Preparation Phase Report

Probabilistic and Geometry-Aware Surrogate Modeling for Steady-State Scrape-Off Layer Plasma Fields

Gabriele Gianuzzo

Eindhoven University of Technology

Contents

Abstract

This thesis investigates data-driven surrogate modeling of steady-state scrape-off layer (SOL) plasma fields in tokamak simulations. First principles solvers accurately capture SOL physics but are computationally expensive, limiting their use in parameter scans, control optimization, and real time applications. The data considered in this work consist of multi physics plasma fields defined on curvilinear, tokamak-specific meshes, which pose challenges for standard machine learning architectures. Current deterministic surrogate models based on multilayer perceptrons fail to generalize due to their inability to encode geometric structure and boundary continuity.

To address these limitations, this work explores a geometry aware reformulation of the problem through mesh unrolling into rectangular strips, enabling the use of convolutional neural networks while preserving one to one correspondence with the original mesh cells. The thesis evaluates alternative unrolling strategies, boundary padding schemes, and conditioning mechanisms, and motivates a transition from deterministic regression to probabilistic generative modeling. In particular, conditional diffusion based models are proposed to capture uncertainty and multi-modality inherent in plasma transport phenomena. This preparation phase report establishes the physical background, data representation choices, modeling assumptions, and research questions that guide the subsequent thesis work.

1 Introduction

The design and operation of magnetic confinement fusion devices rely critically on accurate modeling of plasma behavior in the tokamak edge region. In particular, the scrape off layer (SOL) connects the confined plasma core to material surfaces and governs the exhaust of heat and particles toward divertor targets. Predictive modeling of this region is essential for assessing peak heat loads, plasma wall interaction, material lifetime, and the operational feasibility of future fusion reactors [1].

State of the art SOL simulations are based on coupled, nonlinear partial differential equations describing plasma transport, atomic physics, and plasma surface interactions in complex magnetic geometries. While these first principles solvers provide high physical fidelity, they are computationally expensive and difficult to deploy robustly across large parameter spaces. As a consequence, their direct use in rapid design studies, optimization workflows, or uncertainty quantification remains limited.

Data-driven surrogate models offer an alternative by learning an approximation of the simulation solution operator from precomputed datasets. Once trained, such models enable fast inference, making them attractive for tasks such as parameter scans, sensitivity analysis, and control oriented applications. However, constructing accurate and physically meaningful surrogates for SOL plasma fields remains challenging due to the presence of strong anisotropy, irregular curvilinear geometry, lack of Euclidean symmetries, and sensitivity to boundary conditions [2, 3].

This preparation phase report defines the scope and direction of a master thesis aimed at developing probabilistic, geometry-aware surrogate models for steady state SOL plasma fields. The report serves to clarify the problem formulation, physical context, data representation, limitations of existing approaches, and motivation for the proposed modeling strategy.

2 Problem Formulation

The surrogate modeling task considered in this thesis is formulated as a supervised learning problem mapping global simulation inputs to full spatial plasma fields. Each data sample corresponds to a steady-state solution of a tokamak SOL simulation.

The input to the surrogate model is a vector of eight global scalar parameters,

$$\mathbf{x} \in \mathbb{R}^8,$$

including the tokamak major radius, magnetic field strength, input power, deuterium gas puff rate, deuterium ion flux rate, nitrogen gas puff rate, density transport coefficient, and heat transport coefficient. These parameters define the settings of the simulation and are applied uniformly across the spatial domain.

The output is a collection of two-dimensional plasma fields defined on a fixed-topology curvilinear grid,

$$\mathbf{y}(\mathbf{x}) \in \mathbb{R}^{N_x \times N_y \times C},$$

where $N_x = 104$ and $N_y = 50$ denote the grid resolution and $C = 22$ denotes the number of physical channels. The output channels include electron and ion temperatures, particle densities for multiple ionization states of deuterium and nitrogen, and the corresponding parallel velocities.

The objective of the surrogate model is to approximate the conditional distribution of steady-state plasma fields given the global inputs. Rather than learning a single deterministic mapping, the model is intended to represent

$$p(\mathbf{y} \mid \mathbf{x}),$$

allowing for uncertainty quantification and the representation of multiple physically plausible steady states.

3 Physical and Application Context

Tokamak plasmas are confined by magnetic fields in a toroidal geometry generated by a combination of external coils and plasma current. The resulting helical magnetic field lines form nested flux surfaces in the plasma core. Outside the last closed flux surface, magnetic field lines intersect material surfaces, defining the scrape-off layer.

The SOL plays a critical role in determining plasma wall interaction. It governs the exhaust of heat and particles toward divertor targets and strongly influences peak heat fluxes, erosion rates, and material lifetime. Operational regimes such as attached and detached divertor conditions are determined by SOL physics and are central to reactor viability [1].

Transport processes in the SOL are strongly anisotropic. Parallel transport along magnetic field lines is orders of magnitude faster than cross-field transport. This anisotropy leads to characteristic spatial structures in plasma fields, with smooth variation along field-aligned directions and steep gradients perpendicular to them (see Figure 1). These gradients are particularly pronounced near the separatrix, divertor targets, and magnetic X-point.

Although the simulations considered are steady-state, the resulting solutions are not necessarily unique. Small variations in boundary conditions or transport coefficients can induce qualitative changes in plasma behavior. This sensitivity motivates the use of probabilistic surrogate models capable of representing variability and uncertainty in the solution space.

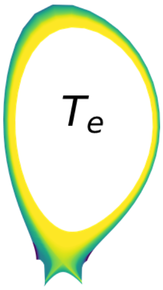


Figure 1: Electron temperature field T_e in the scrape-off layer and divertor region.An enlarged version of these images is provided in Figure 6.

4 Related Work

Machine learning techniques have been increasingly explored for surrogate modeling in fusion research, particularly in the context of plasma exhaust and tokamak edge physics [2, 3]. Existing approaches have demonstrated that data-driven models can significantly accelerate evaluation of edge plasma quantities, though most focus on reduced dimensionality or interpolation tasks rather than full-field prediction.

From a numerical perspective, learning on curvilinear or unstructured meshes poses additional challenges. Standard convolutional neural networks are designed for data defined on regular Cartesian grids and do not directly generalize to irregular geometries without modification. Alternative representations and domain transformations are therefore required to exploit convolutional inductive biases in this setting.

Probabilistic generative models, and diffusion-based methods in particular, have recently shown state of the art performance in modeling high-dimensional distributions in image and physical simulation domains [6, 7, 8]. Their application to fusion plasma modeling remains limited, especially in the context of geometry-aware surrogate models operating on curvilinear simulation grids.

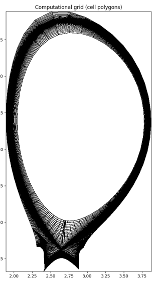
5 Data and Domain Representation

The dataset used in this work consists of 7221 steady-state SOL plasma simulations, divided into 5756 train simulations and 1465 test simulations. These are generated on a two-dimensional curvilinear grid representing a poloidal cross-section of the tokamak. Each grid cell is an irregular quadrilateral with non-uniform size, shape, and orientation.

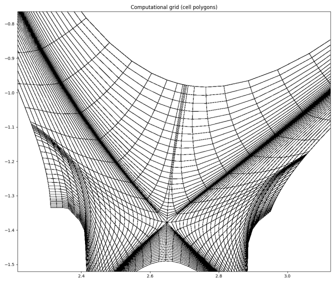
The grid topology is fixed across simulations, meaning that cell connectivity remains unchanged. However, the physical geometry of the grid varies with tokamak parameters. The domain includes an X-point[5] and divertor legs and is inherently asymmetric.

All physical fields are defined on the same grid for all simulations. Strong spatial gradients occur near material boundaries and magnetic separatrices.

The absence of translational, rotational, or reflection symmetries in Euclidean coordinates has significant implications for model design. Standard convolutional architectures operating directly on Cartesian grids cannot be applied without modification, motivating the exploration of alternative domain representations.



(a) Full view of the computational grid used in the simulations, showing the global curvilinear structure and the fixed grid topology across the domain.An enlarged version of these images is provided in Figure 7.



(b) Close-up view of the computational grid highlighting the irregular quadrilateral cell structure and local mesh distortion, especially near the X-point region.An enlarged version of these images is provided in Figure 8.

6 Current Modeling Approach and Limitations

The baseline surrogate model currently available is a deterministic multilayer perceptron (MLP). This model maps the global input vector directly to a flattened representation of the output fields, producing one output neuron per grid cell and physical quantity.

This approach has several practical advantages. It is straightforward to implement, has low inference cost, and can interpolate accurately within the training distribution for smooth regions of the domain. However, it suffers from fundamental structural limitations.

By flattening the grid, all spatial adjacency and geometric information is discarded. The model has no notion of locality, directional transport, or shared structure between neighboring cells. As a result, parameter sharing across spatial locations is impossible, moreover the model size scales linearly with grid resolution.

Furthermore, the deterministic nature of the MLP forces the model to produce a single average prediction for each input. In regimes where multiple physically valid steady states exist, this can lead to unphysical intermediate solutions, particularly in regions with sharp gradients or regime transitions.

Similar limitations of deterministic surrogate models for SOL plasma fields have been observed in previous work, where fully connected architectures were shown to struggle with preserving spatial structure and boundary-consistent behavior across the domain [4].

These limitations are the main motivation for a transition toward geometry-aware and probabilistic modeling approaches.

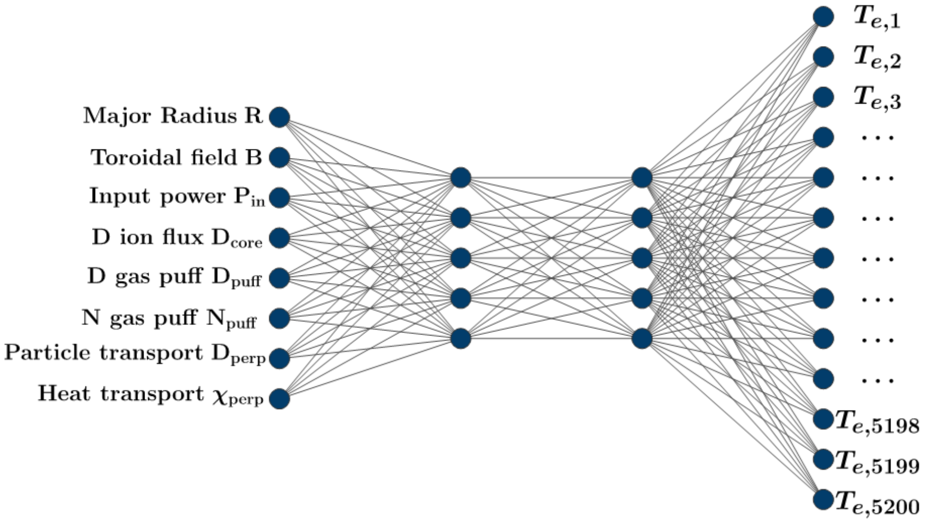


Figure 3: Fully connected neural network mapping global scalar input parameters to flattened spatial field outputs, with one output neuron per grid cell. This architecture does not explicitly encode spatial locality or geometric structure and assumes a single deterministic solution for each set of input conditions.An enlarged version of these images is provided in Figure 9

7 Motivation for Probabilistic and Generative Modeling

Steady-state solutions of SOL plasma simulations exhibit sensitivity to boundary conditions, transport coefficients, and numerical settings. As a result, the conditional distribution of plasma fields given a fixed set of global input parameters may be multi-modal, particularly near regime transitions such as divertor detachment [1].

Deterministic surrogate models are inherently limited to predicting a single point estimate for each input configuration. In regimes where multiple physically plausible steady states exist, this can lead to averaged or unphysical predictions that do not correspond to any realizable plasma state. Probabilistic surrogate models, by contrast, aim to learn the full conditional distribution of outputs, enabling uncertainty quantification and the generation of multiple physically consistent samples.

Among probabilistic approaches, diffusion-based generative models provide a flexible framework for modeling complex, high-dimensional distributions [6, 8]. Their ability to scale to large output spaces and to condition on auxiliary inputs motivates their exploration as the probabilistic backbone of this thesis.

8 Proposed Methodology

The primary modeling direction explored in this thesis is a convolutional neural network-based surrogate operating on an unrolled strip representation of the computational domain. The key idea is to transform the original curvilinear grid into a rectangular grid that is compatible with convolutional architectures.

The unrolled strip is constructed by selecting a connected region of the grid and resampling it into a strip with uniform cell width. This strip is then unrolled into a rectangular two-dimensional array of square cells. Where appropriate, periodic adjacency is enforced between specific boundaries of the strip.

This transformation enables the use of convolutional inductive biases, including locality and parameter sharing. Compared to fully connected architectures, CNNs scale more favorably with grid resolution and require significantly fewer parameters.

The surrogate model is formulated probabilistically. A conditional diffusion-style framework is considered, in which the CNN backbone parameterizes a denoising process conditioned on the global input parameters. The forward process adds noise to physically valid solutions, while the reverse process learns to remove it.

8.1 CNN and Geometry Discussion

The unrolled strip representation is treated as a computational geometry rather than a physically exact embedding. Convolutional kernels operate on the transformed grid, exploiting approximate locality and repeated structural patterns across the domain.

This approach introduces a trade-off between physical fidelity and computational efficiency. True geometric distances, angles, and magnetic field alignment are distorted by the unrolling process. However, the gain in scalability and architectural simplicity is substantial.

Several geometry-handling strategies are currently under investigation. One approach introduces custom padding operations after each convolutional layer, copying pixels across specific boundaries to preserve known adjacencies. Another approach embeds the unrolled domain into a larger rectangular image and applies a binary mask to nullify pixels outside the physical domain.

All architectural choices related to padding, masking, and boundary handling are considered provisional. Their impact on accuracy, stability, and physical consistency will be evaluated during the thesis.

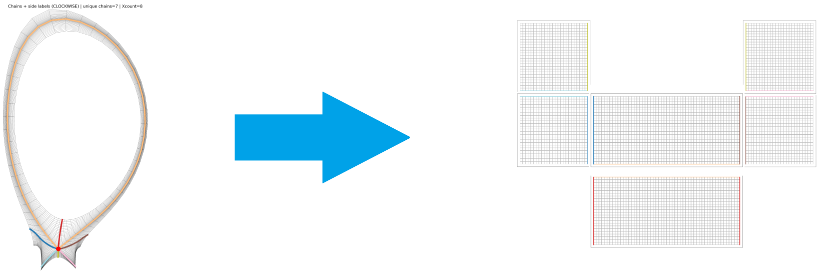

Chains + side labels (CLOCKWISE) | unique chains=7 | Xcount=8

Figure 4: Unrolled strip representation in which physical adjacency across cut boundaries is preserved through explicit padding, at the cost of introducing non-physical pixels that must be masked during training. An enlarged version of these images is provided in Figure 10 and Figure 11

An alternative representation to the one discussed above, illustrated in Figure 4, consists in reordering the unrolled strips in a different manner. Instead of arranging the strips according to a layout that mimics the topology of the original curvilinear mesh, their orientation can be selectively inverted and the strips concatenated to form a single fully rectangular domain, as shown in Figure 5. This construction eliminates empty regions in the rectangular grid and, as a consequence, removes the need for additional masking strategies to exclude non-physical pixels during training.

While this approach simplifies the representation from a computational perspective, it introduces a distortion in the visual and topological adjacency of the domain. In particular, strips that are physically adjacent in the original mesh and share a common boundary may appear separated in the rectangular layout, with the internal mesh region visually interposed between them. As a result, local spatial proximity in the unrolled representation does not always correspond to physical proximity in the original geometry. This representation therefore trades geometric fidelity for compactness and implementation simplicity, and its impact on the learning of boundary-consistent features would need to be assessed empirically.

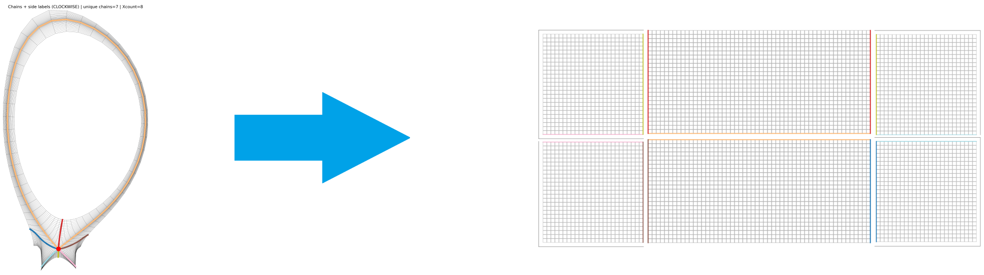

Chains + side labels (CLOCKWISE) | unique chains=7 | Xcount=8

Figure 5: Unrolled strip representation reordered to form a fully rectangular domain without padding or empty regions, where some physically adjacent strips are not directly adjacent in the rectangular layout.An enlarged version of these images is provided in Figure 10 and Figure 12

9 Research Questions

The current surrogate modeling approach for steady-state scrape-off layer plasma fields relies on deterministic regression models that map global input parameters directly to full two-dimensional field outputs. While these models enable fast inference, they do not explicitly encode spatial locality, geometric structure, or uncertainty, and they inherently assume a single deterministic solution for each set of boundary conditions.

To address these limitations, this thesis investigates a probabilistic surrogate modeling framework, with the objective of learning the full conditional distribution of steady-state plasma fields rather than a single point estimate. Such a formulation has the potential to capture solution multiplicity, quantify predictive uncertainty, and improve robustness in regimes characterized by strong nonlinearities and steep spatial gradients. This leads to the first research question:

> **RQ1:** Can a probabilistic surrogate model accurately represent the conditional distribution of steady-state scrape-off layer plasma fields, rather than only their mean behavior?

Introducing a probabilistic formulation alone is not sufficient, the representation of the spatial domain plays a critical role in determining model scalability and performance. The plasma fields are defined on a fixed-topology but geometrically irregular curvilinear mesh, where local interactions between neighboring cells and strong anisotropy induced by the magnetic field are central features of the physics. Preserving this geometric structure can be computationally expensive, while more regular representations may improve efficiency at the cost of physical fidelity. This motivates the second research question:

> **RQ2:** How does the choice of spatial representation affect the trade-off between computational efficiency and physical fidelity in probabilistic surrogate models for scrape-off layer plasmas?

An alternative explored in this work is the unrolled strip representation, which transforms the curvilinear mesh into a rectangular grid suitable for convolutional architectures. This enables strong parameter sharing and efficient training, but introduces design choices related to strip ordering, boundary handling, and spatial adjacency. These choices may influence how well physically meaningful spatial dependencies are captured. This gives rise to the third research question:

> **RQ3:** To what extent can convolutional architectures operating on unrolled strip representations capture physically meaningful spatial dependencies present in the original curvilinear mesh?

Finally, beyond architectural and representation-level considerations, the learning process itself introduces additional modeling assumptions. Standard probabilistic surrogate models rely on generic training objectives and weakly informative priors, which may not reflect known physical properties of scrape-off layer plasmas. Constraints such as anisotropic transport behavior, boundary consistency, and positivity of key physical quantities are typically only satisfied implicitly through the data. Explicitly incorporating such physical knowledge at the level of training objectives, priors, or regularization may improve model robustness and generalization, particularly in regimes that are poorly represented in the training set. This motivates the final research question:

> **RQ4:** Can explicitly enforcing physical constraints during training improve the accuracy and robustness of probabilistic surrogate models for steady-state scrape-off layer plasma fields beyond what is achieved through geometric representation alone?

10 Extended Figures

These figures are included to improve clarity and allow closer inspection of geometric features referenced in the main text.

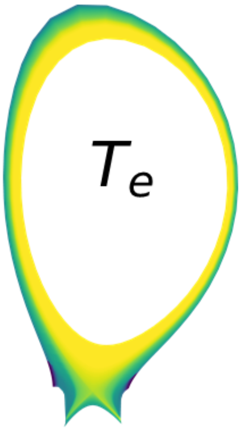


Figure 6: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

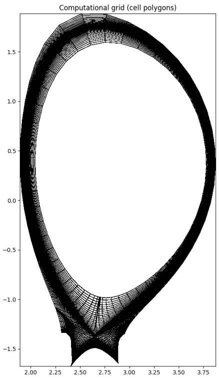


Figure 7: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

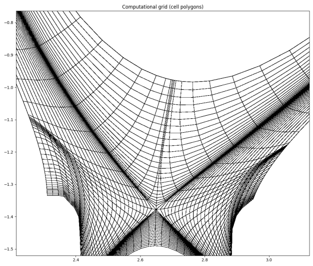


Figure 8: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

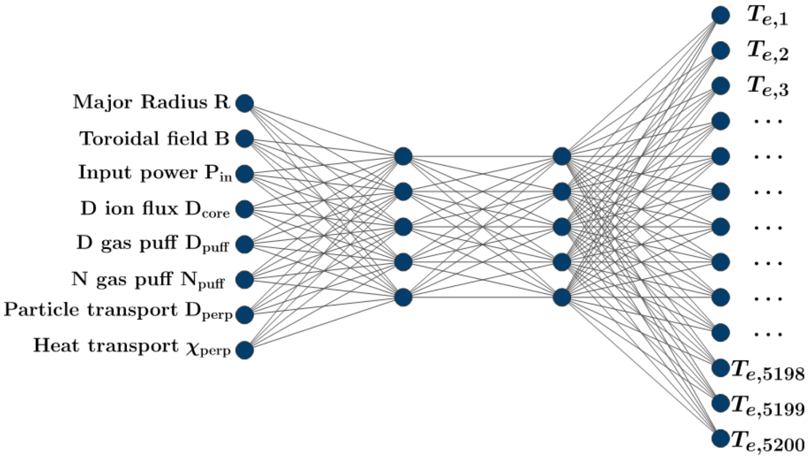


Figure 9: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

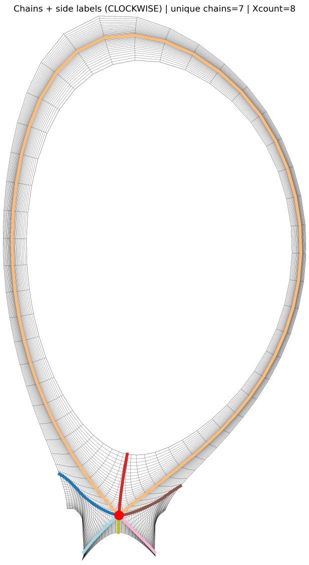

Chains + side labels (CLOCKWISE) | unique chains=7 | Xcount=8

Figure 10: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

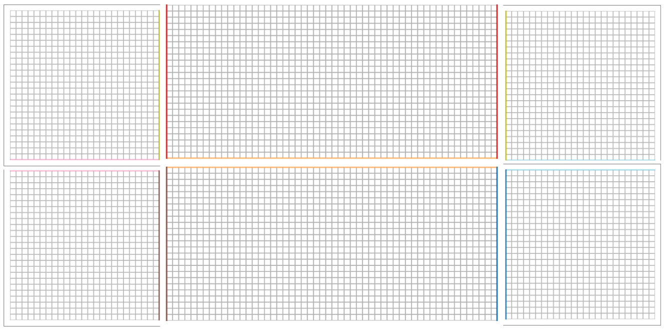


Figure 11: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

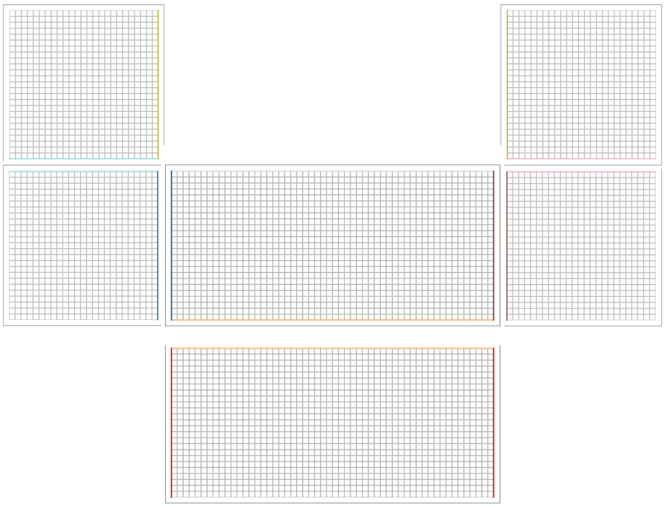


Figure 12: Curvilinear mesh with chain-defined cuts used for the unrolling procedure.

11 References

References

[1] R. A. Pitts et al., “Physics basis for the first ITER tungsten divertor,” *Nuclear Materials and Energy*, vol. 20, p. 100696, 2019. `https://www.sciencedirect.com/science/article/pii/S2352179119300237?via%3Dihub`

[2] S. Wiesen et al., “Data-driven models in fusion exhaust: AI methods and perspectives,” *Nuclear Fusion*, vol. 64, no. 8, p. 086046, 2024. `https://iopscience.iop.org/article/10.1088/1741-4326/ad3f4b`

[3] S. Dasbach and S. Wiesen, “Towards fast surrogate models for interpolation of tokamak edge plasmas,” *Nuclear Materials and Energy*, vol. 34, p. 101396, 2023. `https://www.sciencedirect.com/science/article/pii/S2352179123000352`

[4] S. Wiesen, *Surrogate Modeling of Tokamak Scrape-Off Layer Plasmas*, PhD thesis, Heinrich-Heine-Universität Düsseldorf, 2024. `https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=69084`

[5] M. Bernert et al., “X-point radiation: From discovery to potential application in a future reactor,” *Nuclear Materials and Energy*, vol. 43, p. 101916, 2025. `https://www.sciencedirect.com/science/article/pii/S2352179125000572`

[6] J. Ho et al., “Denoising diffusion probabilistic models,” `https://arxiv.org/abs/2006.11239`

[7] Y. Song et al., “Score-based generative modeling,” `https://arxiv.org/abs/2011.13456`

[8] T. Karras et al., “Elucidating the design space of diffusion-based generative models,” https://arxiv.org/abs/2206.00364