Key Engineering Materials
ISSN: 1662-9795, Vol. 1050, pp 157-169
doi:10.4028/p-5QE7ks

Submitted: 2026-01-19
Revised: 2026-02-17
Accepted: 2026-02-18
Online: 2026-04-16

Predicting Lateral Flow in Hot Sheet Metal Rolling Using Symbolic Regression

Daniel Luis Mederos[1,a], Amirali Hashemzadeh[1,b], Antonella Cometa[1,c*], Celal Soyarslan[1,2,d], Frederic E. Bock[3,e], Benjamin Klusemann[3,4,f], Ton van den Boogaard[1,g]

[1]University of Twente, Faculty of Engineering Technology, Chair of Nonlinear Solid Mechanics, Enschede, The Netherlands

[2]Fraunhofer Innovation Platform for Advanced Manufacturing, University of Twente, Enschede, The Netherlands

[3]Institute of Materials and Process Design, Helmholtz-Zentrum Hereon, Geesthacht, Germany

[4]Institute for Production Technology and Systems, Leuphana University of Lüneburg, Lüneburg, Germany

[a] d.luismederos@student.utwente.nl, [b] a.hashemzadeh@utwente.nl, [c]a.cometa@utwente.nl, [d] c.soyarslan@utwente.nl, [e] frederic.bock@hereon.de, [f]benjamin.klusemann@leuphana.de, [g]a.h.vandenboogaard@utwente.nl

Keywords: Hot Rolling, Lateral Spread, Finite Element Modeling, Symbolic Regression, Data-Driven Modeling.

Abstract. This work investigates the use of symbolic regression (SR) to address the trade-off between predictive accuracy and computational efficiency in modeling physical phenomena by constructing compact, closed-form expressions directly from data. In this study, SR is applied to develop fast and accurate models for predicting lateral spread in the hot rolling of steel slabs. The SR models are trained on high-fidelity finite element simulation data and evaluated against established analytical models. Model selection is guided by a parsimony-based optimization strategy that balances predictive accuracy and expression complexity. The results show that the SR-derived formulations achieve lower prediction errors with reduced complexity compared to traditional analytical models. Moreover, SR maintains strong predictive performance even when trained on limited datasets, demonstrating its robustness. Overall, the findings of this work highlight the suitability of symbolic regression for computationally efficient and accurate modeling of complex physical phenomena.

Introduction

Hot rolling is a key process in steel manufacturing, in which cast slabs are transformed into plates or strips while breaking down the as-cast microstructure and improving microstructural homogeneity [1]. As the thickness is reduced, the metal undergoes elongation in the rolling direction and spreads laterally, the latter being referred to as *lateral flow* or *spread*. Unlike processes where lateral deformation is largely constrained (e.g., cold rolling), lateral spread is a dominant phenomenon in hot rolling and must be accurately controlled to achieve the desired final width. Roughing hot rolling mills play a central role in shaping steel slabs, with rougher rolls reducing thickness and edger rolls controlling width. A schematic representation of a roughing stand is shown in Fig. 1. The final slab geometry and lateral spread at the stand exit result from the combined action of the rougher and edger rolls.
Maintaining tight tolerances over the thickness and width of the rolled slabs is essential to ensure product quality and guarantee consistent performance in downstream operations. Furthermore, precise profile control minimizes product rejection or the need for additional preprocessing. Therefore, accurate prediction of lateral spread is crucial for achieving dimensional control and optimizing the hot rolling process.

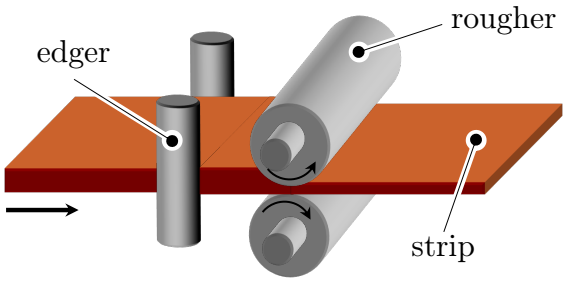


Fig. 1: Schematic of a roughing hot rolling stand, showing the vertical edger rolls and horizontal rougher rolls acting on the strip.

The lateral flow in hot rolling results from complex interactions between material properties, rolling parameters, and temperature, which makes it challenging to model with sufficient accuracy and precision. On the one hand, finite element (FE) simulations can capture the full three-dimensional deformation of the slab and provide detailed insight into material and process behavior [2]; however, despite their accuracy, such models are computationally demanding and thus unsuitable for direct integration into real-time control or optimization frameworks. On the other hand, analytical and semi-analytical models enable a fast prediction of lateral spread and are typically derived from plasticity theory, geometric relationships, and experimental observations. Although computationally efficient, many classical formulations oversimplify the complex physics of rolling; therefore, fail to predict lateral spread with sufficient accuracy, particularly under varying material and process conditions.

To reduce the trade-off between accuracy and efficiency in predicting complex physical phenomena, data-driven strategies have emerged, often in the context of hybrid modeling approaches [3]. These frameworks describe system behavior as the combination of a baseline mechanistic model, capturing established but simplified physical assumptions, and a residual contribution that remains unexplained by the model. Chinesta et al. [4] formalized this idea through the concept of modeler *ignorance*. Modeler *ignorance* denotes an epistemic model-form uncertainty, originally defined as the discrepancy between a physics-based model and the response of the real system; it may also be understood as the systematic discrepancy between a simplified analytical formulation and a more trusted high-fidelity reference description of the process. When experimental measurements are unavailable, this reference may be provided by a detailed FE model. Building on this concept, Bock et al. [5] proposed a predictive framework in which a fast but inaccurate analytical model is corrected towards high-fidelity FE solutions using machine learning to predict residual stresses induced by laser shock peening. A similar hybrid strategy has been applied to the prediction of lateral flow in hot strip rolling by Hashemzadeh et al. [6], who introduced an Analytical Predictor–Machine Learning Corrector framework in which analytical models provide initial width predictions that are subsequently refined using a machine learning model trained on high-fidelity FE data.

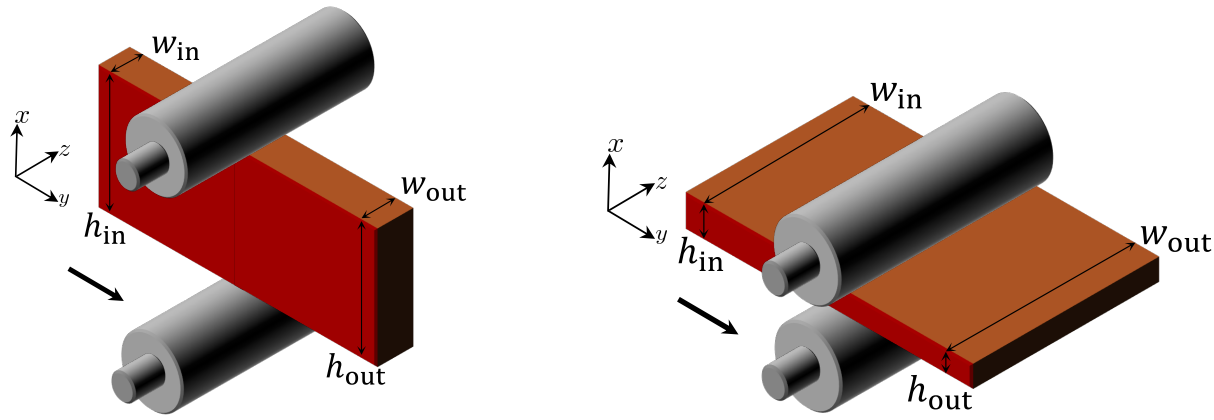
In addition to conventional machine-learning algorithms, symbolic regression (SR) has gained increasing attention as an alternative data-driven modeling approach. SR aims to identify explicit mathematical relationships between input and output variables through an optimization process that balances predictive accuracy and model complexity [7]. Unlike traditional regression methods that rely on prescribed functional forms, SR automatically explores combinations of variables and mathematical operators (such as addition, multiplication, and exponentiation) to discover concise analytical expressions that may support physical interpretation.

Within the broader landscape of data-driven strategies, SR can be employed in two complementary ways: as part of a hybrid modeling framework to learn model discrepancies [8], or as a model-discovery tool to directly derive analytical expressions that represent the underlying physical behavior encoded in the data [9, 10]. In this work, the second approach is adopted and symbolic regression is used to construct a compact analytical model for lateral spread. In the context of hot rolling, direct

experimental measurements of lateral flow are extremely difficult due to the harsh conditions in the rolling stand, including high temperatures and limited accessibility. As a result, high-fidelity FE simulations are employed as a surrogate model for the real physical system.
The remainder of the paper is organized as follows. First, the finite element framework used to simulate the hot rolling process is described, together with the generation of a high-fidelity data set that covers both rougher and edger rolling configurations and the adopted sampling strategy. Next, the symbolic regression methodology implemented using *PySR* [11] is presented, and the procedure for deriving compact analytical expressions for lateral spread from the simulation data is outlined. Finally, the resulting symbolic models are evaluated on an independent validation dataset and compared with established analytical formulations in terms of predictive accuracy and model complexity.

Methodology

FE simulations. Because lateral spread is governed by three-dimensional plastic flow, its accurate prediction requires a fully three-dimensional modeling approach. For this reason, the dataset describing lateral flow as a function of the rolling parameters is generated through three-dimensional FE simulations performed in *Abaqus*. In the numerical model, rolls are represented as rigid cylindrical surfaces, whereas the slab is modeled as a deformable body. Consequently, elastic roll deformations, including roll bending, are not considered. Roll bending modifies the roll gap profile across the strip width, thereby influencing the deformation mechanics and resulting lateral spread. Within the present approach, this effect could be incorporated by extending the FE model, for example through coupling with the influence function method to account for roll deflection [12]. However, this extension falls outside the scope of this study, as the symbolic regression models are assessed against analytical formulations that likewise assume rigid rolls. To capture both rolling configurations of interest in roughing mills, namely the edger and the rougher, the same FE framework is employed to perform a series of independent simulations, each corresponding to a distinct slab cross-sectional geometry. In particular, the width-to-thickness aspect ratio is systematically varied to represent edger-type and rougher-type configurations, rather than modeling their successive occurrence within a single rolling sequence, as shown in Fig. 2. Accordingly, the two rolling stages are treated independently, and deformation history is not transferred from the edger to the rougher (e.g., strain accumulation is not used as input for subsequent passes). This simplification is intentional, as the objective is to generate predictive models for lateral spread under each configuration and to enable direct comparison with analytical formulations, which likewise neglect deformation history between rolling steps. The slab length is chosen to be sufficiently large to ensure that steady-state rolling conditions are achieved, thereby eliminating transient effects near the entry and exit regions and enabling a consistent evaluation of lateral flow.



(a) Edger configuration.

(b) Rougher configuration.

Fig. 2: Edger and rougher rolling configurations considered in the FE model.

The slab has a length of 600 mm and an initial width of 6 mm. It is discretized using hexahedral elements, with the element size selected on the basis of a mesh convergence study. The final mesh resolution is set to a minimum element size of 3 mm in the longitudinal direction, 1.5 mm in the width direction, and 1.28 mm in thickness. This discretization is found to provide an appropriate balance between numerical accuracy and computational cost and is therefore kept constant across all simulations.

The material behavior is described using the Johnson–Cook constitutive model [13], which relates the von Mises flow stress (σ) to the equivalent plastic strain (ε_{p}), the equivalent plastic strain rate ($\dot{\varepsilon}_{\mathrm{p}}$) and the temperature ($T$), according to the following equation:

$$\sigma = \left[A + B\,\varepsilon_{\mathrm{p}}^{n}\right]\left[1 + C\ln\left(\frac{\dot{\varepsilon}_{\mathrm{p}}}{\dot{\varepsilon}_{0}}\right)\right]\left[1 - \left(\frac{T - T_{\mathrm{r}}}{T_{\mathrm{m}} - T_{\mathrm{r}}}\right)^{m}\right]. \tag{1}$$

Table 1: Johnson–Cook material parameters [13].

Parameter	Value	Unit	Description
A	33.901	MPa	Initial yield stress
B	100.18	MPa	Strain hardening coefficient
n	0.4951	–	Strain hardening exponent
C	0.2471	–	Strain-rate strengthening coefficient
m	0.6444	–	Thermal softening exponent
$\dot{\varepsilon}_0$	0.0037	s^{-1}	Reference plastic strain rate
T_{r}	900	°C	Reference temperature
T_{m}	1500	°C	Melting temperature

In this study, an isothermal analysis is performed at a constant temperature of 1100 °C and constant rolling speed of 2.2772 rad/s. These operating conditions are selected in accordance with representative hot rolling conditions reported in the literature [14]. Roll–slab interaction is modeled using a surface-to-surface contact formulation in *Abaqus* with a kinematic constraint and finite sliding. Normal contact is enforced using a hard contact formulation, while tangential behavior follows an isotropic Coulomb friction law with a constant friction coefficient of 0.2. Explicit simulations are employed [15] due to their robustness in handling complex contact conditions typical for rolling processes. To reduce computation time while preserving numerical stability, mass scaling is applied with a factor of 900, ensuring that the kinetic energy remains below 5% of the internal energy and that the response remains quasi-static.

The simulations are executed until steady-state conditions are achieved using the built-in steady-state detection capability in *Abaqus*. The detection criterion is based on the evolution of the equivalent plastic strain (PEEQ in *Abaqus* notation).

Input–output space for data generation. The rolling geometry is primarily characterized by a small set of geometric quantities that govern material flow and lateral spread, namely the slab width (w), slab thickness (h), and roll radius (R). These parameters control the degree of geometric constraint and curvature imposed by the rolls and therefore play a central role in the deformation kinematics. The subscripts *in* and *out* denote entry and exit conditions, respectively, as illustrated in Fig. 2.

To enable robust and scalable data-driven modeling, both input and output quantities are expressed in non-dimensional form. Non-dimensionalization normalizes the parameter space, improves numerical conditioning, and enhances generalization performance in data-driven models [16], while also ensuring dimensional consistency within the symbolic regression framework. Based on the characteristic geometric scales of the rolling process, the following three dimensionless input parameters are

defined:

$$x_1 := \frac{w_{\text{in}}}{h_{\text{in}}}, \quad x_2 := \frac{h_{\text{in}}}{R}, \quad x_3 := \frac{h_{\text{out}}}{h_{\text{in}}}.$$

The parameter intervals are chosen as:

$$x_1 \in [0.5,\, 2], \quad x_2 \in [0.1,\, 0.25], \quad x_3 \in [0.5,\, 1].$$

These ranges are selected to ensure that both rougher and edger configurations are adequately represented within the data set. Specifically, edger configurations correspond to $w_{\text{in}} < h_{\text{in}}$ ($x_1 < 1$), while rougher configurations are characterized by $w_{\text{in}} > h_{\text{in}}$ ($x_1 > 1$). The target output of the model is the exit strip width, which is likewise expressed in non-dimensional form as:

$$y := \frac{w_{\text{out}}}{R}.$$

Sampling method. Once the parameter ranges are defined, the number of simulations that can be performed is limited by computational cost, making it essential to select combinations of input parameters that efficiently cover the design space. This process, referred to as *sampling*, can be carried out using various strategies, including structured grids, random sampling, Latin hypercube sampling, and quasi-random sequences such as Sobol points.

The choice of sampling strategy has a direct impact on the quality and generalizability of the symbolic regression results. A key criterion for evaluating sampling methods is the *discrepancy*, a quantitative measure of how uniformly a set of points fills the parameter space [17]. Lower discrepancy values indicate more uniform coverage and, consequently, more representative sampling. The centered discrepancy is a specific form of discrepancy that emphasizes uniformity around the center of the sampling domain and reduces sensitivity to boundary effects. The centered discrepancy values for the different sampling strategies considered in this study are reported in Table 2.

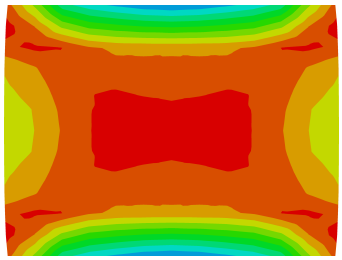
As shown in Table 2, Sobol sampling yields the lowest discrepancy and therefore provides the most uniform coverage of the parameter space. Since Sobol sequences are most effective when the number of samples is a power of two [18], 256 simulations were performed to generate the training dataset, representing a practical balance between computational cost and coverage of the input space. In addition, 128 independent simulations were generated for validation, using the same parameter ranges and Sobol sampling strategy but with a different random seed. The influence of the number of training samples on model accuracy is examined in a subsequent sensitivity analysis.

Table 2: Centered discrepancy values for the different sampling strategies using 256 samples in a three-dimensional input space.

Sampling Method	Centered Discrepancy
Grid	0.014900
Random	0.002068
Latin Hypercube	0.000446
Sobol	0.000045

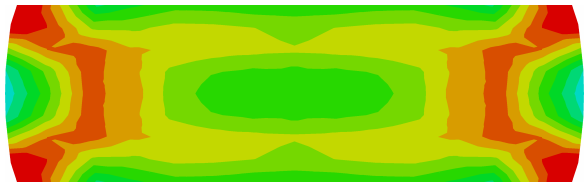
Post-processing and deformed shapes. After completion of the simulations, post-processing scripts are employed to extract the final strip width for training the predictive model. The deformed cross section is evaluated at the location where steady-state conditions are reached, that is also where lateral flow has fully developed. The exit width, denoted as w_{out}, is measured along the centerline of the deformed cross section and serves as a scalar measure of the accumulated lateral spread. Although this constitutes a simplification, since the actual cross section is generally non-uniform due to three-dimensional material flow, this assumption is adopted to ensure consistency with analytical models, which also assume a rectangular cross section.

In addition to width extraction, the steady-state deformed shapes of the slabs are analyzed to assess the influence of the process configuration on lateral flow behavior. Owing to differences in boundary conditions and dominant deformation mechanisms, the rougher and edger configurations exhibit distinct deformation patterns. Figures 3 and 4 show representative deformed slab cross sections for the two configurations, together with contour plots of the maximum equivalent plastic strain (PEEQ, in *Abaqus* notation) for randomly selected samples.



1.46×10^{-1} 2.00×10^{-1}

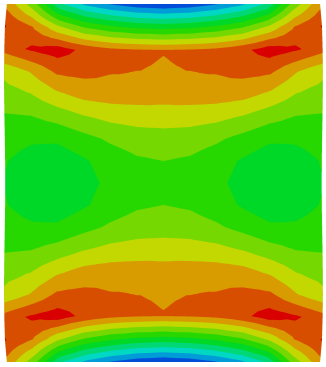
(a) h_{in} = 34.1620 mm, w_{in}= 36.1602 mm, h_{out} = 28.1873 mm, w_{out} = 38.9235 mm.



5.30×10^{-1} 6.42×10^{-1}

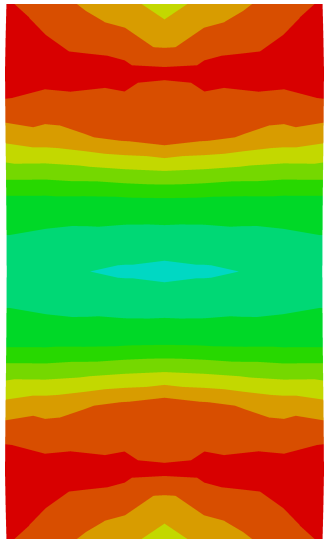
(b) h_{in} = 25.4125 mm, w_{in} = 39.7811 mm, h_{out} = 14.1775 mm, w_{out} = 49.5427 mm.

Fig. 3: Deformed slab cross-sections for representative rougher configurations, showing contours of equivalent plastic strain.



1.32×10^{-1} 1.93×10^{-1}

(a) h_{in} = 48.8902 mm, w_{in} = 35.3641 mm, h_{out} = 41.0318 mm, w_{out} = 37.6633 mm.



3.03×10^{-2} 6.35×10^{-2}

(b) h_{in} = 23.9475 mm, w_{in} = 13.3833 mm, h_{out} = 22.2743 mm, w_{out} = 13.7051 mm.

Fig. 4: Deformed slab cross-sections for representative edger configurations, showing contours of equivalent plastic strain.

Lateral spread originates from the lateral material flow induced by thickness reduction under the action of the rolls. For a given undeformed slab geometry (i.e., fixed width-to-thickness ratio), the magnitude of spread increases with increasing thickness reduction. Conversely, for a fixed amount of thickness reduction, the relative importance of spread increases as the width-to-thickness ratio decreases. For large width-to-thickness ratios, deformation in the central region approaches plane-strain conditions and lateral flow is strongly constrained, resulting in limited spread. As the width-to-thickness ratio decreases, the deformation becomes increasingly three-dimensional and lateral spread becomes more pronounced.

Symbolic regression. Symbolic regression is a machine learning technique that aims to discover analytical expressions that describe physical systems directly from data, without assuming a predefined model structure [19]. In contrast to conventional regression approaches, symbolic regression searches simultaneously for the functional form and the associated parameters, enabling the identification of governing equations from observational or simulation data. In this study, symbolic regression is implemented using *PySR*, a high-performance framework for scientific model discovery [11]. *PySR* represents candidate models as hierarchical expression trees, where mathematical operators form internal nodes and variables or constants appear as terminal leaves. The search over possible expressions is performed using evolutionary algorithms, including selection, crossover, and mutation, to identify models that best map the input parameters to the target output. The search space considered in this work includes the binary operators addition, subtraction, multiplication, and division, as well as the unary operators logarithm, sine, cosine, and exponential, allowing a broad class of nonlinear analytical expressions to be explored. To quantify model complexity, each symbolic expression is represented by its corresponding complexity tree, which explicitly reflects the number and arrangement of elementary operations. As an illustration, Fig. 5 presents the complexity tree corresponding to the following expression:

$$\alpha \, \tanh(\beta \, x + \gamma) \tag{2}$$

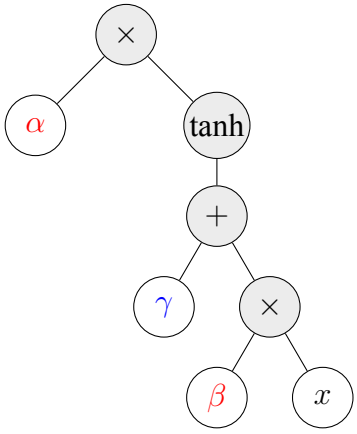


Fig. 5: Complexity tree representation of the symbolic expression in Eq. (2). Multiplicative coefficients are shown in red, the additive bias in blue, and the input variable in black.

For this illustrative example, the resulting complexity is 8, corresponding to the number of elementary operations and terms involved in the expression. Notably, this functional form is equivalent to the computation performed by a single neuron with a hyperbolic tangent activation function in a neural network. Even at this level, the corresponding complexity tree already has a complexity of 8. For a shallow neural network with a single hidden layer and three neurons, the explicit analytical representation would consist of the superposition of three such expressions, resulting in a complexity of 28. For deeper or wider architectures, the complexity grows substantially. In contrast, symbolic regression yields a single closed-form expression whose complexity is explicitly controlled, resulting in compact formulations that retain transparency and may support physical interpretation.
An important feature of *PySR* is its ability to balance model accuracy and complexity by optimizing along the Pareto front, according to the principle of the Occam's razor [19]. This enables the selection of expressions that achieve an optimal trade-off between predictive performance and analytical simplicity. In this study, a total of 256 simulations are used to train the model using *PySR*. The resulting symbolic models are subsequently validated using 128 independent simulations covering the same parameter ranges. Based on this validation set, the expressions are evaluated in terms of both prediction error and model complexity, and the optimal model is selected accordingly.

Analytical models. Three analytical models are used to benchmark the symbolic regression model [14]: the Ekelund, Tselikov, and Helmi–Alexander models.
The Ekelund model computes the exit width w_{out} from:

$$w_{\text{out}}{}^2 - w_{\text{in}}{}^2 = \left[8\delta - 4\ln\left(\frac{w_{\text{out}}}{w_{\text{in}}}\right)[h_{\text{in}} + h_{\text{out}}]\right]\sqrt{R\delta}\left[\frac{1.6\mu\sqrt{R\delta} - 1.2\delta}{h_{\text{in}} + h_{\text{out}}}\right], \tag{3}$$

where $\delta = h_{\text{in}} - h_{\text{out}}$.
The Tselikov model is given by:

$$\Delta w = w_{\text{out}} - w_{\text{in}} = \frac{2c}{\epsilon_{\text{h}}{}^2}\left[\sqrt{R\delta} - \frac{\delta}{2\mu}\right]\left[(1-\epsilon_{\text{h}})^2\ln\left(\frac{1}{1-\epsilon_{\text{h}}}\right) - \epsilon_{\text{h}} + 1.5\epsilon_{\text{h}}{}^2\right], \tag{4}$$

where $\epsilon_{\text{h}} := (h_{\text{in}} - h_{\text{out}})/h_{\text{in}}$ denotes the *relative draught*, the parameter c, known as the *spread factor*, depends on the ratio between the sample width and the projected contact arc (that is, $h_{\text{out}}/\sqrt{R\delta}$) and typically takes values in the range $[0.5,\ 1]$.
Finally, the Helmi–Alexander model is expressed as:

$$\ln\left(\frac{w_{\text{out}}}{w_{\text{in}}}\right) = 0.95\ln\left(\frac{h_{\text{in}}}{h_{\text{out}}}\right)\left(\frac{h_{\text{in}}}{w_{\text{in}}}\right)^{1.1}\exp\left(-0.707\left(\frac{w_{\text{in}}}{\sqrt{R\delta}}\right)\left(\frac{h_{\text{in}}}{w_{\text{in}}}\right)^{0.971}\right). \tag{5}$$

Results and Discussions

SR-derived vs. analytical models. This study evaluates the performance of symbolic regression in comparison with the analytical models established by Ekelund, Tselikov, and Helmi-Alexander. The comparison is conducted using two complementary strategies: (i) assessing the predictive accuracy of symbolic expressions constrained to a complexity comparable to that of the analytical models, and (ii) examining the complexity required by symbolic regression to achieve prediction errors comparable to those of the analytical formulations.
To quantify the structural complexity of the analytical models, each formulation is represented by a *complexity tree*, following the structure illustrated in Fig. 5. The resulting complexity values for the analytical models, expressed as the total number of nodes in the corresponding trees, are reported in Table 3.

Table 3: Complexity of analytical models.

Model	Complexity (nodes)
Ekelund	37
Tselikov	30
Helmi-Alexander	31

The first analysis focuses on the former perspective, in which symbolic regression is constrained to produce a model of similar complexity to the analytical formulations and the resulting predictive performance is assessed. Under this constraint, the *PySR* symbolic regression tool yields the expression reported in Eq. 6, which has a complexity of 29 and is hereafter denoted as SR 29.

$$\frac{w_{\text{out}}}{R} = \left[\left(x_1^{0.9894} - 0.0129\right)x_2\right]\left[\left(x_3^{1.3275} - x_2\right)^{0.0317}\right]\exp\Big(\left(x_2^{x_1}\right)^{x_2}\cos(x_3)\left(1.0324 - x_3\right)\Big). \tag{6}$$

The prediction error of this equation is compared to that of the analytical models using the Mean Absolute Error (MAE) and R^2, both computed based on the predicted slab width. Table 4 summarizes the

results and shows that the *PySR*-derived model achieves an MAE nearly one order of magnitude lower than that of the best-performing analytical formulations (Tselikov's and Helmi–Alexander's models).

Table 4: Performance comparison of models based on MAE and R^2, ranked from worst to best MAE.

Model	**MAE [mm]**	R^2
Ekelund	2.0502	0.9808
Tselikov	0.9123	0.9956
Helmi-Alexander	1.7667	0.9741
SR 29	0.1485	0.9993

The analysis is then extended by relaxing the complexity constraint and identifying a symbolic model with predictive accuracy comparable to that of the analytical formulations. Under this condition, *PySR* yields the expression reported in Eq. 7, which achieves predictive performance similar to that of the best-performing analytical formulation, namely Tselikov's model (MAE = 1.0124 mm, R^2 = 0.9954), while exhibiting a substantially lower complexity of 7.

$$\frac{w_{\text{out}}}{R} = x_1\, x_2\, x_3^{-0.3309} \tag{7}$$

These results demonstrate that symbolic regression consistently achieves a superior trade-off between complexity and prediction error compared to existing analytical models.

Parsimony. A *parsimony* metric is employed to compare *PySR*-generated expressions of varying complexity with the analytical models. Parsimony provides a single scalar measure that balances prediction error and model complexity, defined as:

$$\text{Parsimony} = \text{MAE} + \lambda \text{Complexity}. \tag{8}$$

Here, λ denotes the complexity penalty factor. Lower values of λ emphasize prediction error, whereas higher values penalize model complexity more strongly. In this study, $\lambda = 0.01$ is selected. An overview of the results of the parsimony evaluation is provided in Table 5.

Table 5: Parsimony values for different model complexities with $\lambda = 0.01$.

Model	**Complexity**	**MAE [mm]**	**Parsimony**
Ekelund	37	2.0502	2.4202
Tselikov	30	0.9123	1.2123
Helmi-Alexander	31	1.7667	2.0767
SR 17	17	0.2374	0.4074
SR 19	19	0.2213	0.4113
SR 20	20	0.2184	0.4184
SR 22	22	0.2101	0.4301
SR 23	23	0.1865	0.4165
SR 24	24	0.1630	0.4030
SR 25	25	0.1586	0.4086
SR 26	26	0.1554	0.4154
SR 27	27	0.1534	0.4234
SR 28	28	0.1534	0.4334
SR 29	29	0.1485	0.4385

Among the SR-derived expressions, the model with a complexity of 24 (SR 24 in Table 5) has achieved the lowest parsimony score. The corresponding formulation is reported in Eq. 9.

$$\frac{w_{\text{out}}}{R} = \exp((A - x_3)\cos(x_3)\,(x_2^{x_2})^{x_1}) \left[\sin(x_3 - x_2)^B\,(x_1 x_2)\right]. \tag{9}$$

The fitted parameters in Eq. 9 are $A = 1.0207$ and $B = 0.0478$. The expression in Eq. 9 is identified as the best overall model when both prediction error and complexity are considered jointly.

SR-derived models for varying training dataset sizes. In the previous section, the expressions in Eq. 6, Eq. 7 and Eq. 9 were obtained using *PySR* based on a data set comprising 256 simulations. In this section, the influence of the size of the data set on the performance of the model is investigated by training *PySR* on progressively smaller subsets consisting of 128, 64 and 32 simulations and comparing the resulting models. Table 6 shows a comparison of *PySR*-generated models with a complexity of 29 (SR 29) trained with different numbers of simulations. The lowest MAE is obtained when using the full dataset of 256 simulations. As the number of simulations is reduced, the MAE increases, but remains lower than that of the analytical models for all dataset sizes considered (see Table 4). In addition, prediction errors remain relatively stable when the size of the data set is reduced from 128 to 64 and 32 simulations, indicating that symbolic regression can produce compact and accurate models even with limited training data.

Table 6: MAE of *PySR*-generated models (SR 29) for varying dataset sizes.

Number of simulations	MAE [mm]
256	0.1485
128	0.1700
64	0.1872
32	0.2334

Figure 6 presents the relationship between complexity and prediction error for the *PySR*-generated models trained with different dataset sizes. The graph shows that for model complexities above 20, the lowest error is obtained by the model trained on the full dataset of 256 simulations. Models trained on 128, 64, and 32 simulations exhibit higher errors, which remain comparable among these reduced dataset sizes. Across all cases, the prediction error stabilizes at a complexity of approximately 20, beyond which further increases in complexity result in negligible performance gains.

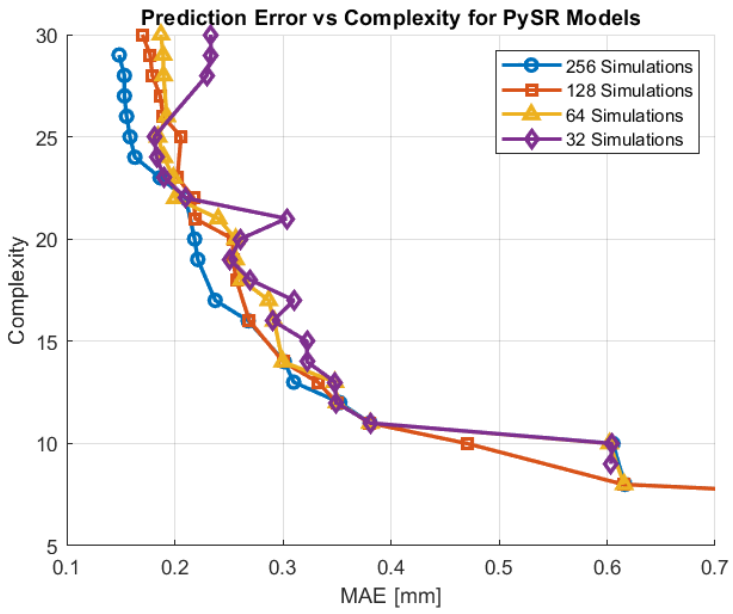


Fig. 6: Comparison of *PySR*-generated models trained using different numbers of simulations.

Impact of rolling configuration on model prediction error. Model performance is further examined by distinguishing between rougher and edger rolling configurations for both analytical and *PySR*-generated models. Although the analytical formulations were originally developed for rougher rolling, their MAE is also reported for edger-only and combined datasets for completeness.
Table 7 summarizes the resulting MAE values. As expected, the analytical models exhibit improved predictive performance when evaluated exclusively on rougher configurations, while their errors increase substantially for edger configurations. In contrast, the *PySR*-generated model with a complexity of 29 (SR 29) achieves consistently low MAE values across all configurations. Notably, the SR 29 model trained on the combined dataset maintains strong predictive accuracy, highlighting its robustness with respect to rolling configuration.

Table 7: MAE in mm for analytical and *PySR*-generated models across rolling configurations.

Model	Rougher	Edger	Both
Ekelund	1.6884	2.6215	1.9982
Tselikov	0.6158	1.3683	0.8657
Helmi-Alexander	0.7529	3.6659	1.7201
SR 29	0.1423	0.0520	0.1485

Finally, Figure 7 shows the relationship between model complexity and prediction error for *PySR*-generated models trained on rougher-only, edger-only, and combined datasets. At a complexity of approximately 20, the combined model achieves low prediction errors that are close to the configuration-specific results and significantly lower than those of the analytical formulations.

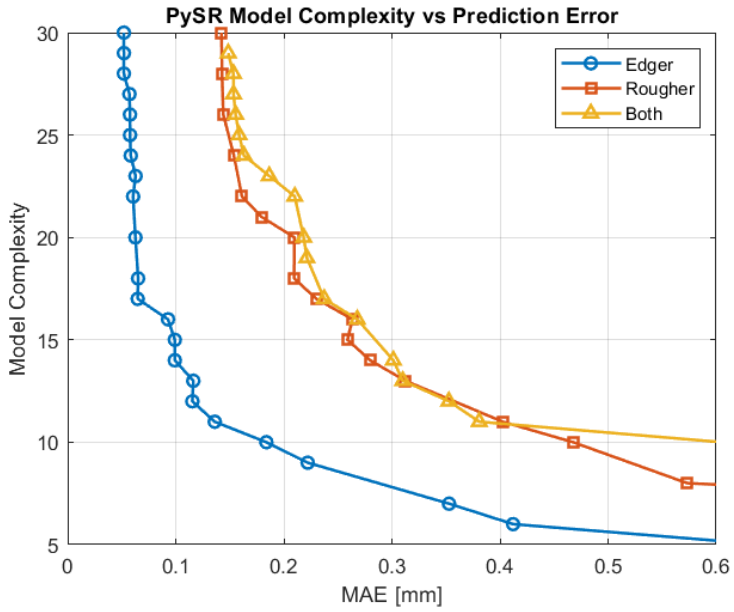


Fig. 7: Trade-off between complexity and prediction error for *PySR*-generated models trained on rougher-only, edger-only, and combined datasets (256 simulations).

Conclusions

This study presents a systematic application of symbolic regression to derive compact analytical models for predicting lateral spread in hot rolling. A structured workflow is employed, beginning with high-fidelity FE simulations and culminating in the identification of concise and accurate symbolic expressions.

A total of 256 FE simulations are performed using Sobol sampling to ensure uniform coverage of the input parameter space. The exit width of the rolled strip, normalized by the roll radius, is used as the target output for training the *PySR* framework. Multiple symbolic expressions are generated and subsequently validated using independent simulation datasets, and their predictive performance is assessed in comparison with established analytical models.

At comparable model complexity, the SR-derived expression achieves nearly one order of magnitude lower prediction error than the best-performing analytical formulation. Conversely, when constrained to match the accuracy of analytical models, symbolic regression yields expressions with substantially lower structural complexity. Furthermore, a parsimony-based analysis shows that the most balanced SR-derived model simultaneously achieves reduced complexity and superior predictive performance relative to all analytical models considered.

Unlike analytical formulations, which are developed specifically for rougher configurations, the symbolic regression models exhibit strong predictive performance across both rougher and edger rolling conditions. In particular, a single symbolic model trained on combined datasets maintains low prediction errors, indicating robust behavior over a wide range of geometrical configurations.

Overall, this work demonstrates that symbolic regression provides an effective alternative to traditional analytical modeling approaches by delivering compact analytical expressions with high predictive accuracy and low computational cost. The resulting formulations, characterized by explicitly controlled complexity, offer a promising basis for fast surrogate modeling of complex physical processes such as hot rolling.

Acknowledgements

This research was carried out under project number T22008 in the framework of the Research Program of the Materials innovation institute (M2i) (www.m2i.nl) supported by the Dutch government. The authors also gratefully acknowledge Tata Steel for valuable discussions.

References

[1] S. L. Semiatin, editor. *ASM Handbook, Volume 14A: Metalworking: Bulk Forming*. ASM International, Materials Park, OH, 2005.

[2] Z. Pater. The application of finite element method for analysis of cross-wedge rolling processes—a review. *Materials*, 16(13), 2023.

[3] Victor Champaney, Francisco Chinesta, and Elias Cueto. Engineering empowered by physics-based and data-driven hybrid models: A methodological overview. *International Journal of Material Forming*, 15(31), 2022.

[4] Francisco Chinesta, Elias Cueto, Emmanuelle Abisset-Chavanne, Jean-Louis Duval, and Fouad El Khaldi. Virtual, digital and hybrid twins: A new paradigm in data-based engineering and engineered data. *Archives of Computational Methods in Engineering*, 27(1):105–134, 2020.

[5] Frederic E. Bock, Sören Keller, Norbert Huber, and Benjamin Klusemann. Hybrid modelling by machine learning corrections of analytical model predictions towards high-fidelity simulation solutions. *Materials*, 14(8), 2021.

[6] A. E. Hashemzadeh, F. E. Bock, C. Hol, K. Schutte, A. Cometa, C. Soyarslan, B. Klusemann, and T. van den Boogaard. An analytical predictor–machine learning corrector scheme for modeling lateral flow in hot strip rolling. In *Proceedings of the 28th International ESAFORM Conference on Material Forming*, volume 54, pages 2002–2011, 2025.

[7] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[8] Delei Zou, Dilyar Thoti, and Zhihui Bao. Study on the application of discrepancy-guided symbolic regression algorithm in analyzing the impact resistance of UHP-SFRC target against high velocity projectile impact. *International Journal of Impact Engineering*, 201:105276, 2025.

[9] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A review. *Artificial Intelligence Review*, 57(1), 2024.

[10] Ismail Alaoui Abdellaoui and Siamak Mehrkanoon. Symbolic regression for scientific discovery: An application to wind speed forecasting. *ArXiv e-prints*, 2021.

[11] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. *ArXiv e-prints*, 2023.

[12] Tao Wang, Hong Xiao, Tie yong Zhao, and Xiang dong Qi. Improvement of 3-d fem coupled model on strip crown in hot rolling. *Journal of Iron and Steel Research, International*, 19(3):14–19, 2012.

[13] Mario F. Buchely, Shouvik Ganguly, David C. Van Aken, Ronald O'Malley, Simon Lekakh, and K. Chandrashekhara. Experimental development of Johnson–Cook strength model for different carbon steel grades and application for single-pass hot rolling. *steel research international*, 91(7):1900670, 2020.

[14] L. G. M. Sparling. Formula for 'spread' in hot flat rolling. *Proceedings of the Institution of Mechanical Engineers*, 175(1):604–640, 1961.

[15] J. O. Hallquist. *LS-DYNA Theory Manual*. Livermore Software Technology Corporation, Livermore, CA, 2006.

[16] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training DNNs: Methodology, analysis and application. *ArXiv e-prints*, 2020.

[17] Peter Shirley. Discrepancy as a quality measure for sample distributions. In *Eurographics*, 1991.

[18] SciPy Developers. SciPy documentation, 2024. Version 1.11.0.

[19] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.