Predicting Absorption Spectra using Artificial Neural Networks

3rd Year Biophysics Project

Author

B. Bert

Supervisor

Dr. J.A. Nöthling

Conducted in the Biophysics research group led by

Prof. T.P.J. Krüger

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Abstract

Photosynthesis serves as a cornerstone of life on Earth by facilitating energy production in plants and other primary producers. This study centers itself on unraveling the quantum mechanical processes involved in photosynthesis through spectroscopy. Central to the project is the creation of a predictive model for the absorption spectra of light-harvesting complexes. The model, rooted in an artificial neural network framework, utilizes the molecular parameters associated with the CP29 light-harvesting complex to predict the spectra associated with the complex. The Machine Learning methods employed in this study are shown to accurately predict the absorption spectra of simplified light-harvesting complexes and can further predict the qualitative features of experimental absorption spectra for the light-harvesting complex CP29. The project serves as a demonstration of the integration of Machine Learning in Biophysics for rapid and accurate prediction of absorption spectra.

Contents

1

Introduction

1.1 The Chloroplast [1]

Chloroplasts, the cellular powerhouses of primary producers, orchestrate the phenomenon of photosynthesis within plant cells. Encased by a protective double membrane, each chloroplast creates a dynamic environment which is ultimately responsible for energy production within the cell. At the heart of this cellular domain stand the grana—assemblies of stacked membranous sacks resembling piles of stacked coins. A single chloroplast may house a myriad of over a hundred grana. Delving deeper, each granum is composed of a few dozen thylakoids. These intricate layers direct the process of photosynthesis. The grana are surrounded by a colourless fluid known as the stroma, in which the Calvin cycle takes place (see section 1.2).

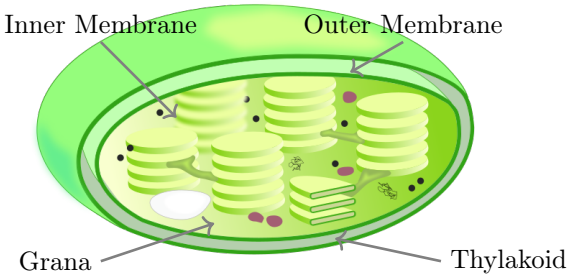

Figure 1: A cartoon of a chloroplast. CC BY-SA 4.0: Karlis Kalviskis (Wikimedia Commons).

1.2 Photosynthesis

Photosynthesis is the mechanism by which plant cells convert energy in the form of electromagnetic radiation from the sun into chemical energy. Photosynthesis takes place in two different phases: the light-dependent and light-independent phases. We will focus on the former, since it is in this phase that the quantum processes giving rise to spectroscopic signals take place.

[1] The main ideas in this section follow from the discussion in [1]

In the light-dependent phase of photosynthesis, photons are absorbed by structures called photosystems. In plants, photosystem I (PSI) and photosystem II (PSII) absorb light continually during the day, but the mechanism of light absorption is more easily described by considering PSII in isolation.

Each of the photosystems is comprised of several tightly bound protein complexes filled with chlorophyll, carotenoid, pheophytin, and quinone pigments. In this project, we will focus on chlorophyll pigments. In the central core of PSII, many light-harvesting complexes (LHCs) encircle a reaction centre, where harvested light energy induces charge separation. A large number of additional LHCs, known as peripheral antenna complexes, are loosely connected to the photosystem. These antenna complexes, in addition to their absorption cross-sections, provide an important adaptation mechanism to the photosynthetic apparatus [2].

Nested within the protein structure of each LHC, the chlorophyll pigments efficiently capture photons. Specifically, these pigments strongly absorb blue and red light from the electromagnetic spectrum. Notably, the pigments' limited absorption of green light gives rise to the characteristic green colour of leaves.

Upon absorption of light energy from the photons, the pigment molecules are excited from their ground states. The energy in the resulting excited states is then directed by the LHCs to the reaction centre within the photosystem, where the electron transport chain is initiated. During the Calvin cycle, which takes place in the stroma, the energy is finally stored as chemical energy [3].

Of particular significance to our investigation is the assembly of pigments constituting the LHC named CP29. This complex plays a pivotal role in the process of dissipating surplus energy that arises during photosynthesis—a function crucial for averting potential harm to the cellular environment. By channeling excess energy away from vulnerable cellular components, CP29 contributes to the delicate equilibrium that governs the cellular processes [4].

1.3 Absorption Spectra

All chlorophyll pigments characteristically absorb certain frequencies of light more readily than other frequencies. By graphing the total intensity of light absorbed by a chlorophyll pigment-protein complex against the frequency[2], an absorption spectrum for the complex is obtained [5].

The structure of the CP29 LHC has proved difficult to investigate, as common techniques that have provided information into the structure of the larger LHC, namely LHCII, have often provided contradictory results [6]. Recent advances in the field have provided insight into the structure of CP29 [7] and the structure of PSII [8]. These advances have shown that CP29 and LHCII both have thirteen to fourteen chlorophyll pigments per site.

[2]In spectroscopy, it is common to use the unit (cm^{-1}) for energy, seeing as the number of wave cycles per unit length is proportional to the photon energy.

2

Theory

2.1 The Hamiltonian

A quantum system's dynamics are governed by Schrödinger's equation

$$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}\Psi \tag{1}$$

where the Hamiltonian $\hat{H}$ encodes the evolution of the wave function ψ and is the observable that corresponds to the total energy of the system.

The Hamiltonian corresponding to an LHC composed of n pigments, can be formally represented as the matrix

$$\hat{H} = \begin{bmatrix} e_1 & J_{12} & \dots & J_{1n} \\ J_{21} & e_2 & \dots & J_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ J_{n1} & J_{n2} & \dots & e_n \end{bmatrix} \tag{2}$$

where e_i is the value of pigment i's transition energy and J_{ij} is the coupling between pigments i and j. The coupling is equal to the strength of the (predominantly dipolar) electrostatic interaction between pigments[3].

By diagonalizing the Hamiltonian as

$$H_D = P^{-1}HP \tag{3}$$

one obtains a new, diagonal, matrix H_D with the exciton energies (see section 2.2) along its diagonal. The matrix P has for its columns a representation of the exciton states while P^{-1} forms the change of basis matrix from the pigment basis to the exciton basis.

The arrangement and orientations of chlorophyll pigments along with the charged amino acids found within an LHC play a crucial role in shaping the immediate electrical properties of the surrounding environment both inside and outside the LHC structure. These properties form the local dielectric environment which determines the range of potential values that the Hamiltonian (and especially the pigment energies) can take on.

[3]Note that the coupling is symmetric (*i.e.*, $J_{ij} = J_{ji}$), such that the matrix H is symmetric.

2.2 Excitons

Pigments are often coupled to one another and thus a linear combination of the pigment states (and in general, not an individual pigment) can be excited upon absorption of a photon. These excited linear combinations of pigment states are called *exciton states* [9]. The exciton states are represented mathematically as eigenstates of the Hamiltonian in eq. (2). Each exciton state can be expressed in the pigment basis as:

$$|\alpha\rangle = \sum_{i}^{n} c_i^{\alpha} |p_i\rangle \tag{4}$$

where $|p_i\rangle$ is a pigment state. To transform between the pigment basis and the exciton basis, one can use the change of basis matrix P^{-1} from eq. (3).

The energy necessary for the formation of an exciton is called the exciton transition energy. There is a strong dependence between the dielectric environment in which the exciton forms and the exciton transition energy.

2.3 Theoretical Calculation of Absorption Spectra

The absorption spectrum, A, for an exciton α can be calculated as:

$$A_\alpha(\omega) = 2|d_\alpha^{eg}|^2 \mathrm{Re}\left[\int_0^\infty dt e^{i(\omega-\epsilon_\alpha)t - g_\alpha(t) - R_\alpha t}\right] \tag{5}$$

where ϵ_α is the exciton transition energy, $g_\alpha(t)$ the so-called lineshape function, R_α the decay rate and d_α^{eg} is the transition dipole moment for the exciton [10].

As its name suggests, the lineshape function serves as a representation of both the amplitude and the overall shape of absorption spectra. This function encapsulates the collective impact of numerous vibrational modes present in the pigment and its protein surroundings. These vibrational modes give rise to fluctuations in exciton energies, leading to the broadening (called homogeneous broadening) observed in the spectra.

The transition dipole moment characterises the dipole moment associated with the transition from the ground state to an excited state, while the likelihood of a transition from the ground state to an excited state is proportional to the square of the transition dipole moment [11].

The rate of decay R_α for exciton α is predominantly determined by the rate at which energy is transferred from exciton α to other exciton states. Although the fluorescence and non-radiative decay rates also cause spectral broadening, their contributions are negligible when compared to the energy transfer rate between excitons. The excitonic energy transfer rates can be quantified through methods such as the Redfield method, the modified Redfield method, the complex time-dependent Redfield method, or the Full Cumulant Expansion method [12] [13].

3

Artificial Neural Networks

3.1 Overview of Artificial Neural Networks

In general, Machine Learning forms a collection of techniques that use statistical methods to teach a computational system to learn the underlying patterns found in a set of data. Artificial neural networks (ANNs) form a particular class of Machine Learning techniques and are characterised by their architecture, which loosely resembles the neural structure found in the human brain [14]. ANNs consist of many interconnected layers of neurons. The first layer, called the input layer, is connected to one or more hidden layers which are, in turn, connected to the output layer [15].

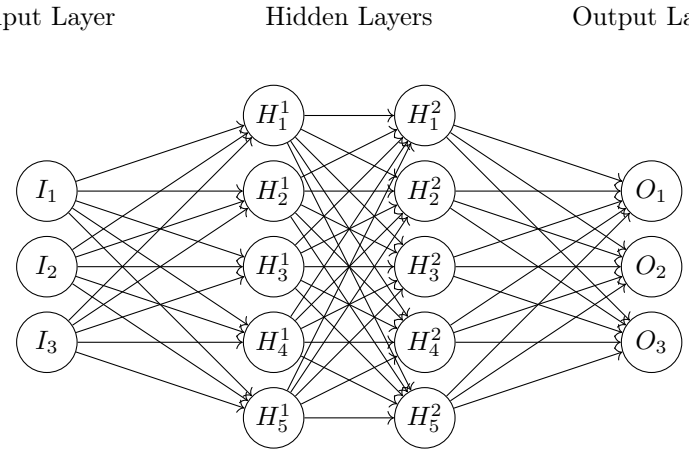

Figure 2: A simple example of a fully-connected neural network. The input vector has three components (I_1 to I_3), each associated with a neuron. The input neurons are connected to the neurons in the first hidden layer. Each hidden-layer neuron (H_i^j) is indexed according to its number in the layer (i) and its depth in the network (j). All the neurons in the last hidden layer are connected to the output layer, which has three components (O_1 to O_3).

In a fully connected Artificial Neural Network (ANN), often referred to as a *dense* ANN, the neurons within each layer establish connections through a non-linear function with all the neurons in the subsequent layer (refer to fig. 2).

In this project, we train our ANNs using the supervised learning paradigm. In this paradigm, an ANN is trained by furnishing it with a dataset containing input values and another dataset containing the corresponding outputs. In the training procedure, a technique known as *backpropagation* is employed to iteratively and incrementally refine the model's parameters. The parameters that are tuned are known as the network's weights and biases. The weights in a network determine how strongly a particular node in some layer will affect another node in the subsequent layer. The biases allow one to characterise the activation sensitivity of a node (this concept is discussed further in section 3.3.1). This refinement aims to systematically minimize the discrepancy (also called the *error* or *loss*) between the outputs generated by the ANN and the accurate outputs in the training dataset, ultimately optimizing the performance of the ANN.

3.2 Feedforward ANNs

A feedforward network is a particular type of ANN that allows the data to flow only in one direction—from the input layer (through the hidden layers) to the output layer. This means that there are no connections that run in reverse in this type of ANN. In contrast, more complex network architectures (like recurrent neural networks [16]) can allow for feedback loops to form.

The Universal Approximation Theorem states that a feedforward ANN with a single hidden layer can approximate any continuous function with arbitrary accuracy, given a sufficient number of hidden-layer neurons [17]. This theorem exemplifies the predictive capabilities of feedforward ANNs—the caveat being that the large number of hidden-layer neurons may not be practically feasible to implement.

3.3 Training Procedure

The performance of ANNs is optimized through the process of training, using the technique of backpropagation discussed above. A typical training regimen is outlined as follows: The network's parameters—encompassing weights and biases—are randomly initialized. The entire training dataset is then presented to the ANN—this forms the first training *epoch*. Many epochs are typically used to train the model. During each epoch, data flows from the input layer, activating neurons as it propagates through the ANN (see section 3.3.1). Typically, the ANN parameters are not updated after every input–output pair provided during training, but only after encountering a larger subdivision of the training data, called a minibatch [18]. For each minibatch, the technique of back-propagation is used to determine how the ANN's parameters are to be updated prior to encountering the next mini-batch (section 3.3.3). Upon completing parameter updates for all minibatches in the training dataset, the following epoch begins. This iterative progression continues until a satisfactory level of predictive accuracy is achieved.

3.3.1 Activation of Neurons

During forward propagation, each neuron (say neuron n) receives input from the previous layer's neurons (every neuron of the previous layer in the case of a fully connected network) and executes a weighted sum of these inputs and adds an additional bias. This value is then fed into a non-linear *activation function*. The value of this activation function on neuron n is the neuron's output.

The activation procedure can be summarized in the following expression:

$$\mathcal{O}_n^l = \mathcal{F}\left(\sum_i W_{ni}^l \mathcal{O}_i^{l-1} + B_n^l\right) \tag{6}$$

where $\mathcal{O}_n^l$ represents the output of the neuron n found in layer l, W_{ni}^l is the weight between n and neuron i in layer $l-1$, which effectively scales the magnitude of $\mathcal{O}_i^{l-1}$ (the output of neuron i in layer $l-1$). The last term in the summation, B_n^l, represents the bias associated with neuron n. The summation result is then fed into the non-linear activation function $\mathcal{F}$.

The non-linearity of the activation function holds pivotal significance, as it equips the artificial neural network with the capacity to discern intricate patterns in the training data. A number of activation functions, such as a hyperbolic tangent function, ReLU, ELU, or SELU are commonly used [19]. In this project, a scaled exponential linear unit (SELU) activation function was utilized. This activation function takes the form

$$\mathcal{F}_{\text{SELU}}(x) = \begin{cases} \lambda\alpha(e^x - 1) & \text{if } x < 0 \\ \lambda x & \text{if } 0 \leq x, \end{cases} \tag{7}$$

where α and λ are specific to the particular problem [20].

3.3.2 Loss Functions

Loss functions play a vital role in the training of ANNs and can affect the trainability of the model [21]. The error of the ANN can be calculated by various loss functions. In this project, we use the Mean Absolute Error (MAE) function which is defined as:

$$\text{MAE} = \frac{1}{n}\sum_i^n |y_{p,i} - y_{a,i}| \tag{8}$$

where $y_{p,i}$ is the predicted output value and $y_{a,i}$ the actual output value from output neuron n.

3.3.3 Backpropagation

The fundamental objective of the backpropagation algorithm is to effectively compute the gradient of the loss function with respect to the various neural network parameters. These learnable parameters are conveniently organized and

stored in a parameter vector often symbolized as W. The primary motivation behind calculating this gradient is to facilitate the iterative refinement of the parameter vector. This iterative refinement process aims to minimize the loss function. To achieve this, the parameter vector is updated, after each mini-batch, a small distance in the direction opposite to the average gradient vector calculated for the minibatch. In this way, the model's parameters are updated multiple times during each epoch, once for each mini-batch.

Once the gradients for all neurons situated within each layer have been computed, these gradients collectively contribute to the adjustment of the parameter vector W. The specifics of this adjustment procedure depend on the particular optimization algorithm chosen for the task. Several optimization algorithms (also called optimizers) are commonly used, such as Adam, Nadam, RMSprop, and standard gradient descent [22]. Based on popularity within the ANN modelling field, we chose the Adam optimizer [23].

3.3.4 Training, Validation and Test Sets

Apart from the training set, two additional datasets serve important roles in the training and testing of ANNs, namely the validation and test datasets. The training, validation, and test sets are usually obtained by calculating a large dataset and, subsequently, partitioning it into three sets. Ratios between the sizes of the training, validation, and test sets are problem-dependent, but a partition ratio of about $80:15:5$, respectively, for these sets is common.

The training set serves as the foundation upon which the ANN acquires knowledge and refines its internal parameters. As discussed in section 3.3.3, this refinement of parameters is achieved by incrementally minimizing the training loss function (the mean absolute error, eq. (8)).

As training progresses, the validation set serves to prevent an undesirable effect known as overfitting. Overfitting occurs when the model becomes excessively tailored to the idiosyncrasies of the training data, which results in poor predictive capabilities of the model when it encounters new, unseen data. After every epoch, the model's performance is tested on the validation set. This dataset therefore acts as an unbiased checkpoint to assess the model's performance on fresh data. This step aids in determining whether the model is acquiring genuine insights from the data or merely memorizing noise.

The ultimate validation for the ANN's efficacy lies in the evaluation phase, during which the model's performance is tested on the, aptly named, test set. The test set, distinct from both the training and validation subsets, presents data that the model has not encountered before. The model's competence in generating accurate predictions based on inputs from this set is the definitive measure of its proficiency. The validation loss is calculated using the validation set. During training, this loss provides an unbiased indication of the model's accuracy and is useful for gauging the extent of training-data overfitting. Although different criteria may be used to identify overfitting, the latter is usually marked by an increase in validation error (even though the training error may still decrease).

3.3.5 Normalization Methods

Weight normalization is a technique used in Machine Learning in which the weights of a neural network are normalized (*i.e.* the magnitudes of the weights range from 0 to 1) [25]. This induces stability in the training process by preventing exploding or vanishing gradients. Such exploding and vanishing gradients can occur in deep ANNs (ANNs with two or more hidden layers). Both problems relate to the gradient of the loss function during backpropagation.

Batch normalization is a method used where the inputs of each layer are normalized [26] using the mean and variance. This is achieved by shifting each input such that the mean is zero and scaling the variance to one. It is important to note that the batch normalization procedure operates differently during training and testing (or validation) of the model. During training, the mean of the currently considered minibatch is calculated for the input to each neuron i in a specific layer. This calculated mean is then subtracted from the inputs to i for each data element in the minibatch. Once the mean has been shifted to zero, the data points are similarly divided by the variance of the mini-batch for each input to neuron i. This scaling process leaves the data (in the considered minibatch) with a variance equal to one. During testing and validation, the average of the means and variances of all the minibatches is used to perform the normalization. The process of batch normalization improves the rate of network optimization by smoothing the loss function landscape. Due to slight noise introduced by using the minibatch (instead of the full training set) statistics for normalization, batch normalization also acts as a so-called regularization method to help prevent overfitting.

4

Methods and Calculations

4.1 Calculation of Absorption Spectra

As a first approximation, a parabolically-shaped lineshape function of the form

$$g_j(t) = \frac{t^2}{\gamma} \tag{9}$$

was used. In order for the smoothness of our modelled spectra to correspond more or less with that of true spectra of light-harvesting complexes, we set $\gamma = 2 \times 10^{-4}$ s^2. For the aims of this project, we also set R_j to 0 and $|d_j^{eg}|^2$ to 1 in eq. (5). These approximations were made in order to improve the speed of calculation, as the calculations are computationally expensive. Although these approximations would not yield correct results for real LHCs, we expect an ANN trained with our data to perform similarly to an ANN trained on more realistic data. The approach therefore serves as a proof of principle that ANNs can indeed be used to calculate absorption spectra.

In plant light-harvesting complexes, the chlorophyll energies range from about 14800 to 15250 cm^{-1} for chlorophyll *a* and 15250 to 16000 cm^{-1} for chlorophyll *b*. To calculate our training dataset, site energies were drawn independently from a uniform distribution in the applicable ranges above.

The integral in the expression for $A_n(\omega)$ in eq. (5) was calculated numerically using trapezoidal integration. We performed all our calculations in Python. A typical calculated absorption spectrum is shown in fig. 3

4.2 The Model Trained

The ANN was coded in Google Colaboratory using the Python-based package Keras. Training of the model was performed using the GPU provided by Google Colaboratory. The model took the form of a feedforward neural network.

The code used to construct and train the neural network can be found in the appendix, and the trained network's specification can be found in table 1. Weight normalization and batch normalization were used in the trained model to improve the speed of the training process and the performance of the model (see section 3.1).

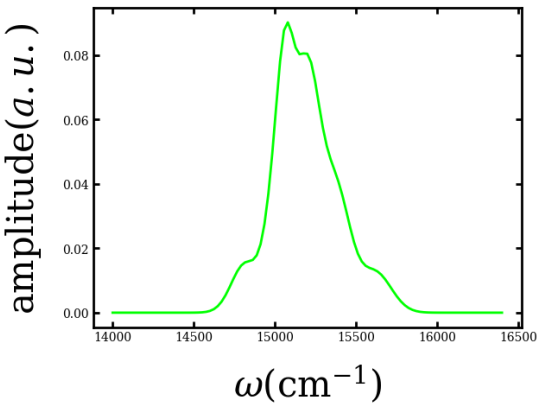

Figure 3: Typical calculated spectrum for a simplified model of a light-harvesting complex with the couplings of CP29. Note that the absorption spectrum is translationally invariant and its exact position on the frequency axis is not important.

4.3 Generating Data for the ANN

In order to generate the data in the training, validation, and test sets, we generated a set of Hamiltonians with the (constant) couplings of the CP29 LHC (obtained from Müh *et al.* [27]) and randomly chosen energies (*i.e.*, any two Hamiltonians had different energies). For each Hamiltonian, a corresponding absorption spectrum was calculated using eq. (5). A training pair consists of a set of thirteen chlorophyll energies, together with the corresponding absorption spectrum. A total of 100 000 energies/spectrum pairs of data were generated using this method and then partitioned as summarized in table 2.

4.4 Performance of Trained Model

The worst prediction made by the model was defined as the value which maximised the expression

$$\sum_i \frac{|\text{Predicted}_i - \text{Calculated}_i|^2}{\max|\text{Calculated}|^2} \tag{10}$$

where the indexing runs over points in the discrete predicted absorption spectrum. This value is defined so as to quantify the least accurate prediction made by the trained model. This allows one to better understand how accurate the overall predictive capabilities are of the ANN. The worst prediction made by the model is plotted against the corresponding ground truth spectrum in figure 5f. Examining this figure, one immediately notices that the predicted absorption is negative over the range 15 700 cm^{-1} to 16 200 cm^{-1}. This points to the general limitation of neural network modelling that predicted results may not

Parameter or Attribute	**Value or Description**
Number of Parameters	2,071,605
Number of Trainable Parameters	1,065,601
Number of Hidden Layers	5
Number of Neurons per Hidden Layer	500
Number of Input Neurons	13
Number of Output Neurons	101
Batch Normalization	before every layer
Weight Normalization	before every layer
Epochs	300
Batch Size	250

Table 1: This table provides the specifications for the trained ANN. Notice the difference between the total number of parameters and the number of trainable parameters. The latter correspond to the weights and biases that are updated during the optimization procedure, whereas the non-trainable parameters, which form part of the total number of parameters, correspond to statistics like the mean and variance found in batch normalization (section 3.3.5). Although these parameters are updated during training, they are not considered as trainable parameters.

	Trained Data	**Training Set**	**Validation Set**	**Test Set**
Size	100 000	90 000	7 500	2 500

Table 2: Partition of training, validation, and test sets for the numerically generated data

be physically allowed, and may therefore not correspond, even qualitatively, to experimental observations.

In table 3, the training, validation, and test losses are presented, along with the time taken to train the model. The losses quantify how accurately the model is predicting for the particular partition of data.

Training Loss	Validation Loss	Test Loss	Time (minutes)
0.0923	0.0760	0.0752	6

Table 3: The final training, validation, and test losses of the model. The time it took to train this specific model, was only six minutes, which suggests that the model certainly trains fast enough for practical purposes

From table 3, one can make comparative assessments on the training, validation, and test losses, for instance—the test loss is significantly lower than the training loss. This is an indication that the model is not overfitting and also elucidates the effect of the regularization due to the batch normalization, as discussed in section 3.3.5.

It is clear from fig. 4 that extending the training time will not significantly improve the performance of the model. The majority of the training (*i.e.* improvement in the predictive performance) takes place during the first 50 epochs. It is also clear from fig. 4 that the rate of improvement diminishes almost exponentially, rendering improvements based solely on longer training times impracticable.

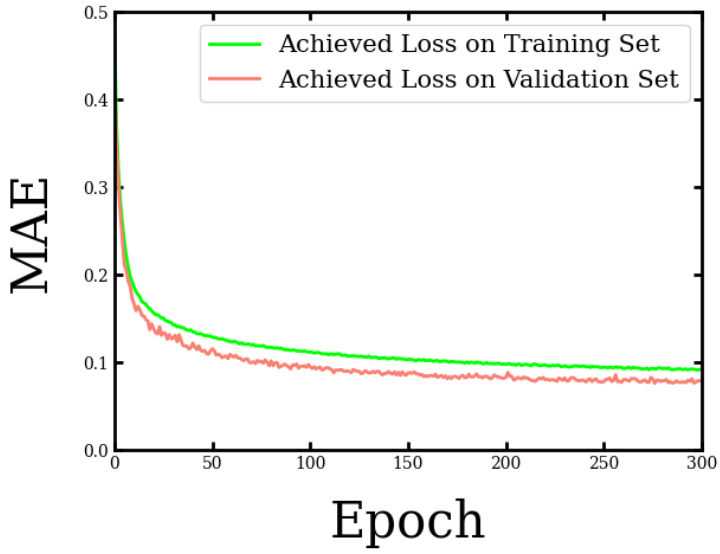

Figure 4: Plot of the mean absolute error (MAE) achieved during the training process. Note that the validation loss decreases along with the training loss during the training process, which suggests that significant overfitting did not occur.

In figure 5, spectra predicted by the ANN from energies in the test set are compared to the corresponding ground truth spectra. The qualitative features

of the ground truth spectra are well matched by the spectra predicted by the ANN. In particular, the number, height, and position of the peaks between the two spectra correspond well. Ultimately, these features of the peaks are what will be of interest in further investigations of absorption spectra.

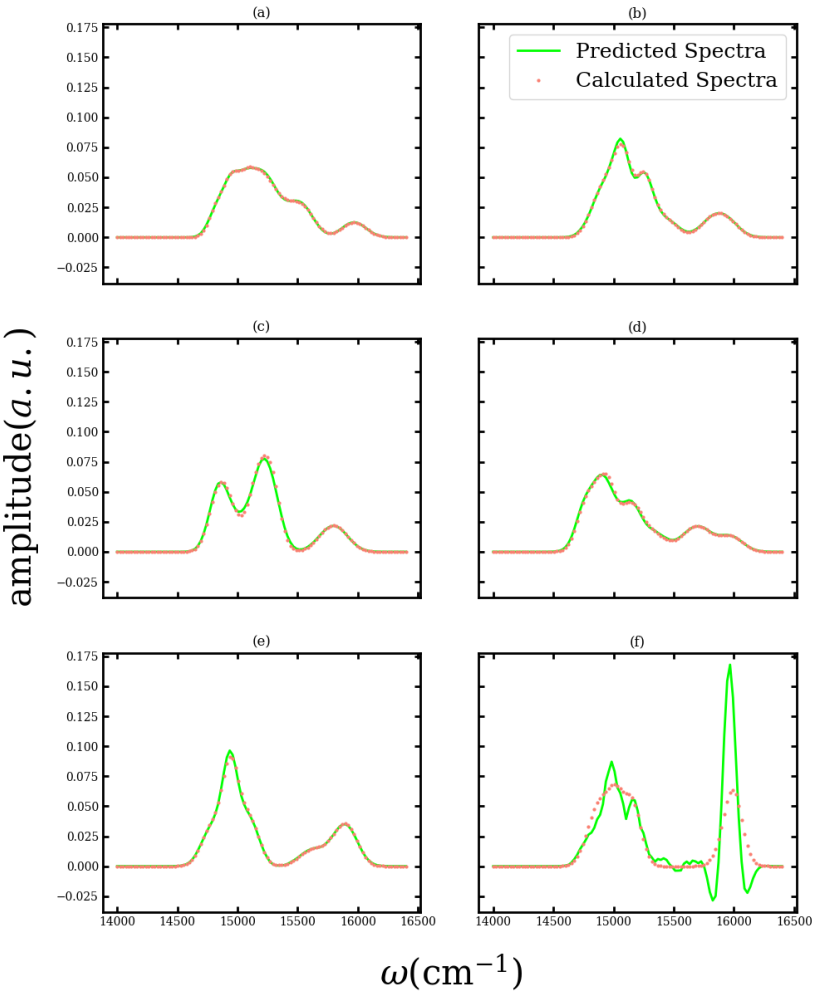

Figure 5: Plots of the spectra predicted by the trained model from test set energies and the corresponding ground truth spectra. The spectra of plots (a) to (e) were randomly selected from the test set data and then compared to the corresponding calculated spectra. Plot (f) is the spectrum which maximized the expression of eq. (10) and thus corresponds to the worst prediction made by the model. The match between the predicted and ground truth spectra can generally be considered as good—in the worst case the predicted spectrum still matched the main features of the calculated qualitatively.

In figure 6 we present the prediction made by the trained ANN for the spectra of the CP29 light-harvesting complex (using the energies of Müh *et al.* [27]). One should note that the model was trained on data that did not take into account any effects due to inhomogeneous broadening. For this reason, the prediction made by the model is expected to differ in terms of its peak widths when compared to a true spectrum. Figure 6 shows that our trained model, despite being trained on data that was restricted to idealized conditions, still manages to predict spectra that qualitatively match the features in a complex LHC like CP29. In particular, we are able to accurately predict the peak positions and relative heights of the chlorophyll a and chlorophyll b pigment clusters. In fig. 6, the experimental spectrum was measured by Pascal *et al.* [28].

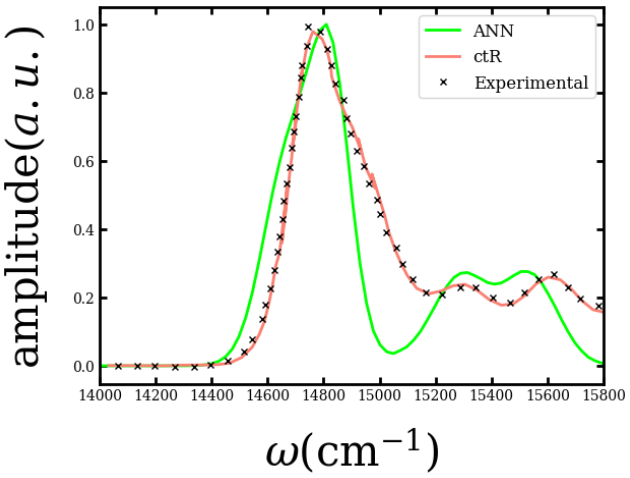

Figure 6: Comparison of absorption spectrum for the CP29 LHC between prediction made by the trained ANN, experimental data of Pascal *et al.*[28] and the analytic calculation using the complex time-dependent Redfield (ctR) method.

4.5 Applications and Improvements

It is clear from fig. 5 that our model generally performs well for predicting simplified absorption spectra. One does notice that the worst prediction (figure 5f) made by the model is a non-physical prediction as the spectrum predicted has a negative absorption amplitude. In future work, this non-physical prediction can be avoided by preventing the model from outputting any negative absorption values.

Furthermore, the model performs well for the prediction of absorption spectra from molecular parameters in the case where the molecular parameters have been simplified to parabolically-shaped lineshape functions, the decay rates are ignored and the transition dipole moments are constant. These simplifications were made as a first step to elucidate the potential capabilities that ANNs have

in predicting absorption spectra. However, even with these limiting simplifications, our model proves to be useful for predicting the main spectral features of the LHC CP29. The same ANN architecture and training regime is therefore anticipated to yield an ANN model that can be used effectively for other LHCs—especially if less limiting simplifications are enforced when calculating the training data. Future iterations of the model should include more realistic training parameters—*i.e.*, more accurate lineshape functions, dipole moments, and exciton decay rates. It should be noted, however, that including these additional parameters will significantly complexify the absorption spectra to be predicted. In particular, the additional parameters will result in broadening of the absorption spectra. These complexities will mean that a larger training set will be required to accurately predict the absorption spectra.

5

Conclusions and Future Work

In conclusion, absorption spectra of light-harvesting complexes (LHCs) found within the chloroplast provide important information about the quantum mechanical processes underlying photosynthesis. The information encoded within these absorption spectra can be computationally expensive to process for intricate LHCs, such as the CP29 LHC.

The ANN model trained in this project was able to make rapid and accurate predictions for simplified LHCs with idealised properties. Remarkably, our rudimentary model was able to make qualitative predictions for the intricate CP29 LHC, that agreed well with the observed experimental absorption spectra.

By employing ANNs, we have successfully demonstrated the potential for predicting absorption spectra of molecular aggregates rapidly and accurately, particularly the intricate LHCs of photosynthetic organisms. Future versions of our model can be used to advance our understanding of photosynthesis by rapidly and accurately predicting spectra for light-harvesting complexes (*e.g.* in an optimisation loop to determine the unknown energies of a complex).

Subsequent iterations of our model could encompass refining its predictive power. This could entail incorporating more accurate lineshape functions, exciton decay rates and transition dipole moments to achieve a greater fidelity in our predictions. Augmenting the training dataset with a larger number of input–output pairs could improve the model's accuracy, as could enhancing its complexity via deeper architectures or a larger number of neurons per layer.

As a part of the discipline of Biophysics, our project serves as a testament to the collaboration of theoretical insight and cutting-edge technology. It specifically underscores the capability of Machine Learning for deciphering complex biological phenomena. Looking ahead, we aim to shed more light on the intricate workings of photosynthesis and open doors to deeper discoveries where biology meets Machine Learning.

References

[1] Mason, K.A., Losos, J.B. and Duncan, T. (2020). Biology. 12th ed. New York, Ny: Mcgraw-Hill Education.

[2] Lokstein, H., Renger, G. and Götze, J.P. (2021). Photosynthetic Light-Harvesting (Antenna) Complexes—Structures and Functions. Molecules, [online] 26(11), p.3378. doi:https://doi.org/10.3390/molecules26113378.

[3] Allen, J.P. and Williams, J.C. (1998). Photosynthetic reaction centers. FEBS Letters, 438(1-2), pp.5–9. doi:https://doi.org/10.1016/s0014-5793(98)01245-9.

[4] Szabó, I., Bergantino, E. and Giacometti, G.M. (2005). Light and oxygenic photosynthesis: energy dissipation as a protection mechanism against photo-oxidation. EMBO reports, 6(7), pp.629–634. doi:https://doi.org/10.1038/sj.embor.7400460.

[5] Garab, G. and van Amerongen, H. (2009). Linear dichroism and circular dichroism in photosynthesis research. Photosynthesis Research, 101(2-3), pp.135–146. doi:https://doi.org/10.1007/s11120-009-9424-4.

[6] Xu, P., Roy, L.M. and Croce, R. (2017). Functional organization of photosystem II antenna complexes: CP29 under the spotlight. Biochimica et Biophysica Acta (BBA) - Bioenergetics, [online] 1858(10), pp.815–822. doi:https://doi.org/10.1016/j.bbabio.2017.07.003.

[7] Pan, X., Li, M., Wan, T., Wang, L., Jia, C., Hou, Z., Zhao, X., Zhang, J. and Chang, W. (2011). Structural insights into energy regulation of light-harvesting complex CP29 from spinach. Nature Structural and Molecular Biology, 18(3), pp.309–315. doi:https://doi.org/10.1038/nsmb.2008.

[8] Wei, X., Su, X., Cao, P., Liu, X., Chang, W., Li, M., Zhang, X. and Liu, Z. (2016). Structure of spinach photosystem II–LHCII supercomplex at 3.2A resolution. Nature, [online] 534(7605), pp.69–74. doi:https://doi.org/10.1038/nature18020.

[9] Knox, R.S. (1963). Solid state physics, suppl. 5: Theory of excitons.

[10] Renger, T. and Marcus, R.A. (2002). On the relation of protein dynamics and exciton relaxation in pigment–protein complexes: An estimation of the spectral density and a theory for the calculation of optical spectra. Journal of Chemical Physics, 116(22), pp.9997–10019. doi:https://doi.org/10.1063/1.1470200.

[11] Chemistry (IUPAC), T.I.U. of P. and A. (n.d.). IUPAC - transition (dipole) moment (T06460). [online] goldbook.iupac.org. Available at: https://goldbook.iupac.org/terms/view/T06460.

[12] J.A. Nöthling. (2022). Modelling of plant light-harvesting spectra. University of Pretoria

[13] J.A. Nöthling, Tomáš Mančal and Tjaart P. J. Krüger (2022). Accuracy of approximate methods for the calculation of absorption-type linear spectra with a complex system–bath coupling. Journal of Chemical Physics, 157(9). doi:https://doi.org/10.1063/5.0100977.

[14] Vanneschi, L. and Castelli, M. (2019). Multilayer Perceptrons. Elsevier eBooks, pp.612–620. doi:https://doi.org/10.1016/b978-0-12-809633-8.20339-7.

[15] Gurney, K. (2018). An Introduction to Neural Networks. CRC Press.

[16] Salem, F.M. (2022). Recurrent Neural Networks. Springer Nature.

[17] Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, [online] 2(5), pp.359–366. doi:https://doi.org/10.1016/0893-6080(89)90020-8.

[18] Aggarwal, C.C. (2023). Neural Networks and Deep Learning. Springer Nature.

[19] Ding, B., Qian, H. and Zhou, J. (2018). Activation functions and their characteristics in deep neural networks. 2018 Chinese Control And Decision Conference (CCDC). [online] doi:https://doi.org/10.1109/ccdc.2018.8407425.

[20] Atto, A.M., Pastor, D. and Mercier, G. (2008). Smooth sigmoid wavelet shrinkage for non-parametric estimation. Proceedings of the ... IEEE International Conference on Acoustics, Speech, and Signal Processing. doi:https://doi.org/10.1109/icassp.2008.4518347.

[21] Dickson, M.C., Anna Sergeevna Bosman and Malan, K.M. (2022). Hybridised Loss Functions for Improved Neural Network Generalisation. Lecture Notes in Computer Science, pp.169–181. doi:https://doi.org/10.1007/978-3-030-93314-2-11.

[22] Doshi, S. (2020). Various Optimization Algorithms For Training Neural Network. [online] Medium. Available at: https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6.

[23] Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Computer Science. [online] doi:https://doi.org/10.48550/arXiv.1412.6980.

[24] Guyon, I. (n.d.). A scaling law for the validation-set training-set size ratio. [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf[Accessed 5 Nov. 2023].

[25] T. Salimans, D.P. Kingma. (2016). Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. arXiv:1602.07868 [cs.LG]

[26] S. Ioffe, C. Szegedy. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs.LG]

[27] Müh, F., Lindorfer, D., Schmidt am Busch, M. and Renger, T. (2014). Towards a structure-based exciton Hamiltonian for the CP29 antenna of photosystem II. Phys. Chem. Chem. Phys., 16(24), pp.11848–11863. doi:https://doi.org/10.1039/c3cp55166k.

[28] Pascal, A., Gradinaru, C., Wacker, U., Peterman, E., Calkoen, F., Irrgang, K.-D., Horton, P., Renger, G., van Grondelle, R., Robert, B. and van Amerongen, H. (1999). Spectroscopic characterization of the spinach Lhcb4 protein (CP29), a minor light-harvesting complex of photosystem II. European Journal of Biochemistry, 262(3), pp.817–823. doi:https://doi.org/10.1046/j.1432-1327.1999.00457.x.