PAPER • **OPEN ACCESS**

Data driven prediction of the neutral gas pressure in the stellarator Wendelstein 7-X

To cite this article: D Angelis *et al* 2025 *Plasma Phys. Control. Fusion* **67** 075004

OPEN ACCESS
IOP Publishing
Plasma Physics and Controlled Fusion
Plasma Phys. Control. Fusion **67** (2025) 075004 (16pp)
https://doi.org/10.1088/1361-6587/ade33e

Data driven prediction of the neutral gas pressure in the stellarator Wendelstein 7-X

D Angelis[1,*], F Sofos[1], S Misdanitis[2], C Dritselis[3], T E Karakasidis[1], D Valougeorgis[2], V Haak[4], D Naujoks[4], G Schlisio[4], S A Bozhenkov[4], V Perseo[4] and W7-X Team[5]

[1] Department of Physics, University of Thessaly, 35100 Lamia, Greece
[2] Department of Mechanical Engineering, University of Thessaly, 38334 Volos, Greece
[3] Department of Civil Engineering, University of Thessaly, 38334 Volos, Greece
[4] Max-Planck Institute for Plasma Physics, Wendelsteinstrasse 1, 17491 Greifswald, Germany

E-mail: dimangelis@uth.gr

Received 28 February 2025, revised 6 June 2025
Accepted for publication 10 June 2025
Published 20 June 2025

Abstract
A machine learning approach, namely symbolic regression (SR), is applied in the stellarator Wendelstein 7-X (W7-X), to investigate the effect of six plasma parameters (line integrated electron density, heating power, toroidal plasma current, fraction of radiated power, core and edge ion temperatures) on the sub-divertor neutral gas pressure. Based on the data from the OP1.2b experimental campaign, closed-form expressions of the neutral gas pressure in terms of the plasma parameters are deduced for the standard, high iota and high mirror magnetic configurations at three different ports of the exhaust system. While common regression schemes assume a predetermined functional form, SR autonomously discovers, via genetic programming, the functional structure of the model, purely from data. In all cases, the optimized data driven SR framework clearly points out that, in estimating the neutral gas pressure, the most dominant parameters are the electron density and the heating power, while the remaining plasma parameters have minor impact, at least from the statistical point of view and may not be included in the correlations. Balancing model generality, complexity(COMP) and accuracy for all considered magnetic configurations and ports, the proposed closed form expressions contain only the product of electron density and heating power raised at some powers, times a constant. The proposed two-parameter symbolic expressions, exhibiting low COMP and excellent accuracy metrics, provide a practical and analytical tool for the acceleration of the neutral gas pressure calculations, that are otherwise computationally very expensive and for the overall performance assessment of the W7-X exhaust system. They may also contribute to more efficient experimental design and operation. performance assessment of the W7-X exhaust system. They may also contribute to more efficient experimental design and operation.

Keywords: symbolic regression, machine learning, fusion energy, W7-X, sub-divertor, neutral gas exhaust system

[5] See Grulke *et al* 2024 (https://doi.org/10.1088/1741-4326/ad2f4d) for the W7-X Team.
* Author to whom any correspondence should be addressed.

1. Introduction

The performance of the exhaust system in fusion devices depends on the core and edge plasma parameters and is characterized by the neutral gas pressure and fluxes at the divertor opening gaps, the pumping speed of the vacuum pumps and the conductance of the sub-divertor domain. Considering that the sub-divertor geometry is subject to numerous operational and maintenance constraints that are not easily altered and since the exhaust rate is the product of the effective pumping speed times the sub-divertor neutral pressure, it is evident that for a given pumping speed the neutral gas pressure is of major importance and should be adequate high to yield exhaust optimization.

Tokamak (poloidal) divertors have been extensively studied to optimize the gas exhaust systems and achieve adequately high neutral gas pressures and divertor fluxes (see e.g. [1, 2] and references therein). On the contrary, in stellarators, where the edge plasma physics is more complex (3D effects are dominant), the divertor research has started more recently and the measured neutral gas pressures remain, in most cases, relatively low (about two orders of magnitude less than in tokamaks). During the OP1.2b Wendelstein 7-X (W7-X) experimental campaign the sub-divertor neutral gas pressure, in several toroidally and poloidally distributed positions, along with the associated parameters in the core and edge of the plasma, have been systematically measured for a large number of discharges in all main magnetic configurations and most of the data have been presented and analyzed in the literature [3–10]. It turns out that the divertor density is approximately directly proportional to the density in the plasma and even in the OP2.1 W7-X campaign with the newly installed cryo-pumps, the neutral gas pressures in the sub-divertor are in the range of only 10^{-2} Pa (max. 0.1 Pa) [11, 12]. Thus, the flow is in the free molecular regime and particle exhaust capabilities remain limited. Only, in the recently discovered low temperature mode of the helical divertor of the Large Helical Device, neutral gas pressures around 1 Pa, combined with particle detachment, have been reported. In this case, the divertor density is proportional, approximately, to the third power of the plasma density, demonstrating that 3D divertors of stellarators may achieve particle exhaust performance, comparable to the one in poloidal divertors [13]. It should be noted however that the closed and compact construction of the LHD divertor (close distance of the X-point to the dome and target plates) [14], differs than that of the open W7-X divertor, which is designed to offer a high degree of flexibility in the selection of different magnetic field configurations.

In parallel, modeling work related to gas exhaust in stellarators has also been performed to further analyze the experimental data. Upstream the divertor ports, the helical Scrape-Off Layer (SOL) has been modeled by simplified approaches, such as the extended two-point model [15–17], as well as by advanced particle codes, such as EMC3-EIRENE (a hybrid fluid–kinetic Monte Carlo code) [18, 19] to interpret the divertor 3D transport mechanisms and analyze the neutral fluxes under attached and detached conditions [20–23]. Additionally, new, more efficient codes (e.g. EMC3-Lite) have been introduced to better examine the observed heat loads at plasma facing components [24]. The sub-divertor flow has been simulated by simplified conductance models in the case of free molecular flow [11], as well as by DIVGAS (a Direct Simulation Monte Carlo code) [25, 26] in flows covering wide gas rarefaction regimes. To simulate the flow through the sub-divertor via some numerical tool (e.g. DIVGAS), it is needed to specify the incoming particle fluxes or alternatively the pressure at the opening gaps. Commonly, this information is obtained by experimental data [25] or by EMC3-EIRENE simulations [26]. Thus, any efforts interrelating the sub-divertor neutral gas pressure with upstream quantities in the divertor may be useful in the sub-divertor modeling.

It is evident that due to the relatively low neutral gas pressures measured in the island divertor of W7-X, further experimental and computational investigation is required to achieve adequate particle exhaust. Since the performance of the exhaust system depends on many upstream and downstream parameters, establishing certain simple and reliable correlations between the magnetic field configuration, the plasma parameters and the exhaust system characteristics will provide a practical and analytical tool in the design of effective particle exhaust systems. Considering that machine learning (ML) is becoming a dominant choice in most prediction-based scientific and technological applications, it is reasonable to investigate the feasibility of ML to satisfactory deduce such correlations.

In fusion reactors, classical ML methods are exploited to create surrogate models, that act as simulation accelerators [27]. They are trained by simulation data and learn to extract interpolated results, bypassing long simulations whenever possible [28]. Research is also driven by quasi-experimental data, preparing the ground for future reactors [29]. In big data applications and multi-dimensional data fields, classical ML algorithms may not fit, and more complex architectures and methods are incorporated. Deep learning techniques based on fully connected neural networks (FCNNs) and convolutional neural networks (CNNs) are popular options, having the ability to deal with 2D and 3D spatiotemporal fields, which are mapped to image representations. Classical ML and deep learning algorithms may be employed in reliable experiment predictions and contribute to the design and operation of experimental campaigns [30].

Alternative approaches that enforce physical laws to the solution are provided by physics informed neural networks (PINNs) [31]. There are only limited references to the application of such methods in magnetic confinement devices. Recent examples include the investigation of blob dynamics in the COMPASS tokamak by processing 2D turbulence images with CNNs [32], infrared movie analysis for analyzing thermal events in the WEST tokamak [33], detection and classification of thermal events from images in tokamaks [34] and mapping infrared images to the surface temperature field of tokamaks [35], with CNN architectures such as U-Net [36]. It has also been found that spectroscopic data can be used to train a deep NN and provide accurate predictions for both the electron density and temperature of the device [37]. Since plasma disruption events in fusion reactors may emerge, monitoring the status of the reactors is of primary importance. Example methods have been implemented with PINNs [38], FCNNs [39], feedforward NNs [40], CNNs [41], and reinforcement learning [42]. Data from internal sensors can be also monitored by long short- term memory (LSTM) neural networks [43], which function as memory elements that learn from previous data and forecast future behavior. In another instance, LSTM cells are exploited to predict the consequent plasma parameters in future frames [44].

In all these cases, prediction and forecasting are possible, but without hand-on equations to describe the phenomena. This fact paves the way to incorporate explainable ML methods, such as symbolic regression (SR), an ML-based approach usually bound to genetic programming (GP), that derives symbolic equations from the available data, reveals hidden relations among system parameters and provides an explainable route of the implied process, bypassing the inherent 'black-box' approach of classical ML models. Magnetic confinement devices like tokamaks and stellarators can profit from SR by enabling the derivation of compact, interpretable equations that reveal insights into plasma behavior and improve predictive capabilities. For example, SR has been applied to identify the safe and disruptive operation in tokamaks [45] or to express the scaling law of the energy confinement time in the stellarator, suggesting a less rigid approach compared to the dominating power law [46]. Also, a new scaling formula for the ITER divertor heat load has been revealed by SR, mainly based on simulation data [47].

From the above review, it is clear that the use of artificial intelligence-based methods in fusion research is relatively recent but growing rapidly. While most of the available literature is in tokamaks, similar work in stellarators remains limited. Additionally, there is a clear need of ML methods that can offer greater interpretability and better generalization of the underlying phenomena.

To address some of these needs, here, an SR approach is presented, exploiting data from the W7-X OP1.2b experimental campaign to examine the impact of six plasma parameters, namely the line integrated electron density (n_e), the heating power (P_heat), the toroidal plasma current (I_tor), the fraction of the radiated power over the heating power (f_rad) and the center (T_c) and edge (T_e) ion temperatures on the neutral gas pressure (p_n) at the sub-divertor. Closed form expressions of the exhaust pressure in terms of the plasma parameters are deduced for the standard, high iota and high mirror magnetic configurations, at three different ports in the low and high iota sections and in each case the effect of the most dominant parameters is identified. The suggested expressions, exhibiting great accuracy and low complexity (COMP), may speed-up the computation of the neutral gas pressure, be useful in the interphases between the plasma edge, divertor and sub-divertor operating regimes and, in general, support the design and optimization of the particle and impurity exhaust system of W7-X.

The rest of the paper is structured as follows: section 2, refers to the implemented methodology. Section 2.1 describes the employed W7-X database and section 2.2 outlines the basic characteristics of the SR method. Then, section 3 presents results and discussion of the deduced correlations for each of the three magnetic configurations under consideration at three different ports. Section 3.1 covers, in detail, the derivation of the closed form expressions for the neutral gas pressure and section 3.2 provides comparisons in graphical form with experimental data. Finally, in section 4, the main concluding remarks are stated.

2. The employed W7-X datasets and SR approach

In the OP1.2b experimental campaign the neutral gas pressure has been measured for a large number of discharges for all main magnetic configurations, while the corresponding available data in the more recent OP2.1 campaign [48] are much less. In the undergoing OP2.2 campaign, large datasets of the neutral gas pressure are expected but they are currently still under development and will be available in the short future. As a result, the present work is based solely on the data of OP1.2b.

In section 2.1, the employed datasets of the OP1.2b experimental campaign of Wendelstein 7-X are defined and then, in section 2.2, the basic characteristic of the SR approach are provided, followed by a description of the imposed SR methodology.

2.1. The employed W7-X datasets

During the OP1.2b experimental campaign of Wendelstein 7-X [3–11], the neutral gas pressure at different locations of the sub-divertor, along with the associated plasma core and edge parameters, have been recorded. In the present work a subset of the database of OP1.2b is employed and refers to electron cyclotron resonance heating, excluding neutral beam injection and the utilization of pellet injection, impurity injection and control coils. In the employed dataset, each data point corresponds to an average over 200 ms in a plasma discharge and over the data from neutral gas pressures gauges at different toroidal locations [11].

More specifically, the employed dataset includes the data of the plasma parameters n_e, P_heat, f_rad, I_tor, T_c and T_e for three magnetic field configurations, namely the standard, high iota and high mirror configurations and the neutral pressure data at the AEH, AEI and AEP ports for partially attached plasma conditions. The former two are in the low iota section of the sub-divertor at the large pumping port and gap respectively, while the latter one is in the high iota section of the sub-divertor at the small pumping gap. The choice of the examined six plasma parameters is mainly based on previous experimental findings [7, 10, 11], where the effect of n_e, P_heat, f_rad

Table 1. Range of measured neutral gas pressure at various magnetic configurations and ports.

Magnetic configuration	Neutral gas pressure p_n (10^{-6} mbar)		
	AEH	AEI	AEP
Standard	2.45–697.11	7.60–1138.04	4.54–296.66
High iota	1.05–106.48	2.23–157.25	28.45–1568.78
High mirror	6.44–155.38	14.16–245.16	3.27–128.62

Table 2. Range of measured plasma parameters at various magnetic configurations.

Magnetic configuration	n_e (10^{19} m^{-2})	P_{heat} (MW)	I_{tor} (KA)	T_c (keV)	T_e (keV)
Standard	0.82–13.52	0.33–7.17	−2.72–10.00	0.33–10.97	0.09–3.16
High iota	0.90–12.01	0.46–6.24	−3.32–1.69	0.28–11.56	0.17–2.72
High mirror	1.37–8.38	0.74–5.25	−0.08–7.92	0.02–9.03	0.00–1.34

Table 3. Data clustering in terms of the fraction of radiated power f_{rad}.

Magnetic configuration	f_{rad}		
	Cluster 1	Cluster 2	Cluster 3
Standard	$0 < f_{rad} < 0.25$ (4860)	$0.25 < f_{rad} < 0.5$ (3114)	$0.5 < f_{rad} < 0.8$ (1599)
High iota	$0 < f_{rad} < 0.3$ (1530)	$0.3 < f_{rad} < 0.5$ (2239)	$0.5 < f_{rad} < 0.8$ (486)
High mirror	$0 < f_{rad} < 0.3$ (887)	$0.3 < f_{rad} < 0.5$ (453)	$0.5 < f_{rad} < 0.8$ (168)

and I_{tor} on the neutral gas pressure is highlighted. The present investigation includes these parameters, as well as temperatures T_c and T_e, which may be important in the estimation of the neutral gas pressure. Other parameters, which may also be influential will be explored in future studies. Detailed descriptions of the W7-X magnetic configurations and divertor geometry, including the exact measurement location of the neutral gas pressure and the strike lines, may be readily found in previous works (see e.g. [10, 11]).

In table 1, the range of the measured neutral gas pressures p_n is provided for all considered magnetic field configurations and pumping gaps and ports. Overall, the highest pressures have been measured at the high iota configuration, close to the AEP port and they are about 0.16 Pa. In the standard configuration the highest pressures are about 0.12 Pa and have been measured at the AEI port, while in the high mirror configuration the highest measured pressures are significantly lower (about 0.025 Pa). The range of the associated measured plasma parameters (n_e, P_{heat}, I_{tor}, T_c, T_e) are shown in table 2. It is noted that the considered data refer only to partially attached plasma discharges, with the fraction of the radiated power $f_{rad} < 0.8$. Plasma discharges with $f_{rad} \geqslant 0.8$ will be considered in future work. The complete dataset may be found in [11].

Here, based on these datasets, closed form expressions of the neutral pressure in terms of the plasma parameters via SR are deduced. Thus, databases are created for every magnetic configuration and port, i.e. 3 configurations times 3 ports equal to 9 datasets. These nine datasets refer to the whole range of the investigated fraction of the radiated power, i.e. $0 < f_{rad} < 0.8$.

In addition, in order to have a more detailed data driven prediction of the neutral gas pressure, a KMeans algorithm [49] is employed to cluster the data in each of the nine aforementioned scenarios, based on the values of f_{rad}. Clustering offers a more targeted approach for analyzing datasets that exhibit varying behaviors across different data regions [50]. In this context, analyzing clusters based on the fraction of radiated power may help interpret the transition behavior from partially attached to near-detached conditions. Therefore, for each one of the magnetic configurations and ports, clustered datasets in terms of f_{rad}, as defined in table 3, are considered and fed into the SR framework. That makes 27 extra datasets (3 configurations $\times$ 3 ports $\times$ 3 fractions of radiated power regions). In table 3, the numbers in parentheses refer to the available data points for each configuration and clustered region of f_{rad}. The largest number of data points is found in the standard configuration, followed by the high iota one, while the number of data points in the high mirror configuration, is relatively small. Also, in each configuration, the data points in cluster 3 are significantly fewer than in the other two corresponding clusters.

Totally, $9 + 27 = 36$ datasets are considered and for each dataset, correlations between the neutral pressure and the plasma parameters are deduced following the SR methodology described in the next subsection. For clarity purposes, the expressions for $0 < f_{rad} < 0.8$ and the ones for the clusters of

f_{rad}, will be named 'overall expressions' and 'clustered expressions' respectively.

2.2. Basic characteristics of the SR method

Symbolic regression is a multifaceted approach to regression analysis that discovers the structure of an expression and its parameter dependencies on the fly. It is superior to traditional least-squares regression methods, especially in cases where the underlying parameter relationship is unknown or non-linear. While common regression schemes assume a predetermined functional form (e.g. linear, polynomial), SR autonomously discovers the functional structure of the model from data, extracting interpretable equations that align with physical or theoretical insights. This flexibility allows SR to capture complex, non-linear, and multi-variable relationships without requiring extensive feature engineering or prior domain knowledge, balancing model accuracy and COMP. Data driven profound expressions may be generated with minimal or even no prior knowledge of the system. More details on SR methods and applications can be found in recent review papers [51, 52].

While there are numerous techniques to perform SR, the most prominent involve ML, where algorithms based on GP [53] concepts raise a vast number of expressions that progressively minimize a loss function. Genetic programming-oriented frameworks make the extraction of a symbolic expression a stochastic procedure, since the expressions are being developed by a random recombination of the constitutive parts via crossover and mutation operations. These expressions are represented in the form of tree-like structures consisting of internal and leaf (or terminal) nodes [50]. The former ones are either mathematical operators $\{+,-,\times,/,\wedge,\exp,\log\}$ or functions, e.g. $\{\exp(x), x^{\alpha}\}$, and the latter ones are either input/output variables or constants. In addition to that, being itself a multi-objective optimization technique, the generated solutions are inherently non-unique. This means that, for a given dataset, one can have expressions of ranging COMP, which refers to the form of the symbolic expression and is proportional to the number of calculation nodes in the tree-like structure. For instance, an expression with nonlinear mathematical functions that reveal dependencies between many input variables is considered more complex compared to a simple addition between some variables. Expressions of increased COMP are more prone to overfitting, so the choice of simple and accurate expressions is preferable. GP frameworks start by creating a vast number of random expressions, which are then optimized throughout the SR procedure and the result is usually exported as a set of potential solutions of varying COMP and accuracy, forming the pareto front.

The present work builds on the SR library package included in the open Python-Julia software PySR [54], employing a mean square error (MSE) based loss function. The implemented SR framework and its subsystems are presented in figure 1. Experimental data is coming from a large number of discharges, functioning in three different configurations (standard, high iota and high mirror) at the AEP, AEH and AEI ports of W7-X (figure 1(a)) and compiled into a data pool for processing through a clustering methodology (figure 1(b)), as described in section 2.1. These datasets are directed in parallel SR computational cells that derive various symbolic expressions driven by the available experimental data (figure 1(c)). In SR, as in most ML methods, data is partitioned by a factor of 70% for train, 20% for validation, and 10% for test [55]. Most of the data is allocated for training, ensuring that the model has sufficient information to discover hidden patterns. The validation set is used for fine-tuning the model, treated as part of the training process during bootstrapping. The test dataset is not considered during training (unseen) but is used to evaluate the discovered expressions. A smaller proportion (10%) is sufficient for evaluation, as testing does not require as much data as training. This approach helps check if the dominant expressions generalize to unseen data and is a key step to avoid overfitting. For each dataset, an optimized equation is proposed, achieving low values of MSE and COMP, as shown in the flow diagram (figure 1(d)).

The available datasets are exploited to predict the neutral gas pressure p_{n} from the input six plasma parameters (n_{e}, P_{heat}, I_{tor}, f_{rad}, T_{c} and T_{e}) at three magnetic configurations (standard, high iota, high mirror), across three different ports (AEH, AEP, AEI). For each investigated case, the SR algorithm performs an equation selection process based on multiple criteria, such as MSE and COMP of the solution, which is a multi-objective optimization problem [56]. Through this process, interpretable expressions emerge with profound generalization. Among the large number of equations proposed by the SR algorithm, selecting simple analytical expressions reduces computational effort and focuses on the objective of formulating equations that capture meaningful physical principles.

In figure 1(e), a space of possible solutions, denoted with circle points balancing model COMP and loss MSE, is shown. The curve composed of the optimal solutions for each COMP level is known as the Pareto front curve, which represents the solutions where a trade-off between equation COMP and accuracy is made. The elbow point (i.e. the red circle point) is often preferred because it represents a balance, where further increasing COMP, yields minor improvement in accuracy. However, if interpretability and generalizability is of importance, simpler models may be prioritized, even if they are slightly less accurate (e.g. the blue circle point). This can be particularly useful, for instance in nuclear fusion experimental design, where simpler equations are more robust to analyze. In other cases, a slightly more complex equation might align better with theoretical insights in the domain (e.g. the green circle point). In this work, the ability of the proposed SR expression to replicate efficiently the experimental data is primarily measured by the coefficient of determination R^2.

Following this procedure, regions where changes in expression COMP significantly impact prediction accuracy or remain stable are being identified. Additionally, to minimize the influence of randomness on the results, a series of runs are performed, each one at a different random state, aiming to identify recurring expression forms to guide the final selection of an expression that remains robust to variations in the training data. Using a different random state at each run allows the algorithm to model a separate portion of the data each time,

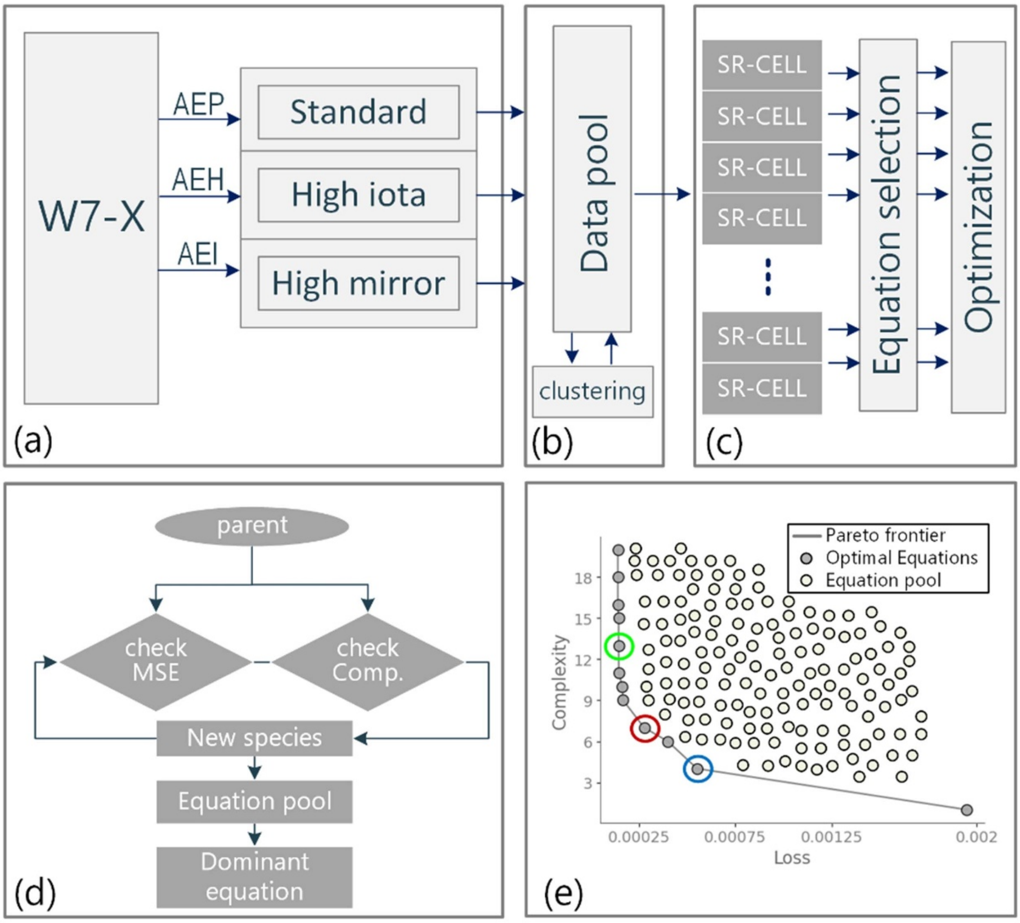

Figure 1. Symbolic Regression model, (a) data acquisition from W7-X, from AEH, AEP and AEI ports and three different magnetic configurations, (b) database creation and clustering driven by f_{rad} values, (c) the final step where parallel SR computations lead to the final, optimized equation, (d) inset of an SR cell, (e) example of the Pareto-optimal solution.

reducing the risk of overfitting and improving the generalizability of the solution.

3. Results and discussion

In section 3.1, the expression selection methodology is outlined and proposed expressions with their accuracy measures are presented. Then, in section 3.2, a detailed comparison of the neutral pressures via the SR expressions versus the corresponding experimental data in graphical form for the clustered and whole f_{rad} range values, is shown.

3.1. Derived closed form expressions via SR

As described in section 2.1, a total of 36 datasets are considered and each dataset always refers to a specific magnetic configuration and port. Also, nine of them refer to the whole range of f_{rad} ($0 < f_{\mathrm{rad}} < 0.8$) examined, while the remaining twenty-seven refer to the clustered regions of f_{rad} specified in table 3. Here, following the SR methodology described in section 2.2, for each dataset the corresponding most suitable closed form expression is deduced.

The impact of all involved plasma variables (parameters) are considered and the generated expressions are of the general form

$$p_{\mathrm{n}} = f(n_{\mathrm{e}}, P_{\mathrm{heat}}, I_{\mathrm{tor}}, T_{\mathrm{c}}, T_{\mathrm{e}}, f_{\mathrm{rad}})\,. \tag{1}$$

After performing extensive parallel SR calculations, the dominant expressions for the whole f_{rad} range, as well as for all clusters of f_{rad}, have been extracted across all ports and magnetic configurations. The quest now is which symbolic expressions to select by compromising between COMP and accuracy to better express the experimental data behavior. In such GP-based procedures, a well-defined selection criterion is the number of appearances in the runs [57], along with the associated accuracy metrics, mainly measured by the coefficient of determination R^2. Here, for values of $R^2 < 0.7$, $0.7 \leqslant R^2 < 0.8$, $0.8 \leqslant R^2 < 0.9$ and $R^2 \geqslant 0.9$, the accuracy is considered low (not acceptable), moderate, high and very high respectively.

The selection methodology is demonstrated by a representative example, focusing on the high-iota configuration and AEP port for $0 < f_{\mathrm{rad}} < 0.8$, where the highest neutral gas pressure has been measured. However, the same analysis holds in all cases. Table 4, presents representative SR expressions

Table 4. Indicative SR expressions for the high-iota configuration at the AEP port ($0 < f_{\mathrm{rad}} < 0.8$).

SR expression	R^2	MSE
$p_n = 69.04 P_{\mathrm{heat}}^{1.45}$ (2*a*)	0.59	18 385
$p_n = 9.88 n_{\mathrm{e}}^{1.99}$ (2*b*)	0.76	10 799
$p_{\mathrm{n}} = 21.63 n_{\mathrm{e}} P_{\mathrm{heat}}$ (2*c*)	0.93	3201
$p_{\mathrm{n}} = 21.64 n_{\mathrm{e}}^{1.10} P_{\mathrm{heat}}^{0.86}$ (2*d*)	0.93	3004
$p_{\mathrm{n}} = 19.54 \left[n_{\mathrm{e}} \left(P_{\mathrm{heat}} + f_{\mathrm{rad}}\right)\right]$ (2*e*)	0.93	2944
$p_{\mathrm{n}} = 20.38 \left[I_{\mathrm{tor}} + \left[n_{\mathrm{e}} \left(P_{\mathrm{heat}} + f_{\mathrm{rad}}\right)\right]\right]$ (2*f*)	0.94	2860
$p_{\mathrm{n}} = 20.37 \left[T_{\mathrm{c}}/T_{\mathrm{e}} + \left[n_{\mathrm{e}} \left(P_{\mathrm{heat}} + f_{\mathrm{rad}}\right) - 1.38\right]\right]$ (2*g*)	0.94	2700

(2*a*)–(2*g*), ranging from simple to complex. The simplest expressions include only one parameter, either P_{heat} (2*a*) or n_{e} (2*b*), the ones with moderate COMP (2*c*)–(2*e*) include three parameters, namely P_{heat}, n_{e} and f_{rad} and the most complex ones (2*f*) and (2*g*) include more than three parameters, with P_{heat} and n_{e} being always present, independent of the expression COMP. For each expression the corresponding accuracy metrics, specifically the R^2 and MSE scores, are included to evaluate its performance. It is clear that the more advanced expressions achieve higher R^2 and lower MSE values, as they reveal more detailed interconnections between the data points. The metrics of the one-parameter expressions (2*a*) and (2*b*) are, in general, low. Equation (2*b*) with the electron density performs much better than equation (2*a*) with the heating power, implying that the influence of n_{e} is more important than P_{heat}, but its accuracy remains relatively low. One-parameter expressions with the remaining variables provide even worse metrics and therefore, are not included in this indicative pool. However, correlations such as (2*b*) remain attractive and will be further examined due to their simplicity. Next, the accuracy greatly improves in the equations with moderate COMP (2*c*), (2*d*) and (2*e*), where R^2 approaches one, while MSE significantly decreases. In all three equations $R^2 = 0.93$ and MSE is around 3000. Finally, in the most complex equations (2*f*) and (2*g*), the accuracy is marginally improved, with $R^2 = 0.94$ and MSE reaching 2700.

It is evident that beyond a certain accuracy threshold, R^2 and MSE show negligible improvement. This threshold can be considered as a decision point to reveal those input parameters that are of importance to the estimation of p_{n}. Based on the data of table 4, that reference point is set to expressions of the form (2*c*) or (2*d*), since it becomes clear that the remaining system parameters can be excluded from p_{n} estimations, as, from a statistical point of view, they do not significantly influence its calculation. The temperature parameters (T_{c}, T_{e}) seem to have no significant impact on p_{n}, since they always appear in highly complex expressions that do not lead to any noticeable improvement in the accuracy metrics. Also, f_{rad} and I_{tor} seem to be more influential than temperatures, providing however marginal gain in accuracy metrics accompanied with higher equation COMP, compared to the two-parameter expressions of n_{e} and P_{heat}.

Having established the key ideas for the extraction of the SR correlations, the same analysis has been repeated for all 36 datasets under consideration. It is noted, that in all magnetic configurations and ports, similar results, with the ones presented in table 4 for the high-iota magnetic configuration at the AEP port, have been obtained. In all scenarios, the SR framework indicates that in the estimation of p_{n}, one-parameter expressions of n_{e}, as well as expressions combining n_{e} and P_{heat}, are the most dominant ones, while expressions containing the other parameters (f_{rad}, I_{tor}, T_{c}, T_{e}), may be excluded from further consideration. In the overall expressions f_{rad} is the third parameter appearing in the correlations (e.g. see equation (2*e*) in table 4), but still its influence on p_{n} remains relatively small compared to n_{e} and P_{heat}.

Consequently, further investigation is focused on one-parameter expressions of the form

$$p_{\mathrm{n}} = c \times n_{\mathrm{e}}^{\alpha}. \tag{3}$$

Since they provide a straightforward correlation between neutral gas pressure and line integrated electron density and mainly, on two-parameter expressions of the form

$$p_{\mathrm{n}} = c \times n_{\mathrm{e}}^{\alpha} \times P_{\mathrm{heat}}^{\beta}. \tag{4}$$

In equations (3) and (4), the coefficient c and exponents α and β have different values for each experimental dataset (i.e. configuration—port—fraction of radiated power regime) in order to capture the minor effects of the other parameters not included in the proposed correlations.

In table 5, the accuracy of the overall SR expressions of the general forms (3) and (4), in terms of R^2, is provided for the standard, high iota and high mirror configurations at the AEH, AEP and AEI ports in the whole range of f_{rad}. The examined overall SR expressions with the two parameters include the specific cases of $\alpha = 1$ or $\beta = 1$ or $\alpha = \beta = 1$. The values of the coefficient c and exponents α and β have been optimized to each of the nine datasets. For clarity it is noted that the tabulated values of R^2 in the high iota configuration at the AEP port, are the same with the ones reported in table 4.

In table 5, it is seen that the one-parameter expression (3) performs much better in the standard configuration ($R^2 > 0.8$) than in the high iota one ($R^2 < 0.8$). Thus, although n_{e} and P_{heat} are profoundly strongly interrelated, still, in the high iota configuration one-parameter expressions with high accuracy metrics are not feasible. Similarly, in the high mirror configuration the one-parameter expression provides adequate accuracy at the AEH and AEI ports, but low accuracy at the AEP port.

Table 5. Values of R^2 for various overall SR expressions in magnetic configurations and ports, where each symbolic expression is applied ($0 < f_{\text{rad}} < 0.8$); the values of c, α and β have been optimized to the specific dataset.

Expression	Standard			High iota			High mirror		
	AEH	AEP	AEI	AEH	AEP	AEI	AEH	AEP	AEI
$p_{\text{n}} = c \times n_{\text{e}}^{\alpha}$	0.88	0.84	0.88	0.79	0.76	0.77	0.85	0.68	0.87
$p_{\text{n}} = c \times n_{\text{e}} \times P_{\text{heat}}$	0.77	0.82	0.81	0.95	0.93	0.93	0.74	0.76	0.59
$p_{\text{n}} = c \times n_{\text{e}}^{\alpha} \times P_{\text{heat}}$	0.80	0.82	0.82	0.95	0.93	0.93	0.76	0.77	0.78
$p_{\text{n}} = c \times n_{\text{e}} \times P_{\text{heat}}^{\beta}$	0.81	0.85	0.85	0.95	0.93	0.94	0.89	0.80	0.92
$p_{\text{n}} = c \times n_{\text{e}}^{\alpha} \times P_{\text{heat}}^{\beta}$	0.89	0.89	0.90	0.96	0.93	0.95	0.90	0.80	0.93

Concerning the two-parameter expressions, it is seen that the correlations with $\alpha = \beta = 1$ or $\beta = 1$, may provide high and very high accuracy, with $R^2 > 0.8$, but only at certain magnetic configurations and specific ports, while in others the accuracy is moderate ($0.7 \leqslant R^2 < 0.8$). On the contrary, the two-parameter expression with $\alpha = 1$, as well as the one of the general form (4), with arbitrary exponents α and β, have always high and very high accuracy, with values of R^2 close to or even higher than 0.9, with the latter one, being significantly more accurate, particularly in the standard configuration. It is also useful to note that in the high iota configuration, all two-parameter expressions exhibit very high accuracy ($R^2 > 0.9$), implying that the presence of P_{heat} in the high iota correlations is more important than in the standard and high mirror ones.

Building on the discussion above and balancing model simplicity, accuracy and generality in the extraction of p_{n}, the functional form of the two-parameter equation (4) is proposed for all 36 datasets, i.e. for the nine datasets corresponding to the whole regime of f_{rad} (overall expressions), as well as for the 27 datasets corresponding to the clustered regimes of f_{rad} (clustered expressions).

In tables 6 and 7, the values of the coefficient c and exponents α and β of the overall and clustered expressions with two-parameters respectively are provided. As mentioned above, the SR framework suggests optimized values for c, α and β within a narrow range, but, oftentimes, different in each evolution run. This is a common feature of GP, since the derived equations start from a random initial parent and evolve to a successive child equation. Since PySR may not yield the perfect equation but rather suggests a set of potential candidates drawn from an infinite pool [58], the tabulated values for c, α and β, in tables 6 and 7, are deduced via an additive optimization step, using scipy's curve fit [59], where the sole criterion is the minimization of MSE.

In addition, to gain a broader perspective on the values of each component and follow a robust statistical basis [60], we have calculated a 95% bootstrap confidence interval for each of the optimized values. This analysis was conducted using only the train dataset, while the performance of all 36 SR expressions has been evaluated on the validation and testing datasets. Each of the optimized values (c, α, β) is presented in the format $X_{\text{Lowerbound}}^{\text{Upperbound}}$, where X represents the mean value of the confidence interval. The accuracy of the overall expressions measured by R^2 for the train, validation and test dataset can be found in figure 2. There, it is clearly observed that the R^2 values are well-balanced for all three scenarios, which is evidence of no overfitting, along with the functional simplicity of the two variable expressions. As it is known, a complex functional has an increased impact on overfitting compared to a simpler one [61, 62]. This is attributed to the fact that a complex functional may fit not only the real pattern in the data but also irrelevant noise, while a simpler expression is more likely to capture the core relationship between variables.

For each experimental dataset the constant and the exponents in tables 6 and 7 capture the variation of the magnetic configuration and port and implicitly incorporate the dependency of the excluded plasma parameters, mainly of f_{rad} and I_{tor}, to better approximate the experimental data. The values of c, α and β of the overall expressions may be considered as weighted averages of the corresponding clustered ones. Furthermore, they are closer to the ones of the clustered expressions for $f_{\text{rad}} < 0.5$, than for $0.5 < f_{\text{rad}} < 0.8$ and this is clearly contributed to the larger number of data points in $f_{\text{rad}} < 0.5$, compared to $0.5 < f_{\text{rad}} < 0.8$. Thus, the expressions proposed in the whole range of f_{rad} are expected to be more accurate in the former than in the latter f_{rad} regime. In all cases examined the exponents α and β are positive, except in the high mirror configuration for $0.5 < f_{\text{rad}} < 0.8$, where they are very close to or less than zero. Also, the relatively higher values of c in all magnetic configurations at the AEI compared to the AEH port, as well as in the high iota configuration at the AEP port, are surely related to the associated higher measured pressures. Finally, it is observed that in all cases the exponent α of n_{e} is larger than the exponent β of P_{heat}, which indicates that the line integrated electron density has a more significant effect compared to the heating power.

The fact that the proposed two-parameter expressions can effectively counterbalance more complex ones suggests that several parameters considered here do not significantly impact the sub-divertor neutral gas pressure, at least from a statistical perspective. In addition, the reduced COMP of the proposed expressions mitigates the risk of overfitting and enhances their generalization potential. The performance of these expressions in relation to the associated experimental datasets is thoroughly examined in the next section.

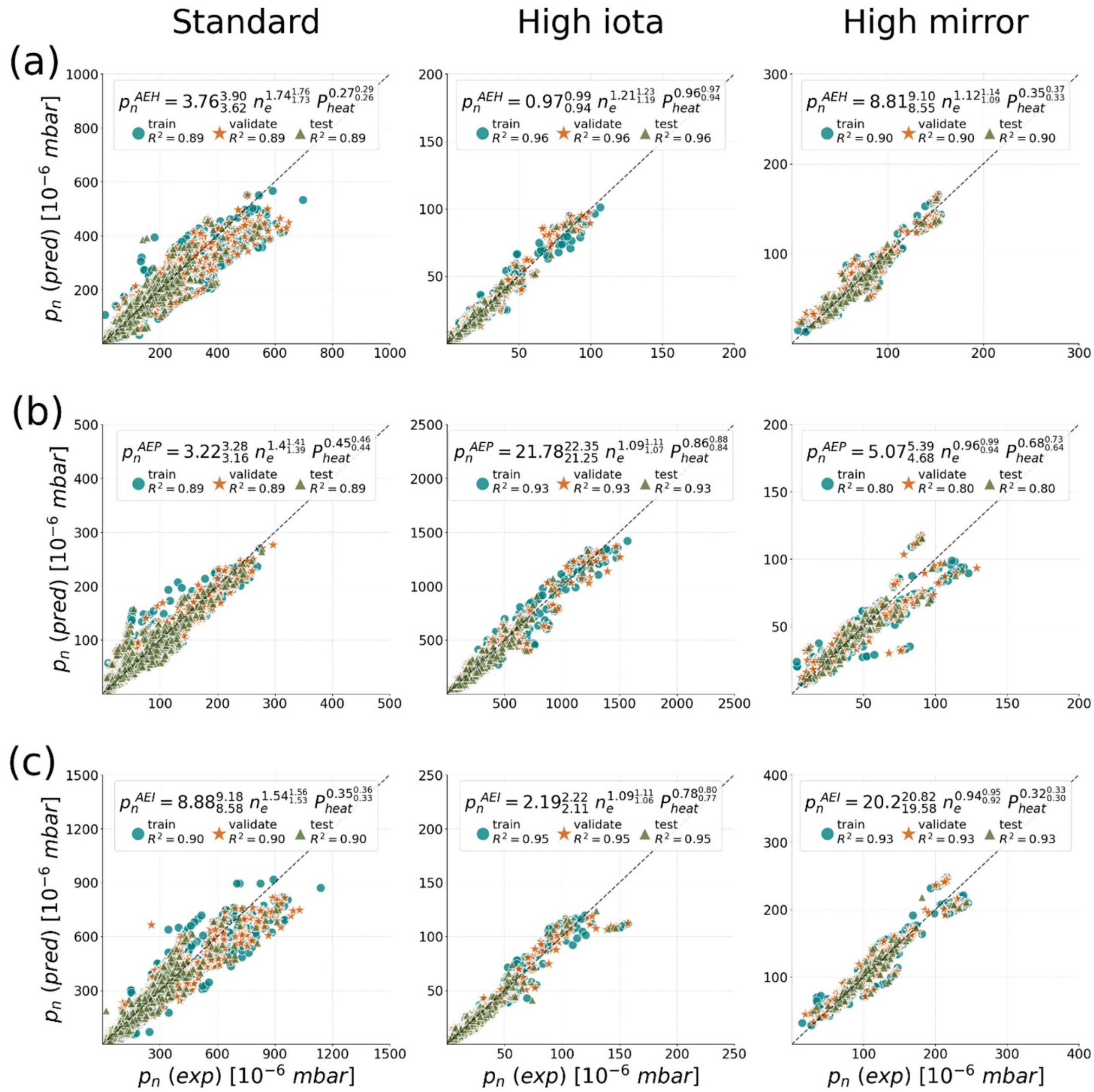

Figure 2. Parity plots of the overall two-variable expressions for p_n (equation (4)) at the AEH (a), AEP (b) and AEI (c) ports for the standard (left), high iota (middle) and high mirror (right) configurations. In each case, the corresponding R^2 value for the train, validation and test dataset is given. Values of the coefficient c and exponents α, β of equation (4) are shown with a 95% confidence interval.

Table 6. Coefficient c and exponents α and β of the overall two-parameter expressions (4) for all magnetic configurations and ports ($0 < f_\mathrm{rad} < 0.8$), with 95% confidence interval .

	AEH			AEP			AEI		
Configuration	c	α	β	c	α	β	c	α	β
Standard	$3.76_{3.62}^{3.90}$	$1.74_{1.73}^{1.76}$	$0.27_{0.26}^{0.29}$	$3.22_{3.16}^{3.28}$	$1.40_{1.39}^{1.41}$	$0.45_{0.44}^{0.46}$	$8.88_{8.58}^{9.18}$	$1.54_{1.53}^{1.56}$	$0.35_{0.33}^{0.36}$
High iota	$0.97_{0.94}^{0.99}$	$1.21_{1.19}^{1.23}$	$0.96_{0.94}^{0.97}$	$21.78_{21.25}^{22.35}$	$1.09_{1.07}^{1.11}$	$0.86_{0.84}^{0.88}$	$2.19_{2.11}^{2.28}$	$1.09_{1.06}^{1.11}$	$0.78_{0.77}^{0.80}$
High mirror	$8.81_{8.55}^{9.10}$	$1.12_{1.09}^{1.14}$	$0.35_{0.33}^{0.37}$	$5.07_{4.68}^{5.39}$	$0.96_{0.94}^{0.99}$	$0.68_{0.64}^{0.73}$	$20.20_{19.58}^{20.82}$	$0.94_{0.92}^{0.95}$	$0.32_{0.30}^{0.33}$

3.2. Comparison of SR expressions with W7-X data

A systematic comparison of the neutral gas pressure p_n obtained by the clustered SR expressions, defined by equation (4) and table 7, with the corresponding W7-X data is performed. The relevant parity plots (SR predicted p_n vs experimental p_n) are shown in figures 3–5 for the standard, high iota and high mirror configurations respectively. Each figure reflects the p_n predictions for the three clustered regimes of f_rad (vertically) at the three considered ports (horizontally), with the corresponding clustered SR expression and the value of R^2 for the testing database being showcased at each of the

Table 7. Coefficient c and exponents α and β of the clustered two-parameter expressions (4) for all magnetic configurations and ports, with 95% confidence interval.

f_{rad}	AEH			AEP			AEI		
	c	α	β	c	α	β	c	α	β
				Standard					
(0,0.25)	$5.53_{5.30}^{5.79}$	$1.63_{1.60}^{1.66}$	$0.15_{0.12}^{0.18}$	$3.17_{3.08}^{3.26}$	$1.48_{1.47}^{1.50}$	$0.28_{0.26}^{0.3}$	$13.37_{12.87}^{14.02}$	$1.27_{1.25}^{1.29}$	$0.4_{0.38}^{0.42}$
(0.25,0.5)	$3.92_{3.69}^{4.19}$	$1.73_{1.7}^{1.76}$	$0.24_{0.22}^{0.27}$	$5.43_{5.28}^{5.58}$	$1.05_{1.03}^{1.07}$	$0.62_{0.61}^{0.64}$	$8.65_{8.21}^{9.12}$	$1.58_{1.55}^{1.6}$	$0.3_{0.28}^{0.32}$
(0.5,0.8)	$2.75_{2.53}^{2.98}$	$1.84_{1.8}^{1.89}$	$0.36_{0.33}^{0.39}$	$9.29_{8.92}^{9.66}$	$0.67_{0.64}^{0.7}$	$0.9_{0.88}^{0.92}$	$5.62_{5.16}^{6.04}$	$1.79_{1.74}^{1.83}$	$0.31_{0.28}^{0.34}$
				High iota					
(0,0.3)	$0.84_{0.81}^{0.86}$	$1.22_{1.2}^{1.24}$	$1.02_{1.0}^{1.05}$	$22.37_{21.79}^{22.98}$	$1.08_{1.06}^{1.10}$	$0.84_{0.82}^{0.87}$	$2.21_{2.16}^{2.27}$	$0.97_{0.96}^{0.99}$	$0.90_{0.88}^{0.92}$
(0.3,0.5)	$1.03_{0.94}^{1.12}$	$1.13_{1.07}^{1.19}$	$1.04_{1.01}^{1.07}$	$21.80_{20.23}^{23.71}$	$0.99_{0.93}^{1.05}$	$1.02_{0.98}^{1.06}$	$2.13_{2.04}^{2.26}$	$1.14_{1.10}^{1.17}$	$0.75_{0.73}^{0.77}$
(0.5,0.8)	$1.68_{1.55}^{1.83}$	$1.05_{1.0}^{1.10}$	$0.79_{0.76}^{0.83}$	$26.54_{22.73}^{29.58}$	$1.20_{1.11}^{1.28}$	$0.58_{0.54}^{0.64}$	$0.50_{0.40}^{0.67}$	$1.88_{1.72}^{2.01}$	$0.64_{0.58}^{0.70}$
				High mirror					
(0,0.3)	$8.98_{8.72}^{9.24}$	$0.99_{0.96}^{1.01}$	$0.5_{0.47}^{0.52}$	$4.68_{4.38}^{4.94}$	$0.84_{0.81}^{0.87}$	$0.81_{0.79}^{0.85}$	$22.86_{22.39}^{23.36}$	$0.79_{0.78}^{0.81}$	$0.38_{0.36}^{0.41}$
(0.3,0.5)	$8.30_{7.94}^{8.69}$	$1.23_{1.20}^{1.26}$	$0.24_{0.21}^{0.26}$	$7.75_{7.43}^{8.12}$	$0.46_{0.44}^{0.49}$	$1.12_{1.09}^{1.15}$	$18.19_{17.72}^{18.70}$	$1.03_{1.01}^{1.05}$	$0.28_{0.26}^{0.30}$
(0.5,0.8)	$3.73_{3.26}^{4.24}$	$1.98_{1.86}^{2.10}$	$-0.2_{-0.29}^{-0.12}$	$2.60_{2.0}^{3.59}$	$1.89_{1.52}^{2.17}$	$0.01_{-0.2}^{0.29}$	$9.0_{8.07}^{10.0}$	$1.70_{1.60}^{1.80}$	$-0.09_{-0.16}^{-0.02}$

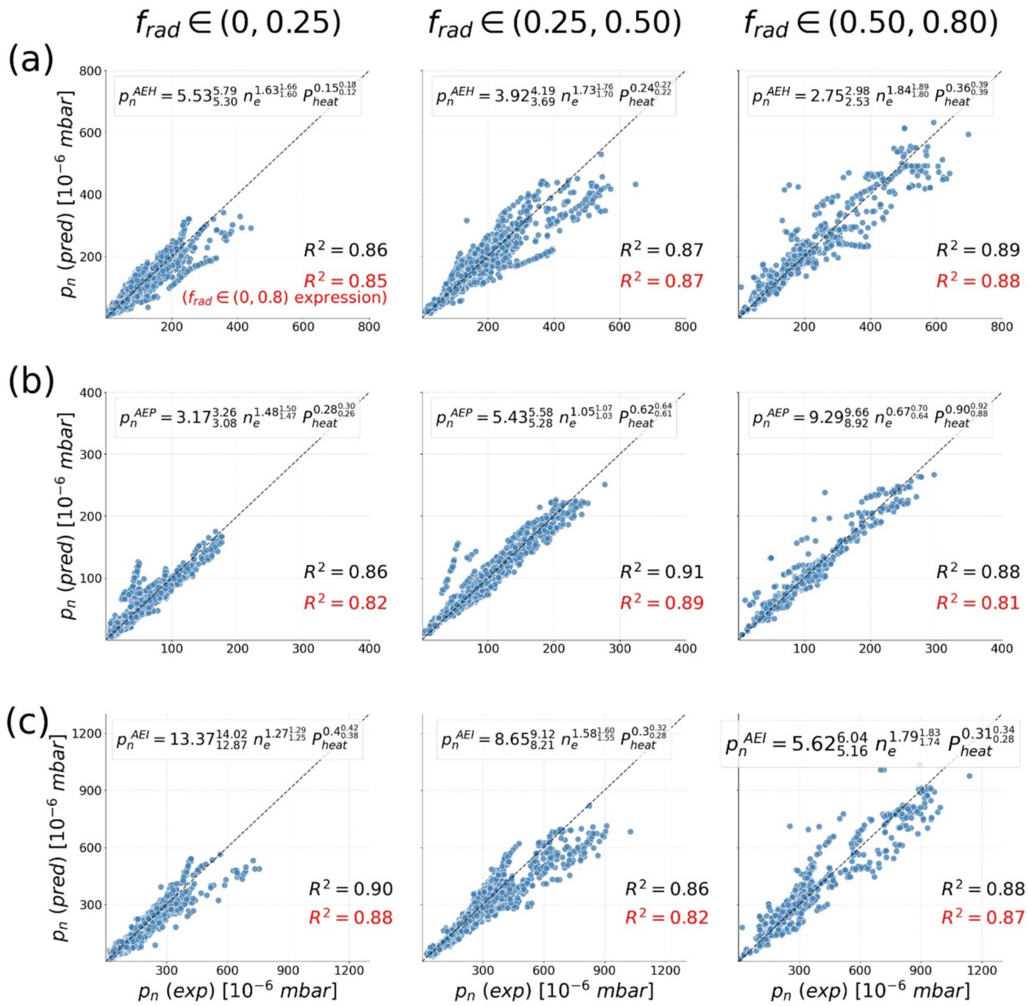

Figure 3. Standard configuration—parity plot of the clustered two-variable expressions for p_{n} (equation (4)) at the AEH (a), AEP (b) and AEI (c) ports. In each case, the corresponding R^2 value of the overall two-variable expression is given in red. Values of the coefficient c and exponents α, β of equation (4) are shown with a 95% confidence interval.

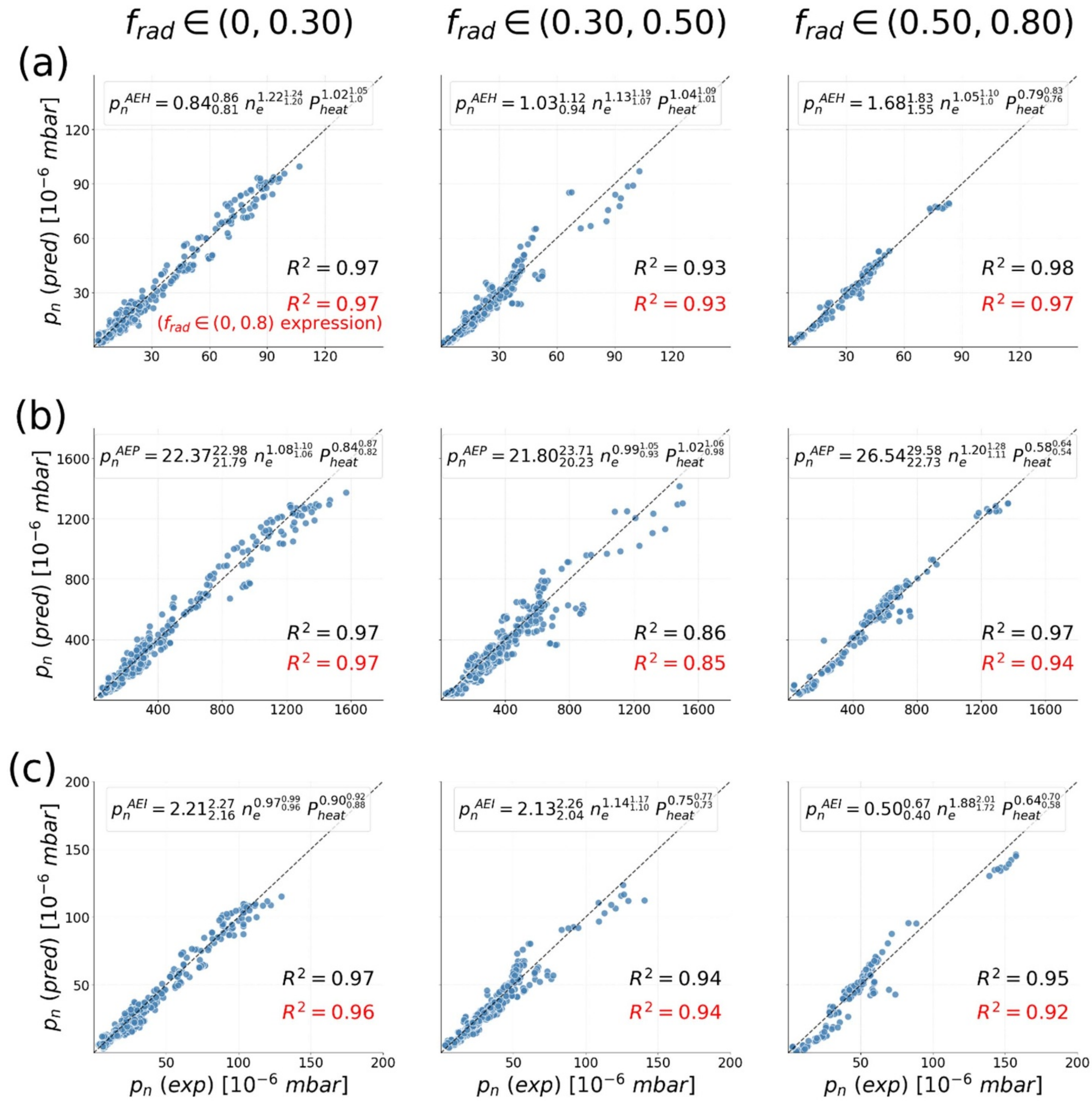

Figure 4. High iota configuration—parity plot of the clustered two-variable expressions for p_n (equation (4)) at the AEH (a), AEP (b) and AEI (c) ports. In each case, the corresponding R^2 value of the overall two-variable expression is given in red. Values of the coefficient c and exponents α, β of equation (4) are shown with a 95% confidence interval.

nine datasets per magnetic configuration. In addition, in each case (magnetic configuration and port), the R^2 value of the overall SR − expression, valid in $0 < f_{\rm rad} < 0.8$ (table 6), is also written in red for direct comparison between the clustered and overall SR expressions. It is noted that in order to perform the comparison on the same basis, the R^2 values of the overall expressions are calculated using the data points of the corresponding cluster and its accuracy, as mentioned earlier, is measured on the test dataset.

It is seen in figures 3–5, that the majority of data points obtained by the two-parameter expressions closely align with the identity line and the R^2 values are typically very close or exceed 0.90, in most cases. It is also seen that the overall expressions also achieve adequate values of R^2, but they underperform when compared to the clustered expressions, as they score lower accuracy metrics. In the clustered regions the underlying physics is more concentrated, compared to the whole $f_{\rm rad}$ region and therefore, the corresponding clustered SR expressions can more effectively adapt their behavior, as defined by the coefficient and two exponents, to implicitly model the effect of $f_{\rm rad}$ and better fit the specific region. On the contrary, the overall SR expressions that have been trained at $0 < f_{\rm rad} < 0.8$ can capture less efficiently the changes in the physical behavior encountered with regard to $f_{\rm rad}$. Thus, they represent correlations that describe the 'average' behavior of the dataset, which may be influenced by the region where the data points are most densely concentrated.

Furthermore, in the standard configuration (figure 3), the maximum pressure at each port increases with $f_{\rm rad}$ and this

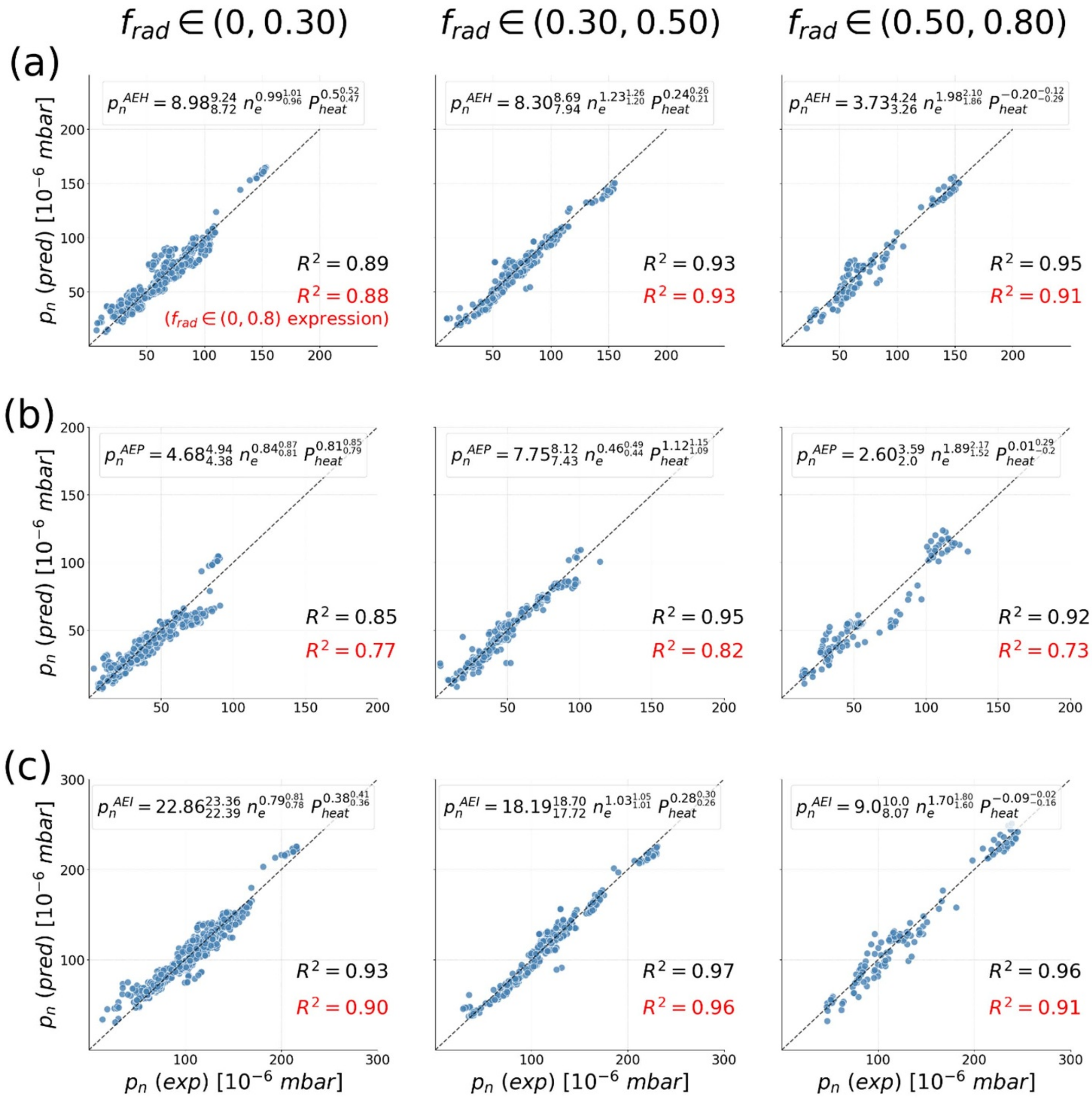

Figure 5. High mirror configuration—parity plot of the clustered two-variable expressions for p_n (equation (4)) at the AEH (a), AEP (b) and AEI (c) ports. In each case, the corresponding R^2 value of the overall two-variable expression is given in red. Values of the coefficient c and exponents α, β of equation (4) are shown with a 95% confidence interval.

is also something reflected by the symbolic expressions. This indicates that the clustered expressions not only capture the overall variability of pressure, but they also maintain range specificity. The values of R^2 of the clustered expressions vary between 0.86 and 0.91, while the corresponding ones of the overall expressions are lower and vary from 0.81 up to 0.89. In the high-iota configuration, the maximum pressure appears to be achievable in all f_{rad} clusters, suggesting a more balanced behavior with respect to f_{rad}. As a result, whole f_{rad} range expression accuracy is more comparable and only slightly behind clustered expressions. In the high iota configuration, the values of R^2 of the clustered and overall expressions are around 0.95, i.e. much higher than in the standard configuration (only at the AEP port and for $0.5 < f_{\text{rad}} < 0.8$ the values of R^2 are around 0.85). In the high mirror configuration, the accuracy metrics vary from 0.85 up to 0.97 in the clustered expressions and from 0.75 up to 0.91 in the overall expressions, with the lowest values for all SR expressions appearing in the AEP port (high iota section). Compared to the other two configurations, the high mirror configuration is characterized by the relatively small number of data points and the low measured neutral gas pressures in all three ports. It is the only configuration, where in the clustered expression for $0.5 < f_{\text{rad}} < 0.8$, SR yields negative exponents β of the heating power, indicating a major change of the effect of P_{heat} on p_n as f_{rad} increases.

A more thorough description on the effectiveness of the SR expressions to capture the behavior of the experimental data is shown in figure 5, where the predicted and experimental neutral gas pressure p_n (y-axis) with respect to the line integrated

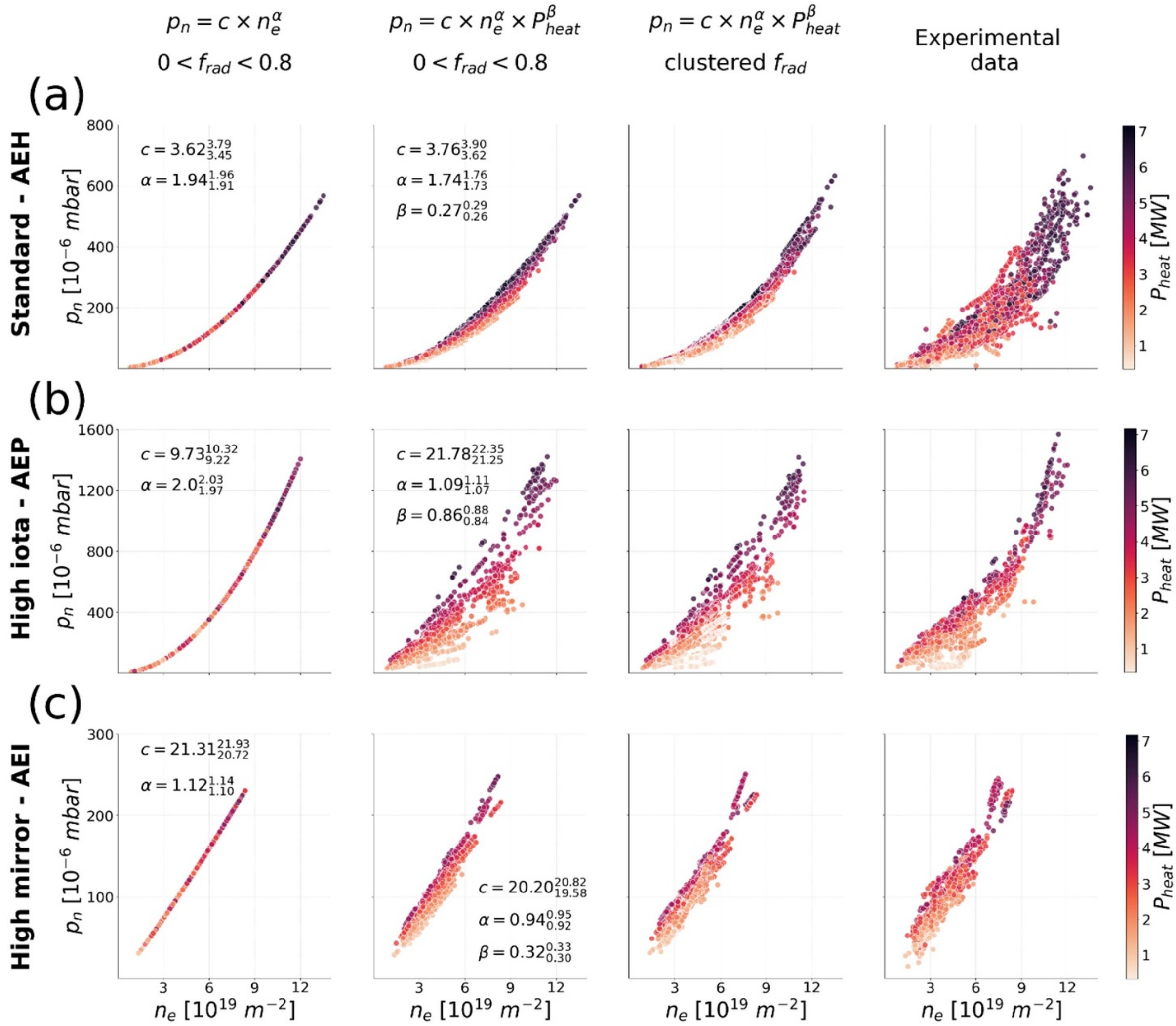

Figure 6. Predicted and experimental neutral gas pressure p_n (y-axis) vs electron density n_e (x-axis) and heating power P_{heat} (color depth) for the standard configuration at the AEH port (a), the high iota configuration at the AEP port (b) and the high mirror configuration at the AEI port (c); the SR predicted data are based on the overall one-variable expressions and the overall and clustered two-variable expressions (the values of c, α, β in the clustered expressions are provided in table 7).

electron density n_e (x-axis) and heating power P_{heat} (color depth) are shown. The comparison is performed for the indicative cases of the standard configuration at the AEH port (up), the high iota configuration at the AEP port (middle) and the high mirror configuration at the AEI port (down). The SR predicted neutral gas pressures include the ones obtained by the overall one-parameter expressions (3), as well as the ones deduced by the overall and clustered two-variable expressions, given by equation (4). The coefficients and exponents of the overall one- and two-parameter expressions are shown in figure 6, while those of the clustered two-parameter expressions are provided in table 7.

All symbolic expressions appear to capture to a great extent a behavior that aligns with the 'average' trend of the experimental data, despite the significant scatter in the latter. The one-parameter expressions in terms of n_e, which as seen in table 5 has high accuracy in the standard and high mirror configurations at the AEH and AEI ports respectively, yield predictions that compared to the scattered experimental data are overly compressed. Thus, although its statistical measures of accuracy may be adequate, the introduction of a second variable, that of P_{heat} is needed to disperse the deduced data points away from the compact pattern and better predict the experimental data. As it is seen in figure 6, this is clearly the case with the two-parameter expressions. Furthermore, clustered symbolic expressions can be more region-specific and deviate from the 'average' behavior, offering a more refined and adaptive prediction. The two-parameter overall expressions trained in the whole range of f_{rad} can make estimations based on a similar p_n behavior as the clustered expressions, but a much more detailed approach is provided by the clustered expressions, which, for instance, may further scatter the predictions. This is clearly observed in all indicative cases presented in figure 6, which in turn, better approach the experimental data. Although this improvement is seemingly minor, the clustered expressions still provide a better alignment with the experimental data, yielding models that more closely match the dataset details.

Closing this section, it may be useful to note that in certain cases, pure data driven approaches, as the present one, even though they achieve satisfactory metrics and predictions, oftentimes, the proposed optimal closed form expressions do not necessarily provide straightforward explanation of the supposed or hidden physical behavior of the system. To address this, additional physical insights should be encompassed to achieve more compact and physics-inspired interpretations. This issue is not considered in the present work. However, having established, the current SR framework, future studies will focus on this aspect.

4. Concluding remarks

Experimental data from the stellarator W7-X OP1.2b experimental campaign for three magnetic configurations (standard, high iota and high mirror) at three ports (AEH, AEP and AEI) of the exhaust system have been employed to correlate the neutral gas pressure p_n at the sub-divertor with various plasma parameters. More specifically, a ML method, namely SR, which is based on GP principles, has been elaborated and optimized to connect and reveal the most important parameters in the estimation of p_n. While common regression schemes assume a predetermined functional form, SR autonomously discovers the functional structure of the model, purely from data.

The considered plasma parameters include the line integrated electron density n_e, the heating power P_{heat}, the toroidal plasma current I_{tor}, the fraction of the radiated power f_{rad} and the center T_c and edge T_e ion temperatures. The investigation is carried out for partially attached conditions with $f_{rad} < 0.8$. For each magnetic configuration and port ($3 \times 3 = 9$), closed form expressions are deduced in the whole range of f_{rad}, as well as well as in three clusters of f_{rad}, accordingly obtained by a K means algorithm. Totally, 36 expressions are deduced: 9 refer to the whole range of f_{rad} and the other 27 to the clusters of f_{rad}.

In each case, starting from a general closed form expression, where p_n depends upon all considered plasma parameters and following a systematic SR purely data driven evolution methodology, yields to a sequence of expressions of certain COMP and accuracy. Balancing model generality, COMP and accuracy along the Pareto front curve, it is finally deduced that the dominant parameter affecting p_n are n_e and P_{heat}, while the other parameters have minor impact, at least from the statistical point of view and may not be included in the expressions estimating p_n. Therefore, for all considered magnetic configurations and ports the proposed closed form expressions are of the form $p_n = f(n_e, P_{heat})$ and more specifically they contain the product of n_e and P_{heat} raised at some powers times a constant (see equation (4)). To have a wider perspective, a 95% bootstrap confidence interval for each of the optimized values of the constants and exponents has been calculated.

It is argued that the functional form of the proposed, purely data driven, SR expressions have certain qualitative resemblance with corresponding expressions, correlating the divertor pressure with plasma parameters, obtained via the extended two-point model [16, 17], as well to corresponding results based on typical regression methods via predefined equations [63]. The similarity with the analytical results is based on the considerations that the sub-divertor neutral gas pressure is proportional to the divertor pressure [13] and the power entering the SOL is proportional to the heating power. Always, the SR expressions have simpler form and are capable to accurately capture the behavior of the experimental data with the well-known moderate dependency of the neutral gas pressure on the upstream density in stellarators, in contrast to much stronger dependencies in tokamak devices.

Overall, the proposed two-parameter SR expressions have certain advantages:

- Suggest a general, simple and accurate approach to obtain p_n over a wide range of system conditions.
- Show only a marginal decrease in accuracy metrics compared to more complex expressions that include additional system parameters.
- Reduce the possibility of over-fitting and offer enhanced generalization potential (due to their simple form).

For each experimental dataset (i.e. configuration—port—fraction of radiated power) the constant and the exponents are accordingly readjusted in order to capture the variation of the experimental conditions. The expressions trained in the whole range of f_{rad}, as well as the clustered ones, can make accurate estimations of p_n, with the latter ones being able to further scatter the predictions, more closely to the experimental data.

The expected benefit of the approach is the assessment of the impact of the plasma parameters on the performance of the exhaust system by simple symbolic expressions, without prior knowledge. The neutral gas pressure calculations that are otherwise computationally very expensive may be substantially accelerated. The proposed simple and reliable correlations between the magnetic field configuration, the plasma parameters and the exhaust system characteristics provides a practical and analytical tool in the development of compatible plasma—exhaust scenarios and the design and optimization of effective particle exhaust systems in W7-X and contributes to more efficient experimental design and operation.

Having the present solid SR methodology in place, future work will include the employment of corresponding data of the OP2.2 and OP2.3 experimental campaigns in W7-X to examine the functioning and applicability of the proposed expressions in different experimental conditions. Assessing the applicability and effectiveness of the proposed predictions to data from other W7-X experimental campaigns constitutes a key element in the wider implementation of the present approach. It is noted that, although the present SR model is based on the limited heating power range, available in OP1.2b, in principal, the proposed SR explicit expressions could also be used for extrapolation outside the investigated data range. However, achieving sufficient predictive accuracy remains an open question and should be carefully examined.

Furthermore, based on predictions via analytical [17, 21] and numerical tools [24, 26, 30], as well as on correlations deduced by typical data regression methods [63], it is planned to enhance the suggested SR methodology with physical insights in order to have a more conclusive quantitative type comparison on the effect of the various plasma parameters on the neutral gas pressure. The investigation may include the examined, as well as additional plasma parameters (e.g. power entering the SOL, radiated power in the SOL and coil control current) and also exploit plasma parameters measured at the plasma edge, which are available in the ongoing campaigns. Since the employed data driven approach shows great potential it may also be used in tokamaks.

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors would like to thank Dr F Reimold and Dr U Wenzel for helpful discussions concerning this work.

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grand Agreement No 101052200—EUROfusion), as well as the National Program of Controlled Thermonuclear Fusion. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This work was supported by computational time granted from the EUROfusion HPC Marconi Fusion.

ORCID iDs

D Angelis https://orcid.org/0009-0001-9565-6277
F Sofos https://orcid.org/0000-0001-5036-2120
S Misdanitis https://orcid.org/0009-0007-2066-2545
C Dritselis https://orcid.org/0000-0002-5907-4579
T E Karakasidis https://orcid.org/0000-0001-9580-0702
D Valougeorgis https://orcid.org/0000-0002-1358-1444
V Haak https://orcid.org/0000-0001-9158-5566
D Naujoks https://orcid.org/0000-0003-4265-6078
G Schlisio https://orcid.org/0000-0002-5430-0645
S A Bozhenkov https://orcid.org/0000-0003-4289-3532
V Perseo https://orcid.org/0000-0001-8473-9002

References

[1] Kallenbach A, Sun H J, Eich T, Carralero D, Hobirk J, Scarabosio A and Siccinio M 2018 Parameter dependence of the separatrix density in nitrogen seeded ASDEX Upgrade H-mode discharges *Plasma Phys. Control. Fusion* **60** 045006

[2] You J H *et al* 2022 Divertor of the European DEMO: engineering and technologies for power exhaust *Fusion Eng. Des.* **175** 1113010

[3] Pedersen T S *et al* 2017 Key results from the first plasma operation phase and outlook for future performance in Wendelstein 7-X *Phys. Plasmas* **24** 055503

[4] Gao Y *et al* 2019 Effects of toroidal plasma current on divertor power depositions on Wendelstein 7-X *Nucl. Fusion* **59** 106015

[5] Lazerson S A *et al* 2019 Tuning of the rotational transform in Wendelstein 7-X *Nucl. Fusion* **59** 126004

[6] Jakubowski M *et al* 2021 Overview of the results from divertor experiments with attached and detached plasmas at Wendelstein 7-X and their implications for steady-state operation *Nucl. Fusion* **61** 106003

[7] Schmitz O *et al* 2021 Stable heat and particle flux detachment with efficient particle exhaust in the island divertor of Wendelstein 7-X *Nucl. Fusion* **61** 016026

[8] Schlisio G *et al* 2021 The evolution of the bound particle reservoir in Wendelstein 7-X and its influence on plasma control *Nucl. Fusion* **61** 036031

[9] Kremeyer T *et al* 2022 Analysis of hydrogen fueling, recycling and confinement at Wendelstein 7-X via a single-reservoir particle balance *Nucl. Fusion* **62** 006023

[10] Wenzel U *et al* 2022 Gas exhaust in the Wendelstein 7-X stellarator during the first divertor operation *Nucl. Fusion* **62** 096016

[11] Haak V, Bozhenkov S A, Feng Y, Kharwandikar A, Kremeyer T, Naujoks D, Perseo V, Schlisio G and Wenzel U 2023 Overview over the neutral gas pressures in Wendelstein 7-X during divertor operation under boronized wall conditions *Plasma Phys. Control. Fusion* **65** 055024

[12] Haak V *et al* 2024 First operation and validation of simulations for the divertor cryo-vacuum pump in Wendelstein 7-X *Fusion Eng. Des.* **208** 114671

[13] Wenzel U, Motojima G, Kobayashi M, Goto M, Hayashi Y, Masuzaki S, Kawamura G, Jagielski B and Haak V 2024 Ultrahigh neutral pressures in the sub-divertor of the Large Helical device *Nucl. Fusion* **64** 034002

[14] Morisaki T *et al* 2013 Initial experiments towards edge plasma control with a closed helical divertor in LHD *Nucl. Fusion* **53** 063014

[15] Kotov V and Reiter D 2009 Two-point analysis of the numerical modelling of detached divertor plasmas *Plasma Phys. Control. Fusion* **51** 115002

[16] Kobayashi M *et al* 2010 Transport characteristics in the stochastic magnetic boundary of LHD: magnetic field topology and its impact on divertor physics and impurity transport *Fusion Sci. Technol.* **58** 220–31

[17] Feng Y, Kobayashi M, Lunt T and Reiter D 2011 Comparison between stellarator and tokamak divertor transport *Plasma Phys. Control. Fusion* **53** 024009

[18] Baelmans M, Reiter D and Börner P 2005 The EIRENE and B2-EIRENE codes *Fusion Sci. Technol.* **47** 172–86

[19] Feng Y *et al* 2014 Recent Improvements in the EMC3-Eirene code *Contrib. Plasma Phys.* **54** 426–31

[20] Feng Y *et al* 2005 Physics of the geometry-related detachment stability in W7-AS *Nucl. Fusion* **45** 89–95

[21] Feng Y *et al* 2021 Understanding detachment of the W7-X island divertor *Nucl. Fusion* **61** 066012

[22] Winters V R *et al* 2021 EMC3-EIRENE simulation of first wall recycling fluxes in W7-X with relation to H-alpha measurements *Plasma Phys. Control. Fusion* **63** 045016

[23] Boeyaert D *et al* 2024 Analysis of the neutral fluxes in the divertor region of Wendelstein 7-X under attached and detached conditions using EMC3-EIRENE *Plasma Phys. Control. Fusion* **66** 015006

[24] Gao Y *et al* 2023 Improvement in the simulation tools for heat distribution predictions and control of baffle and

middle divertor loads in Wendelstein 7-X *Nucl. Fusion* **63** 026031
[25] Varoutis S, Tantos C, Strobel H, Day C, Dhard C P, Haak V, Igitkhanov Y and Naujoks D 2024 Numerical simulation of neutral gas dynamics in the W7-X sub-divertor *Nucl. Fusion* **64** 076011
[26] Varoutis S, Tantos C, Strobel H, Boeyaert D, Igitkhanov Y, Litovoli F, Dhard C P, Haak V and Naujoks D 2025 Numerical analysis of gas exhaust in Wendelstein 7-X using the direct simulation Monte Carlo method *Nucl. Fusion* **65** 076001
[27] Kim S K *et al* 2024 Highest fusion performance without harmful edge energy bursts in tokamak *Nat. Commun.* **15** 3990
[28] Dasbach S and Wiesen S 2023 Towards fast surrogate models for interpolation of tokamak edge plasmas *Nucl. Mater. Energy* **34** 101396
[29] Gąsior P, Kastek M, Ladygina M and Sokulski D 2024 ANN based LIBS models for quasi-experimental spectra relevant for materials for next-step fusion reactors *Phys. Plasmas* **31** 052507
[30] Brenzke M, Wiesen S, Bernert M, Coster D, Jitsev J, Liang Y and von Toussaint U (ASDEX Upgrade Team and EUROfusion MST1 Team) 2021 Divertor power load predictions based on machine learning *Nucl. Fusion* **61** 046023
[31] Cai S, Mao Z, Wang Z, Yin M and Karniadakis G E 2021 Physics-informed neural networks (PINNs) for fluid mechanics: a review *Acta Mech. Sin.* **37** 1727–38
[32] Chouchene S, Brochard F, Lemoine N, Cavalier J, Desecures M and Weinzettl V 2024 Mutual interactions between plasma filaments in a tokamak evidenced by fast imaging and machine learning *Phys. Rev. E* **109** 045201
[33] Grelier E, Mitteau R and Moncada V 2023 Deep learning-based process for the automatic detection, tracking, and classification of thermal events on the in-vessel components of fusion reactors *Fusion Eng. Des.* **192** 113636
[34] Staron C, Le Borgne H, Mitteau R, Grelier E and Allezard N 2024 Detection of thermal events by semi-supervised learning for tokamak first wall safety *IEEE Trans. Instrum. Meas.* **73** 1–9
[35] Juven A, Aumeunier M H and Marot J 2024 U-Net for temperature estimation from simulated infrared images in tokamaks *Nucl. Mater. Energy* **38** 101562
[36] Ronneberger O, Fischer P and Brox T, U-net: convolutional networks for biomedical image segmentation *18th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)* (*Munich, Germany*, *5–9 October 2015*)
[37] Kajita S, Nishijima D, Fujii K, Tanaka H, Vernimmen J, van der Meiden H, Classen I and Ohno N 2024 Machine learning aided line intensity ratio method for helium–hydrogen mixed recombining plasmas *Plasma Phys. Controlled Fusion* **66** 105005
[38] Prantikos K, Chatzidakis S, Tsoukalas L H and Heifetz A 2023 Physics-informed neural network with transfer learning (TL-PINN) based on domain similarity measure for prediction of nuclear reactor transients *Sci. Rep.* **13** 16840
[39] Dong J *et al* 2021 Machine learning application to predict the electron temperature on the J-TEXT tokamak *Plasma Sci. Technol* **23** 085101
[40] Holt G K, Keats A, Pamela S, Kryjak M, Agnello A, Amorisco N C, Dudson B D and Smyrnakis M 2024 Tokamak divertor plasma emulation with machine learning *Nucl. Fusion* **64** 086009
[41] Lennon K, Shand C and Smith R 2024 Machine learning based Compton suppression for nuclear fusion plasma diagnostics *J. Fusion Energy* **43** 17
[42] Seo J *et al* 2024 Avoiding fusion plasma tearing instability with deep reinforcement learning *Nature* **626** 746–51
[43] Wu J *et al* 2024 High spatial resolution and accurate temperature profile measurements in a nuclear reactor core enabled by machine learning *IEEE Sens. J.* **24** 34394–403
[44] Seo J, Na Y-S, Kim B, Lee C Y, Park M S, Park S J and Lee Y H 2022 Development of an operation trajectory design algorithm for control of multiple 0D parameters using deep reinforcement learning in KSTAR *Nucl. Fusion* **62** 086049
[45] Murari A, Peluso E, Craciunescu T, Dormido-Canto S, Lungaroni M, Rossi R, Spolladore L, Vega J and Gelfusa M 2022 Frontiers in data analysis methods: from causality detection to data driven experimental design *Plasma Phys. Control. Fusion* **64** 024002
[46] Murari A, Peluso E, Vega J, García-Regaña J M, Velasco J L, Fuchert G and Gelfusa M 2021 Scaling laws of the energy confinement time in stellarators without renormalization factors *Nucl. Fusion* **61** 096036
[47] Chang C S, Ku S, Hager R, Churchill R M, Hughes J, Köchl F, Loarte A, Parail V and Pitts R A 2021 Constructing a new predictive scaling formula for ITER's divertor heat-load width informed by a simulation-anchored machine learning *Phys. Plasmas* **28** 022501
[48] Grulke O *et al* 2024 Overview of the first Wendelstein 7-X long pulse campaign with fully water-cooled plasma facing components *Nucl. Fusion* **64** 112002
[49] Wu J 2012 *Advances in K-means Clustering: A Data Mining Thinking* (Springer) pp 1–16
[50] Sofos F, Charakopoulos A, Papastamatiou K and Karakasidis T E 2022 A combined clustering/symbolic regression framework for fluid property prediction *Phys. Fluids* **34** 062004
[51] Angelis D, Sofos F and Karakasidis T 2023 Artificial intelligence in physical sciences: symbolic regression trends and perspectives *Arch. Comput. Methods Eng.* **30** 3845–65
[52] Makke N and Chawla S 2024 Interpretable scientific discovery with symbolic regression: a review *Artif. Intell. Rev.* **57** 2
[53] Koza J R 1994 Genetic programming as a means for programming computers by natural selection *Stat. Comput.* **4** 87–112
[54] Cranmer M 2023 Interpretable machine learning for science with PySR and symbolic regression (arXiv:2305.01582)
[55] Lathuilière S, Mesejo P, Alameda-Pineda X and Horaud R 2020 A comprehensive analysis of deep regression *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 2065–81
[56] Jin Y and Sendhoff B 2008 Pareto-based multiobjective machine learning: an overview and case studies *IEEE Trans. Syst. Man Cybern. C* **38** 397–415
[57] Papastamatiou K, Sofos F and Karakasidis T 2022 Machine learning symbolic equations for diffusion with physics-based descriptions *AIP Adv.* **12** 025004
[58] Lemos P *et al* 2023 Rediscovering orbital mechanics with machine learning *Mach. Learn.: Sci. Technol.* **4** 045002
[59] Virtanen P *et al* 2020 SciPy 1.0: fundamental algorithms for scientific computing in python *Nat. Methods* **17** 261–72
[60] Cumming G 2014 The new statistics: why and how *Psychol. Sci.* **25** 7–29
[61] Silva S, Vanneschi L, Cabral A I R and Vasconcelos M J 2018 A semi-supervised genetic programming method for dealing with noisy labels and hidden overfitting *Swarm Evol. Comput.* **39** 323–38
[62] O'Neill M, Vanneschi L, Gustafson S and Banzhaf W 2010 Open issues in genetic programming *Genet. Program. Evolvable Mach.* **11** 339–63
[63] Reimold F *et al* Power and particle exhaust limitations in W7-X and its relation to density build-up in the divertor *26th Int. Conf. on Plasma Surface Interaction in Controlled Fusion Devices* (*Marseille, France*, *12–17 May 2024*)