What are Machines learning? An investigation into the knowledge content of Machine learning algorithms from the perspective of theoretical physics.

Eve Burton

Supervisor Ben Hanson

University of Leeds

Laidlaw foundation

Date 31/08/2026

Abstract

Machine learning models are used every day, but we do not have a clear understanding of how they represent the information they learn internally. To investigate this, I created a simple machine learning model to predict the behaviour of a four-resistor series circuit and examined its internal parameters to investigate how the physical relationship was represented. The model successfully learned the behaviour of the resistor circuit and developed an identity-like matrix structure in its learned weights. However, as the number of neurons increased, this internal structure became less closely aligned with the theoretical physical representation. When the physical system was extended by introducing a transistor, the model was unable to accurately learn the expected behaviour. These results show that as the physical system becomes more complex, the model may no longer predict the behaviour accurately, while still showing similar internal patterns to the simpler model. This suggests that some of these patterns may be related to the network architecture or training process rather than being a direct representation of the underlying physics.

Introduction

For my research, I have been investigating how physics systems can be represented internally in neural networks. But what does this mean?

In today's world, there has been a huge increase in the use of artificial intelligence (AI). AI can produce impressive results and is becoming better over time. But how does it determine these responses? How does it learn and make predictions?

This is where my research begins. I wanted to investigate what happens inside a machine learning model when it learns a known physical relationship. Rather than using a large language model such as ChatGPT, Gemini or Claude, I built my own small machine learning model. This allowed me to examine the model directly and investigate whether physical knowledge and its governing laws could be represented within its internal parameters.

Machine learning (ML) is a branch of artificial intelligence in which models learn patterns from data to make predictions. This research focuses on supervised learning, where a neural network is trained using known inputs and corresponding outputs.[i]

Although neural networks can achieve high predictive accuracy, understanding what they have learned internally can be difficult. They are often described as “black boxes” because the relationship between their inputs, internal parameters and outputs is not always easy to interpret. [ii]A model can therefore produce a correct prediction without necessarily showing whether it has learned the underlying relationship in a way that represents a known physical system and its governing laws.

This raised an important question when machine learning is applied to physical systems with known mathematical laws: **does a neural network learn an internal representation that corresponds to the underlying physics?** A physical system provides a useful environment for investigating this because its behaviour can be mathematically defined and compared with the parameters learned by the model. Examples of physical systems include the solar system, the human body, car engines and electrical circuits.

My research is informed by mechanistic interpretability, which aims to understand what neural networks have learned by examining their internal structure and parameters, such as weights and biases. Instead of only looking at whether the model makes accurate predictions, I wanted to investigate whether its internal parameters could reveal the physical relationship it had learned.

After considering different physical systems, I decided to focus on electrical circuits, as this was an area I was confident in. I chose a controlled physical system consisting of four resistors connected in series. The relationship between resistance, current and voltage is well understood through Ohm’s law, meaning that the expected physical relationships could be directly compared with the internal parameters of the neural network.

To investigate this, I built a simple neural network and trained it using data generated from the four-resistor circuit. The initial investigation examined whether the network could learn the relationship between the resistor values and their corresponding voltage drops. I then increased the size of the neural network by increasing the number of neurons from 1 to 20 to investigate how the internal representation changed as the network became larger than the physical system it represented.

I analysed the learned weights and biases to look for patterns that corresponded to the known physical relationships. Different mathematical measures were used to compare the learned

parameters with the expected physical structure and investigate how these patterns changed as the network increased in size.

Finally, I extended the investigation by introducing a simplified transistor model. This allowed me to test whether the same approach could be applied to a more complex physical relationship and whether the internal patterns observed in the resistor model would persist.

The overall aim of this research was therefore to investigate **whether the internal parameters of a neural network trained on a known physical system contains identifiable representations of the underlying physics, and how these representations change as the complexity of the neural network and physical system increases.**

2. Methodology

2.1 Computational Tools

To support my research, I had to learn how to code Python to generate my dataset and to build and train my models and to analyse my results. To implement my neural network, I used TensorFlow and Keras. These applications are used to build and train machine learning models. I also used NumPy for numerical and matrix calculations, and Matplotlib and Pandas for visualisation and organisation of experimental results. Falstad was used to recreate and verify the circuit models.

2.2 Four-Resistor Model

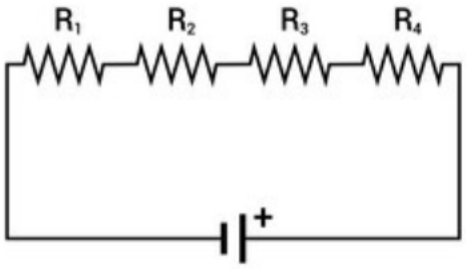

Figure 1:Four Resistor Series Circuit

The initial physical system was a four-resistor series circuit. The current was fixed at 1 A, allowing the voltage across each resistor to be calculated using Ohm's law:

$$V_i = IR_i \quad , \quad (1)$$

Where V_i is the voltage across each resistor, R_i is the associated resistance and the equilibrium current I.

This creates an identity-like relationship between the resistor inputs and corresponding voltage outputs. Random resistor values between 0 and 100 Ω were generated to create the training data.

The neural network consisted of a single dense layer with no activation function, making it a linear model. The number of neurons was varied from 1 to 20. The four inputs to the network were the four resistor values, and the four physical outputs were the corresponding voltage drops across each resistor. When more than four neurons were used, additional non-physical outputs were introduced using zero, random or binary filler values. Twenty independently trained models were produced for each neuron number to assess the consistency of the results.

2.2.1 Non-Physical Output Fillers

When the number of neurons exceeded the four physical outputs, additional outputs were required. These outputs did not represent physical quantities and were therefore treated as filler values. Three filler conditions were investigated: zero, random and binary values.

For the **zero-filler** condition, the additional outputs were fixed at zero. For the **random filler** condition, random values were generated for the additional outputs. For the **binary filler** condition, the additional outputs were assigned binary values of 0 or 1. These conditions were used to investigate whether the choice of non-physical outputs influenced the internal parameters learned by the network.

The same physical resistor outputs were retained in each condition, allowing the effect of the filler values on the learned internal structure to be compared.

2.3 Transistor Model

The transistor model was introduced to test whether the approach used for the resistor system could be applied to a more complex physical relationship. The resistor model was extended by adding a simplified transistor. Four resistor inputs were retained, and an additional gate-voltage input was introduced. A threshold of 33 V was used to define the transistor state:

$$T = \begin{cases} 0, & V_G < 33 \\ 1, & V_G \geq 33 \end{cases} \qquad (2)$$

T is the Transistor state, V_G is the Gate-voltage

When the transistor was OFF, the resistor voltage outputs were set to zero. When it was ON, the resistor voltage drops followed Ohm's law. This created a simple switching behaviour while retaining the original resistor system.

The same single-layer linear architecture was used to investigate whether the model could learn this more complex relationship.

2.4 Neural Network Training

The models were trained using the Adam optimiser with a learning rate of 0.05 and mean squared error (MSE) as the loss function. Training was performed for up to 1000 epochs, with early stopping used when the loss stopped improved.

The learned weights and biases from each model were extracted and saved for analysis. **Weights determine how strongly each input contributes to an output, while biases are additional values that shift the output.** The training loss measures the difference between the model's predictions and the expected outputs and was used to assess how well the model learned the data. These values were then compared across the different models to investigate how their internal parameters changed as the number of neurons increased.

2.5 Physics-Based Parameter Analysis

The learned weight matrices were compared with an ideal identity matrix representing the expected physical relationship. The Frobenius norm was used to calculate the distance between the learned and theoretical matrices' Frobenius norm:

$$D = \| \mathrm{W} - \mathrm{I} \| \qquad (3)$$

where (W) is the learned weight matrix and (I) is the theoretical identity matrix. A smaller value therefore indicates greater agreement with the expected physical structure.

The analysis was performed with and without biases, and for different filler conditions.

2.6 Covariance, Correlation and Diagonalness

Covariance matrices were calculated across the independently trained models to determine how parameters varied together. Correlation matrices were then calculated to examine these relationships independently of parameter scale.

A diagonalness measure was introduced to quantify whether the covariance and correlation matrices were concentrated along their diagonals:

$$D = \frac{1}{N^2}\left(\sum_i C_{ii}^2 - \sum_{i \neq j} C_{ij}^2\right) \qquad (4)$$

Where D is the diagonalness measure, C represents either the covariance or correlation matrix, C_{ii} represents the diagonal elements, C_{ij} represents the off-diagonal elements, and N is the number of parameters in the matrix.

This measure was used to investigate how the internal parameter structure changed as the number of neurons increased.

3. Results

3.1 Neural Network Performance

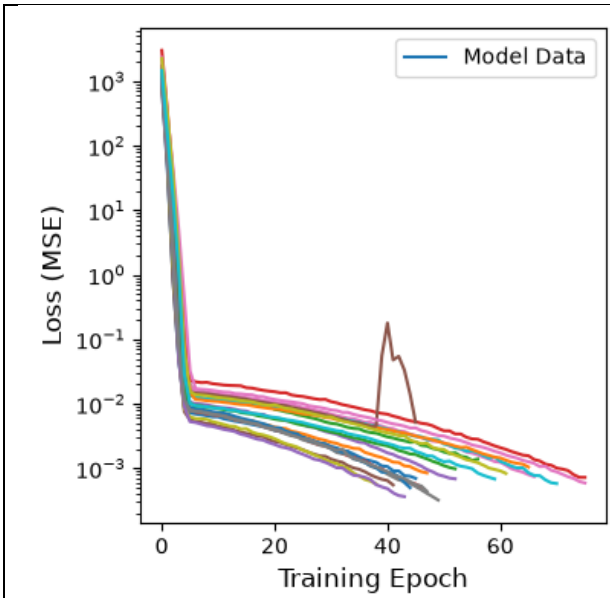

Figure 1: Mean Squared Error (MSE) loss function of 4 neurons with zero fillers

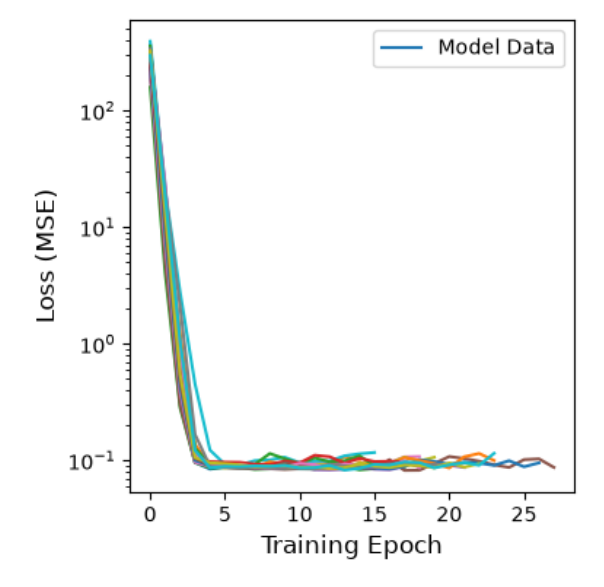

Figure 2: Mean Squared Error (MSE) loss function of 20 neurons with zero fillers

The four-resistor model learned the relationship between resistance and voltage. The training loss, shown by the MSE in Figures 2 and 3, decreased substantially for both the four- and twenty-neuron models. The four-neuron model reached a final MSE of approximately (10^{-3}), while the twenty-neuron model reached approximately (10^{-1}).

Both models were able to reproduce the physical outputs, although the four-neuron model achieved a lower final loss. These results demonstrate that the network was able to learn the resistor-voltage relationship. However, predictive performance alone does not show how this relationship was represented internally. The learned parameters were therefore examined in the following analyses

3.2 Physics-Based Weight Analysis

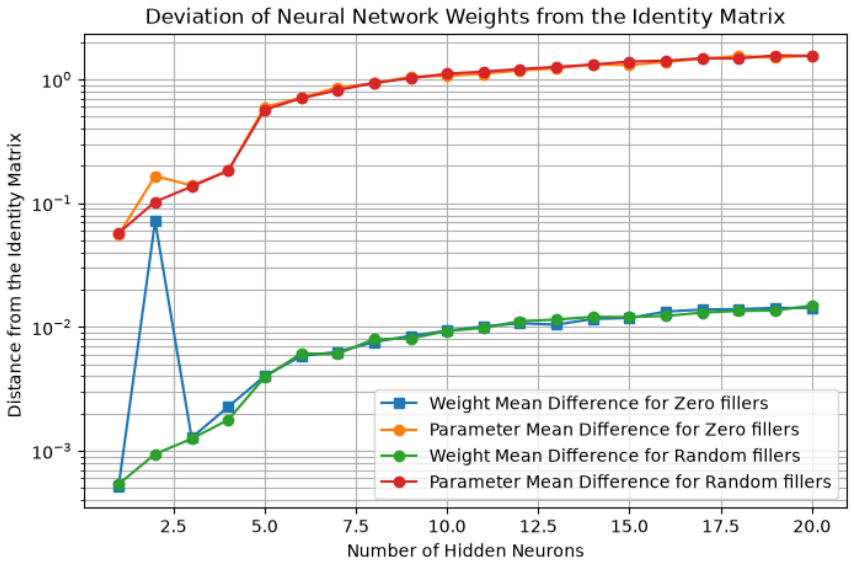

Figure 3:Deviation from theoretical model matrix as defined by Eq. 3

In Figure 4, the learned weights were compared with an identity matrix because the fixed current of 1 A reduces Ohm's law to (V_i=R_i). The learned weights showed an identity-like structure, with the strongest contributions occurring between each resistor input and its corresponding voltage output.

However, the distance from the theoretical identity matrix generally increased as the number of neurons increased. A smaller distance indicates that the learned weights are closer to the expected physical structure, while a larger distance indicates greater deviation from it. Therefore, larger networks generally showed greater deviation from the simple theoretical representation.

The zero- and random-filler conditions showed broadly similar behaviour, with the same overall increase in distance as the number of neurons increased.

3.3 Covariance and Correlation

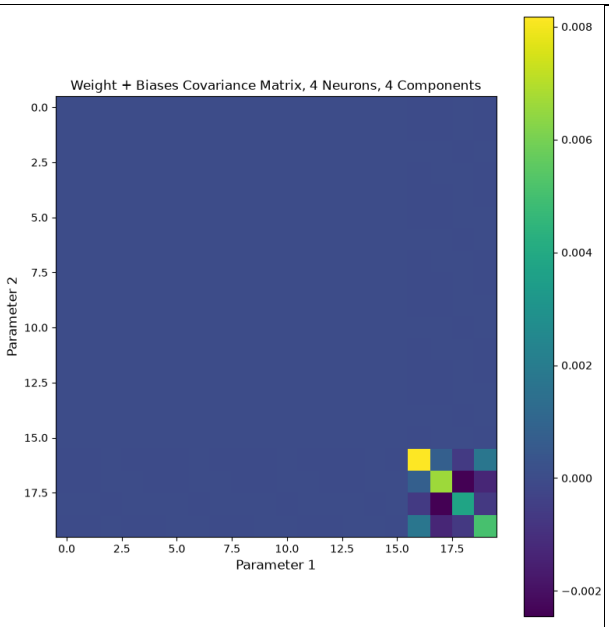

Figure 5 :Weight and bias (parameters) covariance matrix for the 4-neuron resistor model under the zero-filler condition.

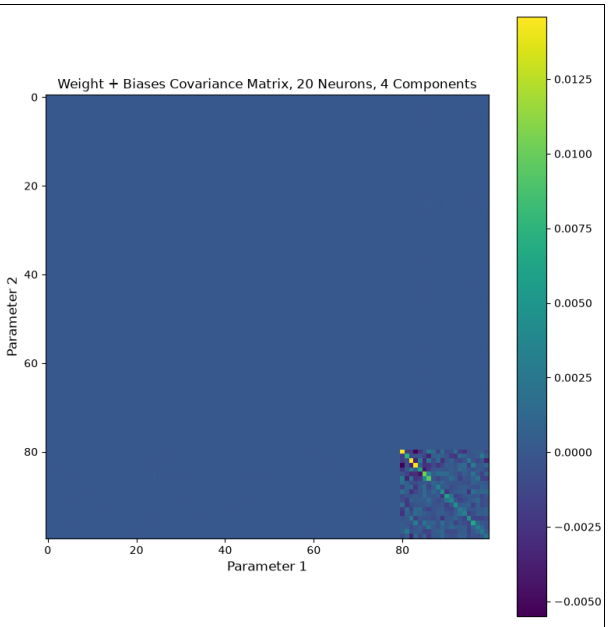

Figure 6:Weight and bias (parameters) covariance matrix for the 20-neuron resistor model under the zero-filler condition

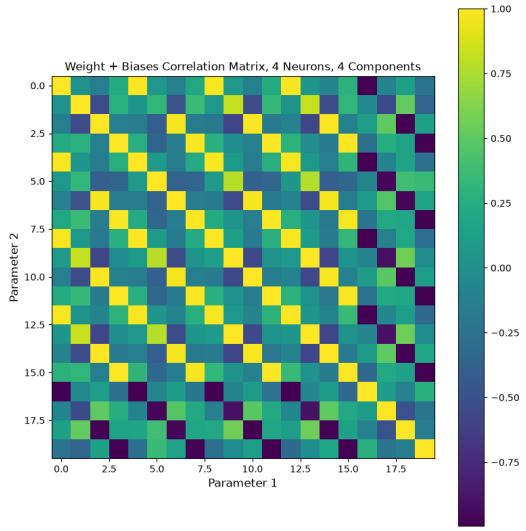

Figure 7: Weight and bias correlation matrix for the 4-neuron resistor model under the zero-filler condition.

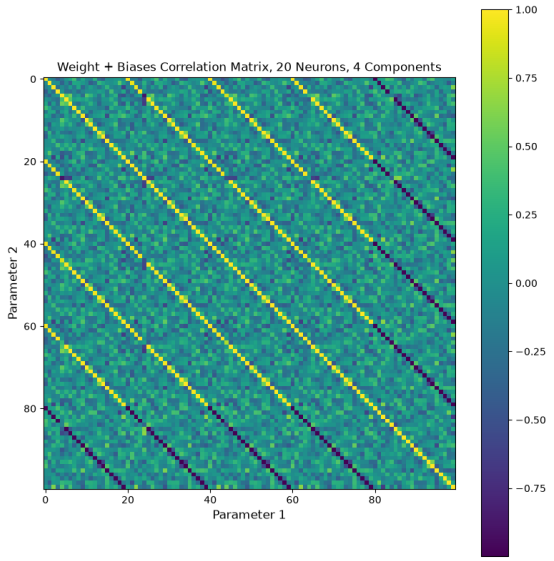

Figure 8: Weight and bias correlation matrix for the 20-neuron resistor model under the zero-filler condition

Figures 5 and 6 show the covariance matrices for the four- and twenty-neuron models. The matrices showed relatively small covariance between most of the weight parameters, while stronger covariance was concentrated around the bias parameters. Covariance describes how two parameters vary together across the independently trained models. Therefore, the low covariance between the weights indicates that they generally varied independently between models, while the stronger covariance between the biases indicates that the biases tended to change together.

Figures 7 and 8 show the corresponding correlation matrices, providing a scale-independent view of these relationships. The weight parameters showed a strong diagonal structure, while the bias parameters showed strong negative correlations with one another. These patterns were broadly consistent across the different network sizes investigated.

3.4 Diagonalness

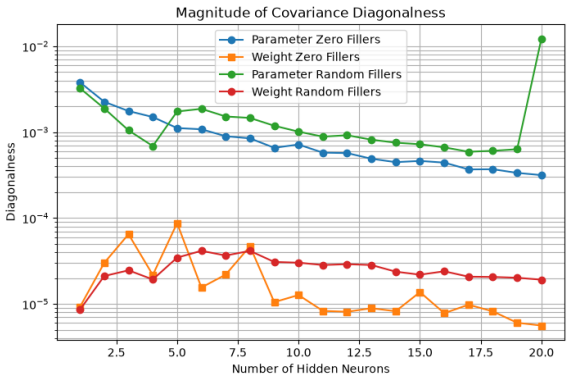

Figure 9: Magnitude of Covariance Diagonalness for the Resistor model

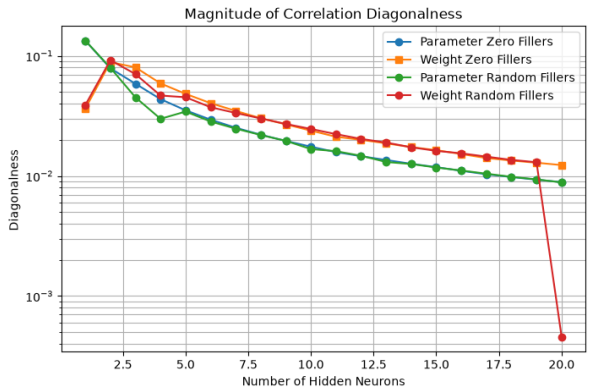

Figure 10: Magnitude of Correlation Diagonalness for the Resistor model

Figures 9 and 10 show the diagonalness of the covariance and correlation matrices for the resistor model. Both measures generally decreased as the number of neurons increased.

The same overall trend was observed under the zero- and random-filler conditions. This shows that the change in diagonalness occurred across different filler conditions rather than being limited to a single filler type.

3.5 Transistor Model

The transistor model was used to test whether the approach could be extended to a more complex physical relationship. Unlike the resistor model, the single linear network was unable to accurately reproduce the transistor behaviour. As shown in Figures 11 and 12, the minimum MSE remained at approximately (10^2), substantially higher than the near-zero error achieved by the resistor model.

Figures 13 and 14 show the covariance and correlation diagonalness for the transistor model. Both measures generally decreased as the number of neurons increased, showing a similar trend to the resistor model. Similar trends were also observed across the different filler conditions, as shown in Figures 15 and 16.

Figures 17 and 18 show the correlation matrices for the four- and twenty-neuron transistor models. Despite the poor predictive performance of the transistor model, similar patterns in the relationships between the learned parameters were observed to those seen in the resistor model.

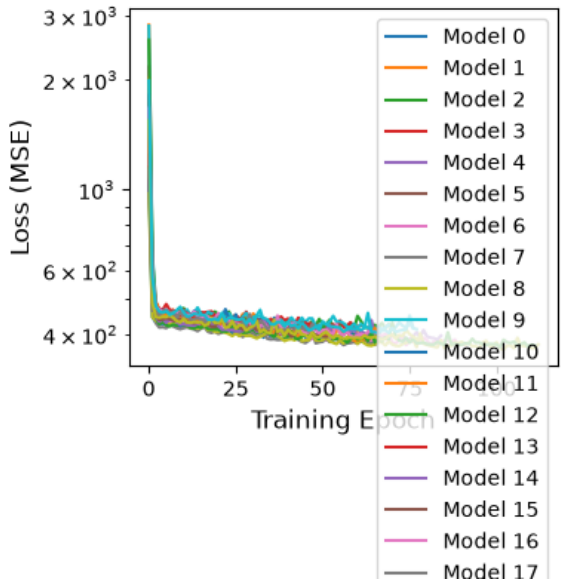

Figure 4: Mean Squared Error (MSE) loss function of 4 neurons with zero fillers for Transistor Model

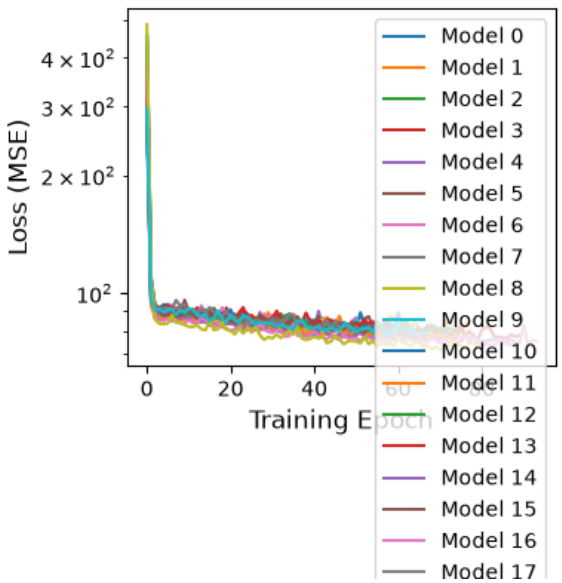

Figure 5: Mean Squared Error (MSE) loss function of 20 neurons with zero fillers for Transistor Model

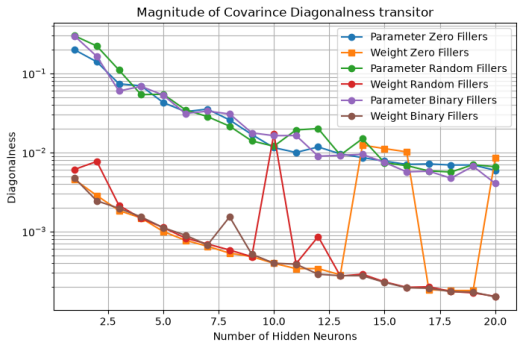

Figure 13 :Magnitude of Covariance Diagonalness for the Transistor model

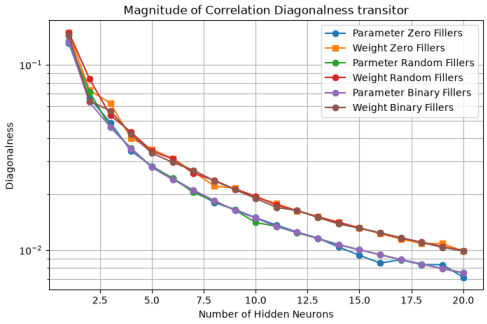

Figure 14:Magnitude of Correlation Diagonalness for the Transistor model

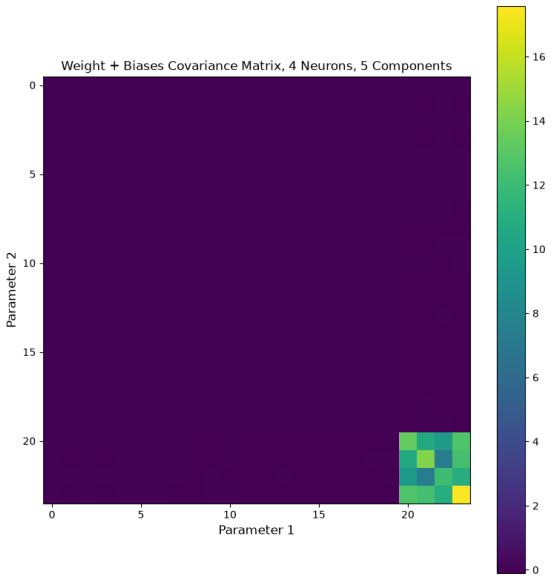

Figure 15: Magnitude of Covariance Diagonalness for the Transistor model for 4 neurons with random fillers

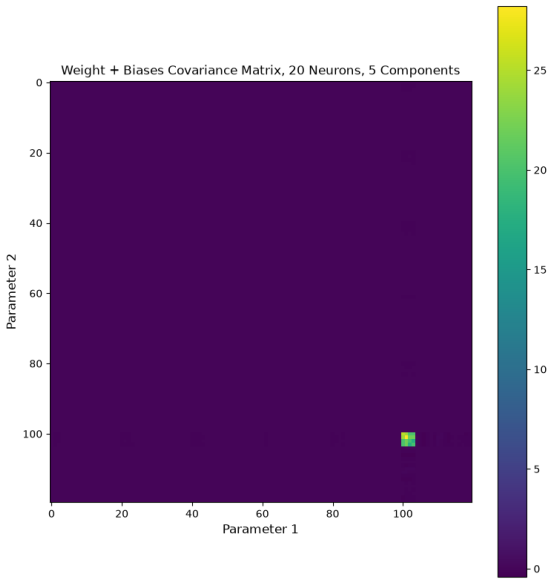

Figure 16: Magnitude of Covariance Diagonalness for the Transistor model for 20 neurons with random fillers

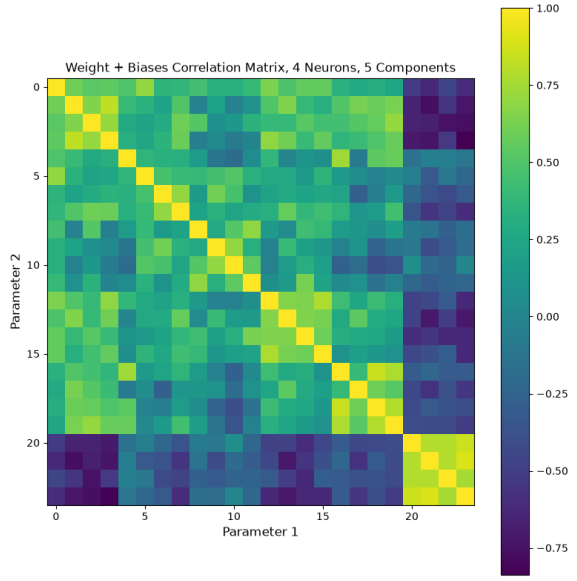

Figure 17: Weight and bias correlation matrix for the 4-neuron transistor model under the zero-filler condition.

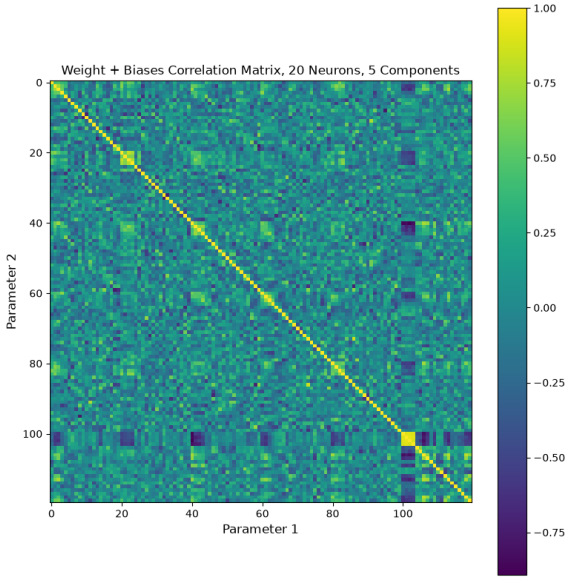

Figure 18: Weight and bias correlation matrix for the 20-neuron transistor model under the zero-filler condition

4. Discussion

The results demonstrate an important distinction between prediction and internal representation. The four-resistor network successfully reproduced the expected physical outputs, but analysis of its parameters revealed additional information about how the relationship was represented internally.

The identity-like structure of the learned weights is particularly significant because the network was not explicitly given the identity matrix or Ohm's law. The structure emerged through training on the physical input-output relationships. This provides evidence that, for a simple linear physical system, the internal parameters of a neural network can contain an interpretable representation of the underlying physics.

However, increasing the number of neurons changed this representation. The decreasing diagonalness and increasing distance from the theoretical matrix suggest that the same physical relationship can be represented differently as additional parameters are introduced. This highlights a limitation of interpreting neural networks purely by looking for a direct correspondence between individual neurons and physical quantities.

The similarity between the zero- and random-filler results also suggests that the observed changes are not strongly dependent on the specific non-physical values introduced. Instead, they may arise from the increasing number of degrees of freedom available to the network.

The transistor experiment provided a useful contrast. The linear network was unable to reproduce the transistor behaviour, demonstrating that the approach that worked for the linear resistor system did not automatically generalise to a more complex physical relationship. Nevertheless, similar trends in diagonalness were observed, raising the possibility that some of the measured structures may be properties of the network architecture or training process rather than direct evidence of learned physical knowledge.

This is an important limitation of the interpretation. A parameter pattern that resembles a physical structure does not necessarily prove that the network has developed a causal or meaningful understanding of that physics. Further experiments would therefore be needed to distinguish genuine physical representations from structures produced by the architecture, dataset or optimisation process.

5. Conclusion

This research investigated whether the internal parameters of a neural network trained on a known physical system could reveal representations of the underlying physics.

The four-resistor model successfully learned the relationship between resistance and voltage, and its learned weights developed an identity-like structure consistent with the expected physical relationship. Covariance and correlation analysis also revealed structured relationships between the learned parameters. These results demonstrate that a simple neural network can contain interpretable internal structures corresponding to a known physical system.

Increasing the number of neurons did not substantially prevent the model from predicting the physical outputs, but it changed the internal representation. The distance from the theoretical matrix increased and diagonalness generally decreased as the network became larger. This

demonstrates that accurate prediction does not necessarily imply a fixed or unique internal representation of the underlying physics.

The transistor model highlighted a further limitation. The linear network could not successfully learn the transistor's switching behaviour, although some parameter patterns remained similar to those observed in the resistor model. This suggests that some observed structures may arise from the network architecture or training process rather than directly representing the physical system.

Overall, the research supports the use of simple, mathematically understood physical systems as controlled environments for studying machine learning interpretability. It also demonstrates that examining internal parameters can reveal information that is not captured by predictive accuracy alone.

The main limitations of the project were the limited research period and my initial lack of experience with machine learning and Python. Future work could investigate whether introducing nonlinear activation functions or additional layers allows the transistor model to learn successfully. This would, however, create a more complex interpretability problem. Further physical systems could also be investigated to determine whether the patterns identified here are specific to simple electrical circuits or represent more general properties of neural networks trained on physical data.

6. Impact of Conducting Research

Throughout this research project, I have developed both technical and personal skills that I did not have at the beginning of the project. One of the most significant areas of development has been learning how to teach myself. I had limited experience with Python and machine learning before starting this research, so I had to learn how to code, read academic literature, investigate unfamiliar concepts, and break down complex ideas into smaller, understandable parts. This process has shown me that I can learn difficult subjects independently, even when I initially have very little understanding of them.

Looking back to the beginning of the project, my perception of artificial intelligence was very different. I initially viewed AI as an almost unknown entity and was interested in the possibility that it could eventually become capable of replacing human abilities. Through this research, I have developed a much more grounded understanding of what a machine learning model does and how it learns from data. Although I still have much more to learn, I no longer view AI as a mysterious or inaccessible technology. I now feel much more confident in questioning how a model has learned something, rather than simply accepting its output.

One of the most challenging aspects of research was the amount of independent work and decision-making required. Unlike teaching university work, there was not always a clear next step or a predefined answer. I had to plan how to use my time, decide what to investigate and regularly redefine my goals. This was particularly difficult during periods where I felt that I was not making much progress. A particularly important moment occurred when I was unsure about what I was doing with my models. My supervisor reminded me that this uncertainty was part of research: the process of investigating, testing, and sometimes not knowing what the next answer will be is research. This changed the way I viewed periods of difficulty. Looking back, I can now recognise that even when I felt I was not progressing, I was developing my understanding and learning how to conduct research.

6.1 My Research Journey

Week 1 – Developing the foundations

During the first week, I spent much of my time understanding what my project was actually investigating. I read academic articles, watched introductory material on machine learning and began teaching myself Python. Python was a particular area I wanted to improve, as I had limited confidence with programming before beginning the project. This week therefore focused less on producing results and more on developing the knowledge and skills required to begin the investigation.

Week 2 – Building my first model

In the second week, I began building my first machine learning model. I initially focused primarily on whether the resistor network could successfully learn to make predictions. This gave me practical experience of constructing, training and evaluating a machine learning model and helped me develop a basic understanding of how the model interacted with the circuit data.

Week 3 – Finding the research question

The third week was one of the most difficult points in the project because I began questioning where I wanted the research to go. My understanding of machine learning and my research experience were still developing, so I found it difficult to identify a question that was both interesting and achievable within the time available. I explored several possibilities involving physics-based models, but many initially appeared too ambitious and closer to the scope of a longer-term research project.

Eventually, I decided to build on what I already understood: electrical circuits. I continued developing the resistor model and considered how this could eventually be extended to a more complex system, such as a transistor model. Although this decision initially felt like narrowing the scope of the project, I now recognise that learning to identify what is achievable is an important research skill.

Week 4 – Moving from building to analysing

By the fourth week, I had developed the models further and began to encounter a new problem: having a successful model was not enough. I needed to understand what the model had actually learned. This led me to investigate the weights and biases within the network and explore different methods of analysing them. Through further research, I eventually focused on covariance and correlation matrices and their relationship to the identity-like structure found within the model.

This was an important change in my approach. Rather than only asking whether the model worked, I was beginning to ask *why* it worked and what information was represented inside it. This shift towards interpretation became one of the most interesting parts of the project for me.

Week 5 – Analysis, automation and the transistor model

The fifth week focused heavily on analysing the results and developing the transistor model. One of the most valuable practical lessons from this stage was the importance of automation. My first analysis took more than two hours because I had to manually load, organise and analyse the data. I realised that repeating this process for every model was inefficient, so I developed code to automate much of the analysis. This reduced the process to approximately twenty minutes per analysis and showed me how programming can be used not only to build models but also to improve the efficiency of research itself.

The transistor model also presented new challenges because of its greater complexity. Building it required me to work through problems with the model structure and coding, including several discussions and debugging sessions with my supervisor. Although this process was frustrating at times, successfully getting the model to run was a useful reminder that research rarely progresses without problems that need to be solved along the way.

Week 6 – Bringing the research together

The final week has focused on analysing and bringing together the results from the different stages of the project. This has been challenging because having the data is only one part of the

research process. I have had to organise the results, compare different models, identify patterns and determine what conclusions can reasonably be drawn from them. This has made me appreciate how much work is involved between obtaining a result and being able to communicate what that result means.

Overall, this research has changed how I view my own ability to learn. I began the project with limited knowledge of machine learning and Python, but I have developed the confidence to approach unfamiliar technical problems independently. I have also developed practical skills in organisation, time management, programming, data analysis and communicating difficulties with my supervisor. Most importantly, I have learned that research does not always feel like progress while you are doing it. Sometimes progress is simply identifying that an approach does not work, asking a better question or understanding a problem more clearly than you did the day before.

7.Leadership Skills

The research project has also developed my leadership skills, particularly my confidence and willingness to take initiative. One area where I have become more confident is organising meetings with my supervisor and actively communicating when I am confused or unsure about something. At the beginning of the project, I was sometimes reluctant to admit that I did not understand something because I felt that I should already know the answer. I have learned that asking for help is not a weakness in research; instead, it allows problems to be addressed more quickly and leads to better discussions.

I have also developed greater determination when facing setbacks. There were several points where my models did not behave as expected or where I was uncertain about the direction of the research. Rather than abandoning the project, I learned to reassess the problem, discuss it with my supervisor, and find another approach.

Another important leadership skill I have developed is learning to balance ambition with what is realistically achievable. At the beginning of the project, I wanted to explore increasingly complex ideas, but I quickly realised that some of these would require far more time and knowledge than I currently had. Learning when to pursue an ambitious idea and focusing on developing my existing work has been an important lesson. I have learned that being ambitious does not always mean taking on the largest possible problem; sometimes it means making the most of the problem you have chosen and finding meaningful questions within it.

8. Dissemination of My Research

I intend to disseminate my research through the Laidlaw research presentation and report, allowing me to communicate both the findings and the process behind them. I am also interested in developing the work into a separate academic-style paper and potentially submitting it to an appropriate student or academic publication. This would give me an opportunity to further develop the research beyond the Laidlaw programme and receive feedback from a wider audience.

An important part of dissemination for me will be explaining the research in a way that is accessible rather than only presenting technical results. One of the lessons I have learned throughout this project is that understanding something well enough to explain it clearly is different from simply being able to produce a result. Communicating with my research will therefore be another way of testing and developing my own understanding.

9. Future and goals

My future and goals, after this project I have learnt so much about how AI works on a fundamental level, but there is so much more to explore. I have potential ideas about going deeper into looking a neural network. Possibly seeing if it's possible to figure out which neurons have the most impact and therefore being able to eradicate some and therefore use less energy. Also, to investigating more systems like my circuit model and seeing if certain components require an extra layer or activation function to work. Therefore, trying to make these models smaller more prices and less data and energy consuming. From learning about machine learning models, I am now curious about large language models work and their different internal structure compared as these are things people use daily. After this project I feel I have an own personal goal to teach people more about AI and ways to use it and how it works and can use your data.

[i] Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav Ther*. 2020;51(5):675-687. doi:10.1016/j.beth.2020.05.002

[ii] Qamar T, Bawany NZ. Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Comput Sci*. 2023;9:e1629. Published 2023 Nov 29. doi:10.7717/peerj-cs.1629