Generative Modeling and Sensitivity Analysis of Gallium arsenide-based Solar Cell Device Processes

Maryam Molamohammadi

Master of Engineering

Department of Mining and Materials Engineering
McGill University
Montreal, Quebec, Canada
April 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering

Abstract

Modeling and process optimization of solar cell devices is a complex procedure due to the high-dimensional parameter space. Moreover, the need for enhancing renewable energy market share is an undeniable fact. Machine learning is rapidly emerging in many fields while providing a promising toolbox for researchers to study complex problems more efficiently. Traditionally, research in materials science has involved a trial-and-error process that often requires various resources. With the help of machine learning, we can explore the high-dimensional problems with improved accuracy and efficiency and fewer user-imposed assumptions.

This thesis presents the modeling and sensitivity analysis of GaAs-based (Gallium arsenide-based) solar cell device processes. We propose a unifying framework for learning a solar cell performance function while providing intuitive interpretations based on the sensitivity analysis of the cell performance with respect to the material variables. In fact, this framework would be a faster and more computationally efficient replacement for its equivalent simulator in a downstream task of sensitivity analysis and consequently design optimization. We use the conditional variational autoencoder and multilayer perceptron for the generative modeling task and Jacobian-based sensitivity indices for the sensitivity analysis task. Furthermore, in the results, we validate our modeling approach with a baseline multilayer perceptron and our sensitivity analysis method with a sampling-based sensitivity analysis approach. We further demonstrate how the sensitivity analysis can be potentially useful from the design perspective as well as interpretability of possible device underperformance.

Abrégé

La modélisation et l'optimisation des processus des cellules solaires est une procédure complexe en raison de l'espace de paramètres de grande dimension. De plus, la nécessité de renforcer la part de marché des énergies renouvelables est un fait indéniable. L'apprentissage automatique (machine learning) émerge rapidement dans de nombreux domaines tout en offrant une boîte à outils prometteuse aux chercheurs pour étudier plus efficacement des problèmes complexes. Traditionnellement, la recherche en science des matériaux a impliqué un processus d'essais et d'erreurs qui nécessite souvent diverses ressources. Avec l'aide de l'apprentissage automatique, nous pouvons explorer les problèmes de grande dimension avec une précision et une efficacité améliorées et moins d'hypothèses imposées par l'utilisateur.

Cette thèse présente la modélisation et l'analyse de sensibilité des processus de dispositifs de cellules solaires à base de GaAs (d'arséniure de gallium). Nous proposons un cadre unificateur pour l'apprentissage d'une fonction de performance de cellule solaire tout en fournissant des interprétations intuitives basées sur l'analyse de sensibilité de la performance de la cellule en ce qui concerne variables de matériaux. En fait, ce cadre serait un remplacement plus rapide et plus efficace en termes de calcul pour son simulateur équivalent dans une tâche en aval d'analyse de sensibilité et par conséquent d'optimisation de la conception. Nous utilisons l'autoencodeur variationnel conditionnel et le perceptron multicouche pour la tâche de modélisation générative et les indices de sensibilité Jacobiens pour la tâche d'analyse de sensibilité. En addition, dans les résultats, nous validons notre approche de modélisation avec un perceptron multicouche de base et notre méthode d'analyse

de sensibilité avec une approche d'analyse de sensibilité basée sur l'échantillonnage. Nous démontrons en outre comment l'analyse de sensibilité peut être potentiellement utile du point de vue de la conception ainsi que de l'interprétabilité de la sous-performance possible du dispositif.

Contributions

The work presented in this thesis is done by Maryam Molamohammadi under the supervision of Professor Nathaniel Quitoriano.

Professor Quitoriano has played an important role through the course of this thesis by supporting the ideas, helping to develop the methodology, giving feedback on the process and results, and editing and reviewing the academic paper related to this work and the thesis itself.

Furthermore, this thesis correlates to the paper presented in the Machine learning for Engineering Workshop at Neural Information Processing Systems (NeurIPS) 2020 conference [1], whose primary author is the author of this thesis; with co-authors Sahand Rezaei-Shoshtari, who helped with the code implementation, and Professor Nathaniel Quitoriano.

Acknowledgements

First, I would like to express my gratitude to my supervisor, Professor Nate Quitoriano, for his guidance throughout the past years in different academic areas including this thesis with his constant support of new ideas while keeping me on track and giving valuable feedback through the process.

I would also like to thank Professor Demopoulos who kindly gave me the permission to work in his laboratory. Also, many thanks to Rana Yekani who brought the solar cell studies to my attention while patiently answering my questions as well as helping me a lot in the laboratory.

I would like to acknowledge my laboratory members, Han Wang and Galih R Suwito who have been very supportive in many ways. Also, I would like to thank all administrative staff of our department, especially Barbara Hanley and Leslie Bernier. Further, I want to show my appreciation to my office mates, Kirklann Lau, Tiffany Turner, Ahmad Saad, Sophia Smith and Kevin Li for all the fun and helpful discussions we had.

More importantly, many thanks to my mom and grandparents, Vida, Hamid, and Ashi, whose vision of life, unconditional love, support and friendship has built and is continuing to build many joyful moments and has shaped me into who I am today. I am deeply grateful for that.

To Sahand, I am very grateful and happy for what we created together through the past years which was not possible without your kindness and constant support which I am very grateful for.

Contents

List of Figures

List of Tables

1

Introduction

Solar energy and solar cell technology play a significant role in shifting the focus from fossil fuels as a primitive source of energy. The global energy consumption in 2019 was reported approximately 158.839 TWh. On the other hand, the average solar power resource on Earth's surface (including water surface area) is 1366 Wm^{-2} [21, 22, 3], therefore if we could efficiently capture only a few percentages, it would vastly contribute to the providence of energy and reducing carbon emissions.

Over the past few years, machine learning has showcased enormous success in a wide variety of domains and in some sense, we are coevolving with it in our lives in many aspects. From an engineering view, machine learning creates a promising toolbox for exploring a chosen problem more efficiently from different aspects, especially from the ones that maybe have not come to researchers' attention yet.

Process design and optimization is a key component in Materials science and engineering research fields. This often requires expensive resources, equipment and time and thus, many approaches such as design of experiments [23], Bayesian optimization [24], genetic algorithms [25] and particle swarm optimization [26] have been utilized in order to efficiently minimize the number of laboratory experiments. Machine learning can study the chosen problem more deeply and can consequently reduce the number of required experiments [27]. Solar cell performance depends on a series of sequential processes, each being a function of various material properties mainly governed by manufacturing parameters, intrinsic characteristics and solar cell architecture. From the device-design point of

view, solar cells consist of a different number of layers stacking together. Therefore, these processes and layers are entangled, and this makes predicting the outcome feasible only through high fidelity and expensive simulators. With a substantial amount of data, machine learning models not only can learn a generative model of the process as a faster and more efficient tool than the simulators but also can capture different aspects of these processes. It is worth noting that the outcome in solar cell processes usually refers to their J-V (current density versus voltage) characteristics.

1.1 Contributions of this Work

This thesis is an attempt at applying machine learning on a GaAs-based solar cell study. Our work proposes a unifying framework for learning a solar cell performance function while providing intuitive interpretations based on the sensitivity analysis of model outputs with respect to its inputs. As mentioned solar cell performance depends on a series of steps each of which is a function of numerous materials variables. Since for the process design, we have to consider all of these steps and their relevant material parameters, the optimization space becomes high-dimensional. In addition, the interdependence of these parameters makes it much complicated. We choose to work with the publicly available dataset [17] with selected five important material descriptors namely donor doping level in the absorber layer, acceptor doping level of the emitter layer, the bulk lifetime, front surface recombination velocity and rear surface recombination velocity.

The first objective of this work focuses on learning a generative model of the solar cell performance as a faster and more computationally efficient replacement for the simulator in a downstream task of design optimization. We propose to use generative models, namely conditional variational autoencoders (CVAE) [28, 29] for this task which is explained in chapters 3 and 4 completely.

The second objective of the thesis is motivated by the fact that in engineering problems, we often need an interpretation and intuition of the system, thus we perform sensitivity analysis. In a simple definition, sensitivity analysis studies the effectiveness degree of each input parameter on output. It gives a better insight into the underlying engineering process that can further help to understand the process, finding the root-cause of the system under-

performance, or design the real-world experiments. In neural networks, sensitivity analysis can provide a toolset for interpreting and understanding the learning process as well.

In our chosen task, we investigated the effectiveness of each of our material variables on the solar cell performance. For this goal, we propose to use the Jacobian of the trained neural network to obtain the derivative of the model outputs with respect to the variables and we show that the Jacobian can be interpreted as the sensitivity index of outputs (solar cell performance) with respect to the inputs (material variables). The methodology used in our proposed framework is completely presented in chapter 4. Needless to say that we validated our approach with a different baseline method in each step as well. Furthermore, we also explore the extended applications of our method while scientifically interpreting the results from our two objectives.

With that said, the main contributions of this work are:

1. Generative modeling of the solar cell device performance for learning a meaningful latent space.
2. Sensitivity analysis of performance with respect to the input parameters (materials variables) through Jacobian analysis of neural networks which provided reliable sensitivity indices. Furthermore, we explored some extended applications of our proposed framework.

1.2 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 presents the background literature relevant to this work. In particular, we review the solar cells working fundamentals and further explain Machine learning definition and general principles as well as its application in materials science.

Chapter 3 provides the theoretical foundations of the machine learning methods used in our proposed framework which we implemented in this thesis. Firstly, we explain Multilayer Perceptrons and then we present the Conditional Variational Autoencoder along with their brief mathematical foundations.

1.2 Thesis Outline

In chapter 4, we describe our proposed methodology and the two main objectives which are unified in an effective framework. The motivation behind the approach, the underlying methodology, and the details of the experiments and evaluation procedure are presented in depth.

Chapter 5 presents the implementation process and the results as well as some interpretations and applications of our proposed method.

Finally, Chapter 6 summarizes the conclusions drawn from this work, and is wrapped up by presenting the challenges faced throughout the course of this thesis while suggesting possible directions for future work.

2

Background

This chapter presents a brief overview of the fundamentals of solar cell devices including photovoltaic principles and common types of solar cells. In addition, it presents a general overview of machine learning and its application in Materials science and engineering and more specifically in solar cell studies, both of which have been studied in this work.

2.1 Solar Cells

From the historical perspective, the discovery of the photovoltaic effect is credited to Edmond Becquerel in 1839 [30] and the first embodiment of modern solar cells, which indeed was a Si-based (p-n) junction solar cell with 6% efficiency, was introduced in 1954 by the Bell Labs [31]. Since then, there have been many efforts going into solar cells studies and recently the highest efficiency reported (39.2% in 1-Sun illumination) namely six-junction inverted metamorphic structured III–V solar cell is introduced [32]. In the following sections, the fundamentals of solar cells are discussed.

2.1.1 Solar Cell Fundamentals

In terms of solar cell architectures, they are usually composed of different layers including, but not limited to: a window, emitter, absorber, back surface and a substrate layer, although they may be referred to by a different terminology for different types of solar cells. In terms of solar cell operation principles or photocurrent generation process, in simple words

includes several steps: light absorption, charge excitation, charge flow (drift and diffusion current), charge separation and charge collection. As a result, the device performance and thus output current and efficiency are affected by all of the aforementioned steps [3].

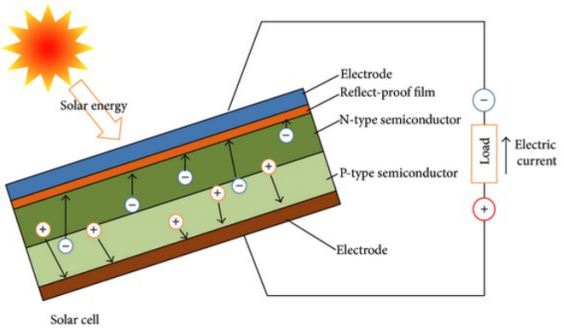


Figure 2.1: Schematic of a basic solar cell [2].

Light Absorption is the first step of the photocurrent generation process. Visible range photons usually have an energy range of (0.6-6) eV and while reaching a matter they mainly interact with the valance electrons. In fact, some photons just reflect from the front surface and do not enter the solar cell at all. For the rest, if the energy of a coming photon is lower than the band gap of the material used in the solar cell, it would be transmitted. If it is larger than the band gap (usually $E > 3 * Eg$) it may result in multiple carrier generation and excitation. Moreover, in general if the energy of a photon is too large in addition to charge generation, it results in the thermalization process and generation of phonons which is considered as a loss in the system. Indeed the effective photons would be those that are absorbed and have a sufficient energy. In this stage, we want to maximize the number of absorbed photons and reduce ones that are reflected or transmitted. This can be mainly obtained by managing the thickness of each layer, their absorption coefficients (by material choice) and interface geometry [3].

We can manage to minimize the number of reflected photons by an engineering parameter called reflectance, R. Indeed, we can describe matter and photon interaction by the index of refraction (n_c, a material property), in which through this the reflectance can be defined. The reflectance equation is shown in Equation 2.2. Thereby, the total number of photons which are actually going through a solar cell would be $(1 - R)\times$ (the input illumi-

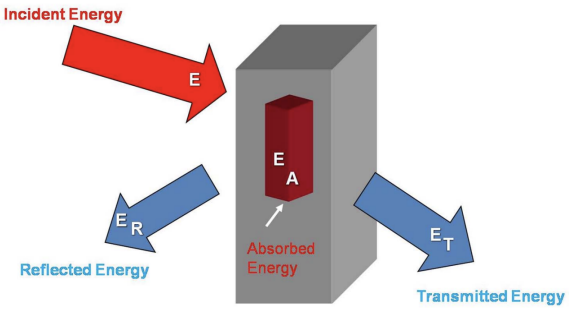


Figure 2.2: Light management[3].

nation intensity). Based on this concept, we can add anti-reflection coatings to the front or back of the device which is commonly used in solar cell architectures. It is worth mentioning that the real component of reflectance indicates phase velocity and the imaginary part or so-called extinction coefficient indicates attenuation of light intensity as it travels inside the material [3].

$$n_c = n + ik \tag{2.1}$$

$$R = \frac{[(n-1)^2 + k^2]}{[(n+1)^2 + k^2]} \tag{2.2}$$

Furthermore, in order to minimize the transmitted photons, we can use the Beer-Lambert law which is described in equation 2.3. This law shows that the photon intensity is decaying exponentially as it travels through a cell based on its absorption coefficient and thickness. Generally speaking, the absorption coefficient is dependant on the wavelength of coming photon. Since the peak intensiry of the solar spectrum is around 550 nm, we can calculate the necessary thickness of a given material and estimate the thickness needed to minimize the transmitted photons. Taking that into account, the thickness and materials cost should be both considered in material selection. The solar spectrum (intensity versus wavelength) is shown in Figure 2.3.

It is worth mentioning that the optical path length is a more important factor than the thickness itself considering that the photon can scatter or bounce within the material [3, 33].

$$I = I_0 \exp(-\alpha.l) \tag{2.3}$$

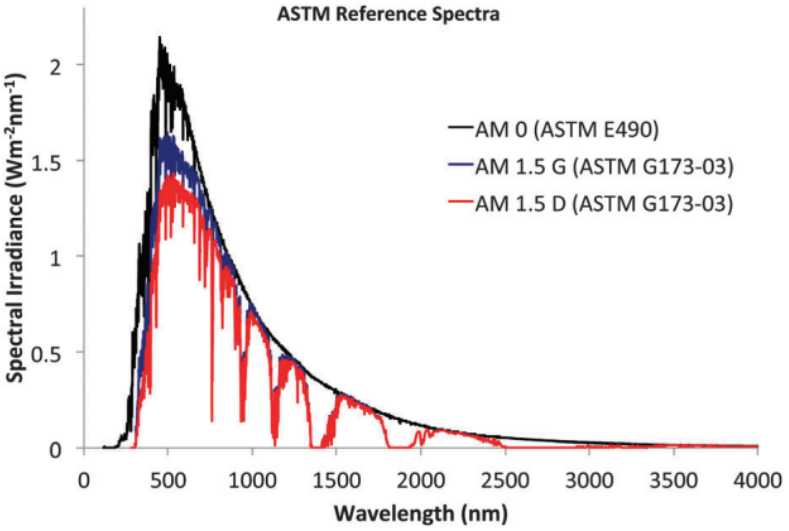


Figure 2.3: Solar Spectra versus wavelength [4, 5]

By considering the aforementioned facts, we can maximize light absorption using several engineering techniques such as:

- **Adding Anti-reflection Coating**- Suppresses the number of reflected photons
- **Texturization**- Increases optical path length
- **Back Surface Reflection**- Increases optical path length
- **Plasmonic Methods**-Increases adsorption

Charge Excitation is the second step towards the photocurrent generation and collection. As mentioned in the last step, there are two possible scenarios for photons which can actually be absorbed; in both they generate carriers, electrons in the conduction band and holes in the valance band. As mentioned, in cases that their energy is too large in addition to the carriers they cause lattice vibration or phonons due to the thermalization process. The thermalization process occurs when a previously excited electron loses its energy and move to the minimum of the conduction band. This energy is disappated as heat and cannot be converted to a current; therefore it represents a loss in the system.

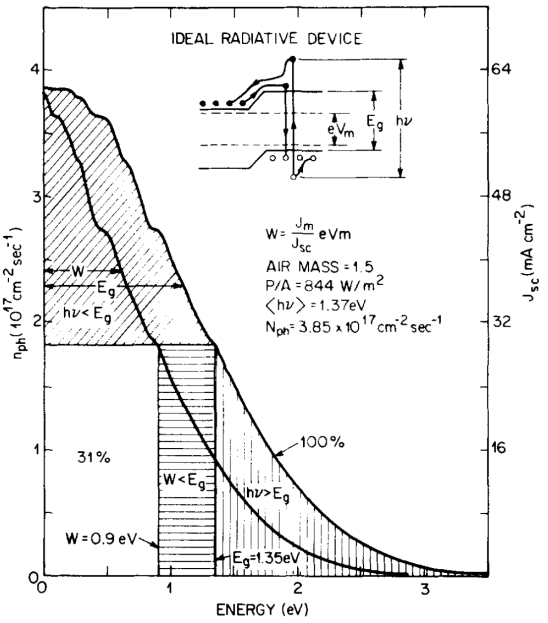


Figure 2.4: Representation of maximum power in single junctions [6].

From a different viewpoint, in terms of increasing our effective photons, the band gap energy of the material used in the solar cell should be optimum. It should not be too small because it will cause more thermalization loss and it should not be too large in order to decrease the number of transmitted photons. The solar spectrum in the visible range and the number of photons in each energy range is shown in Figure 2.4 and the area under the graph indicates the total number of photons. Due to the thermalization loss and the portion of transmitted photons, the optimum available area and thus the potential maximum efficiency would be 31%. The noticeable point here is the fact that this assumption is for a single material that has a specific band gap, if we can include more semiconductors by stacking them together we can capture more of the spectrum with lower thermalization losses, thus achieving higher efficiency. Thereby, the idea of the multijunction solar cells was developed [3, 6].

It is worth noting that the charge generation itself is due to the energy of the coming photons but the driving potential, which leads the charges to flow and ultimately be collected, originates from the built-in voltage in the (p-n) junction.

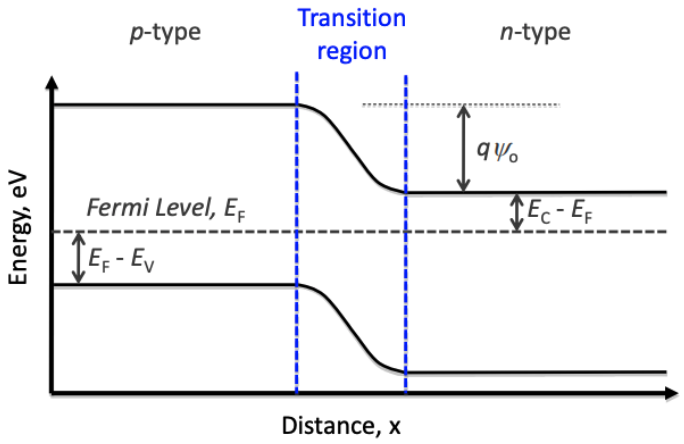


Figure 2.5: Voltage across a non-biased p-n junction[3].

Current Generation and Flow is the third step towards the photocurrent generation and collection. The role of the p-n junction used in solar cells is highlighted here. The built-in voltage in the p-n junction is caused by the balance between the drift and diffusion current flows of electrons and holes. The diffusion current occurs based on the Gaw's law (shown in equation 2.4) and the fact that spatially fixed charges create an electric field and in turn current. On the other hand, the drift current occurs based on the Fick's law (shown in equation 2.5) and the current resulted by the concentration gradient. Drift current is formed by the so-called minority carriers in each side. Generally a p-n junction in dark and under no bias condition, as shown in the figure 2.5, is in equilibrium condition and have zero net current. It is also defined by a depletion region and a built-in voltage. The diode equation, J-V relation, can be driven by numerically solving the combination of the continuity equations and drift and diffusion current equations [3, 33].

$$\frac{d\zeta}{dx} = \frac{\rho}{\varepsilon} \tag{2.4}$$

$$J_e = qD_e\frac{dn}{dx}, J_h = -qD_h\frac{dp}{dx} \tag{2.5}$$

Under illumination, the system is no longer in equilibrium, the electrons and holes generated by photons apply a potential bias to the p-n junction. In fact, there is an additional current flow, an illumination-generated current, on top of the diffusion and drift, thus the

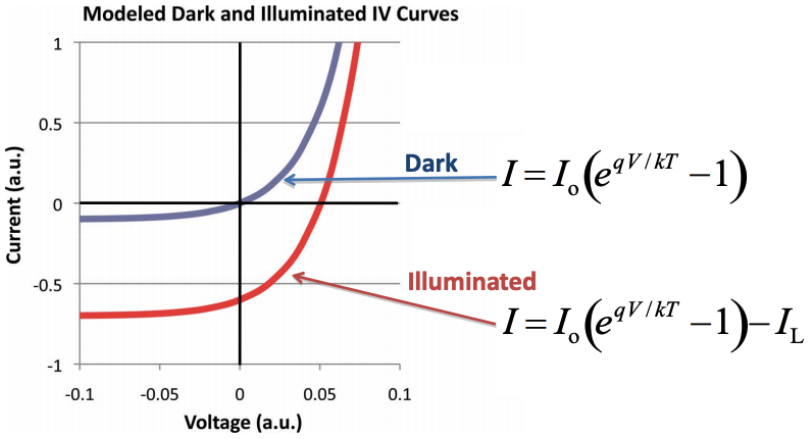


Figure 2.6: Voltage versus current density of a diode in dark and under illumination, I_0 is saturation current and I_L is illumination current (equations are driven based on some assumptions and simplifications) [3].

net current is no longer zero. This is shown in Figure 2.6. The illumination-driven electrons flow in the same direction as the diffusion current and cause more electrons in the n-type side. Therefore, they change the Fermi level on both sides. In fact, illumination imposes a forward bias condition to the solar cell. As there would be higher electron concentration on the n-type side, the electrons travel from the n-type material through the external load in order to recombine with holes in the p-type side. This is done thermodynamically based on their tendency to reduce their free energy and in turn reducing the chemical potential (Fermi level) difference between the n-type and p-type sides. While they are traveling through the external load, we can capture them and deposit their power on a resistor and that is how the solar energy is captured. In addition to a resistor we can have batteries or other solar cell devices attached in series in a real module.

Charge Separation is the fourth step of the photogeneration process. The generated current to be able to travel through the external load, separated and be collected, first carriers have to reach the surface. In other words, the carrier's diffusion length should be enough to move through the thickness and reach the metal junction. The diffusion length is defined as the average radius that a minority carrier moves before recombining. Thus, on the p-type side we pay attention to the electron diffusion length and on the n-type side the holes diffusion length. The diffusion length affects the short circuit current, a longer diffusion

length provides more short-circuit current. The other components which are affected by the diffusion length are the saturation current and thus the open-circuit voltage. Therefore, diffusion length is highly effective on the energy conversion efficiency of the device. It is calculated in equation 2.6 in which D is the diffusivity of the material used in the under studied layer, τ is carriers lifetime and μ is carriers mobility.

$$L_{\text{diff}} = \sqrt{D\tau}, D = \frac{K_b T \mu}{q} \tag{2.6}$$

The diffusion length has a direct relation with the bulk lifetime, temperature and mobility. The bulk lifetime is defined as the average time that a minority carrier can move around before recombination. While the bulk lifetime is more related to the concentration of recombination centers (grain boundaries, defects, etc.), the mobility depends on both the material choice and present defects [34, 3].

In addition to the diffusion length, the surface passivisation is also an important factor in order to determine the charge collection probability.

Now that the overall view on charge separation is established, the resistances or differentiation from the ideality will be presented in the following. If we consider the solar cell device as a circuit, we can have *series resistances* and *shunt resistance* (parallel). In addition to these resistances, we will introduce recombination mechanisms which can occur in a solar cell device.

Series resistance can be caused by movement resistance in different layers of a cell, such as bulk resistance, emitter sheet resistance, contact resistance, and line losses. As is shown in Equation 2.7 and Figure 2.7, with increasing series resistance, the actual J-V curve deviates from the ideal J-V output of the device. The effect of the series resistance is mostly observed in a higher voltage range in a measured J-V curve. Clearly the goal is to reduce this resistance and one solution for that is by designing a solar cell device in which the series resistance becomes effective after the max power point. As expected, series resistance have an indirect effect on Fill factor and the increase of series resistance, results in a drop in the Fill factor [3].

$$J = J_0(\exp{(\frac{q(V - JR_s)}{K_bT})} - 1) - J_l \tag{2.7}$$

Some of the key sources of series resistance are:

- Bulk or absorber resistance depends on the bulk thickness and its resistivity. In fact bulk thickness should be optimized to reduce the resistance and efficiently absorb light. If it is too thin it may not efficiently absorb and if it is too thick it shows high resistance.
- Emitter sheet resistance is defined as the resistance which current is experiencing as it laterally moves towards the contacts. Sheet resistivity has an inverse relation with thickness. Thus, in order to have low resistance it should be thick enough. Conversely, if it becomes too thick it may absorb short-wavelength incident photons and prevent them to go through the bulk thereby reducing the quantum efficiency. Also, if the emitter is too thick it may result in recombination events. The power loss due to the emitter sheet resistance is a function of sheet resistivity and the finger spacing (the space between two contact grids). If we reduce the finger spacing the power loss will be decreased but as we increase the contact grid we are shadowing our device which limit the input light. So finger spacing should be optimized as well. As described there are many parameters in solar cell design that should be co-optimized.
- Contact resistance is defined as the resistance between the metal and emitter semiconductor interface. It can be affected by the atomic interface of the metal and

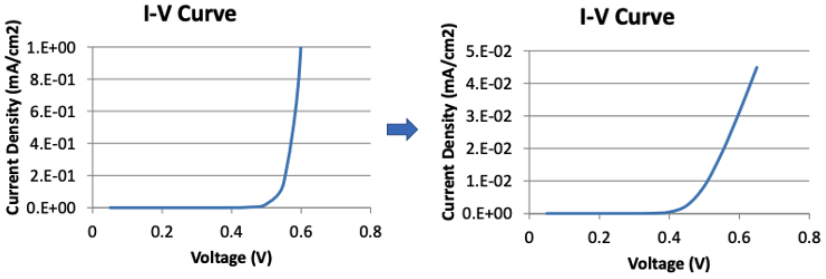


Figure 2.7: J-V of an ideal solar cell on the left and one experiencing series resistance on the right [3].

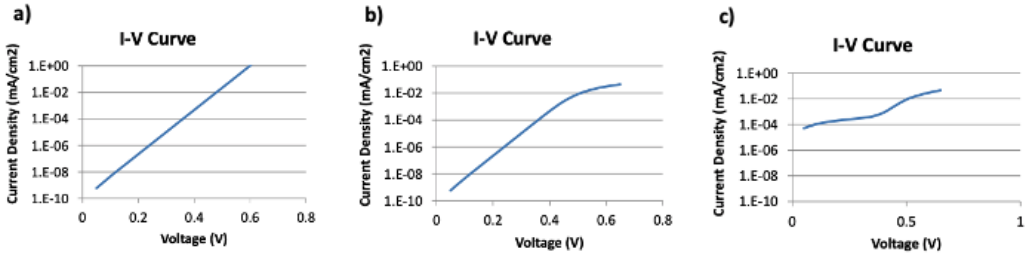


Figure 2.8: J-V in log scale of a) an ideal solar cell b) experiencing series resistance c) experiencing series and shunt resistance [3].

emitter, as well as the dopant's type and density within the emitter.

- Line losses are defined as the losses occurring in the contact metalization lines. Same as the bulk charge movement in which the current is traveling in one direction, in these lines charges travel in one direction as well. Therefore, the material choice for the metal and its resistivity becomes an important factor here[3, 34].

Shunt resistance is created to mainly resist the backflow current inside the device. It means this resistance prevents the situation in which the current instead of going through the external load, flows as the diffusion current backward through the device. The regions in which the current flows backward are called the shunt pathway. The goal is to maximize the shunt resistance or minimize the shunt regions which both reduce the amount of current that flows backwards. This can be done by enhancing the quality and homogeneity of p-n junction. The homogeneous p-n junction provides strong shunt resistance. In other words, shunt pathways are created when there are locally inhomogeounities. It is really important because current crowding can happen in those regions and locally heat the device. This shunt resistance's impact is noticeable by running a J-V measurement at low bias and then compare it to the ideal diode. Also, by increasing the shunt resistance, the Fill factor drops. Equation 2.8 shows the J-V equation correction with the presence of both series resistance and shunt resistance [3].

$$J = J_0(\exp{(\frac{q(V - JR_s)}{K_bT})} - 1) - \frac{(V - JR_s)}{R_{sh}} - J_l \tag{2.8}$$

Recombination Losses

In terms of losses existing in a solar cell device function, there are two possible types of recombination scenarios. In the low bias conditions since there is a noticeable built-in field, some current recombination can occur in the depletion region (space charge region). As bias enhances since the barrier height starts decreasing and makes current flow easier over the junction, there is a higher possibility of current recombination in the bulk. In fact to be exact, we should consider two different saturation currents in our J-V equation.

Generally, recombination is presented by recombination velocity. There are three types of recombination occurring in the bulk, namely Auger, Radiative and Shockley-Read-Hall. Radiative recombination is actually the recombination which involves emitting photons and it can be a limiting factor in direct bandgap materials. Auger recombination stems from injecting high dopant densities. In other words, it is an important factor in determining the suitable dopant concentration. Auger recombination in an n-type material is caused by the simultaneous relaxation of a conduction electron and an excitation of another electron which relaxes to the conduction band minimum, releasing a phonon or lattice vibration. Shockley-Read-Hall recombination can be considered as a model to predict defect-mediated recombination. Since the contamination and defects (impurity point defects to be more specific) introduce midgap trap states, these trap states will act as recombination centers and reduce the collection probability. The Shockley-Read-Hall recombination is often the one which is limiting our bulk lifetime and thus reduces the diffusion length. All recombination mechanisms are a function of different parameters including illumination intensity[3, 34, 20].

It is worth mentioning that Shockley-Read-Hall recombination can be present at the surfaces as well. At interfaces dangling unsatisfied bonds create many trap states within the bandgap. Again, these trap states can provide recombination centers. The surface lifetime depends on sample thickness, recombination velocities and carrier diffusivity. The solution to prevent the creation of trap states in interfaces is to passivate them.

Charge Collection is the last step in solar cells operation processes. It is the step in which carriers would be collected from the device. Contacts are used in charge collection mainly in order to extract current and prevent the back diffusion of carriers into the device. Ma-

terials commonly used for contacts are metals, transparent conducting oxides and heavily doped organics. In terms of properties, it should be electrically conductive. For instance, a transparent conducting oxide like ITO (Indium doped Tin Oxide) is transparent due to its large band gap and is conductive due to the presence of the intermediate energy level created by dopants which makes it suitable for contact application.

Contacts can contribute to series or shunt resistance and thus be effective on J-V characteristics of the device. In fact if a poor contact is made, it will build series resistance. Additionally, if the contact drives too deep in the device it will build shunt resistance. Ideally, a binary compound would form at the interface of the metal and device. There are two classes of contacts, Ohmic and Schottkey which respectively follow Ohm's law and an exponential J-V relation. In reality, besides the choice of materials, the atomic configuration and orientation of the interface is important in determining the barrier height [3].

Solar Cells Characteristics

There are several terms characterizing solar cells. Open circuit voltage (V_{oc}), short circuit current (J_{sc}), Fill factor (FF), Maximum power point (MPP), Energy conversion efficiency (η) and quantum efficiency (QE) are often used for the sake of comparing performance of different solar cell devices. In the following a short definition of each of them is introduced.

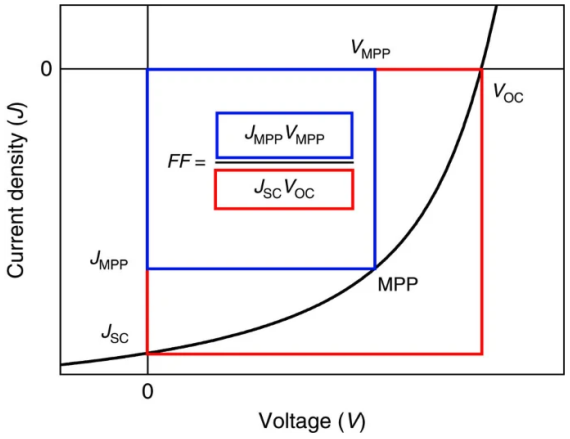


Figure 2.9: General form of J-V graph in solar cells[7].

- **Open Circuit Voltage** is the bias built-in a solar cell device under illumination and in absence of any external load. Open circuit voltage is roughly a function of the interfaces.
- **Short Circuit Current** is a flow of current producing because of the recombination of electrons and holes in p-n junction under illumination when it is short-circuited, i.e., there is no load. Short circuit current is approximately a function of the bulk material quality and illumination condition.
- **Maximum Power Point** is maximum power which can be deposited on the external load. There is maximum power current density and maximum power voltage associated with the maximum power point (J_{mp},V_{mp}). MMP is a function of the ideality of the diode characteristics, V_{oc} and J_{sc} . It is worth mentioning that V_{oc} and V_{mp} are related to the band gap of semiconductors used in the device.
- **Energy Conversion Efficiency** is the ratio of the maximum output power to the input power (total illumination). Typically the efficiency is reported under illumination condition of Air Mass (AM) of 1.5 or 1 Sun at sea level which the input power is 1000 w/m^2. From the engineering point of view, efficiency is important since for the desired power output, it determines the area of the device needed and the cost. The energy conversion efficiency equation is shown in which ϕ is the illumination intensity.

$$\eta = \frac{V_{mp} I_{mp}}{\phi} \tag{2.9}$$

- **Fill Factor** is the ratio of the area obtained from the multiplication of J and V related to maximum power point (V_{mp}, J_{mp}) and J_{sc} and V_{oc}. It usually represents the quality of the p-n junction used in a device, the interfaces, and resistances within the device.

$$FF = \frac{V_{mp} J_{mp}}{J_{sc} V_{oc}} \tag{2.10}$$

- **Quantum Efficiency** is the ratio of the number of output electrons to the number of input photons; this is the external quantum efficiency. The internal quantum efficiency considers the portion of reflected photons and is thus defined as the ratio of the number of output electrons to the number of absorbed photons. These values are

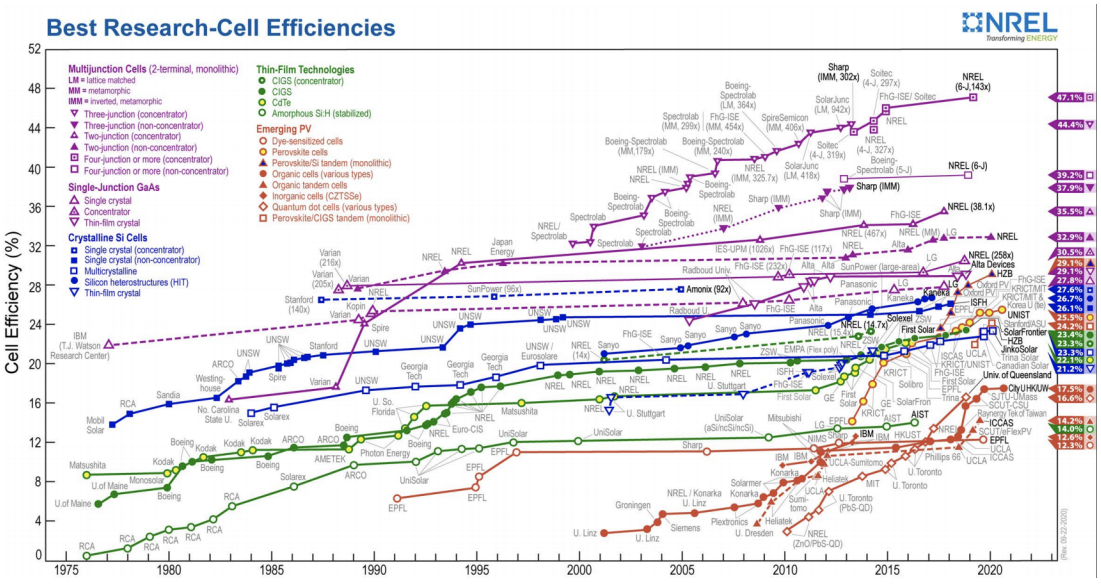


Figure 2.10: Cell efficiencies reported by NREL for different types of cells and their developments throughout the years[8].

wavelength dependant and measured in a given wavelength [3, 20].

2.1.2 Types of Solar Cells

According to the National Renewable Energy Laboratory (NREL), we can categorize solar cells into five main families each of which can be further divided into a variety of subcategories. These categories are:

- Si solar cells
- Single junction GaAs solar cells
- Multijunction solar cells (two junctions or more)
- Solar cells based on Thin-Film Technologies (CIGS, CdTe and Amorphous Si:H)
- Emerging Photovoltaics: Perovskite, Dye-sensitized, Organic, Organic Tandem, Inorganic (CzTSSe) and Quantum dot

Si-based Solar Cells

Crystalline silicon cells are used in the largest quantity of all types of solar cells on the market, representing about 90% of the world total PV cell production [35]. One of the reasons that crystalline silicon is dominating the PV industry is the fact that microelectronics has already been well developed in silicon technology. Secondly, aside from the fact that there is an accumulated knowledge on Si industry, the PV community has also benefited from the silicon feedstock and second-hand equipment have been acquired at reasonable prices [20]. Today, silicon is used in crystalline, multicrystalline and amorphous form. Initially, Si solar cells have been used in monocrystalline structures, but now multicrystalline is more common. A combination of improved material quality and material processing has allowed higher efficiencies at a lower cost.

In terms of Si physics, crystalline silicon has a fundamental indirect band gap (E_g = 1.17 eV) and a direct gap above 3 eV at ambient temperature. It has a low absorption coefficient in the range of photons having an energy range near the band gap. At short ultraviolet (UV) wavelengths of the solar spectrum, the generation of two electron–hole pairs by one photon seems possible, though in small quantities. At long wavelengths, band-to-band movements compete with carrier generation. Recombination in Si cells is usually dominated by Shockley–Read–Hall (SRH) recombinations and thus defects. Also, Auger recombination becomes important at high doping concentrations. Band-to-band direct recombination is a fundamental process but quantitatively negligible in these cells. Silicon crystals with metal content less than 10 ppb(w) have minority-carrier lifetime values as high as 10000 μs[20]. Nowadays, the Czochralski process is commonly used to produce silicon crystals and industrial cells use Czochralski (Cz-Si) wafers because of their availability[36].

GaAs Solar Cells

In terms of optical properties, the absorption coefficient of GaAs, as shown in Figure 2.12, is relatively high (compared to Si) in the visible photons range and particularly at 550 nm which is the peak intensity of the solar spectrum. Optical parameters and dielectric function of GaAs are investigated in [37], which proves the usefulness of GaAs in photovoltaic

Figure 2.11: Daryl Chapin, Gerald Pearson, and Calvin Fuller demonstrated their invention, the first practical solar cell invented in the Bell labs at 1954 [9].

devices.

One of the preferred candidates for window layer or back surface field in GaAs-based cells is GaInP. In fact GaInP/GaAs tandem-cell structures are composed of semiconductors that are all closely lattice-matched and have the lowest interface recombination velocities.

It is worth mentioning that the higher efficiencies and radiation resistance and smaller coefficient of thermal expansion of III-V cells, mainly GaAs-based (multijunctions), have made them an attractive candidate to replace silicon cells on many satellites and space vehicles[20].

Multijunction Solar Cells

As mentioned in the last section, one of the solutions to the fundamental problem arising from the Schottky and Queisser limit, is to use several semiconductors with different band gaps stacking together to convert photons of different energies. The materials used in these multijunction solar cells are mostly III-V materials. The cell configuration and thus band alignments of these semiconductors would be in a specific order so that the upper layer (the layer closest to the light source) has the largest band gap and would let the photons pass through the inner layers which have respectively narrower band gaps. Usually low-

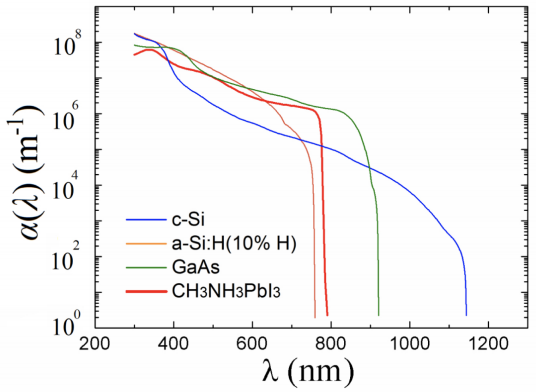


Figure 2.12: Absorption coefficient of different absorber layer versus wavelength [10].

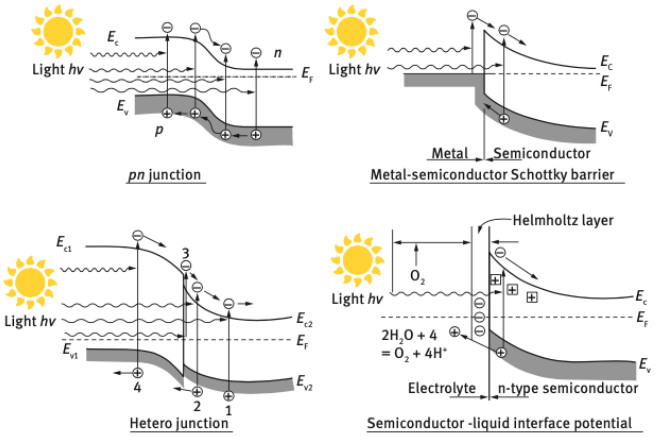


Figure 2.13: Different semiconductor interfaces [11].

energy pass filters are placed between layers so that the reflection threshold of each filter is the band gap of the cell situated above. This helps with light trap management within the device[3].

According to NREL (National Renewable Energy Laboratory) and to the best of our knowledge, the highest converted energy efficiency recorded is related to the six-junction cell with the efficiency of 47.1% measured under concentrated illumination and 39.2% under global 1-sun illumination intensity. The device contains about 140 total layers of various III-V materials to support the performance of these junctions, and yet is three times narrower than a human hair [32].

Thin-Film Technologies

Usually, the thickness calculations should be co-optimized since thicker layers mean more effective light absorption but also thicker layers correlate to more minority carrier recombination so that the thickness should be comparable to the minority carriers diffusion length. Film thickness in thin-film technologies varies from a few nanometers (nm) to tens of micrometers (μm). Generally, a reduction in the cell thickness itself can result in improving the open circuit voltage (V_{oc}) and the Fill factor (FF) of the solar cell. However, as the cell thickness reduces, surface recombination becomes an increasingly important component of the total recombination. Increasing the surface recombination can severely decrease the open circuit voltage. So, thin cells can end up to higher efficiency, higher voltages and higher fill factors if the surface recombination demands are met[3].

There are several types of thin film solar cells namely, CdTe (Cadmium telluride), CIGS or $CuIn_{(1-x)}\ Ga_xSe_2$ (copper indium gallium selenide cells) and Amorphous Si cells. In early studies CdTe p-n junctions were not favourable because of the high rate of surface recombination and strong unwanted optical absorption of the n-CdTe layer. However, what is now referred to as the CdTe cell have n-CdS/p-CdTe structure is more preferred [38] because of its good rectification properties. While the CIGS is a tetrahedrally bonded semiconductor and a versatile material that can be fabricated by various processes and implemented in different forms. Thin films are usually fabricated by evaporation, sputtering, CVD (chemical vapor deposition), plasma decomposition of gases and electroplating [39].

Emerging cells

There are five main types in this category namely, Dye-sensitized, Perovskite, Quantum dots, Organics, and In-organics.

Dye-sensitized cells (DSSCs) have the potential to compete with traditional solar cells. Materials such as TiO2 used in DSSCs are generally inexpensive, abundant and environmentally benign. In comparison with silicon solar cells, they are less sensitive to impurities, which accelerates a transition from the research stage to mass production. Also, from the application point of view, the low weight and flexibility of DSSCs are desirable for portable electronic devices. The schematic of Dye-synthesized cells is shown in figure 2.14. In sum-

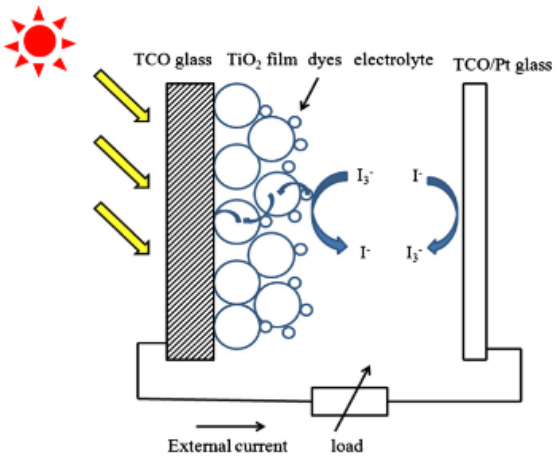


Figure 2.14: Schematic of a Dye-sensitized cell [12].

mary the components and function of DSSCs are brought in the following order:

1. When exposed to sunlight a mesoporous oxide layer deposited on the anode to activate electronic conduction.
2. A monolayer charge transfer dye covalently bonded to the surface of the mesoporous oxide layer to enhance light absorption.
3. An electrolyte containing redox mediator in an organic solvent effecting dye-regenerating.
4. A cathode made of a glass sheet coated with a catalyst to facilitate electron collection [12].

Perovskite cells attract great attention in recent years due to several reasons including low production cost, ease of fabrication and improving device efficiencies [40]. With high efficiencies achieved in lab devices, stability and challenges regarding upscaling were the main issues. From the structural point of view, perovskite solar cell clearly uses perovskite structured material as the absorber layer. The term “perovskite” is referred to all compounds with the same crystal structure as calcium titanite. The perovskite layer has a general formula of ABX_3, where A is an organic cation (e.g., methyl-ammonium CH_3NH_3+), B is a metal cation (e.g., Pb_2+) and X stands for the halide anion (e.g., I-). Perovskite’s largely tunable band gap, high absorption coefficient and long carrier diffusion length made it suit-

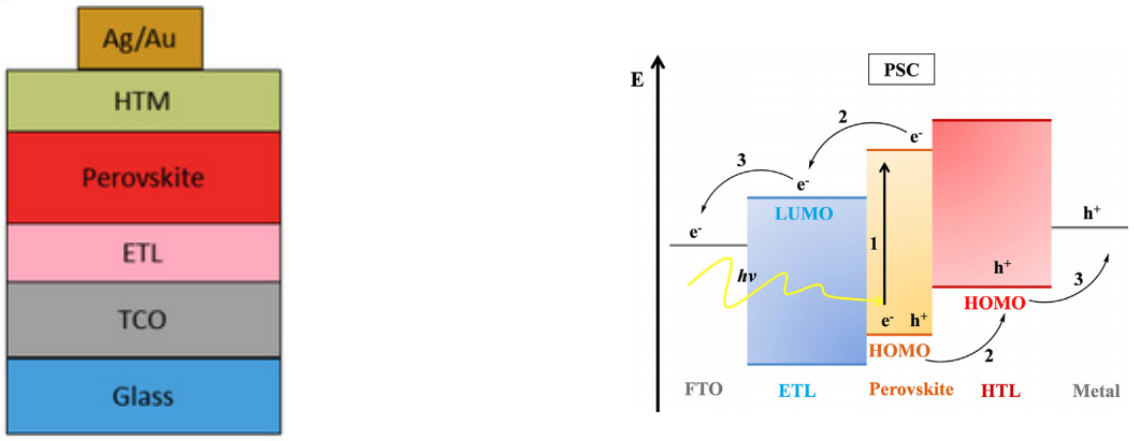


Figure 2.15: a)Schematic of a Perovskite cell[13]. b) Band alignments of a perovskite cel[14].

able for solar cell applications [41]. The general architecture and band alignments of these cells are shown in Figure 2.15.

Quantum dot cells have also attracted much attention for several reasons. Firstly, size quantization allows us to tune the visible response and vary the band offsets to modulate the vectorial charge transfer across different-sized particles. In addition, it opens up new ways to utilize hot electrons or generate multiple charge carriers with a single photon. Impact ionization (or inverse Auger scattering) processes in PbSe nanocrystals have shown that two or more carriers can be generated with a single photon of energy greater than twice the band gap [42, 43].

Now that an overview on solar cell operation fundamentals and its common types have been presented, in the second part of this chapter, an overview on machine learning in Materials engineering will be presented.

2.2 Machine learning

As Tom Mitchell (a computer scientist and professor) stated, machine learning is the study of computer algorithms that allows computer programs to automatically improve through experience.

Machine learning methods can be categorized into four main groups, supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In principle, supervised learning is learning the structure within a labeled dataset and unsupervised learning learns the unlabeled dataset. In other words, supervised learning tries to predict target Y from input X, these targets could be class IDs for a classification task or real numbers as a regression task. In unsupervised learning there is no explicit target to predict, all data has a unified input X. In semi-supervised learning we have a few supervised labeled data in addition to many unsupervised data and it tries to predict target Y from input X [15]. Reinforcement learning on the other hand is a problem that studies intelligent agents with interactive actions in an environment in order to maximize a notion of reward.

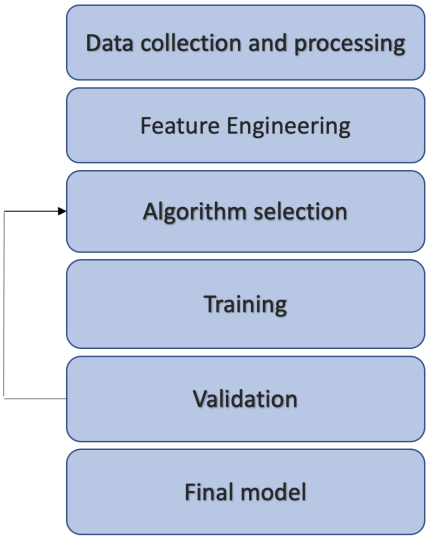


Figure 2.16: General pipeline of applied machine learning.

2.2 Machine learning

From the historical perspective, the journey of machine learning and neural networks began from the theory of Donald Hebb, a psychologist and neuroscientist in the late 1940s. Arthur Samuel of IBM developed a computer program for playing checkers in the 1950s and coined the term "Machine Learning". Then Frank Rosenblatt, a psychologist, invented the concept of perceptrons for pattern recognition by combining Donald Hebb's model of brain cell interaction with Arthur Samuel's Machine Learning efforts. In the 1990s, work on machine learning transferred from a knowledge-driven to a data-driven approach. Computer scientists began creating programs to analyze large amounts of data and draw conclusions or "learn" from the results [44]. After that and today, companies and researchers race to create more algorithms and models which can be generalized to different types of real-world problems.

Figure 2.16 indicates the general process of applying machine learning to a specific problem or other fields. It starts with gathering data, processing and cleaning the data and continues by choosing a particular algorithm which best matches the nature of our chosen problem. Then we train our model based on the collected data. In fact usually a dataset should be divided into three groups, a training set, a validation set and the test set. Respectively, they serve to train the chosen model, to select its hyper-parameters and to estimate the generalization performance. Typically, one uses about 70% of the training data for training and 15% for each validation and test. There is an important concept in machine learning called generalization. Generalization is the behavior of the model on unseen examples, which is indeed the main goal of machine learning. Another commonly used term is Hyper-parameters which are features of most machine learning models that are fixed during the training and are selected by us. These settings control the behavior of the learning algorithm. In fact, we select their optimal conditions based on the validation performance[15].

Every machine learning algorithm functions with the notion of the loss function in which it tries to decrease it through progressive experience. It may seem that by increasing the training time or increasing the number of free parameters in an algorithm, the loss will be decreased, indeed the training loss will decrease but the validation loss almost always decreases and then increases. This brings us to the concepts of underfitting and overfitting or the Bias-Variance trade-off. In an ideal condition, we want to have both low bias and

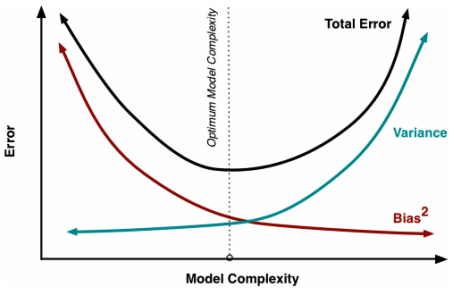


Figure 2.17: Bias-Variance trade-off[15].

variance in our model, but they are not always aligned and thus the trade-off should be optimized [15].

Towards a better understanding of this thesis, we explain following terms often used in machine learning. There are many different approaches that we can categorize learning methods and models within the machine learning field. One of these categories is defined as having a discriminative learning and another is generative learning which are two types of learning processes. In a discriminative learning the model studies the conditional probability $P(X|Y)$ in which it maps X a class label Y. Whereas, generative learning studies the joint distribution of data which is $P(X, Y)$. In this thesis, as it will be discussed thoroughly in the following chapters, we use a generative modeling approach, namely the Conditional Variational Autoencoder (CVAE) as our proposed method for modeling solar cell performance. Since generally the generative modeling approach compared to discriminative models has a higher capacity for learning high-dimensional problems and is thus able to better learn the underlying data distribution. Our results also further confirmed the outperformance of the generative model in comparison to the discriminative one in our task.

It is worth mentioning that, neural networks are currently most common method in machine learning frameworks. Neural networks can have different architectures such as Multilayer perceptron (MLP), Convolutional neural networks (CNN) [45], Recurrent neural networks (RNN) [46] and transformers [47]. In this study, as it is described in the chapters 3 and 4 we use Multilayer perceptron as the architecture for any method that we use.

In principle, machine learning can model complex functions, generalize to unseen samples, and handle high-dimensional spaces and large datasets. We specifically used two machine learning methods in our work, namely Conditional Variational autoencoders (CVAE) and Multilayer perceptrons (MLP). They are fundamentally different methods. In order to better understand each method, refer to the next chapter, Preliminaries. In the following, we introduce a short introduction on the applications of machine learning in materials science and engineering.

2.2.1 Applications of Machine Learning in Material Science

Traditionally research in Materials science and engineering was designed based on a loop of fundamental science study and designing an experiment, then implementing that experiment and effort towards optimizing that with a process of trial-and-error experiments. This often requires resources, equipment and time. The first computational revolution in materials science was strengthened by the advent of computational methods, especially density functional theory (DFT), Monte Carlo simulations, and molecular dynamics, that allowed researchers to explore more efficiently with using the acquired knowledge to design more targeted experiments. Now with the emergence of machine learning, it becomes one of the most useful tools that have emerged in material science research, bridging the gap between multi-level theoretical models and trial-and-error approaches [48]. The main reason why AI and machine learning are particularly apt in material design is due to their inherently strong capabilities in handling large amounts of data as well as high-dimensional analysis. A single material type within a synthesis protocol could contain large intrinsic information, such as various physiochemical properties, chemical structure, and composition information. In addition, adding synthesis condition information such as temperature, pH, pressure and reaction time, the dimensionality of data can create even more dimensions [49].

Recently machine learning has evolved in many different aspects of materials engineering namely: automating materials' characterization and effectively analyzing the characterization dataset, screening the vast material design space (e.g., reducing the prediction time of DFT), predicting properties of complex material systems, mapping high-dimensional synthesis recipes to materials with desired properties, extracting or interpreting generalizable scientific principles from various material systems and discovering new materials

[49]. From another perspective, machine learning can be used to capture the underlying complex functions of computationally expensive simulations or real-world experiments in order to potentially replace them as an optimization function. By providing a variety of architectures and methods it gives us the chance to easily choose a model which is more suitable for our selected problem. In the following, some already successful examples of applied machine learning in materials science are introduced.

Characterization

Characterization and particularly image-based characterization usually contains a large amount of grid-like, high dimensional data. Machine learning in the loop or machine learning, in general, helps to automate the characterization process, leading to a reduction in manual work, improvement of data quality, and discovery of useful hidden information from the high-dimensional data. Traditionally, the process of interpreting information resulted from characterization techniques was done manually or by some computationally intensive processing methods. The lack of automation results in a potential loss in accuracy due to human errors. With advances in computer vision, image-based characterization has been improved significantly.

Online automated processing and reconditioning the microscopy conditions is available by mostly applying convolutional neural networks (CNN) and Kernel Regression to different microscopies. In fact, convolutional neural networks revolutionized the whole notion of computer vision. Kernel regression is utilized to identify the optimum imaging parameters that can be improved by learning. The CNNs have been developed to consequently recognize and recondition the quality of the probe of a scanning tunneling microscope (STM). Apart from that, the discovery of hidden information like determining the crystal formation and growth rate, structure classification and detailed defect detection is further available by CNN-based methods with a combination of different methods or simulators (even when the data is noisy and incomplete) [49].

Property Prediction

The goal of property prediction is to find a fair accuracy function that relates the intrinsic materials' information to the desired functional properties. Machine learning can assist the

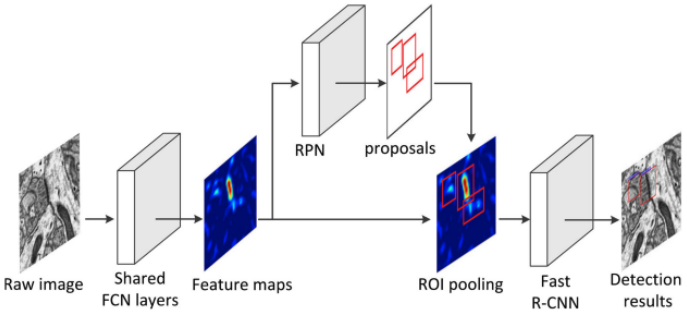


Figure 2.18: Synapse detection done by convolutional neural networks[16]

material property discovery of well-understood systems, this can be done directly by data or coupled with simulation and experiments. There have been great advances in computational material data due to previous efforts from the material genome initiative that provides resources that accelerate materials research [50, 49]. Additionally, machine learning can completely or partially substitute the computationally expensive simulators or increase their accuracy. Since in the case of materials with limited understanding, their properties can be potentially learned from data [49].

For instance, Ya Zhuo *et al.* predicted the band gaps of inorganic solids by support vector machines based on an experimental dataset. In particular, simulation models like density functional theory usually underestimate the band gap compared to real values measured in the lab. Other better functioning simulation methods such as the Becke-Johnson method [51] are either computationally very expensive or not generalizable to other inorganic solid [52]. As another example, the thermodynamic stability of perovskites used in perovskite solar cells is probably the most challenging issue about them and it is hard to estimate. Stability can be calculated using Density Functional Theory (DFT), but it requires a high computational cost when mapping large numbers of compounds. Wei Li *et al.* developed a tree algorithm and kernel ridge regression for predicting the stability of perovskite oxides [53]. In another study, the discovery of a new structural property of glassy liquids, softness, was investigated by training a support vector machine (SVM) with several structure descriptors. This new softness property can be considered as a property that is constructed by a combination of the several structure descriptors[54, 49].

Process Optimization

There are many multiparameter problems in materials science which can be explored by machine learning in order to optimize their synthesis process or interpret and advance investigation of their underperformance. This can either imply the minimization or maximization of a single property or the search for material in the case of requiring multiple objectives [50]. For instance, Kyoungmin Min *et al.* optimized the synthesis parameters of Ni-rich cathode materials that meet electrochemical specifications. For this purpose, they applied the extremely randomized tree model using an experimental dataset[55].

The most usual algorithm choices for process optimization are Bayesian models like Gaussian processes, as they also provide the variance of the predicted function. Gaussian processes have been implemented in different areas of structure optimization and design problems in materials science. A few examples are predicting crystal structures[56], interface configuration[57] and optimizing structural design for optoelectric devices[58]. Other alternative model examples to explore in process optimization space could be support vector regression (SVRs) or decision tree methods[50].

For further study on machine learning applied on materials science their are several review studies including [50] and [49].

2.2.2 Applications of Machine Learning in Solar Cell Studies

The development of a typical solar cell requires three separate sets of parameters which can be optimized independently. These set of research areas are, finding the appropriate PV materials, optimizing the device architecture, and developing the fabrication processes by obtaining optimum intrinsic or fabrication parameters. Many machine learning methods are used for the mentioned aspect of studying solar cells, namely multilayer perceptron, genetic algorithms, particle swarm optimization, support vector machine (SVM), kernel ridge regression (KRR), a randomized trees algorithm, K nearest neighbors (kNN), gradient boosting (GB) and ant colony algorithm (ACA) [59].

For instance, for lead-free perovskites used in solar cells, Jino Im *et al.* applied the gradient-boosted regression trees (GBRT) algorithm to a dataset in order to predict and

2.2 Machine learning

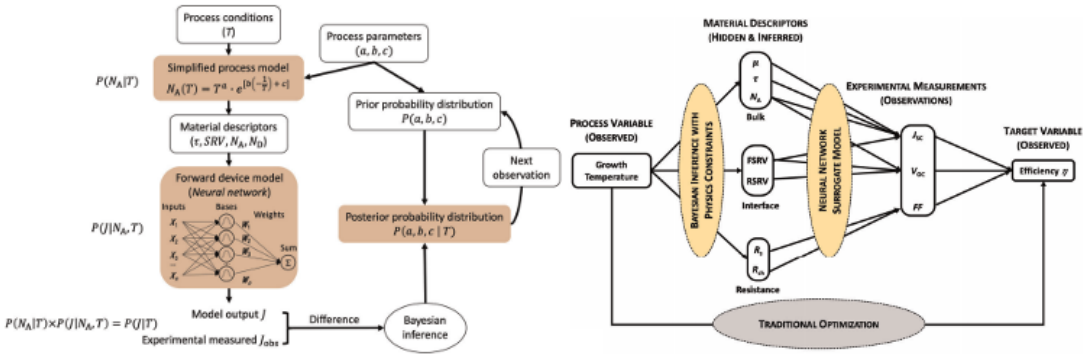


Figure 2.19: The framework that links fabrication parameters to materials properties. [17]

studying their heat of formation and bandgap in which these findings define factors for the discovery of lead-free perovskites [60].

Another interesting study, by Zekun Ren *et al.*, which links temperature (a fabrication parameter) with more intrinsic parameters (such as doping concentrations, carriers lifetime, etc.), uses a Bayesian framework with a neural network using a device-physics notion. This model interprets the possible reasons for device underperformance and identifies optimal process conditions. For this purpose, they used a mixture of simulated and experimental data[17]. The schematic of this study is shown in Figure 2.19.

3

Preliminaries

This chapter presents a brief overview of the fundamentals of the machine learning techniques used in this work, namely Multilayer perceptrons and Variational Autoencoders. These techniques are used in our proposed framework in order to achieve two objectives that we have, which the objectives and framework will be presented in details in the methodology chapter.

3.1 Multilayer perceptron

The multilayer perceptron (MLP) or feedforward neural network or fully connected network is introduced here. The multilayer perceptron is considered one of the simple forms of neural networks and it is a discriminative learning method. It is worth mentioning that in our work we use MLP as a baseline method for comparison with our proposed method.

The goal of the Multilayer perceptron is to approximate some function f. For example, for a classifier $y = f(x)$ is to map an input x to a category y. It defines a mapping $y = f(x; \theta)$ and learns the value of the parameters θ that results in the best function approximation. These models are also called feedforward because information flows through the function being evaluated from x, through the intermediate computations used to define f, and finally to the output y. There are no feedback connections in which outputs of the model are fed back into itself [15].

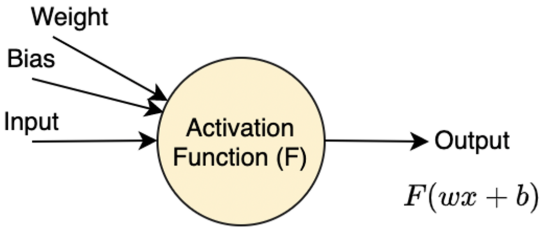


Figure 3.1: A schematic of a neuron.

A single neuron which is the building block of neural networks is defined by its Weight (w), Bias (b) and a non-linear activation function F. It computes the output y by $y = F(wx+b)$. There are several choices for activation function which the most frequently used ones are sigmoid, hyperbolic tangent, rectified linear or ReLU [61], maxout [62] and softmax [63, 64] functions. We select them in the training process mostly based on the nature of the problem that we are trying to solve (e.g., classification or regression)[15]. A schematic of a neuron is shown in Figure 3.1.

The architecture of the MLP is shown in Figure 3.2. It consists of multiple layers each having several neurons, these layers are categorized into the three main layers of input layer, hidden layers and output layer.

The summary of training process is:

1. Initialization step: Assigning random initial weights to begin the process
2. Forward propagation to compute the estimated output
3. Compute the loss function: $L(\hat{Y}, Y)$
4. Obtain derivative of the loss with respect to the weights of the network using chain rule (back propagation): $w = w - \alpha(dL/dw)$ and $b = b - \alpha(dL/db)$
5. Update the weights and bias with gradient descent
6. Go to step 2 and repeat

As mentioned in the background chapter, most of the machine learning technique has several hyper-parameters. Hyper-parameters are fixed features during the training and are selected by us. There are settings that we can use to control the behavior of the learn-

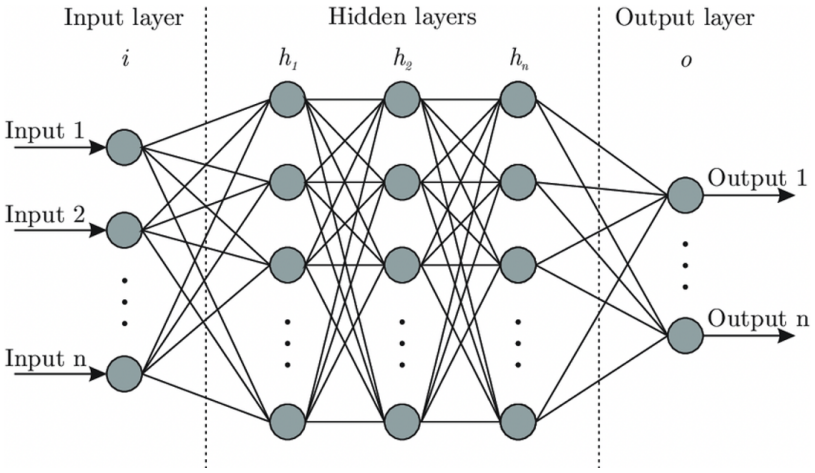


Figure 3.2: A schematic of MLP [18].

ing algorithm[15]. So these parameters are completely non-relevant to the nature of the problem that we are exploring with machine learning, in fact they are directly related to the architecture and principles of the neural network. Some of these hyper-parameters are, number of hidden layers, number of neurons for each hidden layer, learning rate (α), activation function, number of epochs, and batch size, etc. To elaborate more on some of these hyper-parameters, learning rate is the step size while moving toward a minimum of a loss function (see step 4). Number of epochs is the number of iterations through the whole training set.

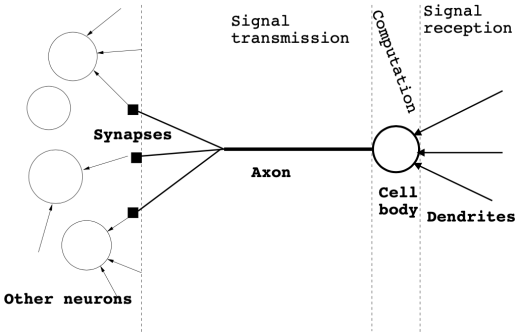


Figure 3.3: A schematic of Synapse, axon and dendrite[19].

It is worth mentioning that the idea of neural networks is heavily inspired by the biological neurons in which we estimate there are around 10^{10} - 10^{11} number of them in a

human brain. Figure 3.3 is a simple schematic of information processing through the brain.

To have more clarified definitions, a few terms are presented here. Firstly, neural networks as defined in the last chapter, are currently the most commonly used method in machine learning due to their recent and significant success. Aside from neural networks, there are other machine learning algorithms such as support vector machines (SVM). Neural networks can have different architectures such as Multilayer perceptron (MLP), Convolutional neural networks (CNN) [45], Recurrent neural networks (RNN) [46] and transformers [47]. Thus, in addition to the fact that MLPs are a specific type of neural network architecture, indeed the simplest one, there are also referred to as fully connected networks. In this thesis, the two terms are used interchangeably. Finally, the architecture of all neural networks used in this thesis is fully connected or MLP. Thus, our proposed framework (CVAE) is a generative approach, and our baseline (MLP) is a discriminative learning approach and both having MLP as their neural network architecture.

3.2 Variational autoencoders

Variational Autoencoders (VAEs), introduced by [28], are a generative learning models. The goal of the generative models is to learn the distribution over $p(x)$ where x is the data and variational autoencoders do so by modeling the joint distribution $p(x, z)$ over unobservable parameters z. The final goal is then to maximize the marginal likelihood of observing $p(x, z)$. The VAE structure and the mathematical theory behind it are introduced in the following.

Standard autoencoders are neural networks trained by reconstructing inputs while projecting them on a lower-dimensional space. This space is called the latent space, as it may contain some information that is not readily observed from the data. Autoencoders consist of two neural networks that are trained simultaneously: an encoder and a decoder. The encoder is a network that maps the inputs to the latent space and reduces the dimensions, while the decoder reconstructs the inputs using the latent space[15].

Generative latent variable models (e.g., VAE) in general learn the joint probability of distribution of the data X and the unobservable latent variables Z, meaning $p_\theta(x, z)$.

The joint distribution is defined as the equation below:

$$p_\theta(x, z) = p_\theta(z)p_\theta(x|z) \tag{3.1}$$

The Equation can be integrated and to obtain:

$$p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz \tag{3.2}$$

Also, the Bayes rule indicates:

$$p_\theta(z|x) = p_\theta(z)p_\theta(x|z) \tag{3.3}$$

The distributions $p(x|z)$ and $p(z|x)$ in Equations 3.2 and 3.3 are intractable, as a solution VAEs introduce a recognition model, a neural network, $q_\phi(z|x)$ as an approximation to the true and intractable posterior $p(z|x)$. Since the encoder network in standard autoencoders' design has similar function with the recognition model, it is referred to as the probabilistic encoder and $p_\theta(x|z)$ is referred to as the probabilistic decoder. It is worth noting that θ and ϕ are the parameters of each network.

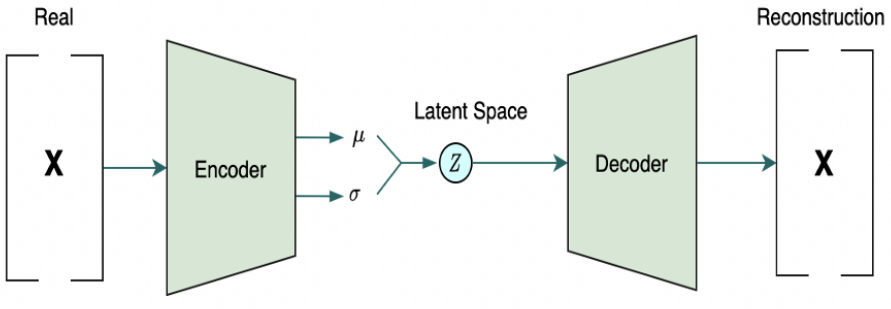


Figure 3.4: A variational autoencoder schematic architecture.

After a series of calculations which are out of the scope of this thesis (refer to [28] for the full derivation), the loss function for training VAEs is obtained as in Equation 3.5:

$$\mathcal{L}(\theta, \phi, \mathbf{x}^{(i)}) = -D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})\big) + \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x}^{(i)})}\big[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\big], \tag{3.4}$$

where the first term $-D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})\big)$ is the Kullback-Leibler divergence between the posterior on the latent space and the prior distribution. In other words, this term enforces the learned latent space to have a Gaussian distribution. The second term is the reconstruction loss defined through both encoder and decoder networks [15, 27] and measures how well the VAE is reconstructing its inputs after it has mapped them on to a lower-dimensional latent space. In the training process we try to minimize this loss function. Indeed, the KL weight (D_{KL}) and the size of the latent space are two of the hyperparameters specific to VAEs.

It is worth mentioning that VAEs can be conditioned on a condition c in order to learn a latent space that is a function of some context [29]. Therefore, the loss function of the conditional VAE (CVAE) for condition c is defined as:

$$\mathcal{L}(\theta, \phi, \mathbf{x}^{(i)}, c) = -D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}, c)||p_\theta(\mathbf{z})|c\big) + \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x}^{(i)},c)}\big[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}, c)\big], \tag{3.5}$$

In this work, we have conditioned our VAE model on the material descriptors of the solar cell in order to give the latent space a more meaningful interpretation.

4

Objectives and Methodology

This chapter presents the description of our machine learning approach for studying a specific architecture of GaAs-based cells. We pursue two main objectives in this work: modeling and sensitivity analysis of solar cell device processes. In the following sections, the details behind the dataset, the objectives and the proposed framework are introduced.

4.1 Dataset

We have selected the simulated GaAs-based solar cell dataset published in [17]. The schematic architecture of the solar cell is shown in Figure 4.1. The absorber or base layer is a 3 μm layer Si-doped GaAs, a 100 nm thick Zn-doped GaAs serves as the emitter, highly Zn-doped 20 nm InGaP is used as the window layer, a 30 nm Si-doped InGaP as the BSF (back surface field) layer, a 100 nm n-doped GaAs back contact layers and C-doped GaAs as the substrate [17].

The dataset contains 20,000 datapoints each presenting five J-V curves (measured under five different illumination intensities) based on the chosen material descriptors. In other words, the material descriptors are mapped into the J-V curves. These material descriptors or parameters are donor doping concentration in the base layer (Si), N_D, acceptor doping concentration in the emitter layer (Zn), N_A, the bulk lifetime in the base, τ, front surface recombination velocity (FSRV) and Rear surface recombination velocity (RSRV). The authors of the paper [17] randomly sampled a set of chosen material descriptors (the five mentioned parameters) from uniform probability distributions. The range of this uniform

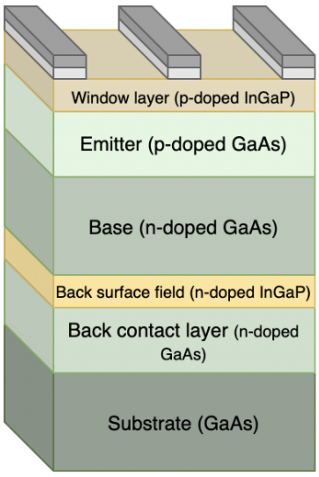


Figure 4.1: The schematic of the solar cell used in the dataset.

distribution is shown in Table 4.1. Then they [17] used scripted PC1D [65, 66] to numerically simulate this dataset which is a one-time implementation process for each material system.

Table 4.1: Parameter (material descriptors) value ranges[17]

Parameters	Range
Zn doping concentration (cm^{-3})	1×10^{16} - 1×10^{20}
Si doping concentration (cm^{-3})	1×10^{16} - 1×10^{20}
Bulk lifetime $[s]$	1×10^{-10}- 1×10^{-6}
Front SRV $[cm/s]$	1×10^{2}- 1×10^{8}
Back SVR $[cm/s]$	1×10^{2}- 1×10^{8}

We further expanded the dataset by calculating and adding two figures of merit in solar cells, the energy conversion efficiency (η) and the Fill factor (FF). They can be easily driven from the J-V curve which is measured under 1-sun illumination (see equations 2.9,2.10). We added these outcome measurements with the incentive of being able to compare the datapoints easier.

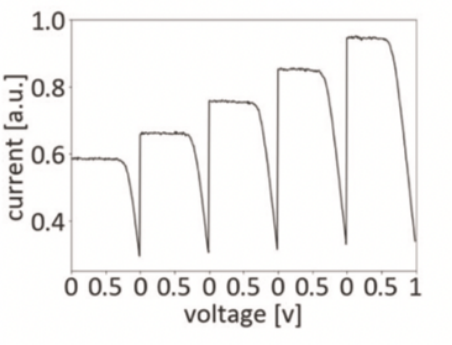


Figure 4.2: A sample of a J-V curve related to a fixed set of material descriptors in our dataset [17].

To wrap this section, there are three groups of data which we worked on, the materials descriptors, J-V curves and the set of efficiency and fill factor which we call the quality vector.

It is worth noting that from the materials point of view as we discussed in the background chapter, the material choices for the designed solar cell architecture (used in our dataset) are good options since it includes GaAs/InGaP layers in which interfaces have a low lattice mismatch.

4.2 Objectives

Machine learning has a great potential for modeling complex models and datasets. As discussed in the background chapter, solar cell performance depends on a series of sequential processes, each being a function of various material properties mainly governed by manufacturing parameters, intrinsic characteristics and solar cell architectures. In addition, from the device-design point of view, as described in Chapter 2, solar cells can consist of a different number of layers; as these processes and layers are entangled in a complicated fashion, predicting the outcome is only feasible through high fidelity and computationally expensive simulators.

4.2 Objectives

We pursue two main objectives in our work:

1. Designing a neural network as a replacement for the computationally expensive simulator with the motivation of having a more efficient and accessible tool.
2. Performing a sensitivity analysis which measures the effectiveness of each of the material descriptors on the efficiency and fill factor of the cell.

Our first objective is the solution to the problem of inaccessibility and computational expense of the common simulators in photovoltaics in the content of our defined task. Although the efforts towards the simulation process are appreciated, it often needs domain expertise, computationally expensive resources and more time even after the simulator is ready and available to the market. On the other hand, in machine learning, after the training process is done, there is no need for necessary domain expertise while being faster and requiring less computational resources. It is worth mentioning that this may differ based on the problem or parameters that one decides to work on, but it is completely valid in our chosen task.

Furthermore, simulators may take into account many assumptions, therefore generally if the dataset used for machine learning is experimental or even a mixture of experimental and simulated data, the final model may outperform the simulator in terms of prediction accuracy.

Moreover, since in our dataset each of the J-V curves consists of 500 points (including five illumination intensities), in fact we are mapping the five datapoints (material descriptors) to 500 datapoints. This is not possible with common simple engineering methods such as regressions.

Our second goal, performing sensitivity analysis of solar cell performance with respect to material descriptors, can give us a more intuitive explanation of the engineering problem. It can also be helpful in identifying the root cause(s) of device underperformance. Sensitivity analysis is defined as how much the uncertainty of the outputs can be apportioned to the uncertainty of the inputs. In our task, this would be how much each of the material descriptors is effective on the quality vector (efficiency and fill factor). Sensitivity analysis becomes especially important in systems where many parameter co-optimizations

are needed. Indeed this happens when some parameters are not aligned together in order to achieve the overall goal which is the same case with solar cells. Additionally from the experimental design perspective, often we have a multitude of parameters (intrinsic properties or fabrication parameters) involved in a task and by using sensitivity analysis we can gain intuition to optimize these parameters.

Although our chosen problem works on five material descriptors (donor doping concentration in the base layer (Si), acceptor doping concentration in the emitter layer (Zn), the bulk lifetime in the base, front surface recombination velocity and rear surface recombination velocity), the framework presented in our work can be readily extended and applied to many more parameters. In fact, we can apply it even to different engineering problems as long as the data exists.

4.3 Framework

The method used in order to deliver the aforementioned objectives is shown in Figure 4.3. The framework includes two parts aligning with the two objectives. We propose a unifying framework for learning a cell's performance function while providing intuitive interpretations based on the sensitivity analysis.

As Figure 4.3 shows, for the first objective, we train the conditional variational autoencoder (CVAE) to generate solar cell J-V curves. Simultaneously, we train a Multilayer

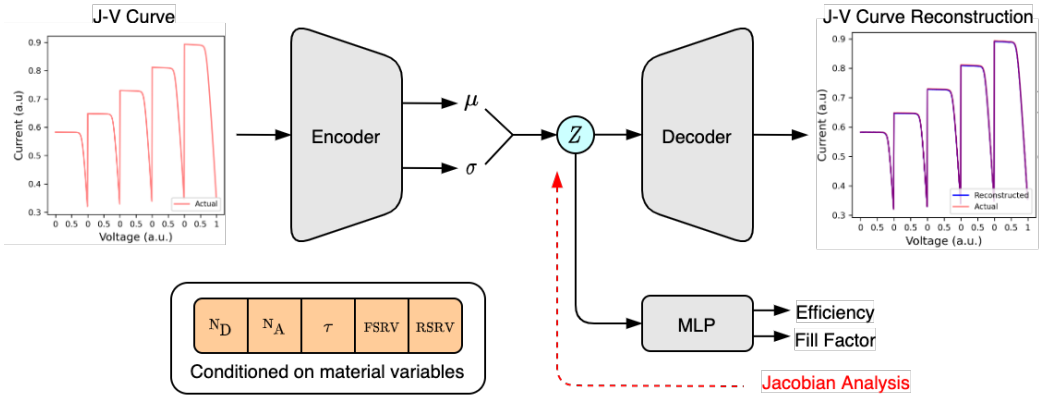


Figure 4.3: The overall proposed framework.

perceptron (MLP) network, on the conditional latent space to predict quality. For achieving the second objective, once the training is done we compute the Jacobian of outputs of the MLP network, which is solar cell quality, with respect to its inputs, which are the latent variables and material descriptors. Then we define the sensitivity indices based on the Jacobian matrices. Each of the mentioned methods and their validation are explained in the following.

Conditional Variational Autoencoder for the First Objective:

For learning to predict solar cell J-V curves and quality vectors, we propose to use the conditional variational autoencoders (CVAE). The fundamentals of variational autoencoders were introduced in the last chapter, very briefly the CVAE has three parts, the encoder, the latent space, and the decoder. The inputs of the encoder are the J-V curves, material descriptors and the quality vectors all together, then the encoder maps the data to a lower-dimensional space or the latent space and then, the decoder maps it back to the original dimensions and reconstructs the J-V curves conditioned on material descriptors. In simple words, CVAE gets J-V curves, material descriptors and quality vector and it reconstructs or predicts J-V curves. Also, to predict the quality vector (efficiency and fill factor) we design a multilayer perceptron that gets the latent space as its inputs and predicts the quality vectors as the output. We will further use this MLP model for our second objective as well.

Besides the CVAE proposed method which is a generative model, we also implement two discriminative models that are two MLP models as our baseline. This enables us to

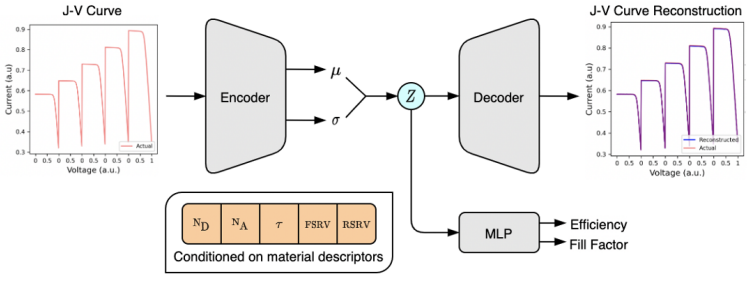


Figure 4.4: Reconstructing J-V curves and quality vectors.

4.3 Framework

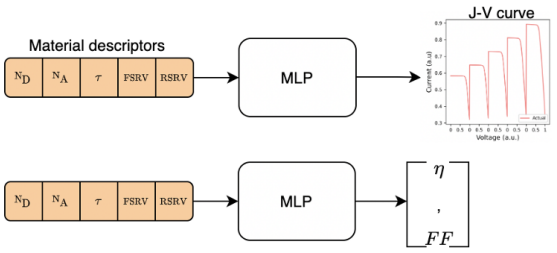


Figure 4.5: Two Multilayer perceptron models as the baseline.

compare our proposed CVAE results with a baseline. As Figure 4.5 indicates, the two MLP models simply get the material descriptors as the input and one predicts the J-V curves as an output and the other predicts the quality vectors (efficiency and fill factor).

Jacobian Matrices as Sensitivity Indices

For performing sensitivity analysis, we demonstrate that Jacobian analysis through the latent space can provide reliable sensitivity indices. Again, the main goal is to explore how much effect each material descriptor has on the quality vector (efficiency and fill factor) by defining the proper sensitivity indices. To do that, we follow three steps:

1. We use the trained MLP network from the previous objective, in which the network's input is the latent variables of the CVAE and its output is the quality vectors.
2. We compute the Jacobian of the quality vectors with respect to the material descriptors. The Jacobian is the partial derivative of the systems outputs with respect to its inputs evaluated at data point $X*$. The Jacobian is defined in Equation 4.1.
3. We define the sensitivity indices as the mean of the square of these Jacobian matrices evaluated over all points.

$$J_{ij}(x^*) = \frac{\partial}{\partial x_j} f_i(x)\bigg|_{x*}. \tag{4.1}$$

In other words, J_{ij}(X*) measures the sensitivity of output i with respect to the input j in the local vicinity of $x*$.

4.3 Framework

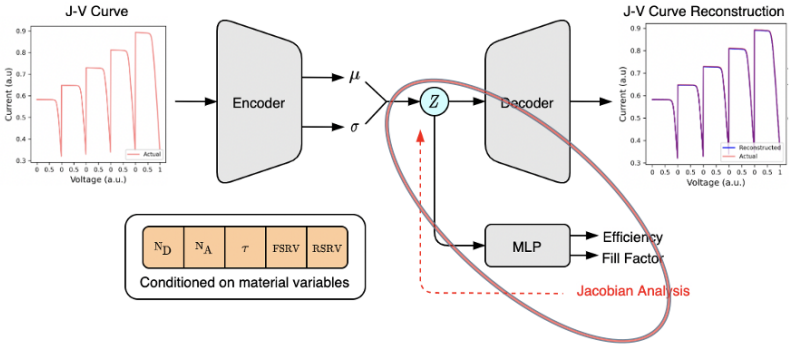


Figure 4.6: Sensitivity analysis is performed on the MLP network that predicts the quality vector based on materials descriptors and the conditional latent space.

This method is more efficient compared to sampling-based sensitivity analysis, since the Jacobian of neural networks is readily available through automatic differentiation. However, the Jacobian matrix only gives local sensitivity analysis, therefore, to obtain a global sensitivity index, we measure the mean of the square of the Jacobian matrices evaluated over all points during the training and testing of the network.

We also validate the sensitivity indices obtained from our approach with a sampling-based method, namely the Sobol method[67] coupled with the Saltelli's sampling scheme[68].

In the next chapter, the results and different implementations of our proposed framework are going to be presented.

5 Results

This chapter introduces the implementation process of our proposed framework for both objectives (J-V and quality prediction as well as sensitivity analysis of efficiency with respect to the chosen material descriptors), their results and some scientific interpretations.

The model implementation includes coding and running the models, tuning the hyper-parameters by doing a grid search, and reporting a model with the best hyper-parameters set (lowest error). The algorithms have been implemented in Python and more specifically using the PyTorch library in which we took advantage of its automatic differentiation [69]. In addition to PyTorch, we used SALib library[70] for our sensitivity analysis baseline.

5.1 First Objective: J-V Curve Prediction

As presented in the last chapter, we compare the performance of our method (the schematic is shown in Figure 4.3) in terms of J-V curve reconstruction and solar cell quality (efficiency and fill factor) prediction with a baseline. The baseline model is a multiplayer perceptron (MLP) that gets material descriptors as input and is trained simply as a discriminative model (the schematic is shown in Figure 4.5).

As a visualization of the training process, we are generating J-V curves which are shown in Figure 5.1. It is worth mentioning that all of the material descriptors are normalized between 0 and 1. Generally, data normalization helps to bring data in the same range and make the learning process more stable.

5.1 First Objective: J-V Curve Prediction

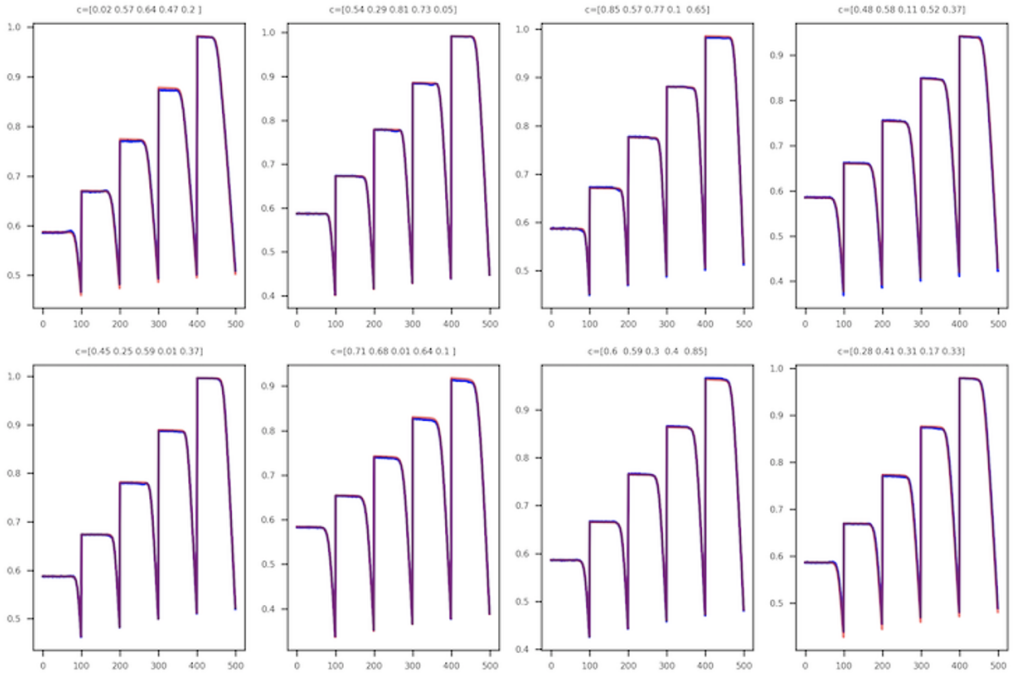


Figure 5.1: Generated J-V curves under five different illumination intensities. Blue and red curves are respectively the predicted and actual J-V curves. Each graph is associated with a set of material descriptors shown at the top of the plots.

5.1.1 Hyperparameter Tuning for CVAE

Here we implement our proposed conditional variational autoencoder (CVAE) network and tune its hyperparameters by grid searching. We choose to work and tune six hyperparameters namely learning rate, KL weight, latent size, batch size, width of hidden layers and number of hidden layers.

Learning Rate

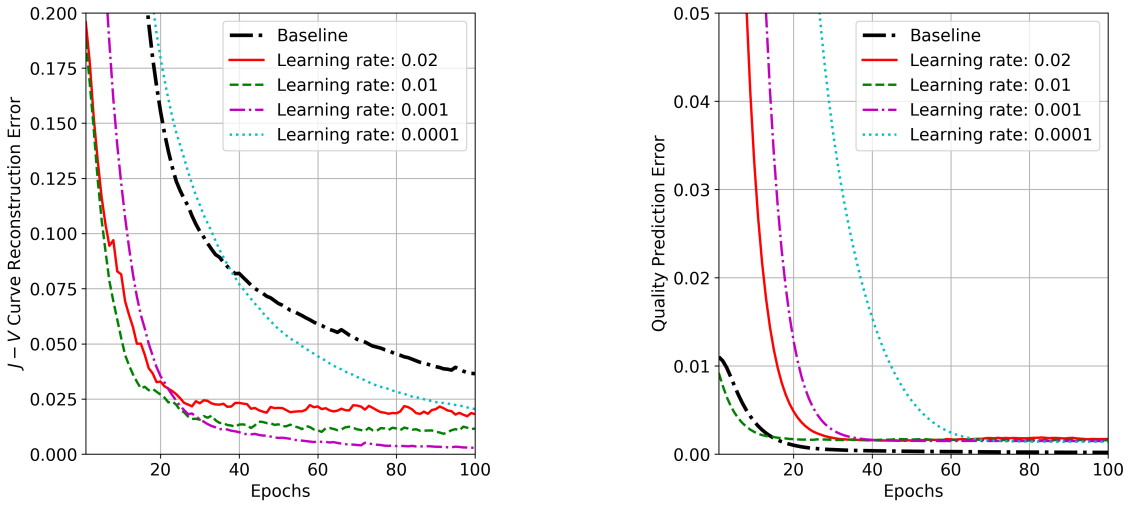
The learning rate is the step size taken while performing the backpropagation algorithm for training the neural network, that is moving towards the minimum of the loss function. We choose four values for the learning rate: 0.0001, 0.001, 0.01 and 0.02.

Figure 5.2 presents the mean squared error of J-V reconstruction and the quality prediction of our proposed CVAE model (under four different learning rate values) and the baseline. As a brief reminder, the baseline is the MLP model. The horizontal axis is the number of epochs which is a hyperparameter that defines as the number of times that the entire dataset is introduced to the system. In our work, we set it at 100.

An important point to notice is that we often report and compare the last epoch's error (100^{th} in our case) in different models because that is when the training process is done. If we can achieve good results, i.e., low errors, we do not continue the training process (i.e., we do not increase the number of epochs). Indeed we are searching for models that can achieve low errors in lower epochs which means they converge faster.

As Figure 5.2 shows in the J-V reconstruction learning rate 0.001 performs better and has the lowest error on the last epoch. learning rate 0.0001 is having slow training process and does not reach a low error at 100^{th} epoch in comparison with others as it is clear in Figure 5.2. We also can conclude that for all values of the learning rate, our CVAE model acts better than the baseline. This validates our approach in J-V reconstruction task in this step.

On the other hand, the baseline slightly outperforms in quality (efficiency and Fill factor) prediction task. This slightly better performance in quality prediction is seen in all of the hyperparameter sets and roots from the fact that there are two MLP models as the



(a) Reconstruction error of *J*-*V* curves.

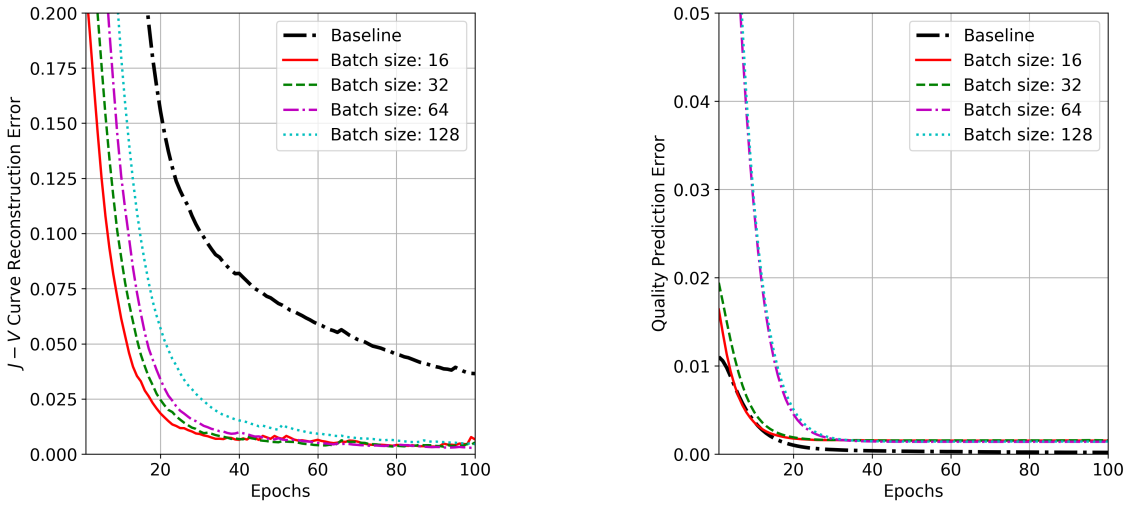
(b) Prediction error of solar cell qualities.

Figure 5.2: Comparison of mean squared error (MSE) of CVAE trained with different learning rates and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.

baseline that independently learn a much simpler task (refer to 4.5 for its schematic). On the contrary, our CVAE model provides a unifying framework for both tasks with competitive performance, thus demonstrating the interpretability of its latent space. It is also worth mentioning that the J-V reconstruction is a much more important and complex task for the model to achieve good performance since that essentially enables it to replace the simulator as a surrogate function and we value J-V reconstruction more.

Batch Size

The batch size is another hyperparameter which is defined as the total number of training examples present in a single batch for a single backpropagation step. Figure 5.3 indicates that CVAE for all values of batch size outperforms the baseline in the J-V reconstruction. For the quality prediction, same as results obtained from exploring the learning rate, the baseline slightly outperforms the CVAE. The most stable learning happens with the batch size of 16, as lower values result in noisy gradients and higher values result in a lower number of backpropagation steps.



(a) Reconstruction error of *J*-*V* curves.

(b) Prediction error of solar cell qualities.

Figure 5.3: Comparison of mean squared error (MSE) of CVAE experiencing different batch sizes and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.

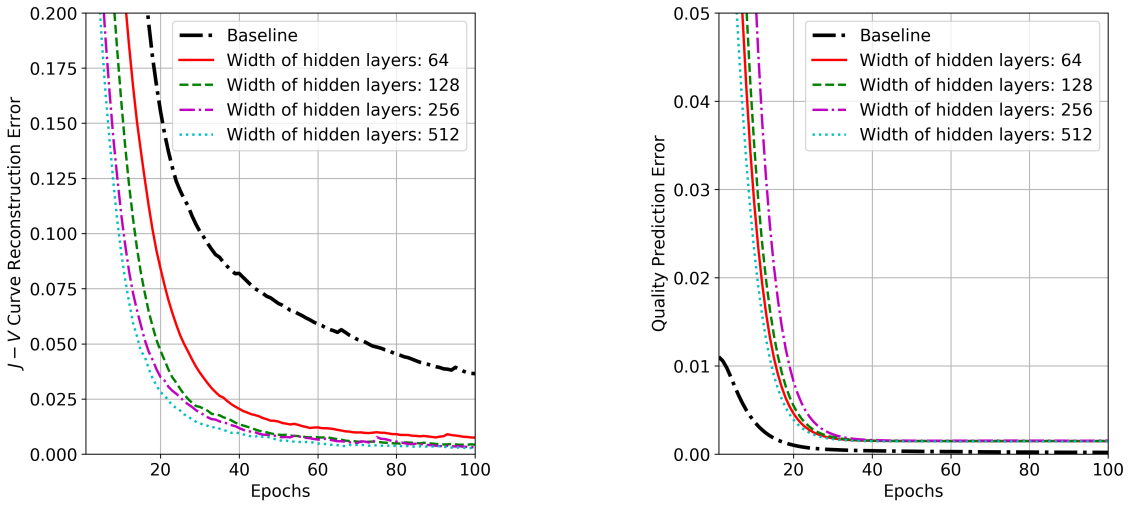
Width of Hidden Layers

The width of hidden layers is a hyperparameter determining the number of neurons in each hidden layer. More neurons would generally result in lower errors but could in principle cause overfitting.

The same trend as other hyperparameters is also shown here where CVAE outperforms the baseline in J-V reconstruction. Figure 5.4 shows width value 512 performs slightly better in comparison with other values of width in the CVAE. On the other hand, same as results from exploring other hyperparameters, in quality prediction task, the baseline has a lower error, thus performing better.

Latent Size

The latent size is a hyperparameter defined as the dimension of the latent space of the variational autoencoders.



(a) Reconstruction error of *J*-*V* curves.

(b) Prediction error of solar cell qualities.

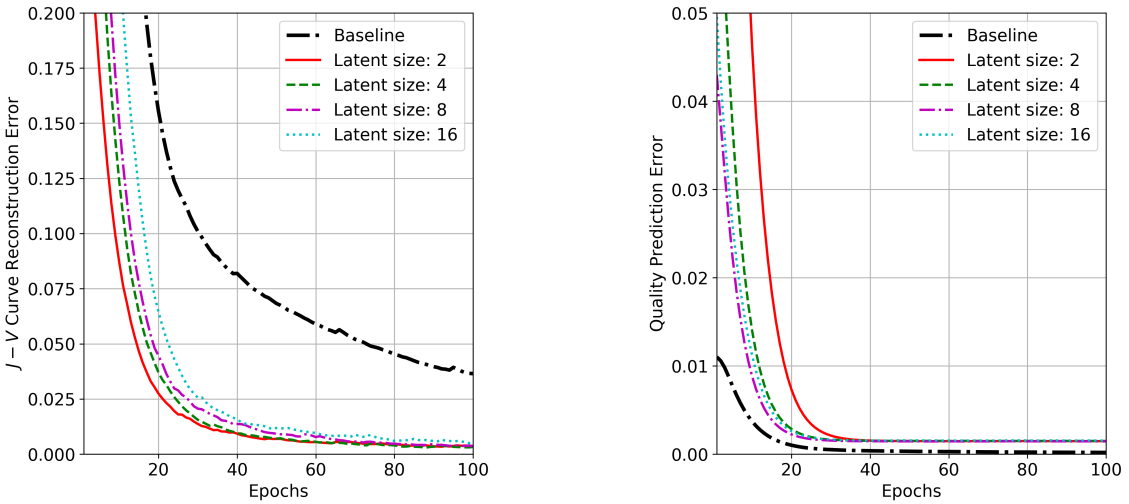
Figure 5.4: Comparison of mean squared error (MSE) of CVAE studying under different width of hidden layers and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.

As Figure 5.5 shows, CVAE outperforms the baseline (MLP) for all values of the latent size and latent size 8 has the best performance and the lowest error at the last epoch (100^{th} epoch). Same as other hyperparameters in quality prediction, the baseline is performing better.

Number of Hidden Layers

The number of hidden layers in the encoder and decoder of the CVAE can have a big impact on the performance, as deep networks usually tend to have better generalization but could possibly overfit the data more easily.

Figure 5.6 demonstrates, three layers of hidden layer shows promising performance and the lowest error at the last epoch (100^{th} epoch). Furthermore, same as other hyperparameters in quality prediction task, the baseline is performing better.



(a) Reconstruction error of *J*-*V* curves.

(b) Prediction error of solar cell qualities.

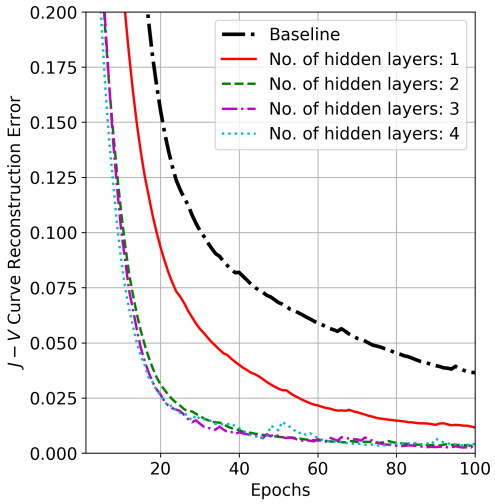
Figure 5.5: Comparison of mean squared error (MSE) of CVAE running under different latent sizes and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.

KL Weight

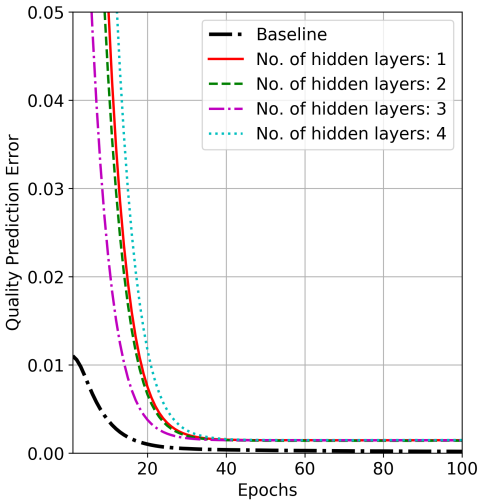
The weight of the Kullback-Leibler (KL) regularizer in the loss function of variotaional autoencoders is another hyperparameter which enforces the learned latent space to have a normal distribution. Increasing the KL weight means that learning focuses more on learning a Gaussian latent space rather than reconstruction of the inputs while decreasing the KL weight could result in a less interpretable latent space.

Figure 5.7 shows, KL value 1 has the best performance and the lowest error at the last epoch (100^{th} epoch). Generally same as results obtained from tuning other hyperparameters, the CVAE outperforms the baseline for all values of KL weight in J-V reconstruction task and in quality prediction, the baseline is performing better.

5.1 First Objective: J-V Curve Prediction

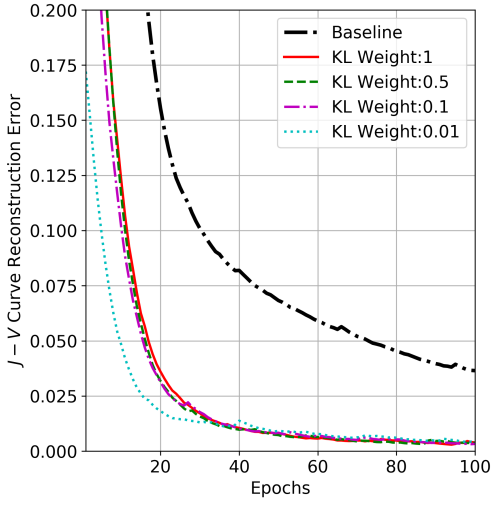


(a) Reconstruction error of *J*-*V* curves.

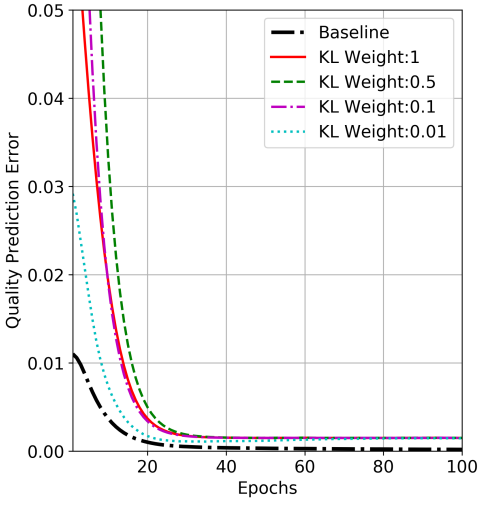


(b) Prediction error of solar cell qualities.

Figure 5.6: Comparison of mean squared error (MSE) of CVAE experiencing different number of hidden layer and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.



(a) Reconstruction error of J-V curves.



(b) Prediction error of solar cell qualities.

Figure 5.7: Comparison of mean squared error (MSE) of CVAE based on different KL weights and the baseline. (a) Reconstruction error of solar cell *J*-*V* curves conditioned on material variables. (b) Prediction error of solar cell qualities (fill factor and efficiency) using the conditional latent space of the CVAE.

5.1 First Objective: J-V Curve Prediction

To conclude from tuning all the hyperparameters, we listed the last epoch's error in a table to decide and finalize the best-performed hyperparameters. As Table 5.1 suggests, the best values for KL weight, learning rate, batch size, width of the hidden layers, number of hidden layers and latent size are respectively 1, 0.001, 16, 512, 3, 8. Therefore, we finalize and train our final model using these values.

	CVAE				MLP
KL Weights	0.01	0.1	0.5	1	
***J-V* Reconstruction Error** $(\times 10^3)$	3.825	3.748	3.746	**2.807**	36.07
Quality Prediction Error $(\times 10^3)$	1.452	1.529	1.478	1.487	**0.178**
Learning rate	0.0001	0.001	0.01	0.02	
***J-V* Reconstruction Error** $(\times 10^3)$	18.963	**2.234**	12.246	16.127	36.07
Quality Prediction Error $(\times 10^3)$	1.401	1.49	1.531	1.788	**0.178**
Batch size	16	32	64	128	
***J-V* Reconstruction Error** $(\times 10^3)$	**2.397**	9.784	2.834	4.013	36.07
Quality Prediction Error $(\times 10^3)$	1.505	1.535	1.383	1.437	**0.178**
Width	64	128	256	512	
***J-V* Reconstruction Error** $(\times 10^3)$	7.042	3.528	3.619	**2.554**	36.07
Quality Prediction Error $(\times 10^3)$	1.459	1.483	1.514	1.425	**0.178**
Numbers of layers	1	2	3	4	
***J-V* Reconstruction Error** $(\times 10^3)$	9.839	2.176	**1.45**	1.785	36.07
Quality Prediction Error $(\times 10^3)$	1.447	1.396	1.451	1.477	**0.178**
Latent size	2	4	8	16	
***J-V* Reconstruction Error** $(\times 10^3)$	3.439	4.035	**2.273**	4.347	36.07
Quality Prediction Error $(\times 10^3)$	1.476	1.484	1.433	1.521	**0.178**

Table 5.1: Comparison of mean squared error (MSE) of CVAE and MLP for *J-V* reconstruction and quality prediction obtained on the final epoch on the validation dataset.

To conclude, the results show that the conditional generative modeling, CVAE, of the process function significantly outperforms the baseline in predicting the J-V curves. This indicates that, in this study, the generative latent variable models, which we proposed to use, have better performance and generalization compared to discriminative models (MLP, the baseline). Since overall we achieve low errors by using machine learning models, we can simply replace the computationally expensive simulator with our faster CVAE network.

5.2 Second objective:Sensitivity analysis

In Chapter 4, we indicated that a second objective we wanted to explore was the degree to which each material descriptor had on the efficiency and Fill factor. We proposed the sensitivity indices which are driven from the Jacobian of the output of the neural network (efficiency and Fill factor) with respect to its inputs (material descriptors). In fact, since we have five inputs and two outputs evaluated on 20000 datapoints, we defined the sensitivity indices as the mean of the square of these Jacobian matrices over all datapoints.

Figure 5.8 presents a visualization of our obtained sensitivity indices in the form of a heat map. It is worth noting that the values are normalized across each row. As 5.8 shows the order of effectiveness of material descriptors on efficiency is donor doping level, front surface recombination velocity, acceptor doping level, carriers bulk lifetime and rear surface recombination velocity. The order of effectiveness of materials descriptors on the Fill factor is slightly different than the efficiency, and the order from the highest to lowest is donor doping level, front surface recombination velocity, bulk lifetime, acceptor doping level, and rear surface recombination velocity.

As a validation of our approach, we implemented the global sensitivity indices computed using the Sobol method coupled with the Saltelli's sampling scheme[67, 68] and compared the obtained values with sensitivity indices obtained from our proposed Jacobian approach. Figure 5.9 demonstrates a visualization of the sobol method driven sensitivity indices, the baseline, in the form of a heat map. The results suggests that the Jacobian-based analysis has successfully computed very similar global sensitivity indices. Furthermore, the order of material descriptors effectiveness aligns with our obtained order as well. This essentially validates our approach.

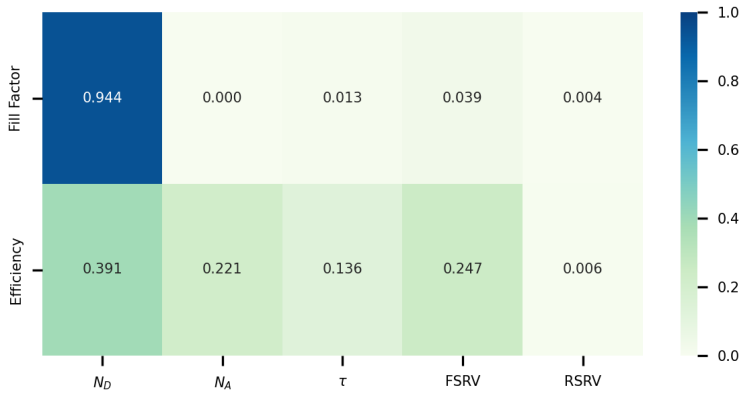


Figure 5.8: A visualization of sensitivity indices obtained by the Jacobian analysis.

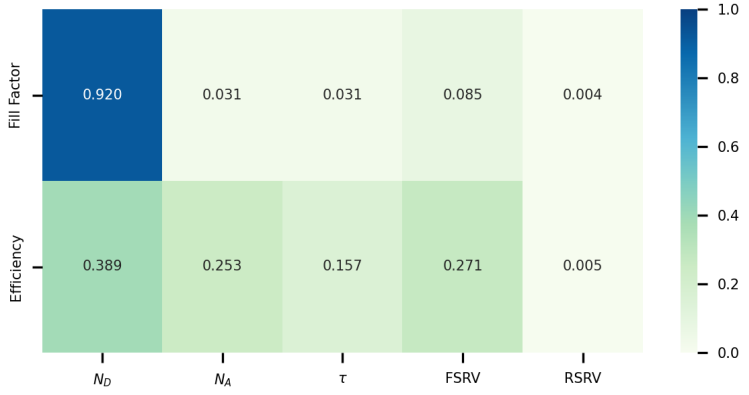


Figure 5.9: A visualization of sensitivity indices obtained by the Sobol method which is the baseline method.

The advantage of our method over the sampling-based sensitivity analysis methods is that it provides a unified framework in which it generates and predicts J-V curves as well as performing the sensitivity analysis. In addition to that, we are benefiting from automatic differentiation which is already built-in in the PyTorch library, this makes our approach computationally efficient.

From the materials engineering point of view, the solar cell efficiency can be affected by J_{sc}, V_{oc} and FF. The donor doping level has an important role in built-in voltage of the p-n junction and thus V_{oc}. It also affects radiative and Auger recombination and therefore bulk lifetime and ultimately J_{sc}. Moreover, the donor doping level effects the series resistance

and the Fill Factor. So, we expected to see a high effect of donor doping level on the efficiency which we observed in our results. The acceptor doping level also has an important role in built-in voltage of the p-n junction and thus V_{oc}. It also affects the emitter's sheet resistance and therefore the FF. The bulk lifetime plays an important role in the diffusion length determination and ultimately J_{sc} and V_{oc}. Furthermore, the surface recombination velocities affect the dark saturated currents and thus J_{sc}. They also effectthe V_{oc}. To conclude, we were aware of these relations and the importance of each material descriptors but obtaining this order of effectiveness which could be different for every solar cell system can give us an insight to interpret our system more precisely or can give us an engineering tool to focus more on most effective parameters. In the next section, we explore some of the other applications of our proposed method.

5.3 Extended Application of Our Proposed Framework

We now show how the proposed framework can give us more intuition about the process by evaluating:

1. The efficiency trend with respect to each material descriptor generated by our MLP model.
2. The partial derivative of efficiency with respect to each material descriptor generated by the Jacobian analysis.

For obtaining the mentioned graphs, since we are working with five material descriptors, these trends can be shown only in six-dimensional space. Therefore, similar to running real experiments in a laboratory, to study each material descriptor trend, we assumed the other material variables to be fixed values. These fixed values are obtained from the highest efficiency datapoint (argmax of efficiency) and its associated material descriptors.

We specifically explored the bulk lifetime and front surface recombination velocity.

Bulk Lifetime

We know that the efficiency can be determined from Equation 5.1:

$$\eta = \frac{J_{sc} V_{oc} FF}{\varphi} \tag{5.1}$$

Therefore, we can interpret the obtained diagrams based on the J_{sc}, V_{oc} and FF.

In Figure 5.10 we can see that as the carrier lifetime increases, the efficiency also increases. As indicated in Chapter 2, the lifetime has an important role in the diffusion length determination ($L_{\text{diff}} = \sqrt{D\tau}$). Diffusion length is the radius that the minority carriers can travel before recombining. Based on Equations 5.2, 5.3 and 5.4 [3], diffusion length directly affects J_{sc} and V_{oc}.

$$V_{oc} = \frac{(K_B T)}{q} \ln\left(\frac{J_{sc}}{J_0} + 1\right) \tag{5.2}$$

$$J_0 \approx \frac{q D n_i^2}{L_{\text{diff}} N} \tag{5.3}$$

$$J_{sc} \sim q G L_{\text{diff}} \tag{5.4}$$

Equations 5.2 and 5.3 show that as the lifetime and diffusion length increase, the J_0 decreases and consequentially V_{oc} increases. In addition to the V_{oc}, Equation 5.4 shows the relation between J_{sc} in which G is the carrier generation rate. By increasing the bulk lifetime and diffusion length, J_{sc} would increase. Thus, according to Equation 5.1 the efficiency would increase as well. This is clearly observed in our obtained results as shown in Figure 5.10.

On the other hand, Figure 5.11 indicates the partial derivative of efficiency with respect to bulk lifetime. All values associated with any bulk lifetime are positive, showing the increasing trend of efficiency with respect to bulk lifetime. This was also clear from Figure 5.10. Figure 5.11 also demonstrates that the partial derivative is decreasing and reaching zero. This shows that as the carrier lifetime becomes longer, its degree of effectiveness on

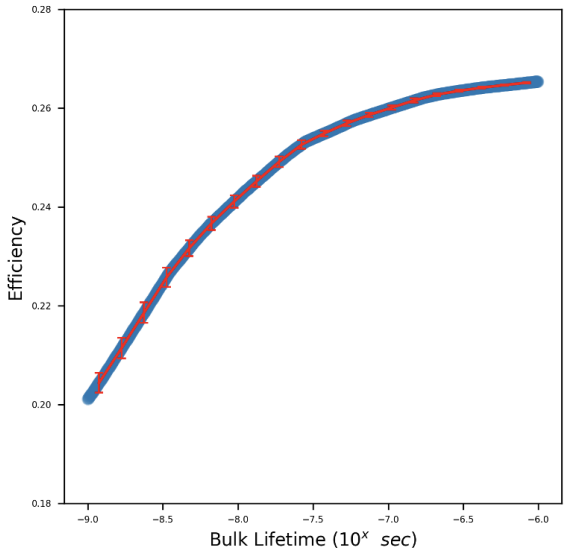


Figure 5.10: Efficiency with respect to bulk lifetime.

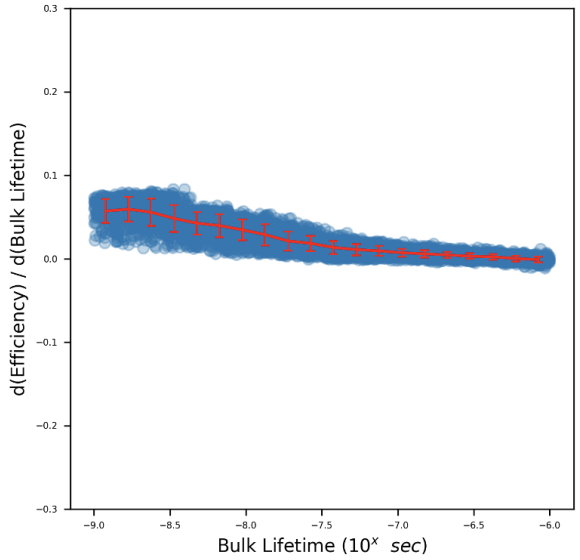


Figure 5.11: Jacobian of efficiency with respect to Bulk lifetime.

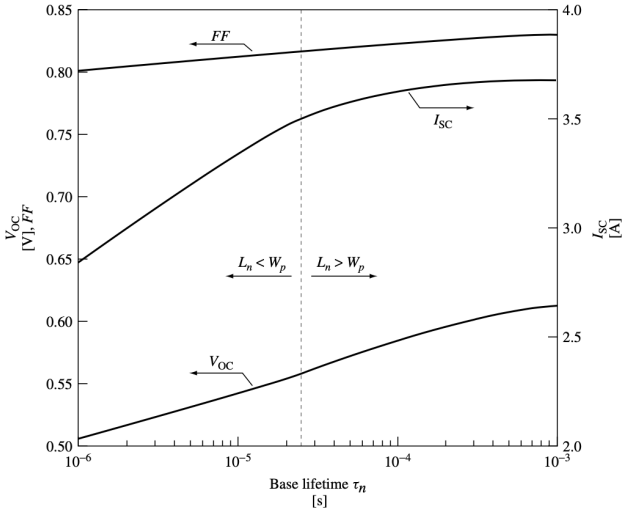


Figure 5.12: Effect of base lifetime on solar cell characteristics. Image taken from [20].

efficiency decreases. This seems scientifically correct, as carrier lifetime becomes longer, the diffusion length also becomes larger. This is generally a positive fact in order to be able to collect carriers, but is not always the optimal solution to design a layer in which carrier diffusion length appears unnecessarily large. The diffusion length should be compared with the thickness of the layers, if it is longer than the thickness, it is sufficiently enough and carriers would pass the thickness and can be able to be collected. We do not need a larger diffusion length or longer lifetime after that point. We can see in Figure 5.11 that the sufficient bulk lifetime, when the graph approaches zero, is approximately around 7.49×10^{-8} seconds for our sample. This notion becomes useful in the device design aspect.

The relation between the bulk lifetime and the solar cell characteristics (I_{sc},V_{oc} and FF) is shown in the Figure 5.12 [20] which further confirms the fact that although by increasing the lifetime and thus the diffusion length, the I_{sc} and V_{oc} are also increasing but the slope of both diagrams is decreasing. This matches our obtained results as well.

Front Surface Recombination Velocity

The front surface in our dataset's cell is referred as the interface between the window layer (p-doped InGaP) and the emitter (p-doped GaAs). Figures 5.13 and 5.14 are the efficiency with respect to the front surface recombination velocity and the partial derivative of the

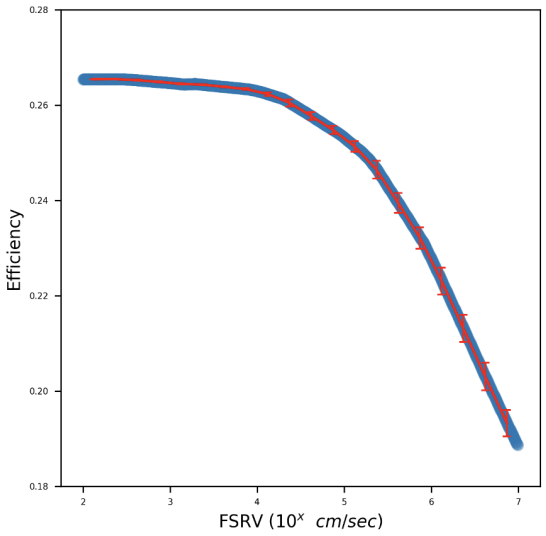


Figure 5.13: Efficiency with respect to Front surface recombination velocity.

efficiency with respect to front surface recombination velocity.

$$I_{01} = qA\frac{n_i^2}{N_A} \times \frac{D_n}{(W_p - X_p)} \times \frac{S_{BSF}}{S_{BSF} + D_n/(W_p - X_p)} \tag{5.5}$$

Equation 5.5 [20] demonstrates the relationship between the saturated current and the surface recombination velocity. The D_n, S_{BSF}, W_p and X_p represent diffusivity, carrier recombination surface velocity, thickness and the carrier movement length. This equation indicates its effectiveness on current density (J_{sc}). On the other hand, usually at the interfaces the density of states and the Dirac-Fermi occupation will change and, consequently, will affect the overall potential, thus the V_{oc}. Based on Equation 5.1, since the efficiency is a function of J_{sc} and V_{oc}, we expected to see a relationship between FSRV (front surface recombination velocity) and efficiency which is also clear in our results in Figure 5.13.

Regarding the trend of graph 5.13, it is expected as we see that the efficiency decreases as FSRV increases. Furthermore, as diagram 5.14 indicates the Jacobian is initially zero and then decreases, which is also clear in graph 5.13 (as the slope is decreasing). This could be useful from an engineering point of view in which we can set a critical maximum acceptable point for FSRV. Indeed the point that after that the Jacobian turns to negative

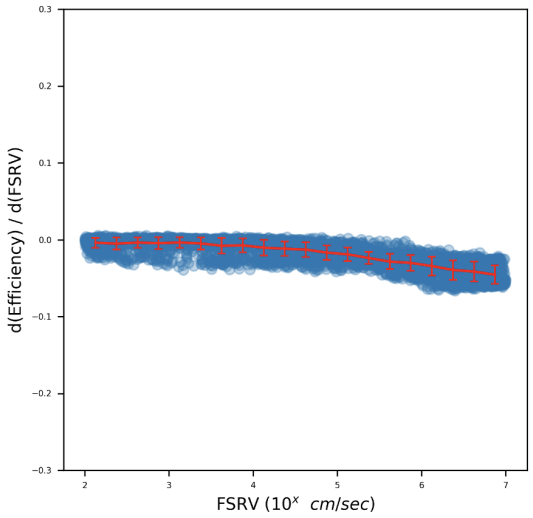


Figure 5.14: Jacobian of efficiency with respect to Front surface recombination velocity.

instead of zero would be a good critical point to set and try to design the interfaces that their surface recombination meets this point. Trying to achieve a lower surface recombination velocity than this point is not an optimum solution because the efficiency is plateaued in that area and would not change drastically. So, it can be wasting resources to achieve lower surface recombination velocities. This optimum point for our cell, obtained from graph

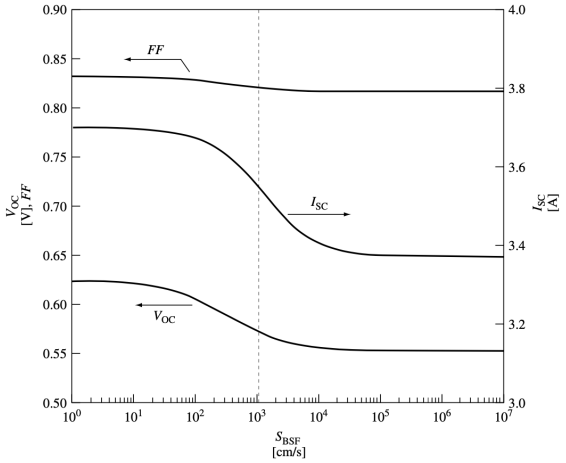


Figure 5.15: Effect of base lifetime on solar cell characteristics [20].

5.14, is 1.7×10^3 cm/sec.

The diagram 5.15[20] confirms our obtained results as well. We can see the decreasing trend as increasing the surface recombination velocity.

It is worth noting that, as indicated in section 5.2 the effectiveness or sensitivity index of the front surface recombination velocity on efficiency is higher than the rear surface. This can be related to the fact that based on Beer-Lambert law (Equation 2.3), photon intensity exponentially decreases while traveling inside a cell and thus the front surface may appear more effective on the overall cell performance.

6

Conclusion

6.1 Summary and Conclusion

This thesis has introduced and explored two main objectives in the context of machine learning applied to a GaAs-based solar cell data. Throughout the thesis, we tried to highlight the importance of the applicability and usefulness of modern machine learning and its various approaches in materials science and more specifically solar cell studies in our case which can be helpful in the sense of generation (prediction) and interpretation of data. This essentially reduces the required resources to conduct research in this field and provides a great toolbox to study more efficiently.

Through Chapters 1-3, we presented the background knowledge in both solar cell field and machine learning methods behind this thesis. Chapter 4 demonstrated our proposed framework which indeed unified our two objectives. We used the CVAE model that generates J-V curves conditioned on our chosen materials descriptors and simultaneously, we trained a multilayer perceptron network, MLP, on the conditional latent space to predict quality (efficiency and Fill factor). Once the training process is done we computed the Jacobian of outputs of the MLP network, which is solar cell quality, with respect to its inputs, which are the latent variables and material descriptors. Then we defined the sensitivity indices based on these jacobian matrices. The main advantage of our proposed model is that makes it possible to explore the two objectives (which are both important tasks in the engineering view) in a unified framework. In addition to that, we took advantage of Python's

(PyTorch library) automatic differentiation to compute the Jacobian which makes it easier to compute sensitivity indices.

In Chapter 5, we presented our detailed results including the hyperparameter tuning and learning process of the CVAE which showed its improved performance compared to discriminative models such as MLP. In addition, the CVAE can definitely be used as a faster and less computationally expensive replacement for the corresponding simulator. Our trained CVAE now allows us to simulate real-world unseen experiments to be conducted.

Furthermore, the comparison of the sensitivity indices with a sampling-based global method demonstrated the validity of our approach. The sensitivity analysis suggested the order of effectiveness on efficiency as donor doping level, acceptor doping level, front surface recombination velocity, bulk lifetime and rear surface recombination velocity. This result was not clear or easy to obtain by doing the theoretical study. Furthermore, even if we could understand this order without machine learning, we should keep in mind that every system (architecture of solar cell) is unique and with having a complex system as solar cells have in which we have huge parameter space that they can be interdependent as well is hard to be able to correctly track the sensitivity analysis without machine learning. In fact, the power of machine learning is in its ability to extract information from the data without specific knowledge-based instructions or hand-engineering the features. The intuition of understanding this order of effectiveness allows us to focus more on parameters that have the higher effectiveness.

In Chapter 5, we brought some scientific interpretations of our proposed method. For instance, as an engineering factor thickness of each layer in solar cell should be co-optimized based on a variety of factors. As indicated in chapter 5, by using the diagram obtained by Jacobian of trained network with respect to the bulk lifetime we can calculate the sufficient lifetime and thus diffusion length and thickness of that layer. As a further effort in the future, we could include absorption coefficient as another material descriptor and explore the Jacobian with respect to both bulk lifetime and absorption coefficient space in order to find the essential (minimum) thickness requirement.

6.2 Future Work

For future research, there are few directions that we can touch up.

Firstly, the ideas and the framework used in this thesis can be readily transferred to other experimental and optimization problems such as LEDs, transistors, etc.

Secondly, our dataset in this work was completely simulated. By choosing experimental or even a mixture of experimental and simulated data, we can explore solar cells in a more realistic way. In that case, sensitivity analysis would be a great tool to find the root-cause of the possible underperformance of the device.

Lastly, Since our approach and sensitivity analysis, in general, are data agnostic, it could be useful for a different set of parameters including fabrication parameters as well. In fact, coupling fabrication parameters (e.g., temperature, pressure) and theoretical parameters (e.g., materials descriptors as carriers lifetime) as parameters set together in order to explore their relationship as well as the performance of the cell would be an interesting area to study especially in newer types of solar cells. It would also highly useful in designing the optimum real-world fabrication procedure.

Publications

M. Molamohammadi, S. Rezaei-Shoshtari, N. Quitoriano, "Jacobian of Generative Models for Sensitivity Analysis of Photovoltaic Device Processes," in *2020 Machine learning for Engineering Workshop, Neural Information Processing Systems (NeurIPS)*, NeurIPS, 2020.

Bibliography

[1] M. Molamohammadi, S. Rezaei-Shoshtari, and N. Quitoriano, "Jacobian of generative models for sensitivity analysis of photovoltaic device processes," *Machine Learning for Engineering Workshop at NeurIPS 2020.*, 2020.

[2] H. Zhang, T. Van Gerven, J. Baeyens, and J. Degrève, "Photovoltaics: reviewing the european feed-in-tariffs and changing pv efficiencies and costs," *The Scientific World Journal*, vol. 2014, 2014.

[3] T. Buonassisi, "2.627 fundamentals of photovoltaics, massachusetts institute of technology: Mit opencourseware." https://ocw.mit.edu.License: CreativeCommonsBY-NC-SA, 2013.

[4] K. A. Mazzio and C. K. Luscombe, "The future of organic photovoltaics," *Chemical Society Reviews*, vol. 44, no. 1, pp. 78–90, 2014.

[5] "Solar spectra versus wavelength." http://rredc.nrel.gov/solar/spectra/am1.5/.

[6] C. H. Henry, "Limiting efficiencies of ideal single and multiple energy gap terrestrial solar cells," *Journal of applied physics*, vol. 51, no. 8, pp. 4494–4500, 1980.

[7] D. Bartesaghi, I. del Carmen Pérez, J. Kniepert, S. Roland, M. Turbiez, D. Neher, and L. J. A. Koster, "Competition between recombination and extraction of free charges determines the fill factor of organic solar cells," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[8] "National renewable energy laboratory efficiency chart 2020." https://www.nrel.gov/pv/cell-efficiency.html.

[9] J. Perlin, "Silicon solar cell turns 50," tech. rep., National Renewable Energy Lab., Golden, CO.(US), 2004.

[10] X. Ziang, L. Shifeng, Q. Laixiang, P. Shuping, W. Wei, Y. Yu, Y. Li, C. Zhijian, W. Shufeng, D. Honglin, *et al.*, "Refractive index and extinction coefficient of ch 3 nh 3 pbi 3 studied by spectroscopic ellipsometry," *Optical Materials Express*, vol. 5, no. 1, pp. 29–43, 2015.

[11] T. Oku, *Solar Cells and Energy Materials*. Walter de Gruyter GmbH & Co KG, 2016.

[12] J. Gong, J. Liang, and K. Sumathy, "Review on dye-sensitized solar cells (dsscs): fundamental concepts and novel materials," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 8, pp. 5848–5860, 2012.

[13] P. Vivo, J. K. Salunke, and A. Priimagi, "Hole-transporting materials for printable perovskite solar cells," *Materials*, vol. 10, no. 9, p. 1087, 2017.

[14] N. Marinova, S. Valero, and J. L. Delgado, "Organic and perovskite solar cells: Working principles, materials and interfaces," *Journal of colloid and interface science*, vol. 488, pp. 373–389, 2017.

[15] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[16] C. Xiao, W. Li, H. Deng, X. Chen, Y. Yang, Q. Xie, and H. Han, "Effective automated pipeline for 3d reconstruction of synapses based on deep learning," *BMC bioinformatics*, vol. 19, no. 1, p. 263, 2018.

[17] Z. Ren, F. Oviedo, M. Thway, S. I. Tian, Y. Wang, H. Xue, J. D. Perea, M. Layurova, T. Heumueller, E. Birgersson, *et al.*, "Embedding physics domain knowledge into a bayesian network enables layer-by-layer process innovation for photovoltaics," *npj Computational Materials*, vol. 6, no. 1, pp. 1–9, 2020.

[18] F. Bre, J. M. Gimenez, and V. D. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings*, vol. 158, pp. 1429–1441, 2018.

[19] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural image statistics: A probabilistic approach to early computational vision.*, vol. 39. Springer Science & Business Media, 2009.

[20] A. Luque and S. Hegedus, *Handbook of photovoltaic science and engineering*. John Wiley & Sons, 2011.

[21] C. Duncan, R. Willson, J. Kendall, R. Harrison, and J. Hickey, "Latest rocket measurements of the solar constant," *Solar Energy*, vol. 28, no. 5, pp. 385–387, 1982.

[22] F. Wang, Z. Mi, S. Su, and H. Zhao, "Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters," *Energies*, vol. 5, no. 5, pp. 1355–1370, 2012.

[23] C. R. Hicks, "Fundamental concepts in the design of experiments," 1964.

[24] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *arXiv preprint arXiv:1206.2944*, 2012.

[25] D. Whitley, "A genetic algorithm tutorial," *Statistics and computing*, vol. 4, no. 2, pp. 65–85, 1994.

[26] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, IEEE, 1995.

[27] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[29] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, pp. 3483–3491, 2015.

[30] A.-E. Becquerel, "Recherches sur les effets de la radiation chimique de la lumiere solaire au moyen des courants electriques," *CR Acad. Sci*, vol. 9, no. 145, p. 1, 1839.

[31] D. M. Chapin, C. Fuller, and G. Pearson, "A new silicon p-n junction photocell for converting solar radiation into electrical power," *Journal of Applied Physics*, vol. 25, no. 5, pp. 676–677, 1954.

[32] J. F. Geisz, R. M. France, K. L. Schulte, M. A. Steiner, A. G. Norman, H. L. Guthrey, M. R. Young, T. Song, and T. Moriarty, "Six-junction iii–v solar cells with 47.1% conversion efficiency under 143 suns concentration," *Nature Energy*, vol. 5, no. 4, pp. 326–335, 2020.

[33] L. A. Green, *Silicon solar cells: advanced principles & practice*. Centre for Photovoltaic Devices and Systems, 1995.

[34] S. R. Wenham, *Applied photovoltaics*. Routledge, 2011.

[35] J. Trube, M. Fischer, G. Erfert, C. Li, P. Ni, M. Woodhouse, P. Li, A. Metz, I. Saha, R. Chen, *et al.*, "International technology roadmap for photovoltaic (itrpv)," *VDMA photovoltaic equipement*, 2018.

[36] D. R. Lide, *CRC handbook of chemistry and physics*, vol. 85. CRC press, 2004.

[37] D. E. Aspnes and A. Studna, "Dielectric functions and optical parameters of si, ge, gap, gaas, gasb, inp, inas, and insb from 1.5 to 6.0 ev," *Physical review B*, vol. 27, no. 2, p. 985, 1983.

[38] T. Baines, T. P. Shalvey, and J. D. Major, "Cdte solar cells," in *A Comprehensive Guide to Solar Energy Systems*, pp. 215–232, Elsevier, 2018.

[39] M. K. Hossain, "Thin film solar cell: Characteristics and characterizations," in *Advanced Materials Research*, vol. 1116, pp. 51–58, Trans Tech Publ, 2015.

[40] L. Yue, B. Yan, M. Attridge, and Z. Wang, "Light absorption in perovskite solar cell: Fundamentals and plasmonic enhancement of infrared band absorption," *Solar Energy*, vol. 124, pp. 143–152, 2016.

[41] Z. Shi and A. H. Jayatissa, "Perovskites-based solar cells: A review of recent progress, materials and processing methods," *Materials*, vol. 11, no. 5, p. 729, 2018.

[42] I. Robel, V. Subramanian, M. Kuno, and P. V. Kamat, "Quantum dot solar cells. harvesting light energy with cdse nanocrystals molecularly linked to mesoscopic tio2 films," *Journal of the American Chemical Society*, vol. 128, no. 7, pp. 2385–2393, 2006.

[43] A. J. Nozik, "Quantum dot solar cells," *Physica E: Low-dimensional Systems and Nanostructures*, vol. 14, no. 1-2, pp. 115–120, 2002.

[44] "A short history of machine learning." https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[46] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[48] F. Häse, L. M. Roch, P. Friederich, and A. Aspuru-Guzik, "Designing and understanding light-harvesting devices with machine learning," *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.

[49] J. Li, K. Lim, H. Yang, Z. Ren, S. Raghavan, P.-Y. Chen, T. Buonassisi, and X. Wang, "Ai applications through the whole life cycle of material discovery," *Matter*, vol. 3, no. 2, pp. 393–432, 2020.

[50] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, pp. 1–36, 2019.

[51] A. D. Becke and E. R. Johnson, "A simple effective potential for exchange," 2006.

[52] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," *The journal of physical chemistry letters*, vol. 9, no. 7, pp. 1668–1673, 2018.

[53] W. Li, R. Jacobs, and D. Morgan, "Predicting the thermodynamic stability of perovskite oxides using machine learning models," *Computational Materials Science*, vol. 150, pp. 454–463, 2018.

[54] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, "A structural approach to relaxation in glassy liquids," *Nature Physics*, vol. 12, no. 5, pp. 469–471, 2016.

[55] K. Min, B. Choi, K. Park, and E. Cho, "Machine learning assisted optimization of electrochemical properties for ni-rich cathode materials," *Scientific reports*, vol. 8, no. 1, pp. 1–7, 2018.

[56] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, "Crystal structure prediction accelerated by bayesian optimization," *Physical Review Materials*, vol. 2, no. 1, p. 013803, 2018.

[57] S. Kiyohara, H. Oda, K. Tsuda, and T. Mizoguchi, "Acceleration of stable interface structure searching using a kriging approach," *Japanese Journal of Applied Physics*, vol. 55, no. 4, p. 045502, 2016.

[58] L. Bassman, P. Rajak, R. K. Kalia, A. Nakano, F. Sha, J. Sun, D. J. Singh, M. Aykol, P. Huck, K. Persson, *et al.*, "Active learning for accelerated design of layered materials," *npj Computational Materials*, vol. 4, no. 1, pp. 1–9, 2018.

[59] F. Li, X. Peng, Z. Wang, Y. Zhou, Y. Wu, M. Jiang, and M. Xu, "Machine learning (ml)-assisted design and fabrication for solar cells," *Energy & Environmental Materials*, vol. 2, no. 4, pp. 280–291, 2019.

[60] J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, and H. Chang, "Identifying pb-free perovskites for solar cells by machine learning," *npj Computational Materials*, vol. 5, no. 1, pp. 1–8, 2019.

[61] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[62] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International conference on machine learning*, pp. 1319–1327, PMLR, 2013.

[63] J. W. Gibbs, *Elementary principles in statistical mechanics: developed with special reference to the rational foundation of thermodynamics*. Dover Publications, 1902.

[64] L. Boltzmann, "Studien uber das gleichgewicht der lebenden kraft," *Wissenschafiliche Abhandlungen*, vol. 1, pp. 49–96, 1868.

[65] D. A. Clugston and P. A. Basore, "Pc1d version 5: 32-bit solar cell modeling on personal computers," in *Conference Record of the Twenty Sixth IEEE Photovoltaic Specialists Conference-1997*, pp. 207–210, IEEE, 1997.

[66] H. Haug, B. R. Olaisen, Ø. Nordseth, and E. S. Marstein, "A graphical user interface for multivariable analysis of silicon solar cells using scripted pc1d simulations," *Energy Procedia*, vol. 38, pp. 72–79, 2013.

[67] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Mathematics and computers in simulation*, vol. 55, no. 1-3, pp. 271–280, 2001.

[68] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer physics communications*, vol. 145, no. 2, pp. 280–297, 2002.

[69] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[70] J. Herman and W. Usher, "Salib: an open-source python library for sensitivity analysis," *Journal of Open Source Software*, vol. 2, no. 9, p. 97, 2017.

Acronyms

CVAE	Conditional Variational Autoencoder
VAE	Variational Autoencoder
MLP	Multilayer Perceptron
GaAs	Gallium arsenide
FF	Fill Factor
CVD	Chemical Vapor Deposition
DSSC	Dye-sensitized solar cell