Thesis for The Degree of Doctor of Philosophy

Interpretable machine learning applied to fusion plasmas

Andreas Gillgren

Department of Physics and Astronomy
Chalmers University of Technology
Gothenburg, Sweden, 2026

Interpretable machine learning applied to fusion plasmas

Andreas Gillgren

ISBN 978-91-8103-424-0
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5881.
ISSN 0346-718X
DOI-ID: 10.63959/chalmers.dt/5881
Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

Department of Physics and Astronomy
Division of Astronomy and Plasma Physics
Chalmers University of Technology
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

Interpretable machine learning applied to fusion plasmas

Andreas Gillgren

Department of Physics and Astronomy
Chalmers University of Technology

Abstract

Magnetic confinement fusion is a field of research that strives to develop an environmental friendly energy source to assist in powering our society. By confining a plasma with magnetic fields, conditions that enable nuclear fusion can be achieved. However, gaining a high efficiency has proven to be a challenging task, which is why fusion research is still at the experimental stage. On another scientific front, machine learning (ML) and artificial intelligence (AI) are advancing with tremendous momentum, and are rapidly being integrated into fusion research. However, the black-box aspect of many popular ML architectures inhibits our ability to interpret what the models have learned during training. Ideally, we would like to understand ML models deployed in fusion research to gain knowledge that may prove useful for advancing the field.

This thesis focuses on exploring interpretable ML methods in fusion research. As a consequence, an interpretable framework called *NeuralBranch* has been developed, which has been applied to two different use cases in fusion. The main application in this thesis relates to the so-called pedestal, which has significance for the energy confinement in fusion experiments. The other, more secondary application in this thesis, relates to the growth rate of plasma instabilities that contribute to heat and particle transport. In summary, the interpretability of the machine learning models deployed reveals intricate parameter relationships in both these applications, beyond what previous traditional data-fitting approaches have been able to reveal.

Keywords

Magnetic confinement fusion, Machine learning, Artificial intelligence, Interpretability, Tokamak, Pedestal

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **A. Gillgren**, E. Fransson, D. Yadykin, L. Frassinetti, P. Strand and JET contributors, *Enabling adaptive pedestals in predictive transport simulations using neural networks*
Nuclear Fusion 62 (2022).

[**Paper II**] E. Fransson, **A. Gillgren**, A. Ho, J. Borsander, O. Lindberg, W. Rieck, M. Åqvist and P. Strand, *A fast neural network surrogate model for the eigenvalues of QuaLiKiz*
Physics of Plasmas 30 (2023).

[**Paper III**] **A. Gillgren**, A. Ludvig-Osipov, D. Yadykin, P. Strand and JET contributors, *Investigating pedestal dependencies at JET using an interpretable neural network architecture*
Nuclear Fusion 65 (2025).

[**Paper IV**] **A. Gillgren**, E. Fransson, A. Ludvig-Osipov, W. Enström, L. Flyckt, M. Green, M. Kvartsén, Y. Liljegren, E. Olsson, A. Orthag, H. Wennberg, and P. Strand, *Investigating characteristics of the growth rates from QuaLiKiz using an interpretable surrogate model*
Physics of Plasmas 32 (2025).

[**Paper V**] **A. Gillgren**, D. Yadykin, P. Strand, and JET contributors, *Interpretability guided transfer learning approaches for tritium pedestal predictions*
Accepted in Plasma Physics and Controlled Fusion (2026).

Other publications

[**a**] C. J. Ham, A. Bokshi, D. Brunetti, G. Bustos-Ramirez, B. Chapman, J. W. Connor, D. Dickinson, A. R. Field, L. Frassinetti, **A. Gillgren**, *Towards understanding reactor relevant tokamak pedestals*
Nuclear fusion 61 (2021).

[**b**] D. R. Ferreira, **A. Gillgren**, A. Ludvig-Osipov, P. Strand and JET contributors, *High temporal resolution of pedestal dynamics via machine learning on density diagnostics*
Plasma Physics and Controlled Fusion 66 (2023).

[**c**] S. Wiesen, S. Dasbach, A. Kit, A. E. Jaervinen, **A. Gillgren**, A. Ho, A. Panera, D. Reiser, M. Brenzke, Y. Poels, *Data-driven models in fusion exhaust: AI methods and perspectives*
Nuclear Fusion 64 (2024).

Acknowledgment

First, I would like to thank my wife Katarina, who, together with our son Henry, are my biggest sources of inspiration and joy in life. I would also like to thank my parents Gunilla and Mårten, and the rest of my family for their ever-constant encouragement throughout the years. Finally, I of course want to thank my supervisors Pär Strand and Dmytro Yadykin for giving me the support and opportunity to explore topics that have intrigued me during my PhD, and my other colleagues at Chalmers for making it such an enjoyable experience. A special thanks to Lars-Göran Eriksson and Hans Nordman for reading my thesis and providing me with very valuable feedback.

I would also like to acknowledge that the work presented in this thesis would not have been possible without support from Vetenskapsrådet (VR, project grant 2020-05465) and from the EUROfusion project ENR-MOD.01.FZJ, "Development of machine learning methods and integration of surrogate model predictor schemes for plasma-exhaust and PWI in fusion".

Contents

Part I

Summary

Chapter 1

Introduction

In the story of human progress, both the topics of energy and artificial intelligence (AI) stand at the forefront. The dream of emulating the energy from the stars, coupled with the rise of AI, has lead to the rapid development of an interdisciplinary field: the application of machine learning in fusion technology. In this chapter, an introduction to both these fields is presented, along with an introduction to the objectives of this thesis.

1.1 Nuclear fusion

Fusion energy is the energy released in the process when two atomic nuclei are merged into a heavier element. This is the counterpart of fission, where heavy nuclei are split into lighter elements. For instance, the hydrogen isotopes deuterium (D) and tritium (T) can fuse into a helium-4 (He^4) nucleus, producing an additional neutron (n) in the process

$$D + T \longrightarrow He^4 + n + \text{energy}. \tag{1.1}$$

Although the helium-4 nucleus is heavier than the deuterium and tritium nuclei respectively, the sum of the mass of the deuterium and tritium nuclei is larger than the sum of the mass of the helium-4 nucleus and the neutron [1]. This is due to the difference in binding energies of the different nuclei [2], and the loss of mass m is converted to kinetic energy E according to Einstein's equation

$$E = mc^2, \tag{1.2}$$

where c is the speed of light in vacuum (299 792 458 m/s) [1]. Since c is a large number, even a small change of mass leads to a large amount of energy. The energy release in chemical reactions is also due to the principle of changed mass. However, chemical reactions are associated with the rearrangement of bonds between outer shell electrons and nuclei. These bonds are carried by the electromagnetic force, which is significantly weaker than the strong nuclear force binding nuclei [3]. Therefore, a smaller change in binding energy in chemical reactions leads to a smaller change in mass and energy in (1.2). As

an illustrative example, a kilogram of fusion fuel can yield nearly four million times more energy compared to the combustion of a kilogram of coal or oil [4].

Nuclear reactions are based on probability rooted in quantum theory [2], which means that we can never be certain if two individual nuclei will fuse. However, when considering a macroscopic scale involving numerous nuclei, the cumulative probability manifests itself in an observable reaction rate [3]. Fundamentally, the probability of a fusion reaction occurring is increased when two nuclei are in a close proximity for an extended duration. However, two positively charged nuclei repel each other according to Coulomb repulsion [5], and to overcome this barrier, the nuclei must have sufficient kinetic energy. That said, too high kinetic energy will also lower the probability of a fusion reaction occurring [3], since the nuclei will pass each other more rapidly, which reduces the time that the nuclei are in a close proximity.

On a macroscopic scale, temperature acts as a measurable quantity for the average kinetic energy of the particles in a system [6]. Therefore, there is an optimal temperature that maximizes the rate of fusion reactions. Due to properties derived from quantum theory, the optimal temperature varies between different pairs of nuclei [3]. This is illustrated in Figure 1.1, where the cross-section σ [3], which is proportional to the reaction probability, is plotted as a function of temperature for different fusion reactions. It is also shown how much energy is converted to kinetic energy in each reaction in mega electron volts (MeV).

It is however not only the temperature that affects the rate of fusion reactions in a system. The density plays a vital part since a more densely populated volume means that a nucleus will cross paths with more nuclei along its trajectory, which increases the probability of a reaction.

In the sun, and other stars, the tremendous gravitational pull leads to sufficiently high temperatures and densities in the core to enable fusion. At these high energy conditions, atoms are in a plasma state, where electrons have been stripped from their nuclei, resulting in a mixture of positively charged ions and free negatively charged electrons. On Earth, we do not have access to this extreme gravity to confine particles and achieve fusion. Therefore, we need to find other ways if we wish to exploit this feature of nature in a controlled manner.

1.2 Fusion on Earth

The two main branches of fusion power on Earth are magnetic confinement fusion (MCF) [4] and inertial confinement fusion (ICF) [8]. In MCF, the confinement is achieved by utilizing magnetic fields that pass through a chamber. Due to the Lorentz force [5], charged particles are confined in a gyrating motion along the magnetic field lines, which will be further described in Chapter 2. As in the sun, the particles in a MCF device are heated to a plasma state to enable the conditions for fusion. This technology has been developed since the 1950s, including configurations such as the mirror machine, the tokamak, and the stellarator. For example, ITER [9] is currently under construction as the

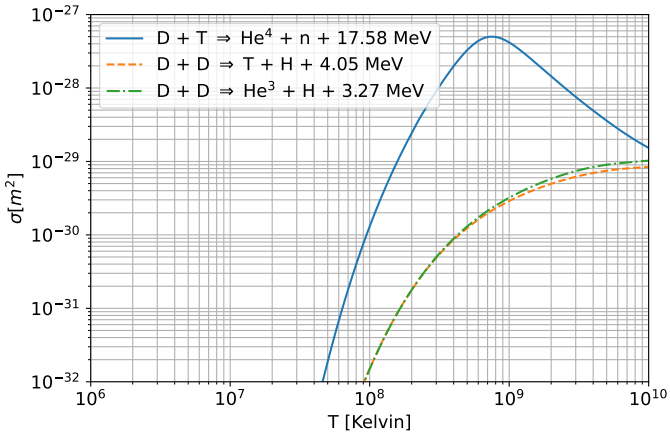


Figure 1.1: The cross section σ of different reactions, which is related to fusion rate, or reaction probability, versus temperature. To achieve a significant rate, temperatures exceeding tens of millions of degrees Kelvin are necessary. The DT reaction provides two advantages compared to the DD reaction as its cross-section peaks at a lower temperature, which would be easier to achieve on Earth, and more energy is converted per reaction. Additionally, out of the two DD reactions, the graph indicates that if a DD reaction occurs, there is approximately a 50% chance to produce a T nucleus, and approximately a 50% chance to produce a He^3 nucleus. Here, H refers to a hydrogen nucleus without neutrons, which is simply a proton. Additionally, it is noteworthy to mention that there are other possible reactions for fusion beyond the ones shown in this figure. The data in this graph was gathered from the International Atomic Energy Agency (IAEA) Nuclear Data Section [7].

largest tokamak to date in Cadarache, France. This project is a collaborative effort involving the European Union, the United States, China, the Russian Federation, Japan, South Korea, and India. As all of the work presented in this thesis is related to MCF, and specifically the tokamak, we will focus on this reactor type later in Chapter 3.

In ICF, a pellet containing the fusion fuel is placed in a chamber. The pellet is compressed by external laser beams, which leads to adiabatic heating such that sufficient conditions for fusion are achieved. This technology has been developed since the 1970s, and the largest operational ICF experiment to date is the National Ignition Facility (NIF) [10] in the United States.

Both MCF and ICF are still at the experimental stage, and research aims to address how these technologies can be developed in order to provide economically viable electricity to the grid.

1.2.1 Fuel

The main fuel expected to be used in the first generation of fusion power plants, with a few exceptions [11], is a combination of deuterium and tritium [4], [10]. Deuterium is not radioactive, and easily accessible and abundant since it can be extracted from water. For instance, relying solely on deuterium-deuterium (DD) fusion reactions to meet the current annual energy demand could potentially provide a supply lasting for billions of years [12]. Tritium however, is radioactive with a half-life of 12.32 years. For this reason, it is not abundant in nature, although it can be produced by irradiation of lithium (Li) with fast neutrons [4]. Lithium is a relatively abundant element in the Earth's crust, but as the demand for lithium for use in batteries for electric vehicles and portable electronics continues to rise, efforts to explore and develop new lithium deposits are also ongoing [13]. In principle, it is possible to only use deuterium as the fuel for fusion, which is a long-term goal. However as illustrated in Figure 1.1, this puts a higher demand on the technology.

1.2.2 Potential benefits with fusion

Beyond the already mentioned benefits about the fuel, fusion is intrinsically environmental friendly and sustainable, since no greenhouse gasses are produced in the reaction. This of course assumes environmental friendly transport of fuel and construction of facilities. That said, even if transport is not fully sustainable, less fuel needs to be transported compared to combustion facilities (coal, oil, biogas), provided the fuel is extracted from the source material before transport.

A fusion reactor will produce some radioactive byproducts through neutron activation of reactor materials, although the hazard differs significantly compared to fission facilities. The most common fission fuels, uranium-235 and plutonium-239, produce highly radioactive products that remain hazardous for thousands of years, which is not the case for fusion [1], [12]. For instance, the fusion reactor material can be chosen such that most of it can be safely recycled or disposed in less than a 100 years after being activated [14].

Moreover, future fusion reactor designs may include a lithium based blanket for tritium breeding in the reactor chamber [4]. If practical challenges related to this is solved, the need to transport or store most of the radioactive tritium would be eliminated. The risk of a nuclear meltdown, for instance due to a natural disaster, is also not a concern in fusion since the reactions are highly dependent on specific conditions that are challenging to maintain consistently. In practice, any major disturbance will immediately terminate the process in relevant designs [4]. At any given time, a reactor contains only a minimal quantity of active materials, which ensures that any potential environmental impact remains localized to the reactor.

Fusion, much like fission and combustion, holds the advantage of not requiring specific weather conditions or location for its operation, which is not the case for solar, wind, and hydropower. It is however important to emphasize that the goal of fusion is not necessarily to replace other renewables. Instead,

fusion may complement existing sustainable energy systems to contribute to a more diverse and resilient energy ecosystem.

1.2.3 Key challenges

Given all the potential benefits, it comes as no surprise that certain challenges have prevented fusion from being realized on a commercial scale yet. For magnetic confinement fusion, which this thesis focuses on, the main obstacles are:

- Confinement - Upholding the desired conditions for high fusion performance is much harder than initially anticipated. This is mainly due to turbulence arising from plasma instabilities, causing particles to escape the magnetic field lines, and energy to leak out through multiple channels that researchers are still working to understand and control. In essence, most other challenges related to MCF compounds from the confinement challenge.

- Scale and complexity - One of the strategies to address the confinement problem is to build larger machines. This helps by increasing the plasma volume relative to the surface area where the losses occur. Moreover, a larger plasma means that it takes longer for energy to leak out from the center. However, large machines imply massive, extraordinarily complex projects, which makes experimentation slow and expensive. For instance, construction of ITER, which will weigh about 23 000 tonnes, started in 2013 and is at present planned to produce its first plasma by 2033-2034, and this is after several delays. The total cost of building ITER is projected to be over 20 billion dollars [15], which makes it more than thrice as expensive as the Large Hadron Collider (around 5.6 billion [16]). Hence, the scale and complexity of larger machines raises questions regarding the economic viability, even in the case where physics related challenges are solved.

- Material limits - The reactor wall will face neutron bombardment, extreme heat fluxes, and plasma interactions that will damage and degrade the material. Beyond research related to mitigation strategies, a key challenge is to identify solutions to how wall material can be effectively repaired or replaced in an economically viable manner over the expected lifetime of a commercial power plant. Potentially, progress in the development of new material that better withstand the necessary plasma conditions may assist in alleviating this problem.

- Tritium breeding - As mentioned, one idea for future power plants is to include a lithium blanket for continuous tritium breeding during operation. However, to date there has been no demonstration that this can be done at the scale and efficiency required for a power plant. Further work is required to confirm that such designs can simultaneously breed tritium, extract heat for power generation, and shield the magnets.

Moreover, regardless if a reactor design includes a lithium blanket or not, initial tritium is required at the starting of the reactor. Further research is needed to confirm that decoupled tritium generation in reasonable proximity to the reactor can be done in an economically viable manner.

1.2.4 Current fusion landscape

Since fusion was first explored in the 1950s, the vast majority of research has been performed in academic environments via governmental funding, often at a high degree of international collaboration. In this time, several tokamaks at variable size have been constructed, such as JET (United Kingdom), ASDEX (Germany), JT60-SA (Japan), EAST (China), and DIII-D (USA) to name a few. Out of these, JET holds the record for the highest scientific Q-value for deuterium-tritium reactions (0.67), which is the ratio of fusion power to the input power. For reference, ITER is expected to achieve break-even with a scientific Q-value in the range of 5-10 [17], although in a future commercial power plant an even higher scientific Q-value will likely be required. This is, for instance, due to energy losses in the process where the fusion energy will be converted to electricity via steam turbines.

Historically, high performance observed in early tokamaks led to a loss of interest in other magnetic confinement concepts. However, stellarators regained attention in the 1990s, partly due to growing problems with the tokamak design, but also due to new methods enabling higher quality in the magnetic field. At present, Wendelstein 7-X (Germany) is the largest stellarator, and in 2025, it broke the record for what is known as the *triple product* for long plasma durations, surpassing those observed in tokamaks [18]. We will return to the triple product later in Chapter 3, but for now it can thought of as a proxy for the goodness of the plasma conditions for fusion. As Wendelstein 7-X has not been operated with a fuel mixture of deuterium and tritium, there is not yet an appropriate Q-value to be compared with those of the highest performing tokamaks.

Inertial confinement fusion has sparked attention in recent years, mainly due to record breaking experiments at NIF (USA). For instance, a target gain of 4.13 was achieved in 2025 [19], which is the ratio of fusion power to the power of the lasers aimed at the fuel pellet. However, ICF faces immense challenges in terms of becoming economically viable. For instance, a much higher power is required to operate the entire facility compared to the power that actually gets delivered to the pellet through the lasers. Moreover, the rate of pellet replacement and laser firing must increase by orders of magnitude for commercial operation, and the pellet manufacturing capabilities must be significantly improved, both in terms of speed and cost. Since this thesis focuses on MCF, we do not elaborate on the challenges of ICF further, but readers more interested in this topic can find useful information in [20].

Despite the majority of fusion research being funded by governmental agencies, the field has expanded notably into the startup domain over the last decade. In general, these startups have one thing in common: they present an alternative concept or variation to previous ideas that aim to solve one or

several of the major challenges in relation to reaching economic viability. For instance, the company Commonwealth Fusion Systems (USA) focuses on using high-temperature superconductors to enable stronger magnetic fields, allowing for a more compact and economical design. Another example is the company Helion (USA), which presents a completely different concept: firing two plasma volumes towards each other at high speeds. An interesting aspect about this design is that the deuterium-He^3 reaction is considered as a candidate, and that the fusion reactions are supposed to induce a current in the reactor magnets for direct extraction, such that no steam turbine needs to be included in the design. A final example is the company Novatron Fusion (Sweden), which presents a design based on the mirror machine, but with a unique magnetic field configuration that is beneficial for confinement and stability. As of 2025, some of these startups have raised more than a 100 million dollars from investors, and most of them present relatively aggressive timelines for when they expect to provide electricity to the grid (mostly in late 2020s or early 2030s) [21]. However, given the significant technical challenges that have historically faced fusion energy, demonstrating economic viability through actual experimental results will likely be necessary to convince the majority of fusion researchers.

1.3 Simulations

A vital aspect of fusion research is to perform simulations of different machine components, regardless of which specific technique is used to achieve fusion. Specifically, simulations help by cutting costs and by allowing a more flexible exploration compared to only relying on results from real experiments. This is particularly important in the design of future, more expensive machines. Moreover, through simulations researchers can investigate how well models based on the current theoretical understanding agrees with results from existing experiments. This is a crucial aspect for the confidence in theoretical extrapolations to future machines, and in general for the capabilities of optimizing reactor performance.

For magnetic confinement devices, such as the tokamak, theoretical approaches rooted in statistical physics are usually employed to model the fusion plasma. We will explore these approaches more in Chapter 2. Beyond the plasma itself, simulations are important for characterizing and capturing effects related to, for instance, the heating and fueling systems, the reactor wall material, the magnets, and the fusion yield.

1.3.1 Integrated modeling

Simulating all aspects of a fusion experiment is no easy task. In fact, only simulating the plasma involves capturing many complex phenomena occurring on different spatial and temporal scales. Therefore, fusion researchers often employ the concept of integrated modeling, which means that multiple modules that are specialized for different aspects are coupled together in a simulation workflow. For instance, it is common to use distinct modules for heat transport

and equilibrium calculation when simulating a tokamak plasma. In such a setup, the user can experiment with, for instance, different combinations of transport and equilibrium codes and settings. Some of the most common integrated modeling frameworks for tokamaks include TRANSP [22], JETTO [23], ASTRA [24], and ETS (European Transport Simulator) [25], of which the last is the main framework used and developed by the research group where this thesis was conducted.

While integrated modeling already constitutes a rigid cornerstone in fusion research, there remains areas of improvement. Specifically, as will be discussed more thoroughly in Chapter 2 and Chapter 3, there are still aspects of the plasma in tokamaks that either are difficult to model or not entirely theoretically understood. Moreover, as also will be discussed in Chapter 2, certain theoretical models are very computationally expensive, which can lead to significant bottlenecks that inhibit exploration. Hence, ongoing efforts strive to explore innovative approaches that both enhance the speed and fidelity of integrated modeling simulations.

1.4 Machine learning / AI

On a different scientific front, machine learning / AI are advancing with tremendous momentum. From achieving handwritten digit recognition in the 1990s [26], [27], to surpassing the world champion in Go in the 2010s [28], and to becoming human assistants with models like ChatGPT and Claude in the 2020s [29], AI researchers have achieved remarkable milestones. Although AI is often mentioned in the context of these popular applications, it has also proven to assist in other research disciplines, such as detection of diseases [30], [31], protein folding [32], environmental modeling [33], [34], but also fusion [35], [36].

AI and machine learning are terms that often are used to describe the same thing. However, machine learning represents a subset of AI that involves the development of algorithms that allow systems to learn from data, identify patterns, and make decisions or predictions without being explicitly programmed for each scenario. A neural network is an example of a machine learning model that we will explore in detail in Chapter 4.

Although the development of new and more powerful machine learning algorithms presents many potential benefits, the technology also poses challenges. For instance, as with many inventions, machine learning, and AI in general, can be utilized in harmful and unethical ways. Another challenge is related to interpretability, which is about understanding the rationale behind a decision or prediction made by a model, which, for instance, holds relevance for the ethical aspect in high-stake applications. That said, there is arguably a difference in the ethical aspect between, for instance, making AI-based judgments in court, and using AI in the development of a clean energy source. However, we would ideally like to understand how AI systems applied in fusion arrive at their predictions, both to ensure confidence in their use and to gain insights that can improve our understanding of fusion plasmas.

1.5 Objectives of thesis

This thesis aims to address the generally unexplored topic of interpretable machine learning applied in fusion research. Specifically, the goal is to leverage interpretability to achieve both predictive capabilities and enhanced understanding of a region in tokamak plasmas referred to as the pedestal [37]. Essentially, the pedestal represents a suppression of energy and particle transport near the edge of the plasma, which leads to overall elevated temperatures and densities. In turn, this leads to improved confinement and performance which, as discussed, is an important consideration for future reactors. Specifically, the pedestal pressure is expected to strongly influence the performance of future power plants [38], [39]. So in a broader context, this thesis aims to contribute to the general effort of addressing the plasma confinement challenge.

Through the machine learning interpretability, the aim is to uncover how the pedestal depends on key tokamak parameters. This is possible because, if we can understand how a model predicts an output from its inputs, we also learn about the relationship between the inputs and the output. Hence, the interpretability offers a pathway to big-data analysis beyond what has been possible with traditional methods [40]. Moreover, as pedestal prediction models will be developed in this process, an additional aim is to offer an alternative to present-day pedestal models for integrated modeling applications [41], [42]. Here, the interpretability matters for transparency and predictability in model behavior, which is important for reliability. As will be discussed in Chapter 3, being able to predict the pedestal accurately is important for the overall fidelity when simulating the performance of fusion plasmas. As will also be discussed in Chapter 3, no existing pedestal model is fully satisfactory, which motivates further development of alternative approaches.

As a consequence of novel interpretable machine learning methods being developed in the scope of this thesis, additional applications in fusion, beyond that of the pedestal, are also included in the appended papers. Specifically, this relates to the growth of plasma instabilities that lead to turbulent transport of heat and particles.

The outline of this thesis is as follows: in Chapter 2, an overview of the theoretical descriptions of plasmas is presented. Chapter 3 focuses on the key aspects of the tokamak that hold the most relevance to the context of the papers appended in this thesis. In the end of Chapter 3, a description of the pedestal is outlined. Moving to Chapter 4, the fundamentals of machine learning are presented, and Chapter 5 outlines the sub-field of interpretability. Chapter 6 summarizes the appended papers, and Chapter 7 concludes the work, which also includes potential avenues for future research.

Chapter 2

Brief overview of plasma physics

As mentioned in the introduction, to create conditions suitable for fusion, it is necessary to heat the particles enough, resulting in the formation of a plasma. Therefore, plasma physics is essential to fusion research, and in this section, a brief overview of plasma physics based on the contents in [4] is presented. The purpose of this chapter is mainly to provide a background for readers not already familiar with plasma physics, since some of the contents presented here are necessary for understanding the discussion of the tokamak and the pedestal in Chapter 3.

2.1 Definition of a plasma

A simple description of a plasma is that the electrons are no longer bounded to the nuclei, which results in a collection of free electrons and ions. Apart from occurring at high temperatures as in fusion plasmas, it can also occur at low densities where electrons are rarely recaptured by ions, such as in astrophysical plasmas.

A more precise description is that a plasma is a quasi-neutral gas consisting of charged and neutral particles that exhibit collective behavior. Quasi-neutrality means that the plasma is neutral in charge on a macroscopic scale but occasionally deviate from neutrality on the microscopic level. However, while these microscopic deviations produce electrical fields that operate on a relatively small spatial scale, they remain long-range compared to the collisional forces between two neutral particles in a conventional gas. Therefore, in contrast to pair-wise collisions in a conventional gas, a particle in a plasma interacts with many particles simultaneously through Coulomb forces, hence the term 'collective behavior'.

One of the main properties of a plasma is the ability to quickly screen out changes in the electric potential associated with a charged particle, which is why deviations in quasi-neutrality occur on a microscopic scale. For instance, if

a positive charge q was to be placed in an otherwise neutral plasma, electrons will rapidly move towards the charge, since they are much lighter than ions and can accelerate faster. They will then arrange such that outside a sphere with the charge q in the center, which is called the Debye sphere, the influence from the charge q becomes negligible. For fusion plasmas, the radius of this sphere, which is called the Debye length λ_D, is on the order of 7×10^{-6} m. The calculation of the Debye length is based on the assumption that the electron density follows a Boltzmann distribution with respect to the electric potential. Since this is a statistical assumption, the number of particles in the Debye sphere must be large. Using this constraint, we can now formulate concrete requirements for the plasma definition to hold:

- The macroscopic spatial dimension of the plasma must be much larger than the Debye-sphere.
- The particle density must be sufficiently high such that many particles populate the Debye sphere.
- Additionally, the frequency of collisions between plasma particles and neutral particles cannot be too high such that it disrupts the plasma dynamics governed by the long-range electromagnetic forces.

2.2 Theoretical descriptions of plasmas

Although the fusion process itself is a quantum mechanical phenomenon, plasma physics is governed by classical physics. In principle, a plasma can be accurately described with the equation of motion for each individual particle in the plasma, together with a self-consistent set of Maxwell's equations for the electromagnetic fields. In practice however, finding a solution to this is not possible, even with supercomputers, due to the large number of particles and the inherent complexity. This has led to the development of different statistical approaches to describe a plasma, which we will summarize in the following subsections. However, we will start by considering the motion of charged particles in electric and magnetic fields, that are assumed to be known, since this can provide valuable information in addition to the statistical approaches we will discuss later.

2.2.1 Particle motion in electric and magnetic fields

Consider a particle with the charge q in an electric field $\vec{E}$ and a magnetic field $\vec{B}$. The acceleration $\vec{a}$ of the particle is governed by the Lorentz force

$$m\vec{a} = \vec{F} = q(\vec{E} + \vec{v} \times \vec{B}), \tag{2.1}$$

where m is the mass of the particle, and where $\vec{v}$ is the velocity of the particle. Now assume a scenario where $\vec{E} = 0$ such that

$$\vec{a} = \frac{q}{m}(\vec{v} \times \vec{B}). \tag{2.2}$$

If $\vec{B}$ is uniform and constant, then the solution to (2.2) is a combination of a circular particle motion, which is referred to as the gyro-motion, perpendicular to $\vec{B}$, and a constant velocity parallel to $\vec{B}$. This is illustrated in Figure 2.1, where the center of the gyro-motion is referred to as the guiding center, or the gyro-center.

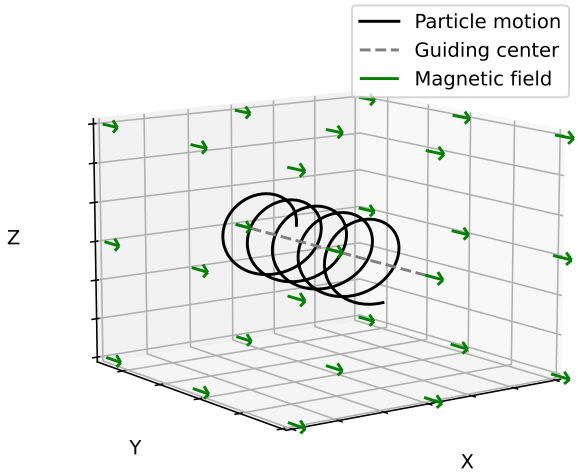


Figure 2.1: The motion of a charged particle in a uniform magnetic field (arrows). X, Y and Z represent spatial coordinates. The particle follows a circular motion around and along the guiding center, which is in the same direction as the magnetic field (Y-direction). Therefore, the particle is confined to gyrate around the field lines.

The solution of (2.2) also provides the gyro-radius, which is called the Larmor radius (which is obtained by balancing the centripetal force with the Lorentz force)

$$r_L = \frac{mv_\perp}{|q||\vec{B}|}, \tag{2.3}$$

where $v_\perp$ is the perpendicular velocity of the particle with respect to $\vec{B}$. Equation (2.3) indicates that ions have a larger Larmor radius compared to electrons since their mass is larger. The gyro-frequency ω, in units radians/seconds, can also be obtained through

$$\omega = \frac{v_\perp}{r_L} = \frac{|q||\vec{B}|}{m}. \tag{2.4}$$

Now consider a scenario where we add an arbitrary force $\vec{F}$ to the equation of motion (2.2), such that

$$\vec{a} = \frac{q}{m}(\vec{v} \times \vec{B}) + \frac{\vec{F}}{m}. \tag{2.5}$$

If $\vec{F}$ has a parallel component with respect to $\vec{B}$, this will lead to acceleration of the particle along the magnetic field lines. The orthogonal component of the force $\vec{F_\perp}$ will lead to a drift across the magnetic field lines, where this particle drift is orthogonal to both $\vec{F_\perp}$ and $\vec{B}$. This can be described by introducing a constant drift velocity $\vec{v_D}$ such that

$$\vec{v_\perp} = \vec{v_D} + \vec{u}. \tag{2.6}$$

By inserting (2.6) into the orthogonal part of (2.5), the solution will show that $\vec{u}$ represents the gyro-motion, or Larmor motion, and that

$$\vec{v_D} = \frac{1}{q}\frac{\vec{F_\perp} \times \vec{B}}{B^2}. \tag{2.7}$$

For instance, an electrical field $\vec{E}$ orthogonal to $\vec{B}$ will generate a force $\vec{F} = q\vec{E}$, which will lead to the drift

$$\vec{v_D} = \frac{\vec{E} \times \vec{B}}{B^2} = \vec{v}_{E\times B}. \tag{2.8}$$

Note that this $\vec{E} \times \vec{B}$ drift is independent of mass and charge, which means that ions and electrons will drift in the same direction with the same velocity. The $\vec{E} \times \vec{B}$ drift is illustrated in Figure 2.2.

The $\vec{E} \times \vec{B}$ drift is however not the only possible drift. By replacing the arbitrary force $\vec{F}$ in (2.5) with, for instance, the centripetal force in a curved magnetic field, we obtain the curvature drift

$$\vec{v_c} = \frac{mv_\parallel^2}{qB^2}\frac{\vec{R}_c \times \vec{B}}{R_c^2}, \tag{2.9}$$

where $v_\parallel$ is the particle velocity parallel to the magnetic field, and where $\vec{R}_c$ is the curvature radius vector. Additionally, a curved magnetic field will not satisfy Maxwell's equations if $\vec{B}$ is homogeneous, which implies that there is a nonzero gradient ∇B. Therefore, over the course of one gyration, a simplified description is that the particle will experience a stronger magnetic field for one half of the gyration compared to the other half of the gyration. Since the strength of the magnetic field affects the Larmor radius (2.3), the radius of one half of the gyration will be smaller compared to the other half of the gyration, which results in a drift. This drift, which is called the ∇B drift, or grad B drift, is illustrated in Figure 2.3.

A derivation of the ∇B drift, which is more complicated compared to the other mentioned drifts, can be found in [4], and it yields the result

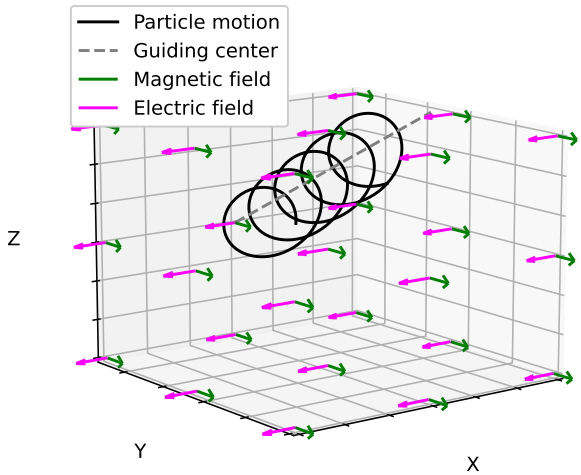


Figure 2.2: The motion of a charged particle in an uniform magnetic field (green arrows in Y-direction) and an uniform electric field (magenta arrows in X-direction) orthogonal to the magnetic field. The particle follows a circular motion, but in contrast to the case illustrated in Figure 2.1, the guiding center, which travels in the Y-direction, now also drifts in an orthogonal direction (Z-direction) compared to both the electric field and the magnetic field as a consequence of the $\vec{E} \times \vec{B}$ drift.

$$\vec{v}_{\nabla B} = \frac{\mu}{q} \frac{\vec{B} \times \nabla B}{B^2}, \tag{2.10}$$

where μ is the magnetic moment of the charged particle

$$\mu = \frac{m v_{\perp}^2}{2|\vec{B}|}, \tag{2.11}$$

which is an adiabatically conserved quantity.

In summary, understanding the drifts discussed in this subsection is essential for confining particles in a magnetic confinement fusion device.

2.2.2 Collisions

In addition to drifts, collisions also lead to particles and energy not being perfectly confined to the magnetic field lines. As mentioned in the beginning of this chapter, these collisions do not occur in a pair-wise manner as in a

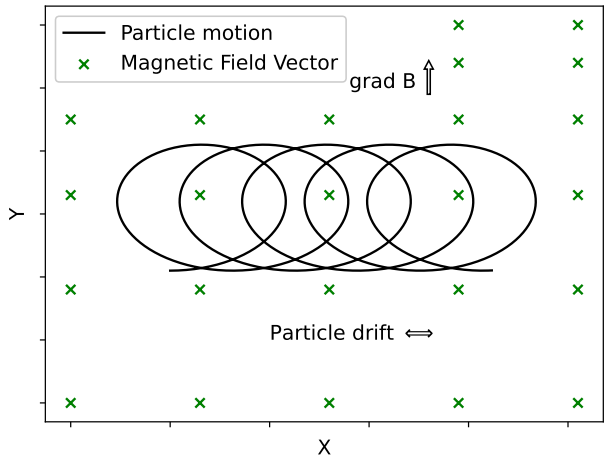


Figure 2.3: Illustration of the grad B drift. The magnetic field lines that are orthogonal to the XY-plane, illustrated by crosses, are closer to each other in the upper part of the figure to illustrate that the magnetic field is stronger for higher Y. This leads to a larger Larmor radius for lower Y according to (2.3), resulting in a drift in the X-direction. The direction of the particle drift depends on the sign of the charge of the particle, as indicated by (2.10), which means that ions and electrons drift in opposite directions for this drift.

conventional gas. Instead, particles deflect due to the combined Coulomb interaction with many particles, which is facilitated by the electromagnetic force. The collision time is defined as the average time it takes for a particle to be deflected 90°, and it is proportional to $T^{3/2}$, where T is temperature, which means that collisions in a plasma occur less frequently at high temperatures.

As will be illustrated in the next subsection, collisions have influence on the equations used for statistical frameworks of plasmas. Moreover, as will be discussed in Chapter 3, the rate of collisions in a plasma also impacts both the plasma resistivity and the transport of particles and energy.

2.2.3 Kinetic description of plasmas

We will now look at statistical approaches to describing a plasma instead of describing the exact position and velocity of every single particle. In the kinetic description, this is done by defining a distribution function $f(\vec{r}, \vec{v}, t)$ that describes the probability of finding a particle at the position $(\vec{r})$ with the velocity $(\vec{v})$ at some time t. For instance, particle density in real space $n(\vec{r}, t)$, which is a measurable quantity, can be obtained by integrating f over the velocity space

$$n(\vec{r},t) = \int d^3v f(\vec{r},\vec{v},t). \tag{2.12}$$

Similarly, the mean velocity of the particles $\vec{u}(\vec{r},t)$ can be defined as

$$\vec{u}(\vec{r},t) = \frac{1}{n}\int d^3v \vec{v} f(\vec{r},\vec{v},t). \tag{2.13}$$

Both n and $\vec{u}$ are called velocity moments of f, where n is the zeroth-order moment (f is multiplied with 1), and where $\vec{u}$ is a first-order moment (f is multiplied with $\vec{v}^1$). Similarly to the equation of motion for individual particles, the kinetic equation determines the distribution function f. A derivation of the kinetic equation can be found in [4]. If collisions between particles are neglected, the kinetic equation is described by the Vlasov equation

$$\frac{\partial f}{\partial t} + \vec{v}\cdot\frac{\partial f}{\partial \vec{r}} + \frac{q}{m}(\vec{E} + \vec{v}\times\vec{B})\cdot\frac{\partial f}{\partial \vec{v}} = 0, \tag{2.14}$$

which applies separately for different species of particles. Here, the acceleration $\vec{a}$ of the particles have been assumed to only be dependent on the Lorentz force. To include collisional effects, a collision operator is added to the right-hand-side of (2.14), which yields the Boltzmann equation

$$\frac{\partial f}{\partial t} + \vec{v}\cdot\frac{\partial f}{\partial \vec{r}} + \vec{a}\cdot\frac{\partial f}{\partial \vec{v}} = \left(\frac{\partial f}{\partial t}\right)_c. \tag{2.15}$$

The collision operator describes how f changes due to collisions. In summary, the research related to kinetic theory of a plasma is related to modeling the Vlasov and Boltzmann equation with different collision operators, and solving them (together with Maxwell's equations) with numerical methods to obtain f, which then can be used to quantify measurable quantities such as the particle density. Out of the statistical approaches to describe a plasma, formulations and models stemming from kinetic theory constitute the most first-principles frameworks, since the approaches that will be described next introduce more assumptions. However, due to there being 6 dimensions (3 spatial dimension and 3 velocity dimensions) plus time, kinetic models are generally computationally expensive.

2.2.4 Gyro-average models

As discussed in Section 2.2.1, a charged particle gyrates around a magnetic field line, a motion characterized by the gyro-frequency (2.4) and Larmor radius (2.3). In reactor-relevant conditions, this gyro-motion often occurs on much shorter timescales and smaller spatial scales than the primary plasma dynamics of interest. By gyro-averaging, the model captures the evolution of the guiding centers rather than the detailed orbital trajectories, effectively reducing the phase-space dimensionality from 6D to 5D. This significantly reduces computational demand. However, in regimes where the characteristic length scales of the plasma phenomena are comparable to the Larmor radius, Finite Larmor Radius (FLR) effects must be accounted for to maintain physical accuracy.

2.2.5 Multi-fluid theory

One approach to reduce the dimensionality of kinetic theory is to treat the plasma as a composite of fluids. Instead of solving equations for the distribution function f, fluid theory strives to solve for the macroscopic quantities directly, such as the particle density n (2.12) and mean velocity $\vec{u}$ (2.13). Therefore, it is not necessary to solve for the velocity space, which reduces the dimensionality by 3.

As mentioned previously, the macroscopic quantities are related to moments of the distribution function, and a moment $\langle\psi\rangle$ is defined as the velocity average of the function $\psi(\vec{v})$

$$\langle\psi\rangle = \frac{1}{n}\int d^3v\psi(\vec{v})f, \tag{2.16}$$

where the zeroth-order moment is obtained by setting $\psi = 1$, and where the first-order and second-order moments correspond to $\psi = \vec{v}$ and $\psi = \vec{v}\vec{v}$ respectively. The general moment equation is obtained by integrating the Boltzmann equation from kinetic theory with respect to velocity, which yields

$$\frac{\partial}{\partial t}(n\langle\psi\rangle) + \nabla\cdot(n\langle\vec{v}\psi\rangle) - \frac{nq}{m}\left\langle(\vec{E}+\vec{v}\times\vec{B})\cdot\frac{\partial\psi}{\partial\vec{v}}\right\rangle = \frac{\partial}{\partial t}(n\langle\psi\rangle)_c. \tag{2.17}$$

For instance, the zeroth order ($\psi = 1$) moment equation is the continuity equation [4]

$$\frac{\partial n}{\partial t} + \nabla\cdot(n\vec{u}) = 0, \tag{2.18}$$

which includes the particle density n and the mean velocity $\vec{u}$. Other macroscopic quantities, such as kinetic pressure (temperature) and heat flux, can be obtained through higher order moment equations. However, a challenge with solving moment equations is that each equation includes the next higher-order moment, which can be seen in the second term in both (2.17) and (2.18). For instance, the continuity equation, which is the zeroth-order moment equation, includes the mean velocity $\vec{u}$, which is a first-order moment. This creates an infinite chain of dependent equations, which must be broken through assumptions and approximations, such that at some point a moment equation is not dependent on the next-order moment. Such approximations are referred to as moment closure since they lead to a finite set of equations to be solved. It is also important to note that each particle species has their own set of moment equations since they are treated as separate fluids.

Similarly to the kinetic theory case, researchers model plasmas on computers with fluid theory and experiment with different closures and numerical methods to solve the moment equations and Maxwell's equations. Fluid theory is not of equally high fidelity compared to kinetic theory, since it introduces more assumptions. However, it is less computationally expensive due to the dimensionality reduction. Moreover, by computing the moments of gyro-kinetic models, it is possible to obtain gyro-fluid models where the dimensionality is reduced further.

2.2.6 Magnetohydrodynamics

Magnetohydrodynamics (MHD) is a theoretical description that modifies fluid theory such that the plasma can be treated as one conducting fluid. For instance, the plasma can be characterized by parameters like the mass density ρ_m and the center-of-mass velocity $\vec{V}$, which are defined as

$$\rho_m := \sum_\alpha n_\alpha m_\alpha, \tag{2.19}$$

$$\vec{V} := \frac{1}{\rho_m} \sum_\alpha n_\alpha m_\alpha \vec{u}_\alpha, \tag{2.20}$$

where the summation over α refers to the sum over the different particle species, and where $\vec{u}_\alpha$ again is the mean velocity, in this case for species α. The MHD equations can be obtained by adjusting the moment equations from multi-fluid theory according to these new MHD parameters. For instance, the first MHD equation can be derived by multiplying the continuity equation with the mass, and by adding the equations for the different species

$$\frac{\partial \rho_m}{\partial t} + \nabla \cdot (\rho_m \vec{V}) = 0, \tag{2.21}$$

which reflects the conservation of mass. Other MHD equations can be obtained by modifying higher order moment equations with MHD parameters and combining them with Maxwell's laws. Inherently, the equations in a specific MHD model depend on the closure that is used in the multi-fluid model that the MHD model is based on. Moreover, different formulations of MHD exist, such as ideal MHD and resistive MHD. The former treats the plasma as, for instance, a perfect conductor, while the latter accounts for resistive effects.

As in the previous cases, MHD equations can be solved with numerical methods for different scenarios. These models are generally computationally cheaper compared to their multi-fluid model counterparts, since fewer parameters are needed to be solved for. Additionally, resistive MHD models are generally more computationally expensive compared to their ideal MHD model counterparts, due to added complexity in the equations. That said, the computational requirements of models can also vary due to other factors. For example, a 'linear' fluid model may be computationally cheaper than a 'nonlinear' MHD model. In the next subsection, we will discuss what is meant by a linear model and a nonlinear model.

2.2.7 Perturbations, linear models, nonlinear models, and plasma instabilities

The theoretical descriptions of plasmas discussed so far can be used, for instance, to calculate the steady state of a plasma. In practice, this means setting $\partial/\partial t = 0$ in the equations employed to characterize the plasma. However, small perturbations from the steady state can lead to waves and instabilities. These can be simulated by introducing perturbations in the plasma parameters when solving a chosen set of plasma equations numerically. Another approach

is to analyze the effect of perturbations analytically. For instance, a plasma parameter $\vec{X}$ can be expanded by assuming that it is a sum of the steady state/background solution $\vec{X}_0$ and a series of perturbations such that

$$\vec{X} = \vec{X}_0 + \epsilon\vec{X}_1 + \epsilon^2\vec{X}_2 + \cdots, \tag{2.22}$$

where ϵ is a small parameter. Let us now consider the continuity equation (2.18) and insert (2.22) with first order perturbations (only $\vec{X}_0$ and $\vec{X}_1$), such that the continuity equation becomes

$$\frac{\partial n_0}{\partial t} + \epsilon\frac{\partial n_1}{\partial t} + \nabla\cdot(n_0\vec{u}_0 + \epsilon n_0\vec{u}_1 + \epsilon n_1\vec{u}_0 + \epsilon^2 n_1\vec{u}_1) = 0. \tag{2.23}$$

Here, the terms that are not multiplied with ϵ represent the steady state solution. We now wish to solve the ϵ^1 equation, where we discard higher order ϵ-terms. We also use that $\vec{u}_0 = 0$ since the average velocity vector of the particles is zero in steady state. This gives the result

$$\frac{\partial n_1}{\partial t} + \nabla\cdot(n_0\vec{u}_1) = 0, \tag{2.24}$$

where we see that the perturbations of the first order, n_1 and $\vec{u}_1$, have been linearized due to the exclusion of higher order ϵ-terms. Let us now, for example, look at monochromatic wave solutions of n_1 and $\vec{u}_1$ such that $\vec{X}_1 \sim \exp[i(\vec{k}\cdot\vec{r} - \omega t)]$, where ω is the angular frequency of the wave, and where $\vec{k}$ is the wave vector. We may then use that $\nabla \sim i\vec{k}$ and $\partial/\partial t \sim -i\omega$, such that (2.24) becomes

$$-i\omega n_1 + in_0\vec{k}\cdot\vec{u}_1 = 0, \tag{2.25}$$

which is an eigenvalue equation. The same principle is applied to all equations in a coupled system, which leads to a set of linear equations that can be solved analytically or numerically. Specifically, we can solve for relations between ω and $\vec{k}$ that provide non-trivial solutions to $\vec{X}_1$. In other words, we seek to find the eigenvalues

$$\omega(\vec{k}) = \omega_r(\vec{k}) + i\gamma(\vec{k}). \tag{2.26}$$

The real part of ω is called the real frequency ω_r and the imaginary part of ω is called the growth rate γ. If γ is negative, the amplitude of the perturbation wave attenuates, and if γ is positive, the amplitude of the wave grows over time. The latter case can be problematic as it may lead to growing instabilities in the plasma that are detrimental for the confinement. Therefore, researchers may perform simulations with perturbations to, for instance, investigate the growth rate of instabilities as a function of the spatial scale dictated by $\vec{k}$.

The difference between linear models and nonlinear models is that nonlinear terms, such as the ϵ^2-term in (2.23), are neglected in linear models. This works well when the associated perturbations are small. However, if there is a positive growth rate, the perturbation will eventually become large enough such that nonlinear terms can no longer be neglected. In reality this leads to coupling

between the parameters which, for instance, can lead to a saturation of the the amplitude of the perturbations. Nonlinear models are more accurate for such cases, but also more costly to solve numerically.

2.3 Plasma physics: Concluding remarks

Understanding theoretical descriptions of plasmas is important for understanding why modeling certain aspects accurately is difficult. For instance, the pedestal, which represents the main area where machine learning is applied in this thesis, is a typical example of this. We will discuss the pedestal more in the next chapter, but in short, challenges arise due to insufficiency in more easily implementable and computationally cheaper frameworks such as ideal MHD. For instance, the pedestal is associated with kinetic effects, FLR-effects, non-linear phenomena, and also coupling between neutral fueling particles and the plasma. By using machine learning, we may bypass some of these challenges by learning from data directly to enable predictive models, and we may use interpretability in an attempt to further understand the different phenomena through data analysis.

Chapter 3

The tokamak

We now turn to the magnetic confinement device that is of relevance in the scope of this thesis: the tokamak [43]. This is the most developed reactor design for magnetic confinement fusion research, and here we focus on its key components and inherent features. Additionally, since the work in this thesis is specifically related to the Joint European Torus (JET) tokamak in Culham, UK, most examples will be based on this device. In the end of this chapter, we describe the pedestal.

3.1 Power plant concept

The goal of a future power plant based on the tokamak concept is to create plasma conditions such that fusion reactions occur at a high rate, and to maintain those conditions for sufficiently long times. As mentioned in the introduction, certain fusion reactions produce neutrons with high kinetic energy. These escape the plasma due to their charge neutrality, and in a future power plant, they will heat a coolant that will drive a turbine, which in turn will generate electricity.

3.2 Plasma geometry

The idea of the tokamak is to create a closed magnetic field geometry, such that charged particles in a plasma are confined in a torus-like shape, or more informally, a donut shape. The geometry of a torus is illustrated in Figure 3.1.

Note that in Figure 3.1, the cross section of the torus is circular to simplify the explanation of the coordinates. In real tokamaks, the shape of the cross section of the plasma can be varied. Therefore, in this thesis the plasma configuration is referred to as 'torus-like'.

To obtain a torus-like shaped plasma, the first requirement is a magnetic field in the toroidal direction B_Φ. This can be achieved by placing several toroidal field coils in a circle, as illustrated in Figure 3.2. In the JET tokamak, a

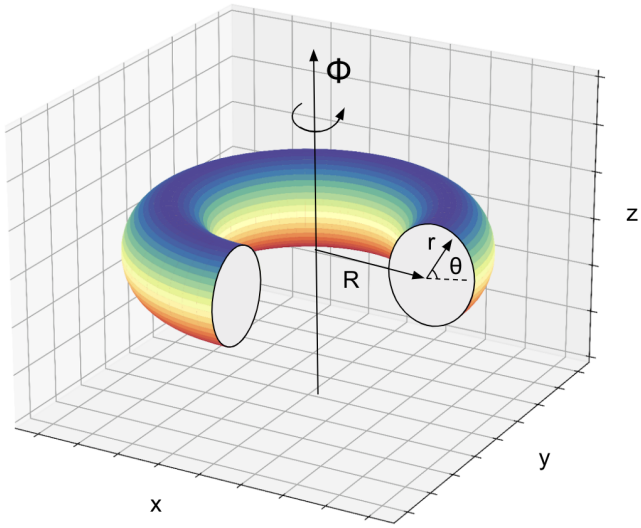


Figure 3.1: The geometry of a torus, which is characterized by toroidal angle Φ, poloidal angle θ, major radius R, and minor radius r. Here, x, y, and z represent Cartesian spatial coordinates.

toroidal field on the order of $B_\Phi = 3.5$ T can be achieved, which is approximately 100 000 times stronger than the magnetic field of Earth at the equator [1].

As discussed in the previous chapter, curved magnetic field lines lead to a nonzero magnetic field gradient, in this case in the R-direction in Figure 3.1. In addition to a curvature drift, this will lead to a ∇B drift of the particles in the z-direction. Specifically, the toroidal field in a tokamak varies as $B_\Phi \propto 1/R$. Since the ∇B drift moves electrons and ions in opposite directions, this will lead to an electric field in the z-direction, which in turn leads to an $\vec{E} \times \vec{B}$ drift in the R-direction. There are also other drifts, such as the curvature drift, that negatively affects confinement in this design. Hence, with all these drifts in mind, we can arrive at the conclusion that a tokamak with only a toroidal field cannot confine particles well.

To solve this issue, a poloidal magnetic field component B_θ can be introduced. In this setup, particles do not only travel around the toroidal axis, but also around the poloidal axis such that the resulting motion of the guiding center exhibits a helical pattern. For instance, if there is a drift in the z-direction upwards in Figure 3.1, then for the upper half of the poloidal orbit, the particle will drift in the outward direction (r increasing), and for the lower part of the poloidal orbit, the particle will drift towards the plasma (r decreasing). In total, the drift approximately cancels out over the course of one full poloidal turn, which leads to improved confinement.

There are different alternatives for creating a poloidal magnetic field component. In tokamaks, this is done by generating a plasma current I_P in the

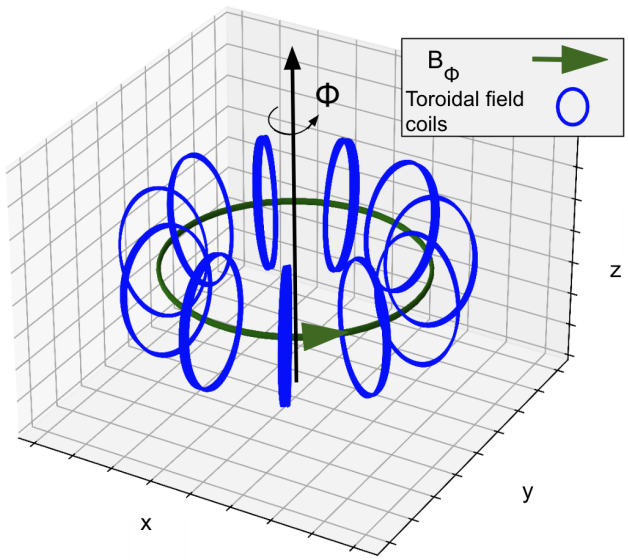


Figure 3.2: An illustration of the toroidal field coils in a tokamak. By running current through the coils, a toroidal magnetic field B_Φ can be generated.

toroidal direction, which is illustrated in Figure 3.3. For instance, one of the methods is to induce a toroidal electric field by gradually increasing a current in a solenoid placed in the middle of the tokamak. Due to the induced electrical field, a toroidal current is generated. Since the solenoid current must keep increasing in order to uphold induction, there is a limit to how long the plasma current can be sustained, which consequently limits the lifetime of the plasma. Therefore, tokamak operations that rely only on an induced plasma current are run in pulses. There are however other non-inductive mechanisms that can contribute to the plasma current, such as the plasma self generating 'bootstrap current' [4], and currents driven by the heating systems that we explore later in this chapter.

While a poloidal field is necessary for good confinement, it is important to emphasize that the strength of the poloidal field in relation to the toroidal field matters to achieving stable plasmas. The safety factor q is used in tokamaks and other magnetic confinement devices to monitor the relation between the poloidal field and the toroidal field

$$q = \frac{r \cdot B_\Phi}{R \cdot B_\theta}. \tag{3.1}$$

Essentially, q describes how many times a magnetic field line wraps around the tokamak the "long way" (toroidally) for each time it wraps around the "short way" (poloidally). For instance, instabilities are associated with rational q-values, but also with low q-values near the plasma edge.

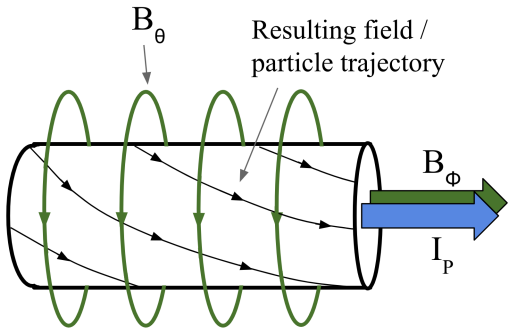


Figure 3.3: A segment of a tokamak plasma (shown as a straight cylinder although it is curved in a tokamak). A toroidal plasma current I_P leads to a poloidal magnetic field B_θ, which together with the toroidal field B_Φ results in helical fields lines. In this illustration, the gyrating Larmor motion of the particles are not drawn.

3.3 Reactor chamber and gas fueling

A tokamak contains the plasma in a chamber, and much like the plasma itself, the chamber adopts a toroidal structure. This configuration is necessary to allocate space for the solenoid and segments of the toroidal field coils in the central region of the torus.

Prior to initiating a pulse, vacuum pumps are used to evacuate the chamber of gasses. In fact, the tokamak is kept close to vacuum the whole time it is operated, since it can take weeks to obtain the required near vacuum conditions. When initiating a pulse, gasses of the desired particle species are injected to create the initial plasma. Additional gas can be injected periodically or continuously as fuel is lost during fusion reactions and due to particle transport. Injection of gasses can also be used to intentionally cool the plasma. For instance, in the case of disruptions [44], where the plasma loses its confinement and stability, it is desirable to inject massive amounts of gas to create a more benign disruption that induces less stress on the tokamak.

Impurities, which are elements with higher atomic numbers than the main fuel ions, can also be injected to the plasma for cooling and stabilization purposes. A useful parameter for quantifying the purity of the plasma is the effective atomic number

$$Z_{eff} = \sum_{\alpha} \frac{n_\alpha}{n_e} Z_\alpha^2, \tag{3.2}$$

where n_α and Z_α are the number density and atomic number of the ion species α respectively, and n_e is the electron number density.

The choice of material for the plasma facing wall is also an important consideration due to the extreme conditions inside the chamber, but also since

particles from the wall can contaminate the plasma, where the impact depends on the particle species. In the JET tokamak, the wall consists of tungsten and beryllium due to properties such as high melting points, heat conductivity, and resistance to erosion. As mentioned in the introduction, future power plants may also include a lithium blanket on the wall in the reactor chamber to enable continuous breeding of tritium during a pulse.

In Figure 3.4, the reactor chamber of the JET tokamak is shown.

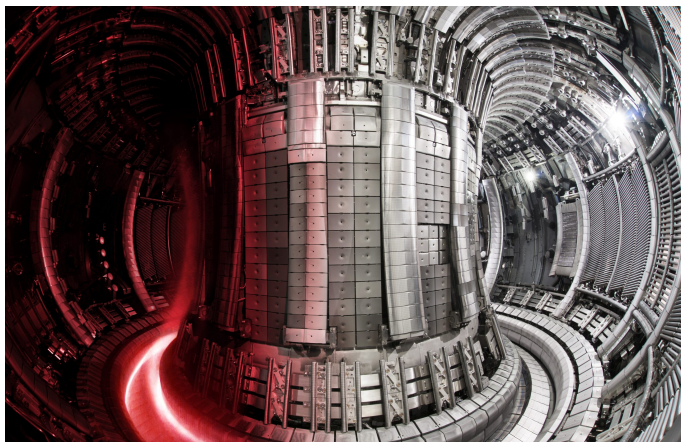


Figure 3.4: The inside of the reactor chamber of the Joint European Torus (JET) tokamak at Culham, UK, with a superimposed image of the hot plasma to the left; credit UKAEA; courtesy of EUROfusion. For size reference, two people standing on top of each other could easily fit in the reactor chamber.

3.4 Plasma heating

Optimally, in future power plants the high kinetic energy of the charged fusion products will be sufficient to keep heating the plasma and trigger new fusion reactions without additional external heating, resulting in 'ignition' of a burning plasma. The conditions needed to achieve ignition is dictated by the Lawson criterion [4], which can be expressed as

$$nT\tau_E \geq \frac{12}{E_{ch}} \frac{T^2}{\langle \sigma v \rangle}, \tag{3.3}$$

where n is the particle density, and where T is the temperature. τ_E is the energy confinement time (measures the rate at which the plasma loses energy). E_{ch} is the average kinetic energy of the charged fusion products in a given reaction, for instance, the He^4-ions (α-particles) in D-T reactions (3.5 MeV). $\langle \sigma v \rangle$ is the velocity average of the product between the reaction cross section and the particle velocity. The left-hand side of (3.3) is the so-called 'triple product' mentioned in the introduction, which fusion researchers strive to increase to achieve an efficient reactor.

However, a burning plasma has not yet been achieved in magnetic confinement fusion, and a reactor that operate below the Lawson criterion with external heating can still be sufficiently efficient if it provides a high Q-value. Moreover, external heating will be necessary regardless in future reactors for control and for heating at the startup stage of a pulse.

3.4.1 Ohmic heating

When currents run through the plasma, heat is generated due to resistivity (collisions), which is therefore called ohmic heating. However, the plasma resistivity η is dependent on the temperature. Specifically, as discussed in Chapter 2, the collision frequency goes down with increased temperature. Hence, when the temperature increases, the resistivity drops, which reduces the effect of ohmic heating. This puts an upper bound on the temperature that can be achieved with ohmic heating alone, and this temperature is not high enough for a sufficiently high rate of fusion reactions. Nevertheless, ohmic heating is still useful, in particular at the initial stages of a pulse.

3.4.2 Neutral beam injection (NBI)

One of the auxiliary heating systems in a tokamak is the Neutral Beam Injection (NBI) system. The concept of NBI is that fast neutral particles are shot into the plasma and ionized through collisions with plasma particles. The injected particles need to be neutral to not be shielded by the magnetic field lines in the tokamak before reaching the plasma. An important aspect of NBI heating, which needs careful consideration, is where the beam deposits its energy in the plasma. For instance, too low energy will lead to collisions and energy deposits only at the plasma edge, and too high energetic beams will shine through the plasma and hit, and potentially damage, the reactor walls [43].

If a NBI source is used for perpendicular injection, along the toroidal axis, it can generate a toroidal plasma rotation. Additionally, when the fast neutral particles are ionized, the ions from the neutral beam have higher momentum in the toroidal direction compared to the electrons from the same neutral beam due to their larger mass. In essence, this impacts the total plasma current since the electrons lose their toroidal velocity more rapidly.

In JET, the NBI system can shoot fast beams into the plasma up to a power of approximately 34 MW [45].

3.4.3 Radio frequency (RF) heating

The other main external heating in tokamaks is via Radio Frequency (RF) waves. RF is based on launching electromagnetic waves into the plasma, which are tuned to be effectively absorbed by a resonance mechanism associated with the motion of the plasma particles. The two principal resonance mechanisms are cyclotron absorption and Landau damping [4]. In the former the wave frequency is tuned to resonate with the cyclotron motion of one or several particle species in the plasma, while Landau damping occurs for particles

traveling with a parallel velocity matching the parallel phase velocity of the waves. The approaches related to cyclotron absorption are called ion cyclotron resonance heating (ICRH) and electron cyclotron resonance heating (ECRH) [43]. Because the magnetic field in a tokamak varies roughly as $1/R$, and the cyclotron frequency is proportional to the magnetic field, one can choose approximately where the power is absorbed. Moreover, schemes based on, for instance, Landau damping are used for current drive applications, such as Lower Hybrid Current Drive (LHCD) and Fast Wave Current Drive (FWCD) [46].

3.5 Diagnostics

Diagnostics play a crucial role in tokamak research by providing measurements for control, optimization, and analysis of plasmas. These include a wide variety of techniques for different aspects, such as Thomson scattering systems [47] and interferometers [48] for temperature and density measurements, and bolometric systems for obtaining spatial distributions of the radiated power of the plasma [49]. There are also, for instance, systems for magnetic diagnostics, neutron diagnostics, and for monitoring the plasma-facing components.

The diagnostics can occupy a substantial amount of space in the design of a tokamak, and future reactors that are optimized for delivering power may not include all of the diagnostics present in current experimental devices. This is also because most current day diagnostics are not sufficiently neutron resistant. Nevertheless, the quality and availability of diagnostic measurements are fundamental for data-driven tokamak research, at least when relying on experimental data rather than model-generated synthetic data.

3.6 Plasma profiles

3.6.1 Two-dimensional profiles

When studying tokamak diagnostics, such as the spatial distribution of plasma temperature and density, the complete three-dimensional representation can often be too comprehensive. Fortunately, due to the toroidal symmetry in a tokamak, one can instead study the two-dimensional (2D) cross section of the plasma.

One can, for instance, study the 2D projection of 'flux surfaces' [4], [43], which are surfaces where the magnetic flux remains constant. An illustration of flux surfaces is shown in Figure 3.5. Here, the field lines are arranged such that the flux surfaces are closed in the core, but open in the outer edge of the plasma. It may seem counter-intuitive to generate open flux surfaces such that particles can escape the plasma and hit *divertor plates*. However this allows for controlled exhaust of particles and energy, which is important both for maintaining the desired conditions in the plasma, as well as for reducing the damage of other plasma facing components. Specifically, the divertor plates are designed to handle larger heat loads compared to other components. The

contour where the flux surfaces go from closed to open is referred to as the Last Closed Flux Surface (LCFS), or the separatrix, and the region outside the LCFS with open flux surfaces is referred to as the Scrape-Off layer (SOL).

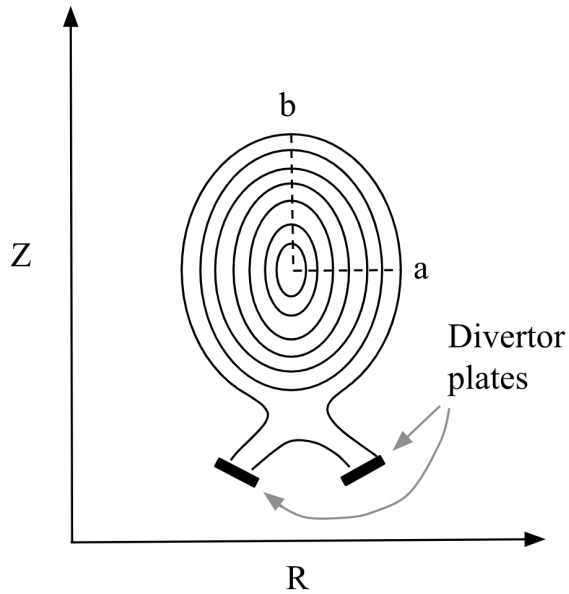


Figure 3.5: Illustration of the flux surfaces in a 2D cross section of a tokamak plasma. The parameters a and b represent the minor radius and height of the plasma respectively. Z and R are the same spatial coordinates as in Figure 3.1.

As illustrated in Figure 3.5, the plasma does not have to be perfectly circular. In this case, the plasma height b is larger than the minor radius a, which leads to an elongated plasma, where plasma elongation is defined as

$$\kappa = \frac{b}{a}. \tag{3.4}$$

The plasma shape is also characterized by triangularity. For instance, the upper triangularity δ_{up} is defined as

$$\delta_{up} = \frac{R_{geo} - R_{upper}}{a}, \tag{3.5}$$

where R_{upper} is the major radius at the highest vertical point of the LCFS. R_{geo} is the major radius at the geometric axis, which is defined as

$$R_{geo} = \frac{R_{max} + R_{min}}{2}, \tag{3.6}$$

where R_{max} and R_{min} correspond to the maximum and minimum major radius of the LCFS. In the example illustrated in Figure 3.5, the upper triangularity is 0 since $R_{geo} \approx R_{upper}$. A triangulated plasma at JET is more shaped like the letter 'D' compared to the elongated cylinder in Figure 3.5.

3.6.2 One-dimensional profiles

Due to the fast motion of particles along the field lines in a tokamak, many properties in a plasma are approximately constant on the flux surfaces described in the previous section [4], [43]. Therefore, it is often sufficient to further simplify the analysis to one spatial dimension (1D) by looking at flux surface averages. For instance, Figure 3.6 illustrates the 1D profile of the flux surface average electron temperature T_e. In such profiles, it is common to have a flux label ψ on the x-axis, and it is normalized to be 1 at the LCFS / separatrix. A simplified description of ψ is that it acts as a proxy for how far we are away from the center of the plasma.

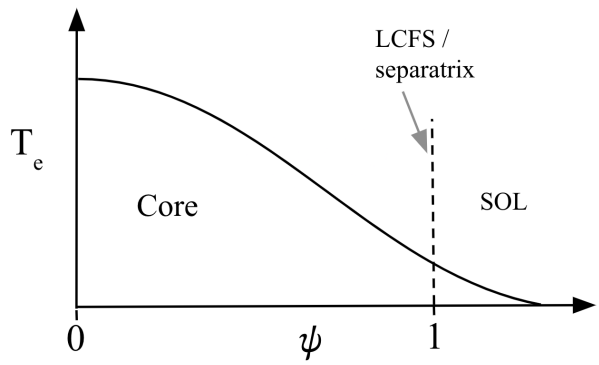


Figure 3.6: An illustration of an 1D flux surface average electron temperature profile. Here, the temperature T_e is drawn to be highest in the core of the plasma, and it decreases towards the SOL region.

It is common to study different quantities in such 1D profiles, such as the safety factor q, ion temperature T_i, electron and ion density n_e and n_i, current density j, electron and ion pressure p_e and p_i. Beyond flux surface quantities, it can also be useful to analyze global quantities, such as the ratio of the volume integrated plasma pressure to the magnetic pressure at the magnetic axis, which is defined as

$$\beta = \frac{p}{B^2/2\mu_0}, \tag{3.7}$$

where μ_0 is the magnetic permeability in vacuum. β is an important parameter for both stability and confinement, and it is also often expressed in terms of its normalized version

$$\beta_N = \beta \frac{aB_T}{I_p}. \tag{3.8}$$

In simulations, it can be of interest to find both steady state solutions and evolutions of 1D profiles since their characteristics reveal how the plasma behaves for different scenarios.

3.7 Heat and particle transport in a tokamak

For most of this thesis, it has been described how particles are confined to magnetic field lines if we neglect drifts. However, this is not the complete picture. In this section, we discuss other mechanisms that lead to heat and particle transport perpendicular to the magnetic field lines, and how they lead to degraded confinement.

3.7.1 Classical transport

This type of transport is based on collisions between charged particles in the plasma. By calculating average collision times and treating the particle movement as a random walk process [6], the total travel distance of heat and particles after a certain time can be calculated if the spatial step size is known. Since particles gyrate in tokamak plasmas, the spatial step size is on the same order as the Larmor radius. Classical transport typically contributes a minor fraction to the overall transport in a tokamak plasma [43], which means that classical transport is not the main obstacle in achieving high confinement.

3.7.2 Neoclassical transport

Classical transport can be extended to neoclassical transport by including effects from the toroidal geometry, such as $B_\Phi \propto 1/R$. When particles travel from a lower magnetic field to a larger field, their perpendicular velocity must increase as the magnetic moment (2.11) of a particle is a conserved quantity. For energy to also be conserved, the parallel velocity must decrease. Some particles with too low initial parallel velocity on the low field side will at some point at higher field strength reach $v_\parallel = 0$, and then bounce back in the opposite direction. This leads to a fraction of particles that are trapped and travel back and fourth on the low field side in orbits referred to as banana orbits due to how they look in a 2D projection. The width of these banana orbits is larger than the regular step size in classical transport, which makes particles collide with other particles that are further out compared to the classical transport case. Hence, compared to classical transport, heat is leaked out more quickly which makes confinement more challenging.

3.7.3 Turbulent transport

Turbulent transport [50] is conceptually different from the previous transport processes as it is not rooted in collisions, but rather arises from the complex behavior of plasma fluctuations. There are different mechanisms that can contribute to turbulent transport. For instance, micro-instabilities can lead to a large heat and particle flux that almost always exceeds the transport contribution from classical and neoclassical transport in current machines. Therefore, understanding and controlling turbulence in tokamaks is a prioritized research topic in MCF.

The turbulent transport from micro-instabilities can be calculated by solving differential equations describing perturbations as discussed in Section 2.2.7.

In simulations, this is done numerically, either through solving the nonlinear equations or by discarding the nonlinear terms, although the previous is far more computationally costly. Certain models, such as EDWM [51], TGLF [52], and QuaLiKiz [53] are quasi-linear models. These calculate the growth rate of instabilities with the linear approach, and then apply a saturation rule to account for the nonlinear interactions resulting in a transport estimation without having to explicitly solve nonlinear equations.

The most common turbulent transport contributing instabilities, which are also referred to as drift waves, are

- Ion Temperature Gradient (ITG) mode - As the name suggests, this instability is driven by gradients in the ion temperature.
- Electron Temperature Gradient (ETG) mode - Similar to the previous but due to gradients in the electron temperature.
- Trapped Electron Mode (TEM) - Arises from the fact that some electrons are trapped in banana orbits and therefore cannot cancel out fluctuations in the electric field to the same extent as if there were no trapped electrons.

3.8 The pedestal

In the 1980s, it was discovered that confinement, and hence performance, suddenly increased at high applied heating power in the ASDEX tokamak, at Garching, Germany [37]. It turns out that transport had been suppressed near the separatrix, which resulted in steep temperature and density gradients in the profiles, as illustrated in Figure 3.7. This operational regime was named the High-confinement mode (H-mode), and the elevated temperature and density near the separatrix led to this region being called the pedestal due to its visual shape. It is believed that the transport suppression in the pedestal arises, for instance, due to stronger shear in the $\vec{E} \times \vec{B}$ drift velocity at the plasma edge [54] (which breaks up turbulent structures), although all details are not fully understood. Nevertheless, in existing devices, the pedestal generally yields a factor of ~ 2 improvement in confinement and stored energy relative to the pedestal-free Low-confinement mode (L-mode) at the same input power [54].

Essentially, the forming of the pedestal consists of two parts; 1) local transport suppression which leads to its build-up; 2) rapid drops in the pedestal top temperature and density due to instabilities called edge localized modes (ELMs) [55]. This results in a cyclic pattern of the pedestal called the ELM-cycle [56], which is illustrated in Figure 3.8.

The ELMs impose an upper limit on the temperature and density at the top of the pedestal, which is influenced by machine parameters such as total plasma current and plasma shape [40]. For integrated modeling applications, these pedestal top values are important as they can be used to act as boundary conditions when simulating phenomena in the core. Consequently, there is a demand for models capable of predicting these pedestal top values from machine parameters.

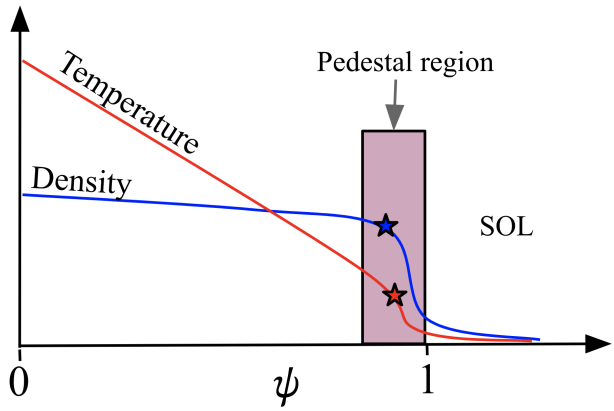


Figure 3.7: An illustration of the pedestal in tokamaks. In the pedestal region, which also can be referred to as the plasma edge, transport is suppressed such that steep gradients are formed. The values at the top of the pedestal are highlighted as stars. The width of the pedestal is the radial distance between the top and the bottom of the steep curve of the pedestal.

The pedestal also has implications for heat load management. Specifically, certain types of ELMs, when triggered, can release a considerable amount of energy in a brief duration [40]. This poses a problem with respect to plasma-facing components, especially the divertor plates. In the context of larger future machines, such as ITER, more energy is going to be stored in the plasma. Therefore, mitigating ELM types associated with excessive energy release is crucial. Optimally, further understanding of the pedestal may help design operational scenarios where a balance in transport is achieved before ELMs are triggered or where less disruptive types of ELMs are intentionally induced [57]. Hence, improved understanding of both the build-up of the pedestal as well as the ELM triggering mechanism will be important.

3.8.1 Pedestal stability

In this discussion about pedestal stability we will focus on *type 1* ELMs, which is the most straightforward ELM type to produce, and hence the most studied. This is also the main ELM type associated with high energy release that will be unacceptable in future devices. We will discuss other ELM types, and how they are characteristically different from type 1 ELMs, later in this chapter.

The general framework for theoretically investigating the stability of type 1 ELMs is MHD. Specifically, steep pressure gradients that form in the pedestal lead to MHD instabilities referred to as ballooning modes. Moreover, high bootstrap current in the pedestal, which is driven by high temperature and density gradients, destabilize MHD modes referred to as peeling modes. Coupling between these two modes results in a stability boundary that can be defined in J_{ped}/p'_{ped} space, where J_{ped} is the current density at the pedestal,

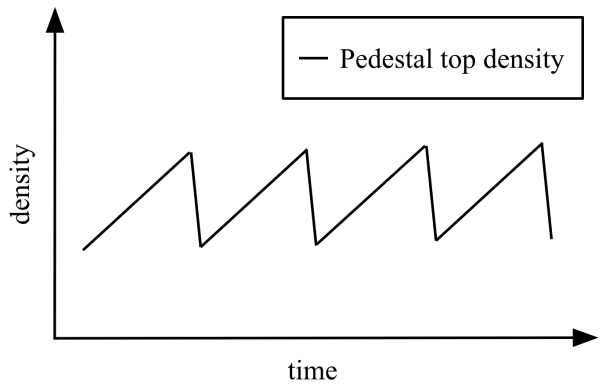


Figure 3.8: Illustration of the pedestal ELM-cycle. Note that the build-up does not necessarily have to be linear as in this illustration.

and where p'_{ped} is the pressure gradient at the pedestal [58]. This is illustrated in Figure 3.9, which shows a so-called peeling-ballooning (PB) mode diagram. The shape of the stability boundary in PB-diagrams depends on various factors, such as the plasma triangularity, the edge safety factor, and the width of the pedestal to mention a few [59].

To obtain a PB-diagram, such as the one illustrated in Figure 3.9, one needs to generate MHD equilibria, which in turn can be obtained by solving a force balance equation

$$\vec{\nabla} p = \vec{j} \times \vec{B} \tag{3.9}$$

where p is the pressure, $\vec{j}$ is the electric current, and $\vec{B}$ is the magnetic field. Specifically, a set of equilibria is obtained by varying J_{ped} and p'_{ped} while keeping other parameters constant. For each pair of J_{ped} and p'_{ped}, the growth rate of the PB-modes γ_{MHD}, which indicates stability or instability, can be calculated. In more detail, γ_{MHD} is compared with diamagnetic stabilization [60], [61], which sets a stability threshold, such as

$$\gamma_{\mathrm{MHD}} > \omega_*^{\max}/4 \tag{3.10}$$

where $\omega_*^{\max}$ is the peak value of the diamagnetic frequency in the pedestal.

When accessible, experimental data can be used as inputs to the MHD equilibria calculation, which gives a type 1 ELM PB-diagram for a specific experimental scenario. If the experimental values of the pressure gradient and the current density also are available, one can pinpoint where the experimental pedestal is situated in the PB diagram. For instance, if the experimental values are taken just before the ELM crash, one can investigate if the point is located near the stability boundary, which would be indicative of agreement between the theory and the experimental data. In general, there has been strong agreement between experimental values and the stability boundary calculated

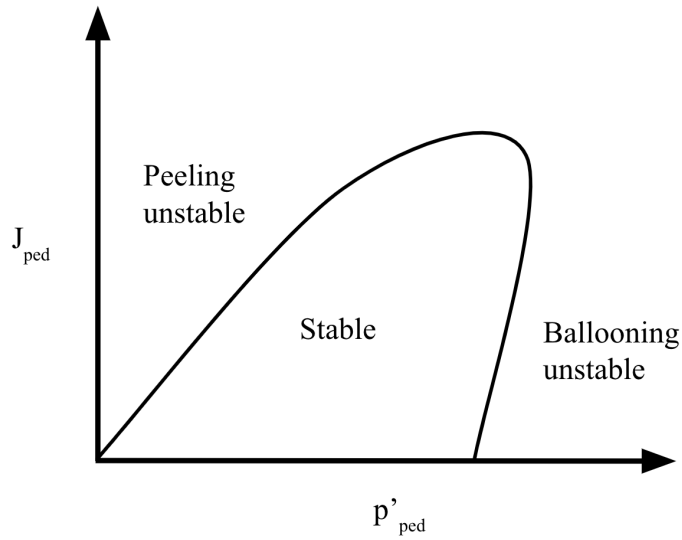


Figure 3.9: Illustration of a peeling-ballooning (PB) mode stability diagram.

with ideal MHD codes, such as MISHKA [62], at different devices [63]–[67]. There has however also been disagreement [68], [69], for instance, when resistive effects are non-negligible. That said, recent work shows promising results in mitigating these experimental and theoretical discrepancies by implementing resistive MHD [70].

From a pedestal prediction perspective, stability analysis using MHD provides a path for predicting the pedestal pressure gradient theoretically given a pedestal width. However, one usually does not know the width beforehand, which means that we have two unknowns. Hence, we need supplementary methods for adding an additional equation or constraint to obtain both the pedestal width and gradient, that also can be used to obtain the pedestal top pressure.

3.8.2 Pedestal transport

A common approach to obtain a second relationship between the pedestal top and width (and hence gradient) is via analysis of transport in the pedestal, or via empirical relationships. Modeling pedestal transport is however not trivial, for instance, due to challenges in separating spatial scales in the pedestal, which causes ordering in MHD and gyro-kinetics to break down [54]. Moreover, the fact that collisionality can vary substantially between the top and bottom of the pedestal, and that electrostatic assumption becomes invalid when the pressure gradient is near the ballooning critical gradient, does not make the theoretical treatment simpler.

However, certain assumptions can be made. For instance, in the pedestal

prediction model EPED [41], [71], the pedestal pressure gradient is assumed to be constrained by ballooning modes from kinetic theory referred to as KBMs. This results in the scaling

$$\Delta \propto \beta_{\theta,\mathrm{ped}}^{1/2} \tag{3.11}$$

where Δ is the pedestal width, and $\beta_{\theta,\mathrm{ped}}$ is the poloidal beta at the top of the pedestal. Since β is proportional to the pressure, the expression provides a relationship between the pedestal top pressure and the width. This relationship can in turn be used together with the PB model described in the previous subsection, where the intersection provides the final pedestal width, gradient, and top pressure.

In the first EPED1 model, the KBM pedestal width expression is further simplified to

$$\Delta = c_1 \beta_{\theta,\mathrm{ped}}^{1/2} \tag{3.12}$$

where c_1 is a constant found by fitting the expression to experimental data. For instance, $c_1 = 0.076$ shows good agreement at the DIII-D tokamak, and for JET pulses with low gas fueling [68], while $c_1 = 0.1$ represents the best fit at the AUG tokamak [72]. There are however also studies where the pedestal width does not show a $\Delta \propto \beta_{\theta,\mathrm{ped}}^{1/2}$ dependency, such as at the TCV tokamak [73].

Other studies have found alternative empirical relationships related to the pedestal structure that has proven useful for pedestal predictions. For instance

$$\langle \nabla T_e \rangle / T_{e,ped} = \text{constant} \tag{3.13}$$

is an empirical relationship that has been exploited in the IMEP framework [74]. Here, $\langle \nabla T_e \rangle$ is the average electron temperature gradient in the pedestal, and $T_{e,ped}$ is the electron temperature at the top of the pedestal. Equation (3.13) has proven to provide accurate pedestal predictions at the AUG tokamak. Moreover, accurate predictions at the JET and C-mod tokamaks have been obtained by modifying the expression to $R\langle \nabla T_e \rangle / T_{e,ped} = \text{constant}$, where R is the major radius of the tokamak [75]. That said, since these expressions are indeed empirical and not based on a complete understanding of the underlying physics mechanism, there is no guarantee that they will scale well to future devices.

3.8.3 Pedestal density prediction

A limitation with PB MHD stability analysis is that it only predicts relationships for the pedestal pressure, and not the pedestal temperature and density explicitly. Hence, in order to obtain, for instance, the pedestal temperature $T_{e,ped}$, one needs the pedestal density $n_{e,ped}$ as an input.

In recent years, progress has been made in predicting $n_{e,ped}$ separately, such that it does not need to be assumed or taken from experiment. For instance, the neutral penetration model is based on combining neutral particle ionization, charge exchange, and particle transport [42], [76]. This has resulted in a

moderately accurate model for predicting $n_{e,ped}$, although at present there is still a noticeable gap between the model and perfect predictions. Nevertheless, the neutral penetration model has previously been coupled with the EPED model to create EuroPED [77], a pedestal prediction model that does not require the pedestal density as an input. However, EuroPED requires other plasma parameters, such as β as an input, which is a parameter that is not always known beforehand.

A feature of the neutral penetration model is that it requires the separatrix density $n_{e,sep}$ as an input. Potentially, if device-independent scrape-off layer models that accurately predict $n_{e,sep}$ are further developed, the neutral penetration model may inherently also become more accurate and generally applicable.

3.8.4 Empirical pedestal scalings

Due to the challenges related to modeling the pedestal theoretically, empirical studies based on experimental data play a crucial role in advancing the understanding of pedestal behavior. Specifically, it is of interest to study how pedestal properties, such as top values and width, correlate with machine and plasma parameters, such as the plasma current and heating power.

Empirical analysis can be done in multiple ways. It is common to select a small set of pulses where most machine and plasma parameters are approximately constant [40], [78]–[81]. By only varying one parameter, its dependency with different pedestal properties can be investigated. However, a caveat with this approach is that the dependency is found for a specific operational scenario. It is not guaranteed that such dependencies remain the same when changing the parameters that were constant in this subset of pulses being analyzed. An optional approach is to investigate the dependency between a machine parameter and the pedestal across a large data set, as in [40]. This method is useful for visualizing general trends that hold for large operational domains. However, as other parameters are not constant across these large data sets, it becomes challenging to isolate dependencies. To counter this issue, multi-variate analysis can be performed by employing curve fitting techniques. Specifically, by assuming a functional mapping between several machine parameters and pedestal properties, it is possible to separate the contribution from different parameters to the pedestal. For instance, in [40], the pedestal is empirically modeled at JET by assuming a power scaling law, which has yielded the results

$$T_{e,ped} = (0.05 \pm 0.03) I^{0.00 \pm 0.2} P^{0.74 \pm 0.12} \delta^{-0.23 \pm 0.15} \Gamma^{-0.16 \pm 0.05} M^{0.3 \pm 0.4}, \quad (3.14)$$

$$n_{e,ped} = (9.9 \pm 0.3) I^{1.24 \pm 0.19} P^{-0.34 \pm 0.11} \delta^{0.62 \pm 0.14} \Gamma^{0.08 \pm 0.04} M^{0.2 \pm 0.2}, \quad (3.15)$$

where $T_{e,ped}$ is the pedestal top temperature in keV, $n_{e,ped}$ the pedestal top density in $\mathrm{m}^{-3}(10^{19})$, I the plasma current in MA, P the total heating power in MW, δ the triangularity which is unitless, Γ the fueling rate in charge per second (10^{22}), and M the effective mass of the plasma, which is unitless.

Both these scalings are reasonably accurate on the large JET pedestal dataset, and they do indeed reveal general dependencies with respect to the machine parameters. However, a caveat with power scalings is the limitation they impose on the functional mapping. For instance, more intricate interactions between the input parameters, that potentially could lead to deeper insights about the pedestal, are not possible to capture with these simple expressions. In fact, the main goal of Paper III in this thesis is indeed to explore more complicated input interactions when predicting the pedestal using interpretable machine learning models.

3.8.5 Other ELM types and operational modes

As mentioned, type-1 ELMs are associated with a high energy loss (on the order of 10% of the stored plasma energy) when triggered. For this reason, type-1 ELMs are also referred to as large ELMs. Moreover, type-1 ELMs are categorized by a higher repetition frequency as the input power gets higher.

There are however also other types of ELMs and operational regimes that have been found in experiments that are associated with a pedestal:

- Various small ELMs regimes, such as grassy ELMs [82] and quasicontinuous exhaust (QCE) H-mode regime (formerly known as type II ELMs) [83]. These are associated with a loss of energy of about 1% or less, and they occur at specific operational conditions. For instance, grassy ELMs occur when there is a low edge safety factor, high triangularity, high β_θ, low pedestal density gradient, and low edge toroidal plasma rotation.
- Quiescent H-mode (QH-mode) [84]. In this mode, the periodic ELMs are essentially replaced by continuous edge transport that is facilitated by edge harmonic oscillations (EHO). In other words, the transport from the EHO holds the pedestal below the PB instability boundary. There is also wide pedestal QH-mode (WPQH) [85], where the EHO is replaced with a broadband turbulence. This allows for a wider pedestal and higher pedestal top pressure. As in the previous case, these modes are accessible at specific conditions, such as high at $E \times B$ velocity shear and high triangularity.
- I-mode is an ELM-free mode characterized by a developed temperature pedestal but no density pedestal [86]. It is frequently associated with Weakly Coherent Modes (WCM) and Geodesic Acoustic Modes (GAMs) that lead to continuous transport, hence hindering the build-up to ELMs.
- Negative triangularity plasmas have recently been shown to achieve high confinement while being ELM-free [87]. As in the other ELM-free modes, moderate transport holds the pedestal below the stability threshold for these scenarios.

The main common trait of avoiding large ELMs, either through small ELMs or transport, makes these operational modes more favorable for future machines. That said, one still needs to understand type 1 ELMs, such that one can operate within its stable domain when accessing the other ELM regimes.

3.8.6 ELM mitigating techniques

There are also ELM mitigating techniques that can be used to impact the plasma such that the pedestal does not reach the large ELM instability threshold. The main approaches are:

- Resonant magnetic perturbations (RMPs) [88]. As the name suggests, magnetic perturbation fields are applied to the plasma edge, which leads to increased transport that hold the pedestal below the PB boundary.
- Pellet injection [89]. In this technique, cold fuel pellets are injected to the pedestal. Due to rapid heating on a time scale faster than dissipation, a local pressure peak occurs, which triggers smaller ELMs. Hence, the build-up of the pedestal is interrupted before it reaches the PB instability boundary.
- Vertical kicks [90]. By modifying the current that flows through stabilization coils, the plasma can be vertically displaced. This leads to increased edge current density, which leads to smaller and more frequent ELMs in a similar way as in the previous technique.

In essence, these techniques may prove to be particularly useful in scenarios where it is more difficult to naturally avoid large ELMs.

3.8.7 Machine learning assisting in pedestal modeling

There are different ways in which machine learning can assist in pedestal modeling and research. However, before discussing this in detail, a short (and very simplified) description of supervised machine learning is provided since its fundamentals are yet to be presented (Chapter 4).

- A machine learning model can be designed to predict some output y from a set of inputs $\vec{x}$. The model is then fed data that contains training pairs of y and $\vec{x}$, such that it can be optimized to achieve as accurate predictions of y as possible.

With this brief description outlined, we can discuss the first area of machine learning application for the pedestal, which is the area of *surrogate modeling*. The main purpose of surrogate modeling is to ease the computational cost of otherwise expensive models. For instance, numerically solving MHD PB stability in the pedestal is computationally expensive, which may create significant bottlenecks in integrated modeling simulations. However, if there exists a database of inputs and outputs from the computationally demanding model, a machine learning surrogate can be trained to mimic it. Post training, the surrogate can provide much faster predictions since a forward pass through machine learning models is much less computationally demanding compared to the cost of the numerical solution in the original model. Indeed, recent studies have demonstrated the possibility of creating surrogates for pedestal modeling, including both complete pedestal prediction tools such as EPED and EuroPED [91], [92], as well as for MHD stability codes such as MISHKA [93].

Using surrogates based on a theoretical model conceptually holds the advantage of not being limited to predictions for present day devices. For instance, an EPED database can be generated for future machines like ITER by selecting combinations of input parameters that align with the expected operational scenarios at ITER. That said, a drawback of surrogate models is that they inherit faulty assumptions made in the original model. For instance, a surrogate trained on data generated by an ideal MHD based model, such as EPED, will not be able to predict the pedestal accurately when resistivity cannot be neglected. Moreover, due to the non-trivial nature of accurately modeling pedestal transport, and the challenges of theoretically predicting the pedestal density, predictions of future device pedestals using theoretically based surrogate models imply non-negligible uncertainties. However, as theoretical modeling becomes more accurate, surrogates will unquestionably play a more significant role in the modeling of the pedestal. Additionally, recent implementations of active learning techniques show promising results in reducing the number of training examples needed from the underlying model [94], which is particularly beneficial when training surrogates for models that currently lack existing datasets.

The other application of machine learning for the pedestal is the approach of training on experimental data. Beyond the papers appended in this thesis, this has been demonstrated with relative success in [95]–[97].

There are both advantages and disadvantages to this approach compared to theoretically based surrogate modeling. For instance, by training on real experimental data, the final model may be more accurate than surrogates in regimes where faulty theoretical assumptions are made. This can in certain scenarios be preferred in applications where prediction accuracy is of highest importance. However, the obvious caveat with the experimental approach is the inability to reliably predict for future devices, since there is no reason for why such models would extrapolate well beyond existing data. Moreover, even for existing devices, the data distributions can in some aspects be limited due to the economical cost of running a wide variety of experiments, and also due to how pulses need to be run to ensure stable plasmas. For instance, at a certain toroidal magnetic field strength, the plasma current cannot be too high since this will lead to a low safety factor, which in turn leads to major plasma instabilities. Hence, in the surrogate approach the exploration is more flexible.

There is however another way in which experimental data can be leveraged using machine learning, namely through interpretability. By training interpretable machine learning models on experimental pedestal data, one may obtain two main benefits:

- Model reliability - by understanding exactly how the model maps the inputs to the output, one can be more confident in employing the model, since there is no ambiguity in how the model arrives at its predictions. For instance, if the pedestal top is predicted from tokamak machine parameters, one can check if the model dependencies are reasonable in relation to established research. Moreover, model interpretability also helps in identifying causes of outlier behavior and malfunctions.

- New data insights - if the model mapping can be interpreted, one may discover new insights into how the pedestal depends on machine parameters, and also how different inputs interact in this mapping. Compared to more simple modeling approaches, such as the prior pedestal power scaling laws (3.14), (3.15), more flexible machine learning models are capable of capturing more complicated parameter dependencies. In best scenario, such findings may guide future research.

These are the motivations for why this thesis focuses on interpretable machine learning for predicting the pedestal, and also because essentially all previous machine learning-based pedestal models are black boxes [91], [93], [95]–[97]. In addition, appended paper IV also demonstrates the use of interpretability for a surrogate modeling application unrelated to the pedestal. Even though the advantages of interpretability may be more obvious for models trained on real experimental data, the benefit of model reliability remains for surrogate models. Moreover, even though the equations of an underlying theoretical model are known, the behavior of the solution is often unknown, which is why costly numerical methods are employed to solve the equations. In other words, an interpretable surrogate model may provide insights into the solution of the underlying model.

The machine learning interpretability methods developed and used in the appended papers are described in detail in Chapter 5, after we have outlined the relevant machine learning fundamentals in the next chapter.

Chapter 4

Machine learning fundamentals

The purpose of this chapter is to provide a background for readers not already familiar with machine learning methodology. Given that the models developed in the appended papers are grounded in neural networks, we will mainly focus on this approach, and gradually introduce their components. To find more detailed explanations of the concepts presented in this chapter, see [98].

4.1 The neural network node

The node is the fundamental building block in neural networks, and it was first implemented in 1958 [99]. As illustrated in Figure 4.1, a node takes a set of inputs, in this example $[x_1, x_2, x_3]$, and predicts an output $\hat{y}$.

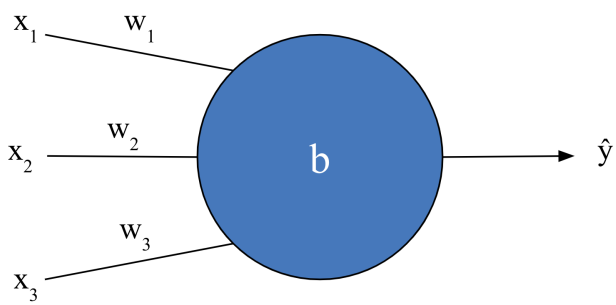


Figure 4.1: An illustration of a node in a neural network.

The computation of $\hat{y}$ is done in two steps. First, the inputs are multiplied with weights, in this example $[w_1, w_2, w_3]$, and a bias b is added such that

$$z = \sum_i w_i x_i + b, \tag{4.1}$$

where z is called the pre-activation value. The output of the node is then obtained by applying a suitable activation function g where

$$\hat{y} = g(z). \tag{4.2}$$

4.2 Example: Linear single-node model

To simplify the concept of machine learning models 'learning from experience', let us now consider an example where we employ a single neural network node as a model. Assume that we want our model to accurately predict y from an input parameter x, where the true underlying relationship is

$$y = 0.8x - 0.5. \tag{4.3}$$

In a real scenario, we would not be aware of the true relationship between the input and output parameters. We might however have access to a dataset with tabular values of y and x that can be used to train our model.

If we use the simplest possible activation function, which is the identity mapping $g(z) = z$, the functional map of our model becomes

$$\hat{y} = wx + b. \tag{4.4}$$

This type of activation is also referred to as a linear activation, as the output now is a straightforward linear function of the input parameter.

4.2.1 The loss and cost function

In machine learning, an initial guess is made for the weight and bias parameters. For our example, let this initial guess be $w = 0.6$ and $b = 0.6$. Consequently, if the first row in a tabular data set is $[x = 1, y = 0.3]$, then our model will predict $\hat{y} = 1.2$, which clearly is wrong since the correct answer is $y = 0.3$ according to the underlying function (4.3). However, we can use loss and cost functions to quantify the error with the aim of guiding the optimization process. Specifically, the loss function L is used to quantify the error of individual predictions. For instance, a common loss function is the squared error

$$L = (y - \hat{y})^2, \tag{4.5}$$

which also can be expressed as a function of the weight and bias parameters by inserting (4.4) into (4.5)

$$L = (y - wx - b)^2. \tag{4.6}$$

The derivative of the loss L with respect to w and b can now be obtained, which we will make use of in the next subsection

$$\frac{\partial L}{\partial w} = 2(y - wx - b) \cdot (-x), \tag{4.7}$$

$$\frac{\partial L}{\partial b} = -2(y - wx - b). \tag{4.8}$$

The difference between a loss function and a cost function is that the latter is the average loss across several data entries, although the two terms are often used interchangeably. The cost function in our example is the mean squared error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{k=0}^{N}(y_k - \hat{y}_k)^2, \tag{4.9}$$

where N is the total number of data entries being evaluated. The derivative of the MSE cost function with respect to, for instance w, can be obtained for our example by expanding (4.7)

$$\frac{\partial}{\partial w}\text{MSE} = \frac{1}{N}\sum_{k=0}^{N} 2(y_k - x_k w - b) \cdot (-x_k). \tag{4.10}$$

4.2.2 Gradient-based optimization

We can now adjust the weight and bias parameters to get a more accurate model. This is possible since we know how the loss and cost depend on the weight and bias. For instance, a data entry that gives $\partial L/\partial w > 0$ indicates that if w is increased, the loss also increases. As the goal is to minimize the loss and cost function, w should instead be decreased. More technically, gradient-based machine learning uses the concept of shifting the trainable parameters in the opposite direction with respect to the sign of the partial derivative of the loss/cost. Gradient Descent is an example of a method that follows this approach, and it defines the update of an arbitrary trainable parameter, that is, a weight or bias parameter θ

$$\theta^{l+1} = \theta^l - \eta\frac{\partial L}{\partial \theta}, \tag{4.11}$$

where η is called the learning rate, which controls the step size of the parameter update. The superscript l refers to the iteration number.

4.2.3 Training procedure and result

In practice, the optimization algorithm occurs iteratively with many updates of the trainable parameters. This is what is referred to as 'training' the model. In our example, we use a dataset with 100 entries generated with the underlying equation (4.3). We use the same parameter initialization as was mentioned before ($w = 0.6$ and $b = 0.6$), and we use Gradient Descent with $\eta = 0.05$ as the optimization algorithm. Instead of computing the weight and bias update for each data entry individually, we calculate the full MSE cost function on the

entire data. This is called full-batch training, although for real applications it is not uncommon to use a smaller batch size.

The training result is shown in Figure 4.2, which shows the evolution of w, b, and the cost function in the iterative training process. After approximately 200 iterations (parameter updates), the model has successfully identified $w = 0.8$ and $b = -0.5$, which has yielded a low cost function value.

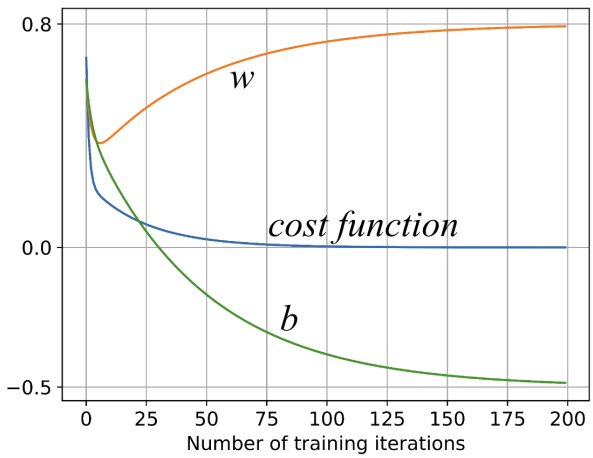


Figure 4.2: The result of the iterative training process. Even though the parameter w decreases at the initial iterations, the algorithm later finds that w needs to increase and then saturate at 0.8 to minimize the cost function.

4.3 Multilayer perceptrons (neural networks)

The one-node model in the previous example was able to find the true underlying relationship between the input and output parameters due to the simplicity of the problem. However, real problems may require the model to capture much more complicated and abstract representations and trends in the data. Hence, to facilitate more intricate functional mappings, one can arrange multiple nodes in several layers to form a neural network (as illustrated in Figure 4.3), which is also called a multilayer perceptron (MLP). In this setup, computing the activation value of each node is the same as in the one-node example. Moreover, since nodes are now connected, the weights provide a quantifiable measure of connection strength between nodes. In that sense, it is the specific connection strength pattern that emerges during training that leads to capable models.

The training procedure of a neural network follows the same principle as for the case with the single node; a cost function is differentiated with respect to all of the trainable parameters $\vec{\theta}$. This differentiation is enabled by the chain rule, which allows the gradient to propagate through the composed functions of the network layers. This process can be performed analytically

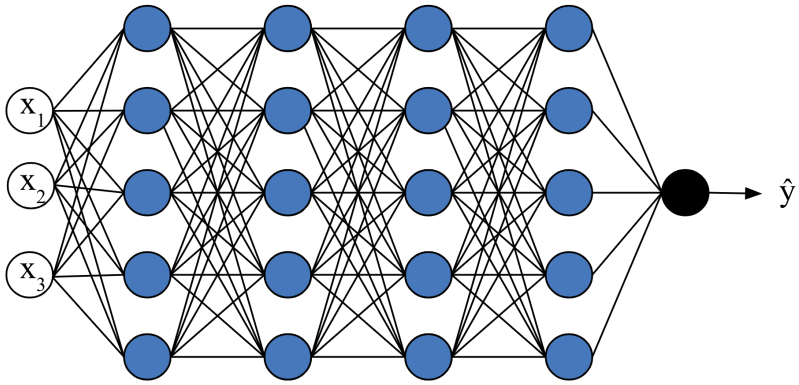


Figure 4.3: A neural network / MLP. Here, the output of a node is forwarded as an input to each node in the next layer (left to right). This particular network consists of 3 input parameters in the input layer (white), 4 hidden layers (blue) with the layer size 5 (5 nodes each), and an output layer with one output node (black).

or through automatic differentiation in a programming script. The trainable parameters are then iteratively updated together through an optimization algorithm like Gradient Descent to minimize the cost function with respect to the dataset. This procedure is often referred to as backpropagation, since the error signal propagates backwards through the model to guide the parameter update. Backpropagation was first introduced in the 1980s [100], where one of the authors was later awarded the Nobel Prize in physics in 2024, largely due to the massive impact that backpropagation had on the field. One of the key features of backpropagation is that it enables parallel computation to update multiple parameters simultaneously, making the training of large-scale neural networks computationally feasible.

4.3.1 Nonlinear activation functions

When the activation function for all nodes in a neural network is set to be the identity function, the overall mapping is effectively reduced to a linear relationship between the inputs and the output. Therefore, to allow for learning of more complicated relationships, nonlinearities need to be introduced. This is done via the activation function applied in the nodes, where common nonlinear alternatives include the Rectified linear unit (ReLU), and the sigmoid function σ, which are defined as

$$g_{\text{ReLU}}(z) = \max[0, z], \tag{4.12}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{4.13}$$

where z again is the pre-activation value in the node (4.1). Both of these activation functions are visualized in Figure 4.4.

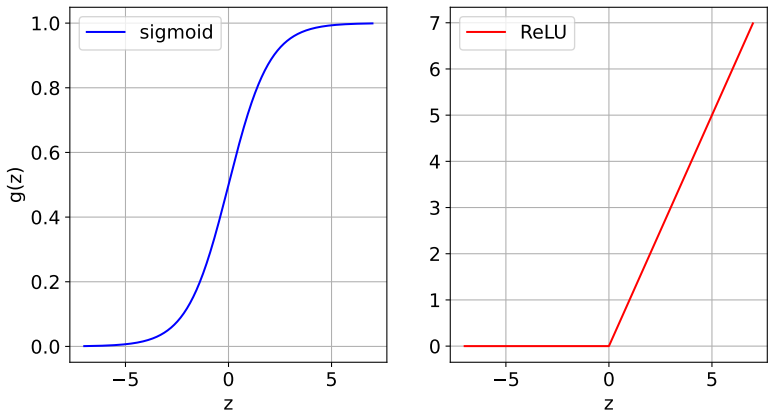


Figure 4.4: The sigmoid activation function (left) is characterized by being bounded to 0 and 1. The ReLU function (right) is characterized as a piece-wise linear function where $g(z) = 0$ when $z < 0$, and where $g(z) = z$ when $z \geq 0$.

Previously, we discussed how the loss function can be expressed as a function of the trainable parameters. This is also true when nonlinear activation functions are implemented, however, the differentiation of the loss function with respect to the trainable parameters must be adjusted based on which activation function is used. Moreover, although nonlinear activation functions are essential for allowing complicated functional mappings to emerge, the output layer of a model that predicts a non-discrete value is usually set to be linear.

4.3.2 Hyperparameters

In machine learning, the term *hyperparameters* refers to settings that are determined before the training procedure starts. For neural networks, hyperparameters include: the number of hidden layers, the number of nodes in each layer, choice of optimization algorithm, learning rate, batch size, choice of loss/cost function, choice of activation function. The number of training iterations is also a hyperparameter, and the number of 'epochs' refers to how many times the full data set is parsed through the model during the training.

4.4 Overfitting and dataset splits

As mentioned, neural networks with many nodes allow for complicated functional mappings to be learned, which makes them useful for solving many challenging tasks. However, if the goal of the model is only to minimize some

error expressed in the cost function, having a large number of tuneable parameters leads to a risk of overfitting. This means that the model might capture noise in the training data, or that it memorizes specific data entries, which both are degrading to the generalization capabilities of the model.

There are different regularization techniques that can be used to mitigate overfitting, such as adding a penalty term in the loss function for large weight values. Moreover, due to the risk of overfitting, it is often meaningless to only evaluate the model on the data that has been used in the training. Instead, the given dataset is usually split into three parts:

- Training set - As the name suggests, this set is used for the training of the model.
- Validation set - This set is held out during training and allows for a more unbiased evaluation when comparing different combinations of hyperparameters to find the most suitable settings. The validation set is also used to monitor when a model shows signs of overfitting, as illustrated in Figure 4.5.
- Test set - To enable a fully unbiased evaluation of the model, a test set is held out both during the training and search for optimal hyperparameters, and is only used for a final evaluation.

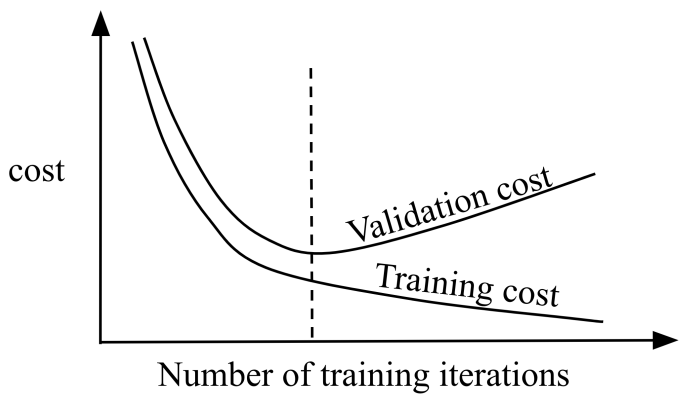


Figure 4.5: An illustration of a cost curve. For the early training iterations, both the validation and training cost decrease. However, at a certain point, which is marked by the dashed line, the validation cost begins to increase while the training cost keeps decreasing. This is an indication that past this point, the model is overfitting to a high degree.

A typical data split would be to allocate approximately 70-80% of the entries for training, 10-15% for validation, and another 10-15% for testing, although it can vary depending on the problem. With very large datasets, a smaller fraction may be used for the validation set and testing set since these will still likely be sufficiently representative.

4.5 Classification models

The introductory example in this chapter illustrated a regression task, since the model was trained to predict a non-discrete output. However, neural networks and other machine learning models can also be used for classification tasks by using a sigmoid activation function in the output layer. As the sigmoid function is bounded between 0 and 1, datasets can be curated to represent classes with ones or zeroes. Moreover, for problems where there are multiple outputs, a softmax [98] activation function can be used to assign probabilities to the different outputs. This is, for instance, used in Large Language Models (LLMs) to produce probabilities of which token comes next in the sequence [101].

Although traditional loss functions, such as squared error, can be used in classification tasks, it is more common to use loss functions that consider probability distribution of predictions and labels, such as the binary cross-entropy loss function

$$L = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})). \tag{4.14}$$

Here, $\hat{y}$ is the predicted output (classification probability between 0 an 1) while y is the true output label (0 or 1).

4.6 Extrapolation and fine-tuning

When tasking a model to make predictions beyond the scope of either its original task or its training distribution, there is naturally no guarantee that the model will behave well. This is often referred to as the extrapolation problem. However, it is entirely feasible to pre-train a model on a large dataset, and then to fine-tune it for a more niche application on a smaller dataset (this is also called transfer learning). The idea here is that models may learn useful characteristics on the larger dataset, such that one does not need to start from scratch for the niche application. For instance, a model trained to classify dogs and cats from images might serve as a decent starting point for training a model to classify cows and pigs. This is because such a model already likely knows how to identify relevant characteristics before the fine-tuning, such as lines and shapes, and even eyes and tails. A more concrete example is ResNet [102], which is an architecture that have commonly been used for transfer learning in computer vision.

4.6.1 Prediction uncertainties

In certain scenarios, it might not be obvious that a model is predicting outside its training domain. As an example, a model trained to predict cancer risk from blood count and age might be used to seeing examples of low blood count and old age, and low age and high blood count, but no examples with low age and low blood count. Hence, since there has been no incentive for the model to learn appropriate trends in the low age / low blood count domain, there is

no reason to trust it for such cases. However, a standard neural network will not indicate that it is unsure as it will simply just predict a value or a class label, which is problematic.

Fortunately, in recent years there has been a rise in strategies aimed to provide prediction uncertainty along with the actual prediction. For instance, Bayesian neural networks learn distributions of weights rather than exact values, such that one can sample predictions and analyze the output distribution [91]. Here, a wider distribution in the output essentially means less certain predictions. Another strategy is to use an ensemble of models instead of just one model. This will also lead to a distribution of outputs that can be analyzed to quantify the uncertainty, similar to the Bayesian approach.

4.7 Beyond dense neural networks

As mentioned, we have focused on dense neural networks in this chapter due to their relevance in the appended papers. However, there are many other methods and variants that are common in the field of machine learning and AI. These include Convolutional Neural Networks (CNNs) [102], [103] that are specialized for tasks where there is a grid-like topology, which makes them useful when dealing with images and other types of problems in computer vision. As an example, the CNN-based model AlexNet published in 2012 [103] was an important piece in sparking the deep learning revolution (the idea of using many neural network layers), when it surpassed the accuracy of image classification models based on manually-designed features. Moreover, there are Recurrent Neural Networks (RNNs) [104], [105] specialized for sequential data, and transformer architectures [101], [106] (that also include neural networks) that have completely revolutionized Natural Language Processing. In particular, models like ChatGPT and Claude depend entirely on the attention mechanism in transformers. As a last example, Graph Neural Networks (GNNs) [107] have turned out to be useful for applications where there are irregular graph structures, such as when dealing with social networks or molecules.

Beyond architectures rooted in neural networks, there are also other types of machine leaning methods that do not rely on gradient-based optimization. For instance, decision trees work by recursively splitting data based on the inputs to make predictions, and support vector machines use kernel functions to implicitly map data into higher-dimensional spaces where it becomes linearly separable [98], [108]. However, at present none of these other techniques genuinely challenge neural network-based methods in terms of reaching more capable systems for more complicated problems.

4.7.1 Self-supervised learning

In the introductory example, we employed traditional supervised learning, which represents the case where there exists a curated dataset with annotated input and output parameters. However, the success of models like ChatGPT is largely due to self-supervised learning, which automatically generates labels

from the data itself rather than requiring manual annotation. Particularly, in LLMs, the prediction of the next token in a sentence makes it possible to implement an algorithm that labels the next token automatically. Since the need for manual annotation is eliminated, it is feasible to pre-train on vast amounts of unlabeled text corpora (which is done before more careful fine-tuning is conducted). It is undeniable that this scalability has been crucial to achieving the impressive capabilities of modern language models [109].

Self-supervised learning is also possible when dealing with image data. For instance, a model can be tasked to predict missing patches in images, where the goal is to learn useful visual representations that can transfer to downstream tasks [110]. Another more recent example is the JEPA concept [111], which essentially makes use of the idea that different parts from the same image, or different images from the same video, should share abstract representations (this makes it possible to automatically generate training examples).

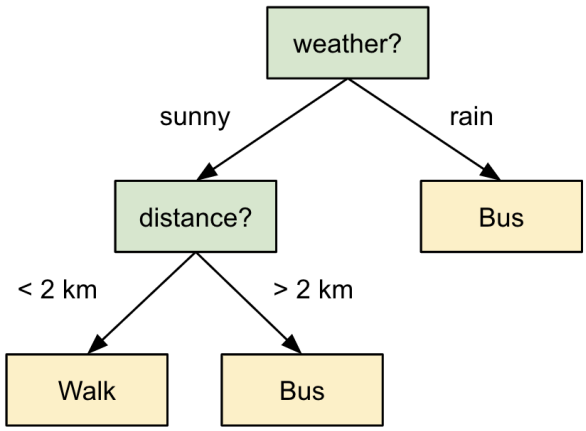
Although the appended papers only deal with supervised learning, it is entirely plausible that self-supervision might find practical use in fusion research in the future. For instance, masking and JEPA could potentially be applied to 1D or 2D plasma diagnostics, which might enable models to capture the underlying patterns and dynamics of fusion plasmas in a similar way that LLMs capture patterns in language.

4.8 Why machine learning models are black boxes

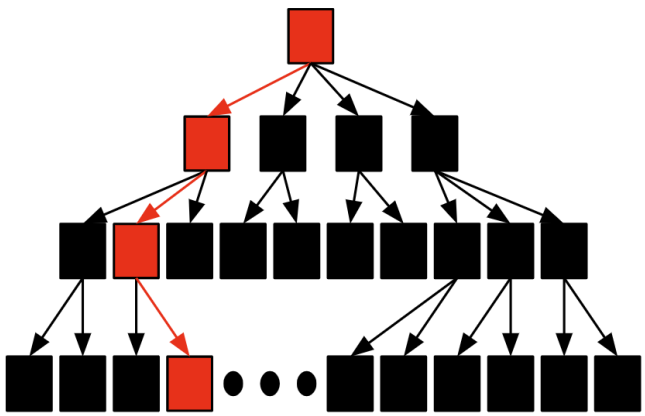
When we say that machine learning models are black boxes, it is not because we do not understand their fundamental components, since these are indeed designed by humans. In fact, with simple configurations, such as the single-node model in the introductory example, there is no black-box aspect. This is because we could easily investigate exactly how the output depends on the input (since we only had two trainable parameters). However, when dealing with hundreds, thousands, or even billions [112] of trainable parameters, challenges arise. In particular, although we have access to the weight values of individual nodes, it quickly becomes unfeasible to grasp the overall qualitative mapping of the full model.

There is also a common misconception that models involving decision trees are interpretable since it is possible to follow the chain of decisions in the specific tree path for an individual prediction. As illustrated in Figure 4.6, this might hold true when there are very few tree-nodes, such that one may grasp how the model will behave for all possible scenarios. However, similarly to neural networks, understanding the global mapping from the inputs to the output becomes infeasible when the number of tree-nodes increases. Decision trees are particularly problematic in this regard, because they might lead to a false sense of model understanding. To have true global interpretability, one needs to understand exactly how the output of the model depends on the inputs, and how the inputs interact, for all relevant scenarios. In other words, one needs to grasp the overall qualitative mapping that has emerged during

training, which is indeed the main goal when applying machine learning to the fusion-related problems in this thesis. In the next chapter, we will discuss different interpretability approaches and their suitability for different scenarios.



(a) A simple and interpretable decision tree.



(b) A decision tree that is deeper and includes more tree-nodes compared to (a). Even though we can follow the path of an individual prediction, it is more difficult to grasp the complete global model behavior.

Figure 4.6: Illustration of how decision trees become less globally interpretable when the depth and number of tree-nodes increase.

Chapter 5

Interpretability

Although the idea of trying to understand and explain the decision making process of AI systems has received more attention in recent years, the field is not new [113]. Already in the 19th century, mathematicians like Gauss and Legendre worked on interpretable linear regression models for fitting data [114], [115], which laid the foundation for many analysis methods still used today, such as Lasso regression [116]. Moreover, in the later half of the 20th century, work was conducted, for instance, on feature importance detection in decision trees [108], and rule extraction from neural networks [117].

However, the rise of more complicated and deep architectures in the last decade, that have enabled more capable systems, has led to an increased demand for new and improved interpretability capabilities. This has become especially important in recent years as AI has rapidly moved into different areas of society. In particular, lack of interpretability can be very problematic in relation to ethics when AI is used in, for instance, medical, juristical, and military applications [118]. Furthermore, lack of interpretability inhibits the user to anticipate the behavior of the model, which is problematic in terms of credibility, control, robustness, and safety. Beyond that, when applied in a scientific environment, lack of interpretability means a loss of opportunity to discover knowledge.

To date, there is no universal method that allows humans to completely understand the internal reasoning of the most capable AI-systems. However, progress is being made, and in this chapter, an overview of leading interpretability techniques is presented. We mostly focus on different approaches for tabular-data problems, since this is the type of problem addressed in the appended papers. That said, we also discuss interpretability for other types of problems, due to their relevance across fusion research more broadly. Beyond established methods, we also present NeuralBranch, an interpretability framework developed during this thesis to address shortcomings of other tabular-data methods.

5.1 Tabular-data problems

As the name entails, tabular problems in machine learning refer to tasks where the data is structured in a traditional table format, with rows representing individual samples and columns representing different known parameters (also referred to as features). In other words, unlike array-structured data, such as images or text, each column typically has a clear, interpretable meaning. Typical tasks in this category include predicting house prices from features such as number of rooms and distance to city center. More relevant to this thesis is predicting pedestal top temperature or density from key tokamak parameters, which is the focus of appended Papers I, III, and V. Moreover, the surrogate modeling approach used for theory-based pedestal models like EPED and EuroPED, that was discussed in Chapter 3, also falls into this tabular-data category [91].

In essence, machine learning models trained for tabular-data problems are narrow in the most fundamental sense (designed for very specific prediction tasks compared to models that can be useful for many different user requests, such as modern chatbots), and more close to traditional multivariate curve-fitting compared to more advanced AI. The main difference compared to traditional curve-fitting is the means by which the sought functional mapping is obtained.

Achieving full global interpretability for tabular-data problems means understanding exactly how each input affects the output, and how the inputs interact in the model mapping. In the following subsections, we will discuss a selection of popular methods aimed to achieve this.

5.1.1 Low-capacity models

The idea of low-capacity models is to restrict the model complexity beforehand and thereby impose intrinsic interpretability. One of the most straightforward examples is linear regression, which assumes the output as a linear combination of the inputs. For instance, when predicting an output y from three inputs: x_1, x_2, and x_3, a linear regression model can be formulated as

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3 \tag{5.1}$$

where c_0, c_1, c_2, and c_3 are the adjustable parameters (also referred to as fitting coefficients) that indicate how each input contributes to the output. For classification tasks, applying a sigmoid function to the right-hand side of Equation (5.1) constrains y to (0,1), which yields the logistic regression formulation. Due to the small number of trainable parameters, linear regression models can often be optimized by analytically finding the global minimum of the loss function directly (although this can be more expensive when there is a large number of features n, as the computational cost scales as $O(n^3)$). It is however also possible to train such models using iterative approaches like gradient descent, which was illustrated in Chapter 4. Moreover, to achieve automatic feature selection (find which inputs that actually are important), one can modify the loss function in linear regression to include a penalty term for the adjustable parameters, where these are encouraged to shrink as

much as possible without degrading prediction accuracy. This is, for instance, the idea behind the Lasso method mentioned earlier, which aims to suppress unimportant inputs using a L1-norm penalty (the sum of absolute values of the fitting coefficients).

Another low-capacity model, which is common in fusion research, is the power scaling formulation. Specifically, if we again consider an example with three inputs, the expression becomes

$$y = c_0 x_1^{c_1} x_2^{c_2} x_3^{c_3}. \tag{5.2}$$

Power scalings have previously been used to model the pedestal [40], as also mentioned in Chapter 3, but also to predict the stored energy in the plasma core [119], the thermal energy confinement time [120], and the L-mode to H-mode threshold power [121].

Beyond being fully interpretable, an advantage of low-capacity models is that they are not prone to overfitting due to the small number of trainable parameters. However, due to the restricted complexity of the models, they are not able to capture more intricate relationships. This limitation can be somewhat alleviated by including more flexible terms, such as interaction terms in linear regression models (for instance $c_{12}x_1x_2$). However, such models are still relatively limited in their expressivity, and with many interaction terms, where the same input appears multiple times in the expression, it can become more difficult to comprehend the overall trends.

Nevertheless, low-capacity models are useful for obtaining a lower benchmark on model accuracy. This lower benchmark can be compared with the accuracy of more complex models, where the difference can be indicative of whether intricate patterns are present in the data. If prediction accuracy is not improved by using more intricate models, there is no reason to go beyond easily interpretable expressions (following the idea of Occam's razor).

5.1.2 Symbolic regression

In symbolic regression [122], [123], the objective is to identify interpretable mathematical expressions that capture how the output depends on the inputs, similar to how global interpretability is achieved in low-capacity models. However, there is a main difference, namely that symbolic regression automatically discovers the appropriate expression instead of having it pre-determined. This can be done in different ways, the most common approach including some form of genetic programming, and the process can be summarized as:

- First, mathematical building blocks (+, -, *, /, sin, cos, log, ...) are randomly combined with coefficients and the inputs to form initial candidate expressions. These expressions are evaluated on the given dataset.

- Through methods such as crossover and mutation, the best performing expressions can iteratively be refined to obtain new, potentially better, expressions.

The main computational challenge in traditional symbolic regression is the exponential growth in exploration space when the number of operations and features in the expressions grow. In recent years, different approaches have been proposed to alleviate this problem. For instance, in [124], a Graph Network with separable internal structures is first trained. Then, symbolic regression is applied to fit the distinct functions learned by the Graph Network. In other words, instead of directly solving the full problem using symbolic regression, it is first broken down to smaller problems that are more easily manageable. Similarly, the AI Feynman symbolic regression method [125] uses physics-inspired properties such as dimensional analysis, symmetries, and separability to recursively break the full problem into simpler ones with fewer variables.

Although symbolic regression is a robust and proven method, it also has drawbacks. Specifically, real empirical data may not always follow simple mathematical patterns like typical physics examples do. For this reason, complex problems may lead to long expressions that are more difficult to comprehend (especially when a feature appears multiple times in the expression). Moreover, there is not always an unique function or mathematical operation that accurately captures the patterns in the data. For instance, the expressions x, $\sin(x)$, and $e^x - 1$ all behave similarly in the region around $x = 0$, which means that if the data is relatively bounded to this region, all three alternatives might be viable. This can lead to over-confidence in resulting expressions that do not necessarily reflect the exact underlying true equations, but rather capture the data sufficiently well in terms of accuracy. That being said, these limitations should not overshadow that symbolic regression can be very useful, in particular when there is moderate complexity, good data coverage, and reasonable physical constraints.

5.1.3 Neural Additive Models

Neural Additive Models (NAMs) [126], introduced in 2020, combines the flexibility of neural networks with the interpretability from Generalized Additive Models (GAMs) [127]. In the GAM formulation, the output depends linearly on smooth functions f_i of the input variables

$$y = c_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m) \tag{5.3}$$

where c_0 is a fitting coefficient, and m is the total number of inputs. The functions f_i can, for instance, follow a specified parametric form (like a polynomial or spline formulation). NAMs differ from GAMs by instead using distinct neural networks to model the functions f_i, where the idea is to enable more flexibility compared to the functions in GAMs.

The reason why NAMs are interpretable is because each distinct sub-network only has one input parameter. This makes it possible to parse the dataset through the sub-networks, and then to plot the output f_i of each sub-network versus its corresponding input x_i, which is illustrated in Figure 5.1. Hence, it is not necessary to analyze the internal weights of each sub-network, we just need f_i and x_i for the visual interpretation for how each input contributes to

the output. In that sense, the interpretation in NAMs is qualitative rather than quantitative. Moreover, although there are distinct sub-networks for each input, the additive aspect in NAMs allows for gradients to flow during backpropagation, such that the sub-networks can be trained jointly. In other words, there is no requirement for prior knowledge of f_i, since these will emerge naturally as the sub-networks are trained to cooperatively predict y.

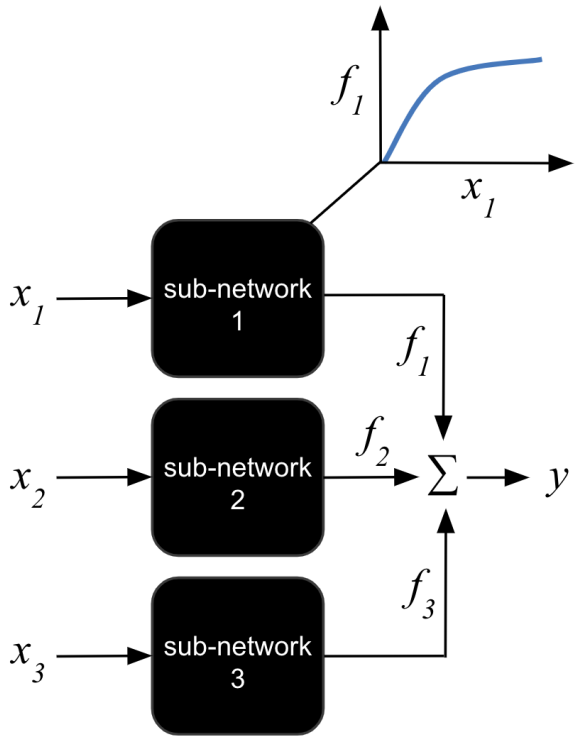


Figure 5.1: Architecture diagram for a NAM with three input parameters. The graph at the top illustrates how interpretability is obtained by visually analyzing how f_1 depends on x_1 (f_1 represents the contribution from x_1 to the output y).

NAMs provide certain benefits compared to expression-based interpretability methods, such as Symbolic Regression. Specifically, the neural networks enable appropriate functions to be learned directly in the optimization process instead of having to try many combinations of pre-chosen functions. Moreover, the interpretable visualizations provided by NAMs may help the user to more easily identify the general dependencies that the model has learned compared to analyzing non-trivial expressions. That being said, this being a benefit depends on the complexity of the problem, and is partly also a matter of subjective preference.

There are however also drawbacks with NAMs, the most obvious being the disentanglement of the inputs, such that no intricate interactions can be learned.

For problems where the underlying data violates this restriction, NAMs will perform worse than black-box models, and also fail in providing the full picture of the true relationships in the data. This limitation has previously been addressed by allowing up to two inputs to be parsed through each sub-network, such that pair-wise interactions can be learned [128]. These NAMs remain interpretable because the second input can be visualized, for instance, using a color scheme, as illustrated in Figure 5.2. However, NAMs still cannot capture higher-order interactions involving three or more features due to the challenge in visualizing more dimensions. Additionally, the additive framework, which assumes that contributions from features (or feature pairs) should be summed to produce the output, inherently restricts the expressive power of the model.

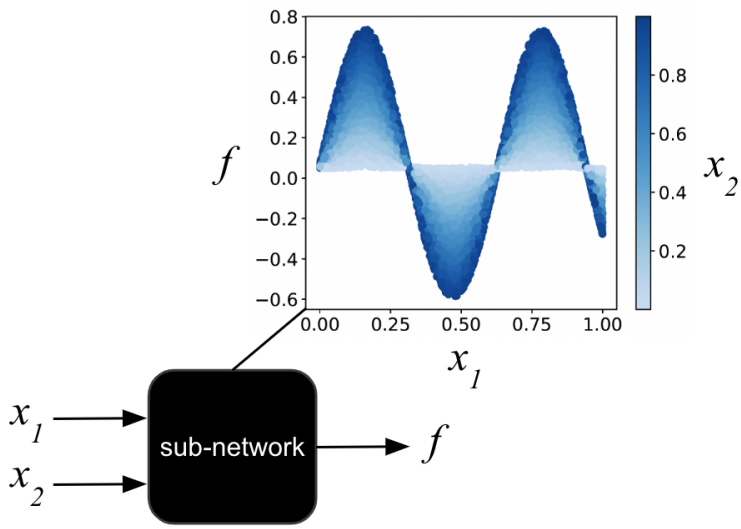


Figure 5.2: Illustration of how sub-networks with only two inputs in NAMs remain interpretable through visualization with a color scheme (which also applies for any model with just two inputs). In this example, the visualization of the points in the dataset reveals that the sub-network has learned a sine relationship between f and x_1, where x_2 appears to control the amplitude. The true underlying equation that was used to generate the data in this example was $f = x_2 \sin(10x_1)$.

5.1.4 Local explainability methods

Local explainability methods are conceptually different from the interpretability methods discussed so far. Specifically, these are generally designed to provide estimations for feature importance for individual instances of predictions (hence the term "local"), rather than providing a transparent picture of the full global behavior of the model. Moreover, local explainability methods are often model

agnostic, which means that they can be applied to any type of black-box model for post-training analysis. In other words, unlike low-capacity models and NAMs, local explainability methods do not require models to be structured in an interpretable format before training.

SHAP (Shapley Additive exPlanations) [129] is a popular local explainability method for tabular problems. Essentially, the idea in SHAP is to analyze how each input contributes to shifting/perturbing the prediction from the baseline (expectation value of the output prediction). For example, assume we have two input features x_1 and x_2. To estimate the contribution of x_2, we fix x_1 to its actual value for the instance being explained, while sampling different values of x_2 from the dataset. By observing how the output of the model changes across these samples compared to the output using the actual value of x_2 in the instance, we can estimate how much x_2 influences the prediction. The shift in the output imposed by varying an input is referred to as a SHAP value.

Although SHAP is primarily a tool for analyzing input importance in individual predictions, it can also be used to get a sense of global patterns by aggregating SHAP values across the given dataset. However, such aggregations are difficult to interpret when there are interactions between the input features, as exemplified in appended Paper III.

LIME (Local Interpretable Model-agnostic Explanations) [130] is another widespread method. In this approach, the idea is to perturb the inputs to impose a change in the output, and then to fit a local interpretable model, such as a linear model. In that sense, LIME offers a heuristic local approximation of the model behavior.

In general, local explainability methods are most useful when the underlying problem is too complex to be modeled accurately using globally interpretable approaches. When globally interpretable models such as NAMs or Symbolic Regression achieve sufficient accuracy, local methods add little value, since the behavior of the model is already explicitly fully transparent.

5.2 NeuralBranch

The NeuralBranch method, introduced in Paper III, was developed during this thesis to address shortcomings of other interpretable frameworks designed for tabular problems, such as Symbolic Regression and NAMs. That said, the NeuralBranch method is greatly inspired by NAMs, in particular the aspect of obtaining interpretability by splitting the full model into smaller sub-networks with only two input features each. The main difference compared to NAMs is that sub-networks in NeuralBranch adapt a computation tree structure, and hence we refer to the sub-networks in NeuralBranch models as *neural branches*.

An example of a NeuralBranch model is shown in Figure 5.3. As can be seen, the output of a neural branch is either forwarded to a sub-sequent neural branch, or set as the final output. The serial aspect of NeuralBranch allows for interactions between more than two inputs, addressing the first main limitation of NAMs. For instance, the output of neural branch 1 (z), which represents the contribution from the inputs x_1 and x_2, can interact with the input x_3 in

neural branch 2, effectively leading to a 3-parameter interaction. Moreover, using neural branches throughout the model, including the final step, alleviates the second major limitation of NAMs: the enforced summation of sub-network outputs. In other words, neural branch 2 in Figure 5.3 is not restricted to learning the addition operation.

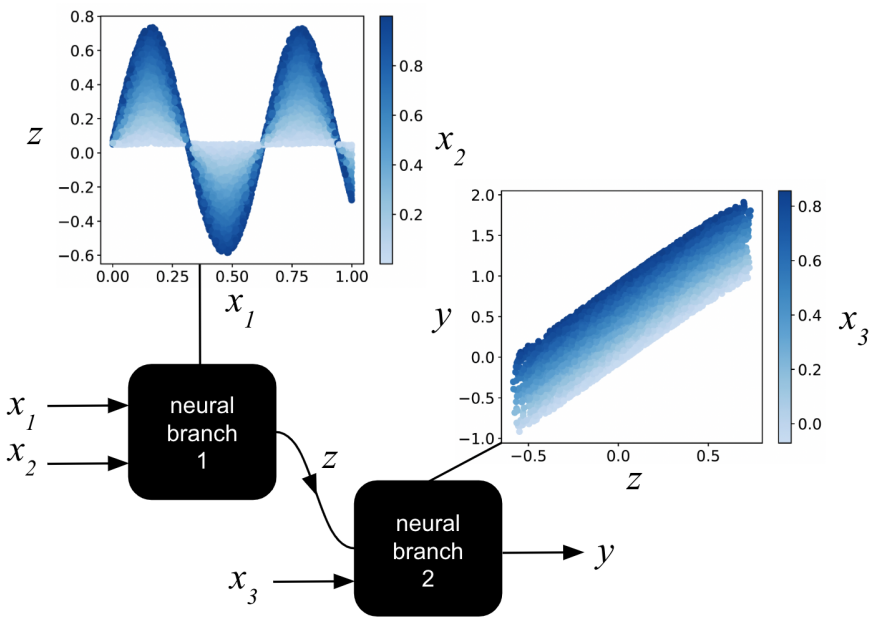


Figure 5.3: Illustration of the NeuralBranch concept. In this example, the true underlying equation is $y = x_2 \sin(10x_1) + x_3$. In the creation of the toy dataset, the three inputs x_1, x_2, x_3 were randomly sampled from a uniform distribution in the interval [0,1]. In essence, the visualizations reveal that neural branch 1 learns the first term ($z = x_2 \sin(10x_1)$), and that neural branch 2 learns to sum z with x_3. Note that global interpretability is obtained through the joint analysis of the different neural branches. Moreover, just as in the case of NAMs, there are simply fully-connected neural network layers inside the neural branches.

As in the case with NAMs, the sub-networks in a NeuralBranch model are trained jointly, which is possible due to gradients flowing through the entire model during backpropagation. Hence, intermediate latent z-variables do not have to be know a-priori, as they are automatically shaped during training.

The main challenge in the NeuralBranch method is to find the appropriate architecture and decide on which inputs should be assigned to which neural branch to best reflect the data. In other words, a method is needed to determine which path each input should take in the computation tree-like structure. In the next section, the method used in the appended papers is outlined.

5.2.1 Finding the appropriate NeuralBranch architecture

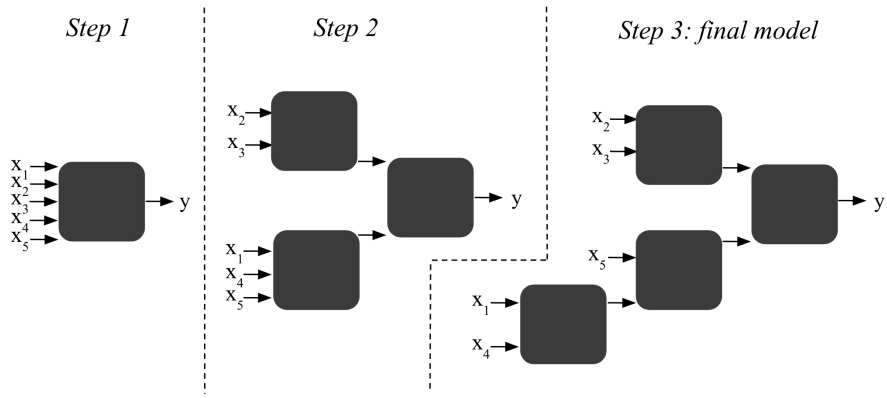


Figure 5.4: Illustration of the recursive input splitting method used to find the final NeuralBranch architecture.

Finding the appropriate NeuralBranch architecture is essentially about finding the correct sequence of pair-wise operations for the problem. This can be done by recursively splitting input parameters into groups, as illustrated in Figure 5.4. The full process includes the following steps:

1. First, a standard dense neural network is used to obtain an upper benchmark for model accuracy.

2. Then, the inputs are split into two groups that are parsed through two distinct neural branches, where the two branch outputs are merged in the final third branch. Here, we need to try different input splits until the benchmark accuracy is reached (a tolerance is applied). In essence, achieving the benchmark accuracy is indicative of a parameter split that respects the relationships in the underlying data. An important detail in this strategy is that every time a new input split is tested, a new model is trained from scratch.

3. The input split process is repeated recursively until the root branches only have two inputs each. This results in a final architecture that can be analyzed to obtain full global interpretability.

The brute force approach of systematically trying different input splits can in theory lead to high computational cost when there are many inputs. However, this is not a practical concern for the input sizes NeuralBranch is designed for. For problems with more than roughly 10 inputs, the model becomes more difficult to interpret due to the complexity of analyzing many interdependent neural branches, rather than being limited by the computation. For the problems considered in the appended papers, where each problem

involves at most five inputs, finding suitable NeuralBranch architectures never took more than a few minutes on a standard laptop.

Another consideration relates to problems where the inputs are not trivially separable in the underlying data. For instance, in the expression $y = (a + b)\sin(b + c)$, b is not easily separable from the other inputs. However, these cases can be handled by allowing one or few parameters to be parsed to both branches if no hard splits yield an accuracy within the tolerance. In fact, this occurred when the NeuralBranch method was used in appended Paper IV. Specifically, a magnetic shear parameter $\hat{s}$ needed to appear as an input to two neural branches in order to achieve the benchmark accuracy (due to $\hat{s}$ interacting with two other inputs in different ways).

5.2.2 NeuralBranch: Main limitations

The NeuralBranch interpretability method is currently only feasible for tabular problems with relatively few inputs, and is generally restricted to narrow tasks. For example, in a tabular problem with 50 inputs, NAMs are likely more useful than NeuralBranch models due to their disentanglement of inputs. Although NAMs may not perfectly capture the true underlying mapping in the data, they can provide an interpretable approximation, which is preferable to a NeuralBranch structure too complex to interpret.

Another limitation arises when working with small datasets, which is a general problem when training neural network based models. For such cases, other interpretability methods that are associated with a lower capacity, such as linear models or Symbolic Regression (when relatively restricted) might be preferable to avoid overfitting.

5.3 Interpretability beyond tabular problems

Beyond the tabular problems addressed in the appended papers, machine learning applications in fusion research increasingly involve problems where inputs are structured as 1D or 2D vectors, time series, or other formats for which tabular interpretability techniques are unsuitable [131]–[133]. Obtaining global interpretability for such problems is generally more difficult, and it becomes increasingly so as models become more complex and capable. Therefore, most current interpretability approaches for problems with high-dimensional inputs rely on local explainability methods, though there are also approaches that aim to identify generally learned patterns or provide more global insights. We will discuss some of these methods in the following subsections as they might be relevant for future work.

5.3.1 Attribution methods

These methods focus on assigning importance scores to input features to explain model predictions. For instance, SHAP falls into this category, but there are also other variants:

- **Occlusion** [134] is a method that determines feature importance by systematically hiding (occluding) parts of the input. For instance, in computer vision, it is common to slide a masking patch (e.g., 15×15 pixels) across an image. Here, regions where occlusion causes large changes in the prediction are considered important, and the result is often shown in a so-called visual saliency map. For problems where there is a temporal structure in the data, it is possible to mask the input at different time steps to identify critical events. The main difference between occlusion and SHAP is that, usually, multiple inputs are masked simultaneously in the occlusion method. The occlusion approach has recently proven useful in fusion research: in 2024, it was employed to explain the decision-making of a model trained to predict disruption events [135].

- **RISE (Randomized Input Sampling for Explanation)** [136] is a variation of occlusion where the masks are randomized and spread out instead of being fixed as, for instance, square patches in images. By sampling many random masks, it is possible to perform a statistical analysis of which inputs that were the most important for a prediction.

- **Gradient-based methods** compute the gradient of the output with respect to the inputs. For images, this reveals which pixels most affect the prediction if changed. GradCAM [137] is a popular variant for convolutional neural networks that computes gradients with respect to the feature maps in the final convolutional layer. This identifies which high-level learned features are important for the prediction, and where in the input these features are spatially activated. Another popular approach is to integrate the gradients, where one starts with a baseline input (such as a completely black image), and then gradually accumulates the gradients on a interpolation path towards the actual input [138]. A benefit with this approach is that it works for any differentiable model (not limited to CNNs).

5.3.2 Concept-based explanations

Concept-based explanation methods aim to interpret model predictions using high-level, human-understandable concepts rather than individual low-level features like pixels or input values. Essentially, the idea is to aligns explanations with how domain experts naturally think about and describe problems. For instance, a fusion physicist might consider ”flat density profiles” or ”edge transport barriers” rather than individual measurement points. Examples of explainability methods that fall into the concept-based category include:

- **TCAV (Testing with Concept Activation Vectors)** [139] quantifies how sensitive the predictions of a model are to user-defined concepts, without requiring the model to be retrained or modified. For instance, assume we have trained a neural network to predict something from temperature and density profiles in fusion research. We can define a ”pedestal” concept by collecting two sets of example profiles: one without

developed pedestals (negative examples) and one with developed pedestals (positive examples). For a chosen layer in the neural network, TCAV trains a linear classifier to distinguish between the activations produced by positive vs. negative examples. The normal vector to this decision boundary becomes the "Concept Activation Vector" (CAV). This is essentially a direction in the internal representation space of the model that corresponds to the presence of a pedestal. TCAV then measures how the predictions of the model change when activations are moved in the direction of the CAV. Through statistical analysis across multiple examples, it is possible to determine how important the high-level concept is for the original prediction task.

- **Concept Bottleneck Models (CBMs)** [140] take a different approach by building interpretability into the model directly in the architecture. These models essentially consist of two sequential components: 1) a concept predictor that maps the raw input to pre-defined concepts, and 2) a label predictor that maps the concept predictions to the final output. This makes it possible to break down the decision-making process of the model, which is useful for debugging purposes. For instance, in a bird species image classification task, reasonable concepts could be: has curved beak, has long tail, has striped wings, etc. If the model predicts "crow," one can see which concepts (e.g., has black wings) drove that prediction. However, CBMs have notable limitations: choosing appropriate concepts is not necessary trivial, and collecting manual concept annotations can be time-consuming.

5.3.3 Probing / representation analysis

Probing (also called representation analysis) [141], [142] is a family of techniques that aim to investigate what information is encoded in the internal representations of a model. For instance, assume a neural network with several hidden layers is trained to predict the presence of MHD instabilities in fusion experiments. To probe what the model has learned, we can run inputs through the trained model and extract activation vectors from a specific layer. We then train additional classifiers (probes) to predict different properties from these frozen layers, such as plasma confinement regime (L-mode/H-mode), q-profile characteristics, edge stability. If a probe achieves high accuracy on these properties, this suggests the layer encodes information that is relevant for these auxiliary tasks. An important note is that the original model remains frozen during probing, such that we are testing what information is already present in the representations rather than learning new features.

The main challenges with standard probing is that it does not reveal how the information is used in the original model. Moreover, it is not trivial to decide how complex a probe should be. Specifically, if the probe is too simple (linear), it might fail to extract encoded information, and if the probe is too complex (deep network), it might learn the task from scratch.

5.3.4 Latent space traversal (counterfactuals)

This approach aims to find interpretable low-dimensional latent representations of the high-dimensional input data, and then to visualize these interpretable features through generative modeling. As an example, in [143], a variational autoencoder (VAE) is trained to find low-dimensional representations z of chest X-ray images, that can be used to reconstruct the input images. By implementing auxiliary classifiers that are trained to predict different medical diagnoses from the latent representation z, such as lung opacity and pleural effusion, human-interpretable features emerge. For instance, when the most important latent variable z_0 for predicting lung opacity is traversed, the counterfactual images produced by the decoder indeed shows that the lungs gradually become more opaque. Hence, it becomes feasible to interpret the learned behavior of the model by qualitatively analyzing what the task-relevant features correspond to.

The main challenge when interpreting latent representations is entanglement. For instance, consider sine waves that can be represented by three dimensions: amplitude, wavelength, and phase shift. When training autoencoders to compress sine waves into three latent variables and reconstruct them, there is no guarantee that latent variable 1 will correspond exactly to amplitude or to any single degree of freedom. In other words, the three dimensions we consider interpretable may become spread across the three latent dimensions in complex combinations, even when orthogonality is encouraged, simply because there is no unique solution to the compression task. However, as exemplified in [143], incorporating auxiliary tasks and encouraging sparsity can help shape the latent space into disentangled, task-relevant features that are more interpretable to humans.

5.3.5 Mechanistic interpretability

Mechanistic interpretability is a research field that has become popular in recent years as a means to understand some of the most advanced AI systems of today. In fact, the interpretability teams at world-leading organizations like Google Deepmind and Anthropic focus on mechanistic interpretability as one of the main approaches to understand their LLMs.

The goal of mechanistic interpretability is essentially to develop causal, algorithmic explanations of how neural networks compute their outputs. This involves identifying minimal computational subgraphs (or 'circuits') consisting of specific neurons, attention heads (which are components in the commonly used transformer architecture), and connections that implement particular capabilities or behaviors. In other words, unlike other interpretability approaches that focus on correlational analyses or input-output relationships, mechanistic interpretability aims to understand the actual computational mechanisms inside the model. For instance, research performed in recent years has shown how sparse autoencoders can be used to identify interpretable features and circuits that describe how the LLM Claude from Anthropic processes information when it produces a certain word from a prompt [144]–[146]. Another example is the use of mechanistic interpretability on the image classification model AlexNet,

where researchers were able to identify circuits for detecting specific objects such as dogs and houses [147].

Due to mechanistic interpretability approaches varying depending on the architecture and problem, we will not discuss these in more detail here. It remains to be seen whether mechanistic interpretability will find practical applications in fusion research, as this likely depends on whether the learned representations of plasma behavior can be decomposed into conceptually understandable features for human experts.

5.4 Interpretability: Concluding remarks

This chapter has covered interpretability methods across the spectrum from tabular to high-dimensional problems. The appended papers focus on tabular interpretability, where the presented NeuralBranch method has been used to provide robust explanations for problems with meaningful scalar inputs. However, as the field is increasingly adopting deep learning techniques for high dimensional data, such as profiles and time series, new challenges are presented. Here, attribution methods, concept-based explanations, and mechanistic interpretability represent potential paths forward alongside new techniques that may be developed in coming years.

Chapter 6

Summary of Appended Papers

6.1 Enabling adaptive pedestals in predictive transport simulations using neural networks

This paper, published in 2022, investigates whether predicting the pedestal from key tokamak parameters using machine learning is feasible (i.e., whether it constitutes a well-posed machine learning problem). After confirming that this is indeed feasible, the paper then demonstrates the implementation of the model (referred to as PENN: PEdestal Neural Network) into the integrated modeling framework ETS.

To be more specific, we trained neural networks on the JET pedestal database [40] (mostly deuterium plasmas, but also a few hydrogen plasmas) to predict the electron temperature and electron density at the top of the pedestal in tokamaks from 12 input features, such as plasma current I_P and the plasma parameter β_N. The philosophy was to include as many potentially relevant input parameters as possible to maximize prediction accuracy. However, later analyses revealed that several inputs were redundant, which suggests that this strategy was somewhat naive. Nevertheless, we were able to achieve high prediction accuracies on held out test sets ($R^2 = 0.93$ for the pedestal temperature, and $R^2 = 0.91$ for the pedestal density, where R^2 is the coefficient of determination). Moreover, we used an ensemble of models to obtain estimates for prediction uncertainty.

Since this work was done before interpretability became the main research direction of this PhD project, the models presented in this paper are black-boxes. However, the paper includes a few case studies that demonstrates that the models behave reasonably with respect to previous research regarding how the pedestal depends on key tokamak parameters. For instance, the paper shows examples where the predicted pedestal temperature $T_{e,ped}$ increases with increased input power, but decreases with increased gas fueling.

In the ETS demonstration, two pulses with different NBI power were

simulated, where results showed that PENN was able to replicate the difference in the pedestal region due to the variation in NBI power.

An additional test was done to investigate whether full edge profiles could be predicted, rather than only predicting the pedestal top values. This extended prediction includes the pedestal width, position, and core slope inside the pedestal top. Results showed that accurate predictions of the entire edge profiles could be achieved. However, this accuracy was largely attributable to the relatively narrow distributions of pedestal width, position, and core slope in the database. In other words, the model performed well primarily by predicting values close to the database mean for these parameters, provided the pedestal top prediction was accurate. This conclusion is supported by relatively low R^2-values (≈ 0.5) for the individual predictions of pedestal width, position, and core slope. Thus, while predicting the pedestal top constitutes a well-posed machine learning problem for this dataset, the same conclusion cannot be made for the pedestal width, position, and core slope.

6.2 A fast neural network surrogate model for the eigenvalues of QuaLiKiz

This is the first QuaLiKiz paper included in this thesis, and it was published in 2023. Given that less have been said about QuaLiKiz in the chapters of this thesis, a quick introduction to this project is warranted: QuaLiKiz [53] is a quasi-linear model that calculates turbulent transport in a plasma in two steps; 1) calculation of eigenvalues of micro instabilities using linear theory; 2) a saturation rule is applied for the eigenvalues to obtain the transport fluxes. QuaLiKiz can be used in integrated modeling frameworks, although it can be time consuming during long simulations or extensive analysis. Therefore, a machine learning surrogate model for QuaLiKiz (QLKNN [148]) was previously developed. However, a caveat with QuaLiKiz is that the saturation rule is calibrated using experimental data, which makes it challenging to predict for future machines since there is no guarantee that the saturation rule translates well.

The purpose of this paper is to investigate the feasibility to create a surrogate model for the part of QuaLiKiz that is robust and translates between machines, which is the calculation of the linear theory to obtain the eigenvalues of the instabilities. This is also the most computationally costly part of QuaLiKiz. Specifically, our objective is to explore the key considerations that must be taken into account from a machine learning perspective when solving this problem. A large dataset generated by the QuaLiKiz model was available for the training and evaluation of the model. The output in this problem consists of the growth rate and real frequency of instabilities at 18 different spatial scales, where the shortest scales generally correspond to the ETG-mode, and where the longer scales generally correspond to the ITG-mode and the TEM-mode. The inputs consists of 15 different plasma parameters, such as the ion temperature gradient.

In the data set, the growth rate (one of the outputs) is either positive

(unstable mode), or zero (stable or damped mode). Results showed that splitting the model into a stable/unstable classifier and a regression model for the unstable entries, yielded an accurate surrogate model as a whole. The accuracy could be further improved by using a weighted loss function for the classifier due to class imbalance between the stable/unstable classes.

An additional test was performed to investigate whether it is beneficial to first classify the instability type on the ion scale (ITG/TEM), and then use distinct regressions models for these two cases. However, this led to overall worse accuracy, which is mainly attributed to the regression models making poor predictions on wrongly classified entries. In other words, when the regression model trained only on ITG instabilities was fronted with a TEM example (wrongly classified as ITG), it performed worse than a regression model trained on both ITG and TEM data (even when the multi-instability regression model is not explicitly aware that there are different instability types).

The main conclusion of this paper is that the problem of predicting the eigenvalues in QuaLiKiz is well-posed (which is expected since the surrogate is based on a theoretical model where there exists an underlying mapping between the inputs and the outputs), and that unstable/stable classification, followed by a single regression model for all instabilities, resulted in the most accurate model.

6.3 Investigating pedestal dependencies at JET using an interpretable neural network architecture

In this paper, published in 2025, we introduce the interpretable NeuralBranch framework while demonstrating it on the pedestal prediction application. Besides introducing the method, the goal is to investigate in detail how the pedestal depends on key tokamak parameters across deuterium plasmas in the JET pedestal database using the machine learning interpretability. An additional goal is to provide transparent alternatives to the black-box pedestal prediction models we trained in the paper from 2022.

The main result of this paper is that we successfully reveal global intricate parameter dependencies and interactions that cannot be represented by simpler models like power scalings (that have previously been used to empirically investigate how the pedestal depends on key tokamak parameters). Specifically, key findings are:

- While both input power and plasma current are positively correlated with pedestal top pressure and temperature, NeuralBranch reveals an attenuating interaction. This means that increasing power weakens the impact that current has on pedestal pressure and temperature, and vice versa.

- There is an amplifying interaction between plasma current and triangularity, where higher triangularity amplifies the effect of plasma current

on pedestal density, and vice versa.

- Higher input power lowers the pedestal density, but only in the input power range $\approx 5 - 20$ MW. For higher input power, the pedestal density remains approximately constant when the power is further increased.

Moreover, the interpretable NeuralBranch models fully matched the accuracy of black-box models trained on the same data, which means that prediction transparency can be obtained without sacrificing accuracy for this application.

Furthermore, in this paper, we perform a more rigorous input parameter importance analysis compared to the 2022 paper. This analysis reveals that the following inputs generally contribute to higher prediction accuracy: plasma current I_P, total input power P_{tot}, traingularity δ, gas fueling rate Γ, and separatrix density $n_{e,sep}$. The inputs that turned out to be redundant were: toroidal magnetic field B, effective atomic number Z_{eff}, minor radius a, elongation κ, plasma volume V, safety factor q_{95}. Note that the plasma parameter β_N was not considered in this work as we chose to focus on key engineering parameters that can be determined before running the experiments (with the exception of $n_{e,sep}$, but we also created models where $n_{e,sep}$ was excluded).

6.4 Investigating characteristics of the growth rates from QuaLiKiz using an interpretable surrogate model

This paper, published in 2025, is a follow-up to the previous QuaLiKiz-paper from 2023. Essentially, the goal of this paper is to open up the black box to understand how the growth rates depends on the inputs of QuaLiKiz. Specifically, we chose to focus on the growth rate of the ITG instabilities, which is the primary contributor to turbulent transport in the plasma core.

There are several reasons for why interpreting surrogate models is relevant, for instance:

- While the system of equations in QuaLiKiz is known, the exact analytic solution remains unknown, which is why numerical methods are used. An interpretable surrogate model trained to replicate the original model could offer a transparent view of how the solution depends on the inputs. In other words, interpretability can enhance the understanding of the behavior of the surrogate model and the original model (QuaLiKiz).

- Even when there is a general understanding of the behavior of a model, interpretability remains important for validating that the surrogate behaves as expected. This is important for building trust in machine learning-based surrogates.

However, before implementing interpretable methods, we first performed an input feature importance analysis (this was not done in the QuaLiKiz paper from 2023 as we simply gave the model all of the 15 QuaLiKiz inputs available).

Results showed that only four inputs are necessary for the stable/unstable ITG classification sub-task, namely: ion temperature gradient R/L_{T_i}, magnetic shear $\hat{s}$, ion and electron temperature ratio τ, and electron density gradient L_{n_e}. For the growth rate regression sub-task, adding $E \times B$ shearing rate γ_E as an input to the four inputs already mentioned yielded over 90% of the accuracy compared to when using all inputs. Hence, for the regression sub-task, we chose to focus the analysis on these five most important inputs.

The NeuralBranch method was then used to create interpretable surrogate models for the two sub-tasks. Key findings were:

- There is an interaction between $\hat{s}$ and R/L_{n_e}. Specifically, by increasing R/L_{n_e}, the least stabilizing $\hat{s}$-value shifts from $\hat{s} \approx 0$ to $\hat{s} \approx 1$ (both sub-tasks).
- The stabilizing effect of γ_E is completely suppressed at low $\hat{s}$ in the regression sub-task.
- Low values of R/L_{T_i} also appear to suppress the stabilizing effect of γ_E in the regression sub-task.

In summary, this paper demonstrates that methods that enable interpretability can assist in providing deeper insights into the behavior of models like the QuaLiKiz eigenvalue solver and assist in making machine learning-based surrogate models more transparent and, therefore, more trustworthy.

6.5 Interpretability guided transfer learning approaches for tritium pedestal predictions

This last paper related to the pedestal have been submitted for publication in 2026. The goal of this paper is to extend the applicability of pedestal models previously trained on deuterium (D) plasmas (Paper III) to tritium (T) and DT mixtures. Enabling and understanding pedestal predictions for DT and T plasmas is important since a DT fuel mixture is a leading candidate for future fusion experiments.

Our approach was to use model interpretability to guide our choice of transfer learning strategy, and to investigate what impact the isotope composition has on the pedestal top. To accomplish this, we had access to 71 experimental DT/T samples from JET.

We first tried the transfer learning approach of fine-tuning the weights of the pre-trained pedestal models to fit the T/DT data. However, this resulted in severe overfitting due to multicollinearities and clusters in the T/DT data, which was revealed by the interpretability aspect of the models. Therefore, we instead deployed a more simple and robust calibration technique, where the pre-trained models were frozen and power scalings were fit on top of them to incorporate the effect of isotope mass. This resulted in improved accuracy compared to uncalibrated models, and isotope scaling consistent with previous research: pedestal density scales positively with increased isotope mass, and

pedestal temperature shows a weak negative scaling with increased isotope mass. However, as the pre-trained models are more accurate on the D data (before calibration), we concluded that the simple output calibration technique captures a decent approximation of the isotope impact on the pedestal, but not the full picture.

In summary, this work demonstrates the importance of interpretable models for understanding model behavior in transfer learning tasks, which in our case guided us to use a more simple scaling strategy. Given that sparse data with multicollinearities are encountered in magnetic confinement fusion research, this approach may be important for other tasks within the field beyond pedestal studies.

Chapter 7

Conclusion and outlook

7.1 Conclusion

The work included in this thesis demonstrates that it is entirely feasible to interpret and understand how the output of a machine learning model depends on the model's inputs, in particular for tabular problems in fusion with relatively few input parameters. The novel NeuralBranch method has shown to provide more global insights compared local explainability methods like SHAP or LIME, while maintaining high model expressivity such that accuracy is not sacrificed. In this thesis, the method revealed new insights about empirical relationships between the pedestal and machine parameters, as well as relationships between ITG growth rates and plasma parameters in the QuaLiKiz model. However, as the method is not limited to a specific use case, it may also find use elsewhere, even beyond the field of fusion.

In addition to the specific results presented, a goal of the work has been to spread awareness that we indeed can and should attempt to understand machine learning models deployed in fusion (and in other fields). As we as a species move forward, it may become even more important that we use and improve on interpretability techniques that make model behavior comprehensible to humans. Achieving interpretability for tabular problems may seem like just a small puzzle piece in the big picture, which it to a large extent is. However, as long as humans are around, it is likely that the idea of breaking down overwhelming data and tasks into key parameters will remain.

7.2 Potential avenues for future work

7.2.1 Pedestal predictions

The findings in this thesis related to the presented pedestal models are limited to predictions and analysis for the JET tokamak. It will be necessary to create sufficiently large and diverse pedestal datasets for other devices to investigate how global empirical pedestal dependencies translate between machines. This is also important for extrapolations of how the pedestal might behave at future

machines like ITER. Moreover, due to less available ion temperature and density diagnostics, it remains to explore how the ion pedestal deviates from the behavior of the electron pedestal.

It also remains to investigate how pedestal dependencies vary when comparing interpretable models trained on real experimental data versus models trained on synthetic data generated from theoretical models, such as EPED. Such comparisons may reveal which qualitative global patterns that theoretical models fail to capture. This is also of high relevance for future machine predictions, since surrogate modeling is one of the main viable options for cases where there is no real experimental data yet.

Finally, it would be valuable to perform an interpretability-facilitated pedestal dependency analysis using dimensionless plasma parameters rather than the engineering parameters used in the presented papers. Dimensionless parameters such as plasma collisionality ν^* and normalized gyroradius ρ^* enable more robust comparisons across different machines, as they capture the underlying physics independent of device size and engineering specifications. Consequently, insights derived from such analyses would potentially be more readily extrapolated to future machines.

7.2.2 Update other power scalings

The interpretable methods presented in this thesis are not limited to predicting the pedestal. Potentially, they could be used to provide more detailed insights into other properties that have previously been predicted using simple power scalings, such as the L-H mode transition power, the stored energy in the plasma core, and the thermal energy confinement time.

7.2.3 Surrogate modeling

As surrogate modeling gradually becomes more prominent in fusion research, interpretability may play an important part in deepening our understanding of the models we deploy. Moreover, there is no downside to making surrogate models more interpretable to ensure that their learned patterns are regularized and reasonable.

7.2.4 Physics-informed neural networks

The physics-informed neural network (PINN) approach is conceptually different from the classical supervised learning approach. Specifically, PINNs are used to parametrize the solution in differential equations, where the equation residual is used to represent the loss. As an example, recent work has demonstrated that PINNs can be useful for discovering new mathematical singularities in fluid equations [149], which is a step towards solving the challenge of finding any singularity in the Navier-Stokes equations (one of the six famous Millennium Prize Problems that are still unsolved).

Future work includes exploring how interpretability may assist in understanding PINNs, in particular when they are used in fusion research.

7.2.5 High-dimensionality problems

As mentioned in Chapter 5, fusion research increasingly deploy machine learning for problems where inputs are structured as 1D or 2D vectors, time series, or other formats for which tabular interpretability techniques are unsuitable. As such, future work involves exploring how new interpretability techniques, such as attribution methods and latent traversal methods may fit for different high-dimensional problems in fusion.

An ambitious idea would be to train models on large amounts of high-dimensional data, potentially with self-supervised techniques, to investigate if a general understanding of fusion plasmas can be learned (that may prove useful for several downstream tasks). Interpretability may assist in understanding such models and, optimally, lead to knowledge discovery that helps us move forward towards realizing the promise of fusion as an environmentally friendly and practically limitless energy source.

Bibliography

[1] C. Nordling and J. Österman, *Physics Handbook*, 9th ed. Studentlitteratur, 1980 (cit. on pp. 3, 6, 26).

[2] J. J. Sakurai and J. Napolitano, *Modern Quantum Mechanics*, 2nd ed. Cambridge University Press, 1985 (cit. on pp. 3, 4).

[3] C. G. Tully, *Elementary Particle Physics in a Nutshell*, 1st ed. Princeton University Press, 2011 (cit. on pp. 3, 4).

[4] F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion*, 3rd ed. Springer, 2016 (cit. on pp. 4, 6, 13, 16, 19, 20, 27, 29–31, 33).

[5] D. G. Griffiths, *Introduction to Electrodynamics*, 4th ed. Cambridge University Press, 1999 (cit. on p. 4).

[6] F. Mandl, *Statistical Physics*, 2nd ed. Wiley, 2023 (cit. on pp. 4, 34).

[7] IAEA, *Iaea nuclear data services*, `https://www-nds.iaea.org/`, Accessed: 2023-11-20, Year Published: 2007/ Last Updated: 2023 (cit. on p. 5).

[8] K. A. Brueckner, *Inertial Confinement Fusion*, 1st ed. Springer, 1998 (cit. on p. 4).

[9] S. Geng, "An overview of the iter project," *Journal of Physics: Conference Series*, vol. 2386, no. 1, p. 012 012, 2022. DOI: `10.1088/1742-6596/2386/1/012012`. [Online]. Available: `https://dx.doi.org/10.1088/1742-6596/2386/1/012012` (cit. on p. 4).

[10] E. Moses and the NIC Collaborators, "The national ignition campaign: Status and progress," *Nuclear Fusion*, vol. 53, no. 10, p. 104 020, 2013. DOI: `10.1088/0029-5515/53/10/104020`. [Online]. Available: `https://dx.doi.org/10.1088/0029-5515/53/10/104020` (cit. on pp. 5, 6).

[11] R. M. D. Kirtley, "Fundamental scaling of adiabatic compression of field reversed configuration thermonuclear fusion plasmas," *Journal of Fusion Energy*, vol. 42, no. 30, 2023. [Online]. Available: `https://transformer-circuits.pub/2025/attribution-graphs/biology.html` (cit. on p. 6).

[12] J. Lilley, *Nuclear Physics: Principles and Applications*, 1st ed. Wiley, 2001 (cit. on p. 6).

[13] A. Chagnes and J. Swiatowska, *Lithium Process Chemestry: Resources, Extraction, Batteries and Recycling*, 1st ed. Elsevier, 2015 (cit. on p. 6).

[14] A. Markandya and P. Wilkinson, "Electricity generation and health," *The Lancet*, vol. 370, no. 9591, pp. 979–990, 2007, ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(07)61253-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673607612537 (cit. on p. 6).

[15] M. O. M. Holt T. Kuiken, "Iter—an international nuclear fusion research and development facility," 2025. [Online]. Available: https://www.congress.gov/crs-product/R48362 (cit. on p. 7).

[16] CERN, "Facts and figures about the lhc," [Online]. Available: https://home.cern/resources/faqs/facts-and-figures-about-lhc (cit. on p. 7).

[17] R Aymar, P Barabaschi and Y Shimomura, "The iter design," *Plasma Physics and Controlled Fusion*, vol. 44, no. 5, p. 519, 2002. DOI: 10.1088/0741-3335/44/5/304. [Online]. Available: https://doi.org/10.1088/0741-3335/44/5/304 (cit. on p. 8).

[18] F. Fleschner, "Wendelstein 7-x sets new performance records in fusion research," 2025. [Online]. Available: https://www.ipp.mpg.de/5532945/w7x (cit. on p. 8).

[19] "Achieving fusion ignition," 2025. [Online]. Available: https://lasers.llnl.gov/science/achieving-fusion-ignition (cit. on p. 8).

[20] J. H. Nuckolls, "Grand challenges of inertial fusion energy," *Journal of Physics: Conference Series*, vol. 244, no. 1, p. 012 007, 2010. DOI: 10.1088/1742-6596/244/1/012007. [Online]. Available: https://doi.org/10.1088/1742-6596/244/1/012007 (cit. on p. 8).

[21] T. D. Chant, "Every fusion startup that has raised over $100M$," 2025. [Online]. Available: https://techcrunch.com/2025/09/01/every-fusion-startup-that-has-raised-over-100m/ (cit. on p. 9).

[22] Štancar, K. Kirov, F. Auriemma *et al.*, "Overview of interpretive modelling of fusion performance in jet dte2 discharges with transp," *Nuclear Fusion*, vol. 63, no. 12, p. 126 058, 2023. DOI: 10.1088/1741-4326/ad0310. [Online]. Available: https://doi.org/10.1088/1741-4326/ad0310 (cit. on p. 10).

[23] G. Cenacchi, A. Taroni and R. I. ENEA, "Jetto a free boundary plasma transport code," 1988 (cit. on p. 10).

[24] D. Lopez-Bruna, F. Castejon, J. M. Fontdecaba and M. S. CENTRO DE INVESTIGACIONES ENERGETICAS MEDIOAMBIENTALES Y TECNOLOGICAS (CIEMAT), "Transport with astra in tj-ii," 2004 (cit. on p. 10).

[25] D. P. Coster, V. Basiuk, G. Pereverzev *et al.*, "The european transport solver," *IEEE Transactions on Plasma Science*, vol. 38, no. 9, pp. 2085–2092, 2010. DOI: 10.1109/TPS.2010.2056707 (cit. on p. 10).

[26] Y. LeCun, B. Boser, J. Denker *et al.*, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2, Morgan-Kaufmann, 1989 (cit. on p. 10).

[27] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278 –2324, Dec. 1998. DOI: 10.1109/5.726791 (cit. on p. 10).

[28] D. Silver, A. Huang, C. Maddison *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, Jan. 2016. DOI: 10.1038/nature16961 (cit. on p. 10).

[29] T. Brown, B. Mann, N. Ryder *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901 (cit. on p. 10).

[30] A. Esteva, B. Kuprel, R. A. Novoa *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3767412 (cit. on p. 10).

[31] V. Gulshan, L. Peng, M. Coram *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, ISSN: 0098-7484. DOI: 10.1001/jama.2016.17216. eprint: https://jamanetwork.com/journals/jama/articlepdf/2588763/joi160132.pdf. [Online]. Available: https://doi.org/10.1001/jama.2016.17216 (cit. on p. 10).

[32] J. Jumper, R. Evans, A. Pritzel *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. [Online]. Available: https://doi.org/10.1038/s41586-021-03819-2 (cit. on p. 10).

[33] O. Dehzangi and M. Farooq, "Ssvep recognition using discriminative score fusion and transformation for portable brain computer interface," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2018, pp. 1–4. DOI: 10.1109/BHI.2018.8333355 (cit. on p. 10).

[34] J. F. Gemmeke, D. P. W. Ellis, D. Freedman *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261 (cit. on p. 10).

[35] J. Kates-Harbeck, A. Svyatkovskiy and W. Tang, "Predicting disruptive instabilities in controlled fusion plasmas through deep learning," *Nature*, vol. 568, Apr. 2019. DOI: 10.1038/s41586-019-1116-4 (cit. on p. 10).

[36] S. Dasbach and S. Wiesen, "Towards fast surrogate models for interpolation of tokamak edge plasmas," *Nuclear Materials and Energy*, vol. 34, p. 101 396, 2023, ISSN: 2352-1791. DOI: https://doi.org/10.1016/j.nme.2023.101396. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352179123000352 (cit. on p. 10).

[37] F. Wagner, G. Becker, K. Behringer *et al.*, "Regime of improved confinement and high beta in neutral-beam-heated divertor discharges of the asdex tokamak," *Phys. Rev. Lett.*, vol. 49, pp. 1408–1412, 19 1982. DOI: 10.1103/PhysRevLett.49.1408. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.49.1408 (cit. on pp. 11, 35).

[38] J. Kinsey, G. Staebler, J. Candy, R. Waltz and R. Budny, "Iter predictions using the gyro verified and experimentally validated trapped gyro-landau fluid transport model," *Nuclear Fusion*, vol. 51, no. 8, p. 083 001, 2011. DOI: 10.1088/0029-5515/51/8/083001. [Online]. Available: https://doi.org/10.1088/0029-5515/51/8/083001 (cit. on p. 11).

[39] P. Rodriguez-Fernandez, N. T. Howard, M. J. Greenwald *et al.*, "Predictions of core plasma performance for the sparc tokamak," *Journal of Plasma Physics*, vol. 86, no. 5, 2020. DOI: 10.1017/S0022377820001075 (cit. on p. 11).

[40] L. Frassinetti, S. Saarelma, G. Verdoolaege *et al.*, "Pedestal structure, stability and scalings in jet-ilw : The eurofusion jet-ilw pedestal database," *Nuclear Fusion*, vol. 61, no. 1, 016001, 2021, QC 20210113. DOI: 10.1088/1741-4326/abb79e (cit. on pp. 11, 35, 36, 40, 59, 71).

[41] P. B. Snyder, R. J. Groebner, A. W. Leonard, T. H. Osborne and H. R. Wilson, "Development and validation of a predictive model for the pedestal heighta)," *Physics of Plasmas*, vol. 16, no. 5, p. 056 118, May 2009, ISSN: 1070-664X. DOI: 10.1063/1.3122146. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.3122146/14032267/056118_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.3122146 (cit. on pp. 11, 39).

[42] S. Saarelma, J. Connor, P. Bilkova *et al.*, "Testing a prediction model for the h-mode density pedestal against jet-ilw pedestals," *Nuclear Fusion*, vol. 63, no. 5, p. 052 002, 2023. DOI: 10.1088/1741-4326/acc084. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/acc084 (cit. on pp. 11, 39).

[43] J. P. Freidberg, *Plasma Physics and Fusion Energy*, 1st ed. Cambridge University Press, 2007 (cit. on pp. 25, 30, 31, 33, 34).

[44] D. S. D. George V. Pereverzev and T. P. Group, *Plasma Disruptions in Tokamaks*, 1st ed. Springer, 2004 (cit. on p. 28).

[45] H. Xie, V. S. Chan, R. Ding *et al.*, "Evaluation of tritium burnup fraction for cfetr scenarios with core-edge coupling simulations," *Nuclear Fusion*, vol. 60, no. 4, p. 046 022, 2020. DOI: 10.1088/1741-4326/ab742b. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/ab742b (cit. on p. 30).

[46] G. A. Houlberg, *Radio Frequency Plasma Heating*, 1st ed. CRC Press, 1993 (cit. on p. 31).

[47] H. Fajemirokun, C. Gowers, P. Nielsen, K. Hirsch, T. Kajiwara and H. Salzmann, "High resolution lidar thomson scattering at jet," p. 31–50. 1991 (cit. on p. 31).

[48] A Boboc, J Macdonald, R Felton *et al.*, "Jet far-infrared interferometer/polarimeter diagnostic system—40 years of lessons learned," *Plasma Physics and Controlled Fusion*, vol. 66, no. 8, p. 085 011, 2024. DOI: 10.1088/1361-6587/ad5376. [Online]. Available: https://doi.org/10.1088/1361-6587/ad5376 (cit. on p. 31).

[49] K. F. Mast, H. Krause, K. Behringer, A. Bulliard and G. Magyar, "Bolometric diagnostics in jet," *Review of Scientific Instruments*, vol. 56, no. 5, pp. 969–971, May 1985, ISSN: 0034-6748. DOI: 10.1063/1.1138007. eprint: https://pubs.aip.org/aip/rsi/article-pdf/56/5/969/19116345/969_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.1138007 (cit. on p. 31).

[50] E. Fransson, "Doctoral thesis," Ph.D. dissertation, Chalmers University of Technology, 2023 (cit. on p. 34).

[51] J. Weiland, *Stability and Transport in Magnetic Confinement Systems*, 1st ed. Springer, 2012 (cit. on p. 35).

[52] G. M. Staebler, J. E. Kinsey and R. E. Waltz, "Gyro-Landau fluid equations for trapped and passing particles," *Physics of Plasmas*, vol. 12, no. 10, p. 102 508, 2005 (cit. on p. 35).

[53] C. Bourdelle, X. Garbet, F. Imbeaux *et al.*, "A new gyrokinetic quasilinear transport model applied to particle transport in tokamak plasmas," *Physics of Plasmas*, vol. 14, no. 11, p. 112 501, Nov. 2007 (cit. on pp. 35, 72).

[54] M. Fenstermacher, L. Baylor, E. de la Luna *et al.*, "Progress in pedestal and edge physics: Chapter 3 of the special issue: On the path to tokamak burning plasma operation," *Nuclear Fusion*, vol. 65, no. 5, p. 053 001, 2025. DOI: 10.1088/1741-4326/adb1f3. [Online]. Available: https://doi.org/10.1088/1741-4326/adb1f3 (cit. on pp. 35, 38).

[55] A. W. Leonard, "Edge-localized-modes in tokamaksa)," *Physics of Plasmas*, vol. 21, no. 9, p. 090 501, Sep. 2014 (cit. on p. 35).

[56] C. Maggi, L. Frassinetti, L. Horvath *et al.*, "Studies of the pedestal structure and inter-elm pedestal evolution in jet with the iter-like wall," *Nuclear Fusion*, vol. 57, no. 11, p. 116 012, 2017 (cit. on p. 35).

[57] L. Gil, C. Silva, T. Happel *et al.*, "Stationary elm-free h-mode in asdex upgrade," *Nuclear Fusion*, vol. 60, no. 5, p. 054 003, 2020. DOI: 10.1088/1741-4326/ab7d1b. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/ab7d1b (cit. on p. 36).

[58] J. W. Connor, R. J. Hastie, H. R. Wilson and R. L. Miller, "Magnetohydrodynamic stability of tokamak edge plasmas," *Physics of Plasmas*, vol. 5, no. 7, pp. 2687–2700, Jul. 1998, ISSN: 1070-664X. DOI: 10.1063/1.872956. eprint: https://pubs.aip.org/aip/pop/article-pdf/5/7/2687/19160451/2687_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.872956 (cit. on p. 37).

[59] P. B. Snyder, H. R. Wilson, T. H. Osborne and A. W. Leonard, "Characterization of peeling–ballooning stability limits on the pedestal," *Plasma Physics and Controlled Fusion*, vol. 46, no. 5A, A131, 2004. DOI: 10.1088/0741-3335/46/5A/014. [Online]. Available: https://doi.org/10.1088/0741-3335/46/5A/014 (cit. on p. 37).

[60] K. V. Roberts and J. B. Taylor, "Magnetohydrodynamic equations for finite larmor radius," *Phys. Rev. Lett.*, vol. 8, pp. 197–198, 5 1962. DOI: 10.1103/PhysRevLett.8.197. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.8.197 (cit. on p. 37).

[61] P. B. Snyder, H. R. Wilson, J. R. Ferron *et al.*, "Edge localized modes and the pedestal: A model based on coupled peeling–ballooning modes," *Physics of Plasmas*, vol. 9, no. 5, pp. 2037–2043, May 2002, ISSN: 1070-664X. DOI: 10.1063/1.1449463. eprint: https://pubs.aip.org/aip/pop/article-pdf/9/5/2037/19228425/2037_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.1449463 (cit. on p. 37).

[62] X. Q. Xu, B. Dudson, P. B. Snyder, M. V. Umansky and H. Wilson, "Nonlinear simulations of peeling-ballooning modes with anomalous electron viscosity and their role in edge localized mode crashes," *Phys. Rev. Lett.*, vol. 105, p. 175 005, 17 2010. DOI: 10.1103/PhysRevLett.105.175005. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.105.175005 (cit. on p. 38).

[63] G. Huysmans, T. Hender and B. Alper, "Identification of external kink modes in jet," *Nuclear Fusion*, vol. 38, no. 2, p. 179, 1998. DOI: 10.1088/0029-5515/38/2/303. [Online]. Available: https://doi.org/10.1088/0029-5515/38/2/303 (cit. on p. 38).

[64] M Bécoulet, G Huysmans, Y Sarazin *et al.*, "Edge localized mode physics and operational aspects in tokamaks," *Plasma Physics and Controlled Fusion*, vol. 45, no. 12A, A93, 2003. DOI: 10.1088/0741-3335/45/12A/007. [Online]. Available: https://doi.org/10.1088/0741-3335/45/12A/007 (cit. on p. 38).

[65] J. Ma, X. Xu and B. Dudson, "Linear peeling–ballooning mode simulations in snowflake-like divertor configuration using bout++ code," *Nuclear Fusion*, vol. 54, no. 3, p. 033 011, 2014. DOI: 10.1088/0029-5515/54/3/033011. [Online]. Available: https://doi.org/10.1088/0029-5515/54/3/033011 (cit. on p. 38).

[66] N. Aiba, M. Furukawa, M. Hirota *et al.*, "Mechanisms of plasma rotation effects on the stability of type-i edge-localized mode in tokamaks," *Nuclear Fusion*, vol. 51, no. 7, p. 073 012, 2011. DOI: 10.1088/0029-5515/51/7/073012. [Online]. Available: https://doi.org/10.1088/0029-5515/51/7/073012 (cit. on p. 38).

[67] A. Burckhart, M. Dunne, E. Wolfrum *et al.*, "Elm behaviour and linear mhd stability of edge ecrh heated asdex upgrade plasmas," *Nuclear Fusion*, vol. 56, no. 5, p. 056 011, 2016. DOI: 10.1088/0029-5515/56/5/056011. [Online]. Available: https://doi.org/10.1088/0029-5515/56/5/056011 (cit. on p. 38).

[68] C. Maggi, S. Saarelma, F. Casson *et al.*, "Pedestal confinement and stability in jet-ilw elmy h-modes," *Nuclear Fusion*, vol. 55, no. 11, p. 113 031, 2015. DOI: 10.1088/0029-5515/55/11/113031. [Online]. Available: https://doi.org/10.1088/0029-5515/55/11/113031 (cit. on pp. 38, 39).

[69] M. Beurskens, L. Frassinetti, C. Challis *et al.*, "Global and pedestal confinement in jet with a be/w metallic wall," *Nuclear Fusion*, vol. 54, no. 4, p. 043 001, 2014. DOI: 10.1088/0029-5515/54/4/043001. [Online]. Available: https://doi.org/10.1088/0029-5515/54/4/043001 (cit. on p. 38).

[70] H. Nyström, L. Frassinetti, S. Saarelma *et al.*, "Effect of resistivity on the pedestal mhd stability in jet," *Nuclear Fusion*, vol. 62, no. 12, p. 126 045, 2022. DOI: 10.1088/1741-4326/ac9701. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/ac9701 (cit. on p. 38).

[71] P. Snyder, R. Groebner, J. Hughes *et al.*, "A first-principles predictive model of the pedestal height and width: Development, testing and iter optimization with the eped model," *Nuclear Fusion*, vol. 51, no. 10, p. 103 016, 2011. DOI: 10.1088/0029-5515/51/10/103016. [Online]. Available: https://doi.org/10.1088/0029-5515/51/10/103016 (cit. on p. 39).

[72] M. N. A. Beurskens, T. H. Osborne, P. A. Schneider *et al.*, "H-mode pedestal scaling in diii-d, asdex upgrade, and jet a)," *Physics of Plasmas*, vol. 18, no. 5, p. 056 120, May 2011, ISSN: 1070-664X. DOI: 10.1063/1.3593008. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.3593008/15959602/056120_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.3593008 (cit. on p. 39).

[73] U. A. Sheikh, M Dunne, L Frassinetti *et al.*, "Pedestal structure and energy confinement studies on tcv," *Plasma Physics and Controlled Fusion*, vol. 61, no. 1, p. 014 002, 2018. DOI: 10.1088/1361-6587/aae7bd. [Online]. Available: https://doi.org/10.1088/1361-6587/aae7bd (cit. on p. 39).

[74] T. Luda, C. Angioni, M. Dunne *et al.*, "Integrated modeling of asdex upgrade plasmas combining core, pedestal and scrape-off layer physics," *Nuclear Fusion*, vol. 60, no. 3, p. 036 023, 2020. DOI: 10.1088/1741-4326/ab6c77. [Online]. Available: https://doi.org/10.1088/1741-4326/ab6c77 (cit. on p. 39).

[75] T Luda, C Angioni, M. G. Dunne *et al.*, "Validation of imep on alcator c-mod and jet-ilw elmy h-mode plasmas," *Plasma Physics and Controlled Fusion*, vol. 65, no. 3, p. 034 001, 2023. DOI: 10.1088/1361-6587/acb011. [Online]. Available: https://doi.org/10.1088/1361-6587/acb011 (cit. on p. 39).

[76] S. Saarelma, J. Connor, P. Bílková *et al.*, "Density pedestal prediction model for tokamak plasmas," *Nuclear Fusion*, vol. 64, no. 7, p. 076 025, 2024. DOI: 10.1088/1741-4326/ad4b3e. [Online]. Available: https://doi.org/10.1088/1741-4326/ad4b3e (cit. on p. 39).

[77] S. Saarelma, L. Frassinetti, P. Bilkova *et al.*, "Self-consistent pedestal prediction for jet-ilw in preparation of the dt campaign," *Physics of Plasmas*, vol. 26, no. 7, p. 072 501, Jul. 2019, ISSN: 1070-664X. DOI: 10.1063/1.5096870. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.5096870/16687623/072501_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.5096870 (cit. on p. 40).

[78] C. Challis, J. Garcia, M. Beurskens *et al.*, "Improved confinement in jet high plasmas with an iter-like wall," *Nuclear Fusion*, vol. 55, no. 5, p. 053 031, 2015. DOI: 10.1088/0029-5515/55/5/053031. [Online]. Available: https://dx.doi.org/10.1088/0029-5515/55/5/053031 (cit. on p. 40).

[79] E. Stefanikova, L. Frassinetti, S. Saarelma *et al.*, "Effect of the relative shift between the electron density and temperature pedestal position on the pedestal stability in jet-ilw and comparison with jet-c," *Nuclear Fusion*, vol. 58, no. 5, p. 056 010, 2018. DOI: 10.1088/1741-4326/aab216. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/aab216 (cit. on p. 40).

[80] C. Bowman, D. Dickinson, L. Horvath *et al.*, "Pedestal evolution physics in low triangularity jet tokamak discharges with iter-like wall," *Nuclear Fusion*, vol. 58, no. 1, p. 016 021, 2017. DOI: 10.1088/1741-4326/aa90bc. [Online]. Available: https://dx.doi.org/10.1088/1741-4326/aa90bc (cit. on p. 40).

[81] J. Cordey, for the ITPA H-Mode Database Working Group and the ITPA Pedestal Database Working Group, "A two-term model of the confinement in elmy h-modes using the global confinement and pedestal databases," *Nuclear Fusion*, vol. 43, no. 8, p. 670, 2003. DOI: 10.1088/0029-5515/43/8/305. [Online]. Available: https://dx.doi.org/10.1088/0029-5515/43/8/305 (cit. on p. 40).

[82] N. Li, X. Xu, Y. Wang *et al.*, "Characteristics of grassy elms and their impact on the divertor heat flux width," *Nuclear Fusion*, vol. 62, no. 9, p. 096 030, 2022. DOI: 10.1088/1741-4326/ac83d9. [Online]. Available: https://doi.org/10.1088/1741-4326/ac83d9 (cit. on p. 41).

[83] M. Faitsch, M. Dunne, E. Lerche *et al.*, "The quasi-continuous exhaust regime in jet," *Nuclear Fusion*, vol. 65, no. 2, p. 024 003, 2025. DOI: 10.1088/1741-4326/adaa86. [Online]. Available: https://doi.org/10.1088/1741-4326/adaa86 (cit. on p. 41).

[84] L. Meier, M. Hoelzl, A. Cathey *et al.*, "Mhd simulations of formation, sustainment and loss of quiescent h-mode in the all-tungsten asdex upgrade," *Nuclear Fusion*, vol. 63, no. 8, p. 086 026, 2023. DOI: 10.1088/1741-4326/acd5e2. [Online]. Available: https://doi.org/10.1088/1741-4326/acd5e2 (cit. on p. 41).

[85] K. Burrell, X. Chen, C. Chrystal *et al.*, "Creation and sustainment of wide pedestal quiescent h-mode with zero net neutral beam torque," *Nuclear Fusion*, vol. 60, no. 8, p. 086 005, 2020. DOI: 10.1088/1741-4326/ab940d. [Online]. Available: https://doi.org/10.1088/1741-4326/ab940d (cit. on p. 41).

[86] X. Zhong, X. Zou, A. Liu *et al.*, "I-mode plasma confinement improvement by real-time lithium injection and its classification on east tokamak," *Nuclear Fusion*, vol. 64, no. 12, p. 126 040, 2024. DOI: 10.1088/1741-4326/ad80a8. [Online]. Available: https://doi.org/10.1088/1741-4326/ad80a8 (cit. on p. 41).

[87] H. S. Wilson, A. O. Nelson, J McClenaghan, P Rodriguez-Fernandez, J Parisi and C Paz-Soldan, "Characterizing the negative triangularity reactor core operating space with integrated modeling," *Plasma Physics and Controlled Fusion*, vol. 67, no. 1, p. 015 026, 2024. DOI: 10.1088/1361-6587/ad9be5. [Online]. Available: https://doi.org/10.1088/1361-6587/ad9be5 (cit. on p. 41).

[88] X. Wu, Y. Sun, Q. Ma *et al.*, "Influence of n=4 rmps on pedestal structure and stability in east," *Nuclear Fusion*, vol. 65, no. 7, p. 076 031, 2025. DOI: 10.1088/1741-4326/ade268. [Online]. Available: https://doi.org/10.1088/1741-4326/ade268 (cit. on p. 42).

[89] P. Lang, G. Conway, T. Eich *et al.*, "Elm pace making and mitigation by pellet injection in asdex upgrade," *Nuclear Fusion*, vol. 44, no. 5, p. 665, 2004. DOI: 10.1088/0029-5515/44/5/010. [Online]. Available: https://doi.org/10.1088/0029-5515/44/5/010 (cit. on p. 42).

[90] E. de la Luna, I. Chapman, F. Rimini *et al.*, "Understanding the physics of elm pacing via vertical kicks in jet in view of iter," *Nuclear Fusion*, vol. 56, no. 2, p. 026 001, 2015. DOI: 10.1088/0029-5515/56/2/026001. [Online]. Available: https://doi.org/10.1088/0029-5515/56/2/026001 (cit. on p. 42).

[91] A Panera Alvarez, A Ho, A Järvinen *et al.*, "Europed-nn: Uncertainty aware surrogate model," *Plasma Physics and Controlled Fusion*, vol. 66, no. 9, p. 095 012, 2024. DOI: 10.1088/1361-6587/ad6707. [Online]. Available: https://doi.org/10.1088/1361-6587/ad6707 (cit. on pp. 42, 44, 53, 58).

[92] J. Abbate, R. Conlin and E. Kolemen, "Data-driven profile prediction for diii-d," *Nuclear Fusion*, vol. 61, no. 4, p. 046 027, 2021. DOI: 10.1088/1741-4326/abe08d. [Online]. Available: https://doi.org/10.1088/1741-4326/abe08d (cit. on p. 42).

[93] A. M. Bruncrona, A. Kit, A. E. Järvinen, S. Saarelma, L. Frassinetti and J. Contributors, "Machine learning surrogate model for ideal peeling–ballooning pedestal mhd stability," *Physics of Plasmas*, vol. 32, no. 9, p. 092 501, Sep. 2025 (cit. on pp. 42, 44).

[94] L. Zanisi, A. Ho, J. Barr *et al.*, "Efficient training sets for surrogate models of tokamak turbulence with active deep ensembles," *Nuclear Fusion*, vol. 64, no. 3, p. 036 022, 2024. DOI: 10.1088/1741-4326/ad240d. [Online]. Available: https://doi.org/10.1088/1741-4326/ad240d (cit. on p. 43).

[95] A Kit, A. E. Järvinen, L Frassinetti, S Wiesen and J. Contributors, "Supervised learning approaches to modeling pedestal density," *Plasma Physics and Controlled Fusion*, vol. 65, no. 4, p. 045 003, 2023. DOI: 10.1088/1361-6587/acb3f7. [Online]. Available: https://doi.org/10.1088/1361-6587/acb3f7 (cit. on pp. 43, 44).

[96] J. F. Parisi, J. G. Clark, J. W. Berkery *et al.*, *Hiped: Machine learning framework for spherical tokamak pedestal prediction and optimization*, 2025. arXiv: 2504.19861 [physics.plasm-ph]. [Online]. Available: https://arxiv.org/abs/2504.19861 (cit. on pp. 43, 44).

[97] E. U. Zeger, F. M. Laggner, A. Bortolon *et al.*, "Prediction of diii-d pedestal structure from externally controllable parameters," *IEEE Transactions on Plasma Science*, vol. 49, no. 10, pp. 3212–3227, 2021. DOI: 10.1109/TPS.2021.3114608 (cit. on pp. 43, 44).

[98] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org (cit. on pp. 45, 52, 53).

[99] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–408, Nov. 1958. DOI: 10.1037/h0042519 (cit. on p. 45).

[100] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, ISSN: 1476-4687. DOI: 10.1038/323533a0. [Online]. Available: https://doi.org/10.1038/323533a0 (cit. on p. 49).

[101] A. Vaswani, N. Shazeer, N. Parmar *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762 (cit. on pp. 52, 53).

[102] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385 (cit. on pp. 52, 53).

[103] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (cit. on p. 53).

[104] W. Zaremba, I. Sutskever and O. Vinyals, *Recurrent neural network regularization*, 2015. arXiv: 1409.2329 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1409.2329 (cit. on p. 53).

[105] I. Sutskever, O. Vinyals and Q. V. Le, *Sequence to sequence learning with neural networks*, 2014. arXiv: 1409.3215 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1409.3215 (cit. on p. 53).

[106] A. Rush, "The annotated transformer," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, E. L. Park, M. Hagiwara, D. Milajevs and L. Tan, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 52–60. DOI: 10.18653/v1/W18-2509. [Online]. Available: https://aclanthology.org/W18-2509/ (cit. on p. 53).

[107] B. Sanchez-Lengeling, E. Reif, A. Pearce and A. B. Wiltschko, "A gentle introduction to graph neural networks," *Distill*, 2021, https://distill.pub/2021/gnn-intro. DOI: 10.23915/distill.00033 (cit. on p. 53).

[108] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324 (cit. on pp. 53, 57).

[109] J. Kaplan, S. McCandlish, T. Henighan *et al.*, *Scaling laws for neural language models*, 2020. arXiv: 2001.08361 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2001.08361 (cit. on p. 54).

[110] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, *Masked autoencoders are scalable vision learners*, 2021. arXiv: 2111.06377 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.06377 (cit. on p. 54).

[111] Y. LeCun, *A path towards autonomous machine intelligence*, 2022. [Online]. Available: https://openreview.net/forum?id=BZ5a1r-kVsf (cit. on p. 54).

[112] T. B. Brown, B. Mann, N. Ryder *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.14165 (cit. on p. 54).

[113] C. Molnar, G. Casalicchio and B. Bischl, "Interpretable machine learning – a brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops*, I. Koprinska, M. Kamp, A. Appice *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 417–431, ISBN: 978-3-030-65965-3 (cit. on p. 57).

[114] C. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (Carl Friedrich Gauss Werke). Sumtibus F. Perthes et I.H. Besser, 1809. [Online]. Available: https://books.google.se/books?id=ORUOAAAAQAAJ (cit. on p. 57).

[115] A. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805. [Online]. Available: https://books.google.se/books?id=Ia8WAAAAQAAJ (cit. on p. 57).

[116] "Regression shrinkage and selection via the lasso," vol. 58, no. 1, pp. 267–288, 1996, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2346178 (visited on 21/10/2025) (cit. on p. 57).

[117] W. W. Cohen, "Fast effective rule induction," in *Machine Learning Proceedings 1995*, A. Prieditis and S. Russell, Eds., San Francisco (CA): Morgan Kaufmann, 1995, pp. 115–123, ISBN: 978-1-55860-377-6. DOI: https://doi.org/10.1016/B978-1-55860-377-6.50023-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781558603776500232 (cit. on p. 57).

[118] L. L. D. O. P. Svenmarck and U. W. Bolin, "Explainable artificial intelligence: Exploring xai techniques in military deep learning applications (foi-r–4849–se)," 2020 (cit. on p. 57).

[119] J. Cordey, for the ITPA H-Mode Database Working Group and the ITPA Pedestal Database Working Group, "A two-term model of the confinement in elmy h-modes using the global confinement and pedestal databases," *Nuclear Fusion*, vol. 43, no. 8, p. 670, 2003. DOI: 10.1088/0029-5515/43/8/305. [Online]. Available: https://doi.org/10.1088/0029-5515/43/8/305 (cit. on p. 59).

[120] I. P. E. G. on Confinement, Transport, I. P. E. G. on Confinement Modelling, Database and I. P. B. Editors, "Chapter 2: Plasma confinement and transport," *Nuclear Fusion*, vol. 39, no. 12, p. 2175, 1999. DOI: 10.1088/0029-5515/39/12/302. [Online]. Available: https://doi.org/10.1088/0029-5515/39/12/302 (cit. on p. 59).

[121] Y. R. Martin, T Takizuka and (andthe ITPA CDBM H-mode Threshold Database Working Group), "Power requirement for accessing the h-mode in iter," *Journal of Physics: Conference Series*, vol. 123, no. 1, p. 012 033, 2008. DOI: 10.1088/1742-6596/123/1/012033. [Online]. Available: https://doi.org/10.1088/1742-6596/123/1/012033 (cit. on p. 59).

[122] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009. DOI: 10.1126/science.1165893. eprint: https://www.science.org/doi/pdf/10.1126/science.1165893. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1165893 (cit. on p. 59).

[123] W. L. Cava, P. Orzechowski, B. Burlacu *et al.*, *Contemporary symbolic regression methods and their relative performance*, 2021. arXiv: 2107.14351 [cs.NE]. [Online]. Available: https://arxiv.org/abs/2107.14351 (cit. on p. 59).

[124] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia *et al.*, *Discovering symbolic models from deep learning with inductive biases*, 2020. arXiv: 2006.11287 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2006.11287 (cit. on p. 60).

[125] S.-M. Udrescu and M. Tegmark, *Ai feynman: A physics-inspired method for symbolic regression*, 2020. arXiv: 1905.11481 [physics.comp-ph]. [Online]. Available: https://arxiv.org/abs/1905.11481 (cit. on p. 60).

[126] R. Agarwal, L. Melnick, N. Frosst *et al.*, "Neural additive models: Interpretable machine learning with neural nets," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=wHkKTW2wrmm (cit. on p. 60).

[127] T. Hastie and R. Tibshirani, *Generalized Additive Models* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Taylor & Francis, 1990, ISBN: 9780412343902. [Online]. Available: https://books.google.se/books?id=qa29r1Ze1coC (cit. on p. 60).

[128] C.-H. Chang, R. Caruana and A. Goldenberg, *Node-gam: Neural generalized additive model for interpretable deep learning*, 2022. arXiv: 2106.01613 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2106.01613 (cit. on p. 62).

[129] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1705.07874 (cit. on p. 63).

[130] M. T. Ribeiro, S. Singh and C. Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, 2016. arXiv: 1602.04938 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1602.04938 (cit. on p. 63).

[131] J. Kates-Harbeck, A. Svyatkovskiy and W. Tang, "Predicting disruptive instabilities in controlled fusion plasmas through deep learning," *Nature*, vol. 568, no. 7753, pp. 526–531, 2019 (cit. on p. 66).

[132] J. Degrave, F. Felici, J. Buchli *et al.*, "Magnetic control of tokamak plasmas through deep reinforcement learning," *Nature*, vol. 602, no. 7897, pp. 414–419, 2022, ISSN: 1476-4687. DOI: 10.1038/s41586-021-04301-9. [Online]. Available: https://doi.org/10.1038/s41586-021-04301-9 (cit. on p. 66).

[133] X. Wei, S. Sun, W. Tang, Z. Lin, H. Du and G. Dong, "Reconstruction of tokamak plasma safety factor profile using deep learning," *Nuclear Fusion*, vol. 63, no. 8, p. 086 020, 2023. DOI: 10.1088/1741-4326/acdf00. [Online]. Available: https://doi.org/10.1088/1741-4326/acdf00 (cit. on p. 66).

[134] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, 2013. arXiv: 1311.2901 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1311.2901 (cit. on p. 67).

[135] L. Bonalumi, E. Aymerich, E. Alessi *et al.*, "Explainable artificial intelligence applied to algorithms for disruption prediction in tokamak devices," *Frontiers in Physics*, vol. Volume 12 - 2024, 2024, ISSN: 2296-424X. DOI: 10.3389/fphy.2024.1359656. [Online]. Available: https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2024.1359656 (cit. on p. 67).

[136] V. Petsiuk, A. Das and K. Saenko, *Rise: Randomized input sampling for explanation of black-box models*, 2018. arXiv: 1806.07421 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1806.07421 (cit. on p. 67).

[137] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, 336–359, Oct. 2019, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7 (cit. on p. 67).

[138] M. Sundararajan, A. Taly and Q. Yan, *Axiomatic attribution for deep networks*, 2017. arXiv: 1703.01365 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1703.01365 (cit. on p. 67).

[139] B. Kim, M. Wattenberg, J. Gilmer *et al.*, *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*, 2018. arXiv: 1711.11279 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1711.11279 (cit. on p. 67).

[140] P. W. Koh, T. Nguyen, Y. S. Tang *et al.*, *Concept bottleneck models*, 2020. arXiv: 2007.04612 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2007.04612 (cit. on p. 68).

[141] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. [Online]. Available: https://aclanthology.org/N19-1419/ (cit. on p. 68).

[142] Y. Belinkov, *Probing classifiers: Promises, shortcomings, and advances*, 2021. arXiv: 2102.12452 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2102.12452 (cit. on p. 68).

[143] R. Harkness, A. F. Frangi, K. Zucker and N. Ravikumar, *Learning disentangled representations for explainable chest x-ray classification using dirichlet vaes*, 2023. arXiv: 2302.02979 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2302.02979 (cit. on p. 69).

[144] T. Bricken, A. Templeton, J. Batson *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits Thread*, 2023, https://transformer-circuits.pub/2023/monosemantic-features/index.html (cit. on p. 69).

[145] A. Templeton, T. Conerly, J. Marcus *et al.*, "Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet," *Transformer Circuits Thread*, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html (cit. on p. 69).

[146] E. Ameisen, J. Lindsey, A. Pearce *et al.*, "Circuit tracing: Revealing computational graphs in language models," *Transformer Circuits Thread*, 2025. [Online]. Available: https://transformer-circuits.pub/2025/attribution-graphs/methods.html (cit. on p. 69).

[147] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, *Network dissection: Quantifying interpretability of deep visual representations*, 2017. arXiv: 1704.05796 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1704.05796 (cit. on p. 70).

[148] A. Ho, J. Citrin, C. Bourdelle *et al.*, "Neural network surrogate of QuaLiKiz using JET experimental data to populate training space," *Physics of Plasmas*, vol. 28, no. 3, p. 032 305, Mar. 2021, ISSN: 1070-664X. DOI: 10.1063/5.0038290. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/5.0038290/12361366/032305_1_online.pdf. [Online]. Available: https://doi.org/10.1063/5.0038290 (cit. on p. 72).

[149] Y. Wang, M. Bennani, J. Martens *et al.*, *Discovery of unstable singularities*, 2025. arXiv: 2509.14185 [math.AP]. [Online]. Available: https://arxiv.org/abs/2509.14185 (cit. on p. 78).