (ii) *JSC* presents a marked increase for Cu2O thicknesses above 500 nm (where *JSC* equals ) as seen in Figure 12A and D. Similar trends (yet with less sharp transitions) are also seen for *IQE* and *VOC* (Figure 12B and E and Figure 12C and F, respectively). Interestingly, Cu2O thicknesses above 500 nm where hardly explored by the original library. (iii) The nature of the back contact (Ag vs. Ag|Cu) has the largest effect on the dependence of *JSC* on the thickness of the Cu2O layer (compare Figure 12A and D) which is followed by *VOC*(compare Figure 12B and E). In contrast, the dependence of IQE on the thickness of the Cu2O layer is the least affected by the back contact (compare Figure 12C and F). (iv) Certain combinations of and are predicted to have both high *JSC* and *VOC* values. These trends are largely in accord with previous conclusions on these systems deduced from experiments and other data mining approaches.*(*[*23*](#_ENREF_23)*)*

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Figure 15: Virtual cells based on the with Ag back contacts (A: *JSC* *(μA/cm2*); B: *VOC* *(V);* C: IQE (%)) and With Ag|Cu Back Contacts (D: *JSC* *(μA/cm2*); E: *VOC* *(V);* F: IQE (%)) solar cells libraries. The white regions are outside of the model’s applicability domain.

### 3.4.2. virtual library

In a similar manner, another virtual library was constructed for the MOs composition. In the original library, the thicknesses of the different layers ranged from 289.9 nm to 354.8, nm, from 30.7 nm to 245 nm and from 38.9 nm to 61.7 for TiO2, Co3O4 and MoO3,respectively. In the virtual library, these ranges were increased to 30 nm to 500 nm and to 40 nm to 100 nm for the Co3O4 layer and MoO3 layer, respectively (50 bins for each range) while the TiO2 layer was kept at a constant value of 340 nm. This led to a virtual library consisting of 2,500 cells.

For this particular library, Koushik et al.*(*[*5*](#_ENREF_5)*)* shows that IQE is mainly affected by the thickness of both the and layers. This conclusion was further supported by a computational analysis.*(*[*25*](#_ENREF_25)*)* Figure 13 shows that RANSAC’s prediction is in line with this proposition (i.e., to achieve relatively high IQE values, the thickness of the layer must be low – smaller than 150 nm and this property is also influenced by the thickness of the layer). In addition, RANSAC’s models point to an inherent problem in producing solar cells with both high *JSC* and IQE values for this MOs combination since the former seems to yield maximum value at layer thickness at the 500 nm region while the latter, yields its global maxima at the 30 nm region. Finally, Figure 13 suggests possible combinations for additional experiments that may lead to high IQE values, for example small thicknesses of both and layers.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 16: Virtual cells based on the library (A: *JSC* *(μA/cm2*); B: *VOC* *(V);* C: IQE (%)). The white regions are outside of the model’s applicability domain.

# 4. Conclusion

In this work, we combined several data mining methods into a unified DSS for PV solar cells libraries. Furthermore, the DSS was equipped with a simple, user friendly GIU. To the best of our knowledge, this is the first example of a completely automated tool that applies Data mining in the field of material informatics. The ability to use such tools for the analysis of solar cell libraries opens new opportunities for understanding the factors affecting solar cells performances and for designing new solar cells. Clearly, this research should be conducted in close collaboration with experts in the field to both provide physics/chemistry based explanation to the observed trends and to capitalize on the results. We expect that the tools and methods implemented in this work will further be used in material science researches.

# 5. List of Publications

* Yosipof, A.; Kaspi, O.; Majhi, K.; Senderowitz, H., Visualization Based Data Mining for Comparison Between Two Solar Cell Libraries. *Molecular Informatics* **2016**, 35, 622-628.

# References

1. English, R.; Lebovitz, Y.; Griffin, R., Institute of Medicine (US) Forum on Drug Discovery Development and Translation. Transforming clinical research in the United States. Challenges and opportunities: workshop summary. *Washington (DC): National Academies Press (US); PubMed PMID: 21210556* **2010**
2. Joseph A. DiMasi, Tufts Center for the Study of Drug Development, Cost of Developing a New Drug
3. H Geerts; A Spiros; P Roberts and R Carr, Has the Time Come for Predictive Computer Modeling in CNS Drug Discovery and Development? CPT: Pharmacometrics and Systems Pharmacology · November 2012
4. Breakthrough Business Models: Drug Development for Rare and Neglected Diseases and Individualized Therapies: Workshop Summary, National Academies Press
5. An analysis of the attrition of drug candidates from four major pharmaceutical companies: Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace & Alex Weir.