

Interpretable Hybrid Surrogate Modeling for Diameter Deviation Prediction in Injection Molding Using Symbolic Regression and Gaussian Process Residual Correction

Amine Aboutaib

mramine18@gmail.com

Universite Sultan Moulay Slimane Faculte des Sciences et Techniques <https://orcid.org/0009-0005-4304-0584>

Hanaa OUAOMAR

Universite Sultan Moulay Slimane Faculte des Sciences et Techniques

Souad Touairi

Universite Sultan Moulay Slimane Faculte des Sciences et Techniques

Mourad KADIRI

Universite Sultan Moulay Slimane Faculte des Sciences et Techniques

Research Article

Keywords: Injection molding, symbolic regression, PySR, Gaussian Process, residual learning, process-informed feature engineering, small-data modeling, surrogate modeling, process optimization, uncertainty quantification

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11271710/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Interpretable Hybrid Surrogate Modeling for Diameter Deviation Prediction in Injection Molding Using Symbolic Regression and Gaussian Process Residual Correction

Amine ABOUTAIB, Hanaa OUAOMAR, Souad Touairi, Mourad KADIRI

Laboratory of Intelligent Systems, Advanced Mechanics and Renewable Energy
FST Beni Mellal
Sultan Moulay Slimane University
Beni Mellal, Morocco

Corresponding author: Amine ABOUTAIB — mramine18@gmail.com

Abstract

Dimensional quality in injection molding is hard to model when process variables act through coupled pressure, flow, and cycle-history effects and the experimental budget is small. This study develops an interpretable hybrid surrogate for that setting. The method combines process-informed feature engineering, symbolic regression with PySR, Gaussian Process (GP) residual correction, and surrogate-based optimization. Two design-of-experiments campaigns produced 127 runs, which were consolidated into 27 representative operating conditions by averaging repeated or overlapping conditions. Four controllable variables were retained (holding pressure, counter-pressure, cycle time, and injection speed) and expanded into twelve model inputs through pressure ratios, normalized pressure contrasts, multiplicative interaction terms, and rate-like combinations. A compact symbolic equation selected from the search served as the deterministic baseline. It reached $R^2 = 0.8665$ and $RMSE = 9.37 \mu\text{m}$ on the full dataset, and $R^2 = 0.8423$ and $RMSE = 10.19 \mu\text{m}$ under leave-one-out cross-validation (LOOCV). A GP with an RBF + White kernel, trained on the symbolic residuals, lowered the LOOCV RMSE to $9.32 \mu\text{m}$ and raised R^2 to 0.8681. Embedding the two surrogates in a Differential Evolution optimizer over the experimentally explored ranges gave predicted minima of about $33.25 \mu\text{m}$ for the symbolic model and $23.73 \mu\text{m}$ for the hybrid. These are model-based predictions and were not verified experimentally. The contribution is methodological: a transparent way to pair equation discovery with probabilistic residual refinement when only a few representative operating conditions are available. The framework keeps an explicit analytical core and adds a flexible correction layer and predictive uncertainty for engineering decision support.

Keywords: Injection molding; symbolic regression; PySR; Gaussian Process; residual learning; process-informed feature engineering; small-data modeling; surrogate modeling; process optimization; uncertainty quantification.

Nomenclature

Symbol / term	Meaning	Unit	Role
P	Holding pressure	bar	Primary process input
C	Counter-pressure	bar	Primary process input
T	Cycle time	s	Primary process input
V	Injection speed	mm/s	Primary process input
Def	Diameter deviation	mm (reported in μm for evaluation)	Response
SR	Symbolic regression	–	Interpretable baseline
GP	Gaussian Process	–	Residual correction and predictive variance
LOOCV	Leave-one-out cross-validation	–	Primary validation strategy
DE	Differential Evolution	–	Surrogate optimization algorithm
$\tilde{p}$	Min–max normalized holding pressure	–	Input of Eq. (1)
$\tilde{\varphi}_{\Delta P}$	Normalized pressure contrast, $(P - C)/P$	–	Input of Eq. (1)

Symbol / term	Meaning	Unit	Role
$\tilde{\varphi}_{CVT}$	Normalized interaction term, C·V·T	–	Input of Eq. (1)
$\hat{y}_u$	Normalized symbolic prediction	–	Output of Eq. (1)
$y_{\min}, y_{\text{range}}$	Minimum and range of the target used for scaling	mm	Scaling in Eq. (2)
$\hat{y}_{\text{SR}}, \hat{y}_{\text{hybrid}}$	Symbolic and hybrid predictions of Def	mm	Model outputs
$r = y - \hat{y}_{\text{SR}}$	Residual of the symbolic baseline	mm	Target of the GP

1. Introduction

Injection molding is a high-throughput process in which dimensional consistency has to be maintained despite strong coupling between pressure, filling behavior, thermal history, material response, and cycle timing. In production, the quality indicator of interest is rarely a single machine setting; it is the geometric outcome of how the settings interact. Diameter deviation is a good example. A change in holding pressure can shift the final dimension by different amounts depending on the counter-pressure and the preceding injection conditions, and cycle time can act through the evolving state of the material. Predicting such a response is therefore a process-level problem and not a one-factor regression task.

Classical process development still rests on Design of Experiments (DoE), response analysis, and careful adjustment of machine settings. DoE matters here because it decides how a limited experimental budget is spread over the design space. Comparative work on injection molding has shown that the choice of experimental design affects the performance of the machine-learning models trained afterwards, so data coverage is itself a modeling decision [1]. Reviews of machine learning in injection molding likewise point to the variety of quality-prediction tasks and to the lasting importance of data quality and feature representation [2].

Machine learning has widened the range of available predictors. Published studies use conventional learners, sensor-based models, neural networks, multitask architectures, feature-selection frameworks, and transfer learning to predict part quality under increasingly varied process conditions [3–12]. Models built on process settings and sensor signals have produced useful quality predictions from fairly compact experimental datasets [3,4], and recent attention-based and multitask architectures combine time-series process variables with non-time-series settings to predict several quality characteristics at once [7]. Other work addresses limited data directly through deep transfer learning and multi-feature fusion [8,12].

These advances come with a practical trade-off. A more flexible model may predict better, but it is harder to inspect and harder to explain to a process engineer. Feature-selection and explainability studies go some way toward closing that gap, yet a post hoc explanation is not the same as an explicit process relationship [9,17]. A model can flag a feature as important without showing how it enters the response, and different explainability methods can attribute the same black-box model in conflicting ways [17]. For process tuning, maintenance, and quality assurance this matters, because engineers usually want a prediction together with a short account of why it changes when a setting is moved.

The data regime of industrial experimentation points toward a different strategy. Machine time, material, part inspection, and scheduling all cap the number of experiments that can be run. Recent injection-molding studies have turned to transfer learning, small-sample deep learning, and mixed-effects modeling to get more out of limited observations [8,12,13]. Gaussian Process Regression (GPR) suits this setting well because it offers nonlinear regression in a probabilistic formulation with a predictive variance [13,22]. Bayesian optimization has also been used to find process settings on experimentally expensive response surfaces [14], and learning-based control studies show the growing role of surrogate and data-driven models in real-time process management [15,16].

Symbolic regression offers a complementary strength. Rather than fixing a parametric architecture in advance, it searches for an explicit mathematical expression. The idea goes back to genetic-programming-based equation discovery [18] and has become a practical scientific-computing tool through free-form law discovery, AI Feynman, and PySR [19–21]. The result is an analytical expression that can be inspected,

evaluated, differentiated, and placed directly inside an optimization routine. Its weakness is just as clear: a single expression may capture the dominant trend and still miss local nonlinear structure caused by measurement variability or by effects that the chosen inputs do not represent.

This motivates the hybrid formulation developed here. Instead of choosing between symbolic regression and a Gaussian Process, we give each a different job. Symbolic regression supplies the interpretable baseline and the dominant relationship. The GP is then trained only on the residual left by that baseline, so it acts as an adaptive correction layer and does not replace the equation with a second black box. The formulation is related to residual and delta-modeling approaches in surrogate modeling, where a low-complexity baseline is refined by a second model [23,24]. The difference in our case is that the baseline is itself a data-driven symbolic model and not a simulator or an externally prescribed mechanistic law.

A second methodological choice is central to the study: we treat feature engineering as part of symbolic discovery. Instead of giving PySR only the four raw process variables, we construct pressure ratios, normalized pressure contrasts, multiplicative pressure–speed–time terms, and rate-like combinations. These features do not define a unique physical law. They make process-relevant interactions directly available to the symbolic search, so the resulting equation should be read in the context of the engineered representation and not as an unrestricted expression chosen from raw variables alone.

With this background, the objective of the present work is to develop and evaluate an interpretable, small-data surrogate for diameter deviation in injection molding. The scope is deliberately local. We do not aim for a universal law of injection molding; we aim for a transparent predictive model over the experimentally explored operating region, and we ask whether a GP residual layer changes the resulting optimization landscape.

1.1 Research questions and contributions

- Can a process-informed engineered representation enable PySR to recover a compact symbolic relationship for diameter deviation from only 27 representative operating conditions?
- Does Gaussian Process residual learning provide measurable predictive refinement beyond the selected symbolic baseline while preserving an interpretable analytical core?
- How does the residual correction change the model-based optimum in the experimentally explored parameter domain, and what can the GP variance add to the interpretation of that recommendation?

The work makes four contributions. First, it defines a sequential symbolic-plus-GP surrogate architecture for small-data molding experiments. Second, it shows a feature-engineering strategy in which pressure ratios, normalized contrasts, and interaction terms are built on purpose to guide symbolic equation discovery. Third, it evaluates the standalone symbolic and hybrid models under LOOCV and reports both R^2 and RMSE in engineering units. Fourth, it links prediction to surrogate optimization and predictive-uncertainty analysis, while keeping model-predicted improvements clearly separate from experimentally demonstrated ones.

2. Related Work and Research Gap

2.1 DoE and data-efficient process modeling

The quality of a learned surrogate is bounded by the quality and geometry of its training data. In injection molding, DoE is therefore more than a preliminary statistical exercise, because it determines which process interactions can be identified. Heinisch et al. compared experimental-design strategies for neural-network modeling of injection molding and showed that the chosen design affects model performance [1]. For small-data modeling the consequence is direct: when only a few dozen representative conditions are available, the experimental design is effectively part of the modeling pipeline.

Response-surface and optimization methods remain relevant because they offer structured ways to study factor effects with few runs. Taguchi designs, Response Surface Methodology, and population-based

optimization have been combined for injection-molding settings [5]. We keep this process-development logic but replace the fixed response-surface form with an equation-discovery stage followed by a probabilistic residual model.

2.2 Machine learning for injection molding quality prediction

Machine-learning research on injection molding covers both conventional and deep models. The review by Selvaraj et al. documents applications in quality prediction, process modeling, and parameter optimization across several machine-learning families [2]. Párizs et al. demonstrated an Industry 4.0 quality-prediction workflow built on machine data [3], and Schulze Struchtrup et al. described a holistic part-quality prediction strategy that integrates several process variables [4]. Fernández et al. compared linear regression with neural networks for quality assessment and showed the value of nonlinear learners for process–quality relationships [6].

More recent work puts the emphasis on richer representations rather than simply larger networks. Kim et al. developed a deep feature-selection framework for quality prediction [9]. Qian et al. combined a DL-BiGRU architecture, multi-feature fusion, and deep transfer learning to handle small-sample quality prediction [8]. Muaz et al. introduced a multitask encoder–decoder model that merges time-series and non-time-series process parameters to predict several quality characteristics together [7], and Liparoti et al. used neural networks to predict flow length under different machine settings [10]. These studies give strong evidence that nonlinear interaction structure matters, but their solutions also demand more architecture and more data than an explicit equation does.

Our approach does not compete with these architectures on model capacity. It targets a different goal: a compact local surrogate, built from very few representative operating conditions, whose analytical form stays inspectable. The question is not whether a deep network can fit the response surface. It is whether a carefully structured equation with a local residual correction is enough for the intended operating domain.

2.3 Interpretability and process-informed representation

Interpretability has become more important in manufacturing because quality prediction is often just a step toward diagnosis or parameter adjustment. Muaz et al. showed that different explainability methods can give conflicting feature attributions in injection molding, which affects root-cause analysis [17]. This separates two things: explaining a black-box function after training, and building a model whose mathematical structure is explicit by construction.

Feature engineering sits between raw data and symbolic discovery. In sensor-based molding studies, pressure descriptors and other stage-specific summaries can encode process mechanisms that are hard to capture from raw measurements [3,4]. Here the engineered features have a narrower purpose. They expose selected interaction structures to the symbolic search so that PySR can combine them directly. The contribution is not simply more features but a representation designed to make particular process interactions discoverable.

2.4 Symbolic regression and physics-aware learning

Symbolic regression searches over mathematical structures rather than tuning parameters inside a fixed architecture. Genetic programming set out the general framework for this kind of equation discovery [18], and Schmidt and Lipson later showed that compact physical relationships can be recovered automatically from experimental data [19]. Al Feynman introduced a physics-inspired symbolic-regression strategy [20], and PySR with `SymbolicRegression.jl` provides practical tools for large-scale interpretable equation search [21].

We use the term physics-informed with care. The study does not impose a governing differential equation, a conservation law, a mechanistic simulator, or a hard physical constraint. Process-informed feature construction only biases the symbolic search toward combinations that have a clear engineering meaning. The model is therefore a process-informed symbolic-regression model and not a physics-constrained governing-equation model. Work on physics-informed Gaussian Processes shows that domain knowledge can enter through priors, kernels, mean functions, or data structure [25], and those mechanisms should not be confused with the feature-engineering strategy used here.

The distinction bears on the claimed novelty. What we claim is the sequential coordination of three components in a small-data molding setting: process-informed feature engineering, symbolic equation discovery, and GP residual learning. We do not claim a new governing physical model.

2.5 Gaussian Processes and hybrid surrogate modeling

Gaussian Processes define a distribution over functions through a mean function and a covariance kernel, which gives a flexible nonparametric regression framework [22]. For small datasets this is attractive: predictions come with a variance, and the kernel can represent smooth nonlinear relationships. In injection molding, Luo et al. used mixed-effects GP modeling to account for structured variability [13]. Jung et al. applied Bayesian optimization and constrained inverse design to process optimization [14], Ren et al. developed a learning-based model predictive control strategy for injection-speed tracking [15], and Cheng et al. demonstrated data-driven online prediction and control of molding quality [16].

Hybrid surrogates are common in engineering optimization because a low-order baseline can capture the broad trend while a second model learns the remaining discrepancy [23,24]. In our framework the symbolic model plays the baseline role. Let $r(x)$ be the residual between the measured response and the symbolic prediction; the GP learns $r(x)$ from the original four process variables. This lightens the GP's task, since it models only what the analytical baseline leaves unexplained and does not have to reconstruct the full response surface.

2.6 Research gap

The literature therefore offers strong and increasingly sophisticated solutions for injection-molding quality prediction, but most are organized around one dominant paradigm: statistical response modeling, black-box machine learning, deep representation learning, symbolic regression, Gaussian Processes, or online control. The combination studied here is narrower. We ask whether an interpretable symbolic equation can serve as the main surrogate, and whether a GP trained only on its residuals can add local probabilistic refinement when only 27 conditions are available. A second gap concerns feature construction. Instead of using engineered variables only to raise accuracy, our workflow treats them as a deliberate way of designing the search space for equation discovery.

3. Methodology

3.1 Overview of the modeling workflow

The workflow is sequential and not end-to-end. The experimental data are first consolidated into a representative set of operating conditions. The four retained process variables are then transformed into a twelve-variable process-informed representation, and symbolic regression identifies a compact analytical baseline from it. A Gaussian Process models the baseline's residuals using the original four variables. Finally, the symbolic and hybrid surrogates are evaluated by LOOCV and, after evaluation, used as objective functions for Differential Evolution optimization. Figure 1 summarizes the workflow.

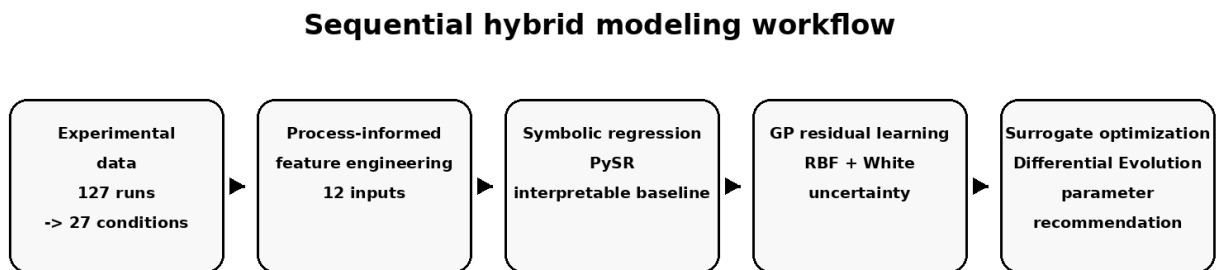


Figure 1. Sequential hybrid workflow used in this study. The symbolic model captures the dominant relationship; the GP is restricted to residual correction and uncertainty estimation.

3.2 Experimental design and data consolidation

The data come from two experimental campaigns. The first gave an initial characterization of the process, but its variables did not provide a sufficiently informative final representation for our modeling objective. The experimental strategy was therefore revised for the second campaign: two variables from the original configuration were held at nominal levels while alternative variables, more directly related to the target response, were explored. The two campaigns are thus complementary rather than simply additive.

In total, 127 experimental runs were carried out across the two campaigns. Several operating conditions occurred in both, so repeated observations were present. We consolidated these overlapping conditions by averaging the repeated measurements, which gave 27 representative operating conditions. The reduction comes from aggregating repeated or overlapping conditions, not from discarding observations. It provides one representative response per operating condition and reduces within-condition measurement variability before symbolic model discovery.

The four variables retained for the final model were holding pressure P_{maintien} (P), counter-pressure C_{pression} (C), cycle time T_{cycle} (T), and injection speed $V_{\text{injection}}$ (V). The response is the diameter deviation Def , recorded in millimeters and converted to micrometers when reporting prediction errors. Optimization was restricted to the experimentally explored ranges: P in [280, 320] bar, C in [50, 70] bar, T in [30, 60] s, and V in [35, 45].

3.3 Process-informed feature engineering

The four raw variables describe the process compactly, but the symbolic search space stays highly combinatorial unless interactions are exposed explicitly. We therefore built twelve inputs before applying PySR. Table 1 lists the complete representation used in the symbolic stage.

Feature	Definition	Interpretive role
P_{maintien}	P	Raw holding-pressure effect
C_{pression}	C	Raw counter-pressure effect
T_{cycle}	T	Cycle-history contribution
$V_{\text{injection}}$	V	Flow-rate contribution
$P_{\text{over } C}$	P / C	Relative pressure balance
$P_{\text{minus } C \text{ over } P}$	$(P - C) / P$	Normalized pressure contrast
$V_i \times T$	$V \times T$	Kinematic-time interaction
$P_{\text{over } C} \times V_i T$	$(P / C) \times V \times T$	Coupled ratio-interaction
$PmC \ P \times V_i T$	$((P - C) / P) \times V \times T$	Pressure-contrast interaction
$V_i \text{ over } T$	V / T	Rate-like dynamic term
$P \times V_i T$	$P \times V \times T$	Pressure-speed-time coupling
$C \times V_i T$	$C \times V \times T$	Counter-pressure-speed-time coupling

Table 1. Raw and process-informed engineered features supplied to the symbolic-regression stage.

Three design principles guided the engineered representation. First, pressure ratios and normalized contrasts expose the relative balance between holding pressure and counter-pressure. Second, multiplicative terms make interactions between pressure, injection speed, and cycle time directly available to PySR. Third, V/T provides a rate-like descriptor that complements raw speed and cycle time. We do not claim that any engineered term is a complete physical law. The features encode process hypotheses in a form that an equation-discovery algorithm can test and combine.

One statistical consequence deserves attention. The twelve engineered inputs are not twelve independent measurements; they are deterministic transformations of four underlying process variables. Their value is representational, because they change which algebraic structures the symbolic search can express compactly. With only 27 representative conditions, this choice is a real part of the model design and should be reported as such, not treated as incidental preprocessing.

Within each validation fold, the engineered inputs were min–max scaled and the target was scaled to the unit interval, using the training observations of that fold only. Predictions were then mapped back to physical units. The selected equation is thus defined in normalized feature space and not directly in raw engineering units, so the scaling constants must be kept together with the equation when it is used.

3.4 Symbolic regression and equation selection

Symbolic regression was run with PySR [21]. The operator set included standard arithmetic and nonlinear transformations, and parsimony pressure favored compact expressions over needlessly complex ones. Candidate expressions generated during development were compared by leave-one-out cross-validation. We also explored AI Feynman [20], but it gave weaker predictive behavior under the same validation criterion, so it was not retained.

The final expression is written in normalized feature space. Let $\tilde{P}$ denote the normalized holding pressure, $\tilde{\varphi}_{\Delta P}$ the normalized pressure contrast $\varphi_{\Delta P} = (P - C)/P$ (feature P_minus_C_over_P in Table 1), and $\tilde{\varphi}_{CVT}$ the normalized interaction $\varphi_{CVT} = C \cdot V \cdot T$ (feature C_x_ViT). Tildes indicate min-max normalization. The selected equation is:

$$\hat{y}_u = 0.600834\tilde{P}[(0.36742553 - \tilde{\varphi}_{CVT}) \cdot 8.060018 \cdot (\tilde{P} - 1.0755086) + \tilde{\varphi}_{\Delta P}] - 0.24252947\sqrt{|\tilde{\varphi}_{\Delta P}|} + 0.33141157 \quad (1)$$

Predictions in physical units are obtained by rescaling:

$$\hat{y}_{SR} = y_{\min} + y_{\text{range}} \cdot \hat{y}_u \quad (2)$$

Here $y_{\min}$ and $y_{\text{range}} = y_{\max} - y_{\min}$ are the minimum and the range of the target used for scaling in Eq. (2). Equation (1) is compact in the engineered representation, but its coefficients alone are not enough for direct use in raw process units. The feature definitions and the normalization parameters are part of the model: an expression that depends on normalized variables is fully specified only when the preprocessing is supplied with it.

The equation uses only three of the twelve inputs: the raw holding pressure and two engineered features, $\varphi_{\Delta P}$ and φ_{CVT} . Cycle time and injection speed therefore enter only through the product $C \cdot V \cdot T$. The search did not need every engineered term to reach a compact response relation. The expression should still not be read as a mechanistic governing law; it is a data-derived local surrogate found within the explored operating domain.

3.5 Gaussian Process residual learning

Once the symbolic expression was selected, its predictions defined the residual target. For each operating condition x_i , the residual is $r_i = y_i - \hat{y}_{SR}(x_i)$. A Gaussian Process was trained on these residuals using the original four process variables, $x = [P, C, T, V]$. The hybrid predictor is:

$$\hat{y}_{\text{hybrid}}(x) = \hat{y}_{SR}(x) + \hat{y}_{GP}(x) \quad (3)$$

Here $\hat{y}_{GP}(x)$ is the GP posterior mean for the residual. The split gives the two components distinct responsibilities. The symbolic expression carries the dominant, globally interpretable trend, and the GP handles whatever structured discrepancy remains. Because the GP does not have to reconstruct the full response surface, its target is typically smaller in amplitude and more localized than the original response.

We evaluated four kernel families: Matérn 5/2 + White, RBF + White, Rational Quadratic + White, and linear-augmented variants. The RBF + White kernel gave the best LOOCV performance and was retained for the final hybrid model. Hyperparameters were optimized by likelihood-based fitting with multiple random restarts, following the standard GPR formulation of Rasmussen and Williams [22].

3.6 Validation protocol

We used leave-one-out cross-validation because the final modeling set contains only 27 representative operating conditions. Each condition is left out once, the model is fitted to the remaining 26, and the left-out condition is predicted. The pooled predictions are then used to compute R^2 and RMSE:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

and the root-mean-square error is

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

where y_i is the measured response for condition i , $\hat{y}_i$ the corresponding prediction (under LOOCV, from the model fitted without condition i), $\bar{y}$ the mean measured response, and $n = 27$ the number of conditions. RMSE is reported in micrometers because that unit is directly meaningful for dimensional-quality assessment. We also report full-data performance for the standalone symbolic equation. The two sets of numbers answer different questions: the full-data fit describes how well the selected equation reproduces all representative conditions, while LOOCV is the primary out-of-sample assessment.

One limitation of this protocol should be stated plainly. Feature scaling was refitted inside each fold, but the symbolic expression and the GP kernel were selected by comparing LOOCV scores on the same 27 conditions. The reported LOOCV values should therefore be read as cross-validated estimates conditional on the selected model structure, and not as fully independent test performance. A fully nested analysis, in which expression search and kernel selection are repeated inside every fold, would remove this dependence.

3.7 Surrogate-based process optimization

After model evaluation, the symbolic and hybrid surrogates were embedded as objective functions in a Differential Evolution (DE) routine, which minimizes the predicted diameter deviation over the experimentally explored domain:

$$\min_{x \in \Omega} \hat{y}_{\text{model}}(x) \quad (6)$$

where Ω is the box defined by $280 \leq P \leq 320$ bar, $50 \leq C \leq 70$ bar, $30 \leq T \leq 60$ s, and $35 \leq V \leq 45$. For the hybrid surrogate the GP provides both a predictive mean and a predictive variance, but in this study the variance was used only to interpret candidate solutions; it was not included as a penalty or acquisition term in the objective. The method is therefore not uncertainty-aware in the optimization sense: the objective is a predicted response, and the uncertainty is examined alongside the candidate solution to judge how far it can be trusted. A Bayesian-optimization formulation with an acquisition function could use the uncertainty directly, but that lies outside the scope of this study.

4. Results and Discussion

4.1 Predictive performance of the symbolic model

The selected symbolic model reached $R^2 = 0.8665$ and $RMSE = 9.37 \mu m$ on the full consolidated dataset. Under LOOCV it reached $R^2 = 0.8423$ and $RMSE = 10.19 \mu m$. The gap between the two is expected, since LOOCV predicts every observation from a model that did not see it during fitting. The results show that the symbolic equation captures a substantial part of the response structure within the explored domain.

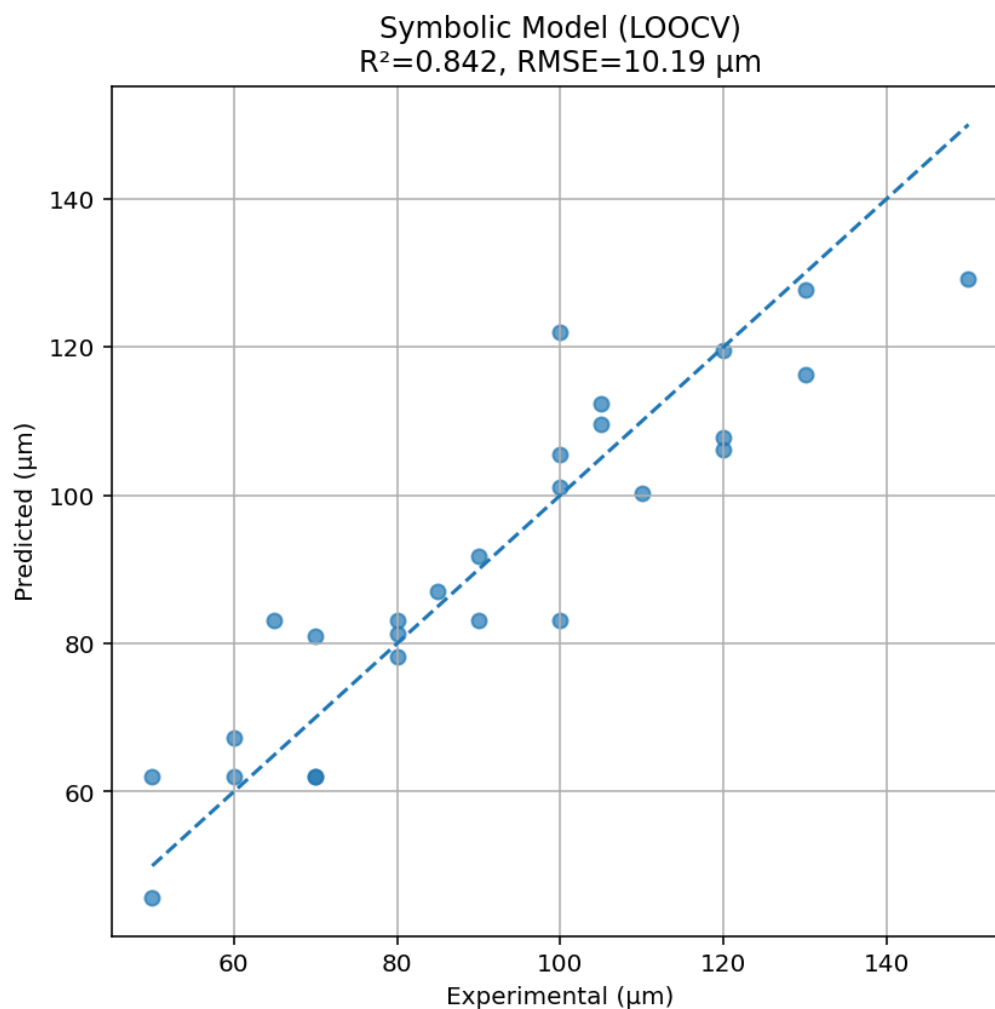


Figure 2. LOOCV parity plot for the standalone symbolic regression model ($R^2 = 0.8423$, $RMSE = 10.19 \mu m$).

The predictions scatter around the one-to-one line across the observed response range, with larger departures at some of the lowest and highest response values. In a dataset this small, single observations can strongly influence the aggregate metrics, so the plot should be read together with the LOOCV protocol and not taken as evidence of performance outside the explored region.

4.2 Comparison of symbolic and hybrid predictive performance

Model	Evaluation	R^2	RMSE (μm)	Interpretation
Symbolic regression	Full dataset	0.8665	9.37	Descriptive fit on all 27 conditions

Model	Evaluation	R ²	RMSE (μm)	Interpretation
Symbolic regression	LOOCV	0.8423	10.19	Primary out-of-sample result
Hybrid SR + GP	LOOCV	0.8681	9.32	Best cross-validated result

Table 2. Predictive performance of the selected symbolic and hybrid surrogates.

4.3 Hybrid GP residual learning

The RBF + White GP residual model raised the LOOCV R² from 0.8423 to 0.8681 and lowered the RMSE from 10.19 to 9.32 μm, a change of 0.0258 in R² and 0.87 μm in RMSE. The gain is measurable but modest. With only 27 representative conditions, it should not be read as proof that GP residual correction is generally superior. What it shows is that the residual layer improved the cross-validated reconstruction for this dataset.

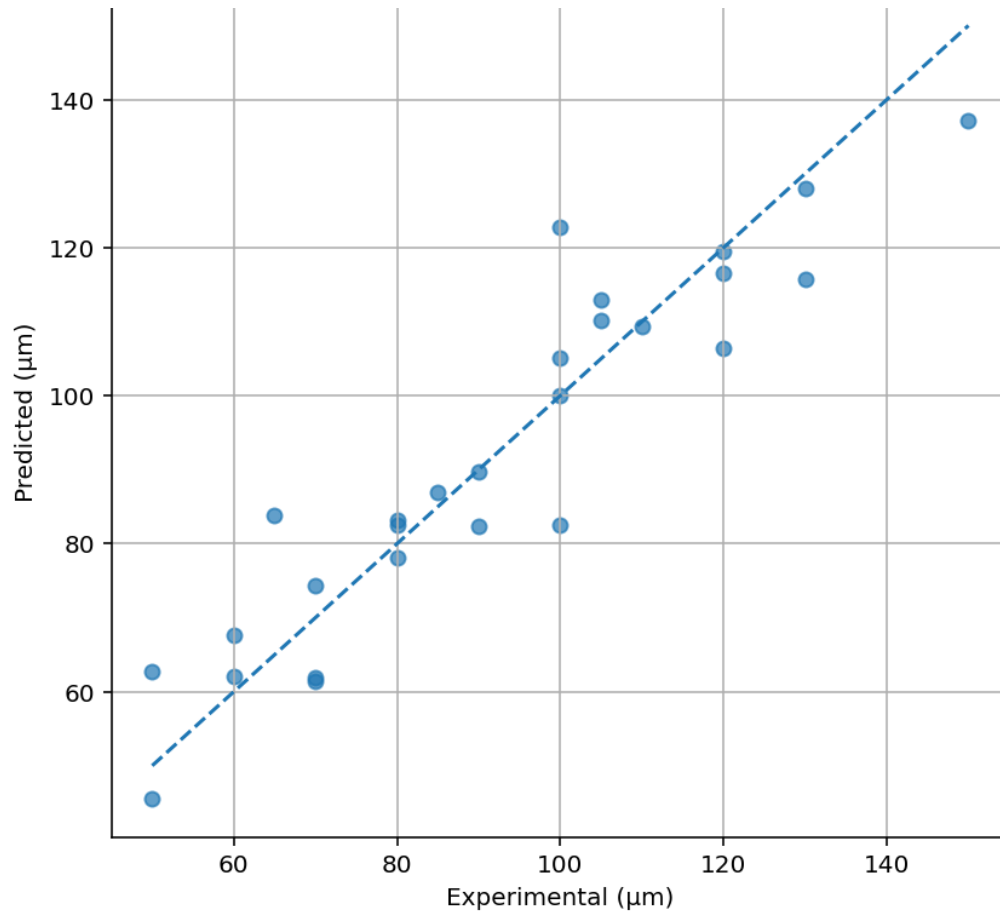


Figure 3. LOOCV parity plot for the hybrid symbolic regression + GP model (R² = 0.8681, RMSE = 9.32 μm).

The hybrid parity plot (Figure 3) remains broadly aligned with the experimental measurements. As in Figure 2, a few individual conditions account for the larger remaining errors.

4.4 Residual behavior and interpretation of the hybrid architecture

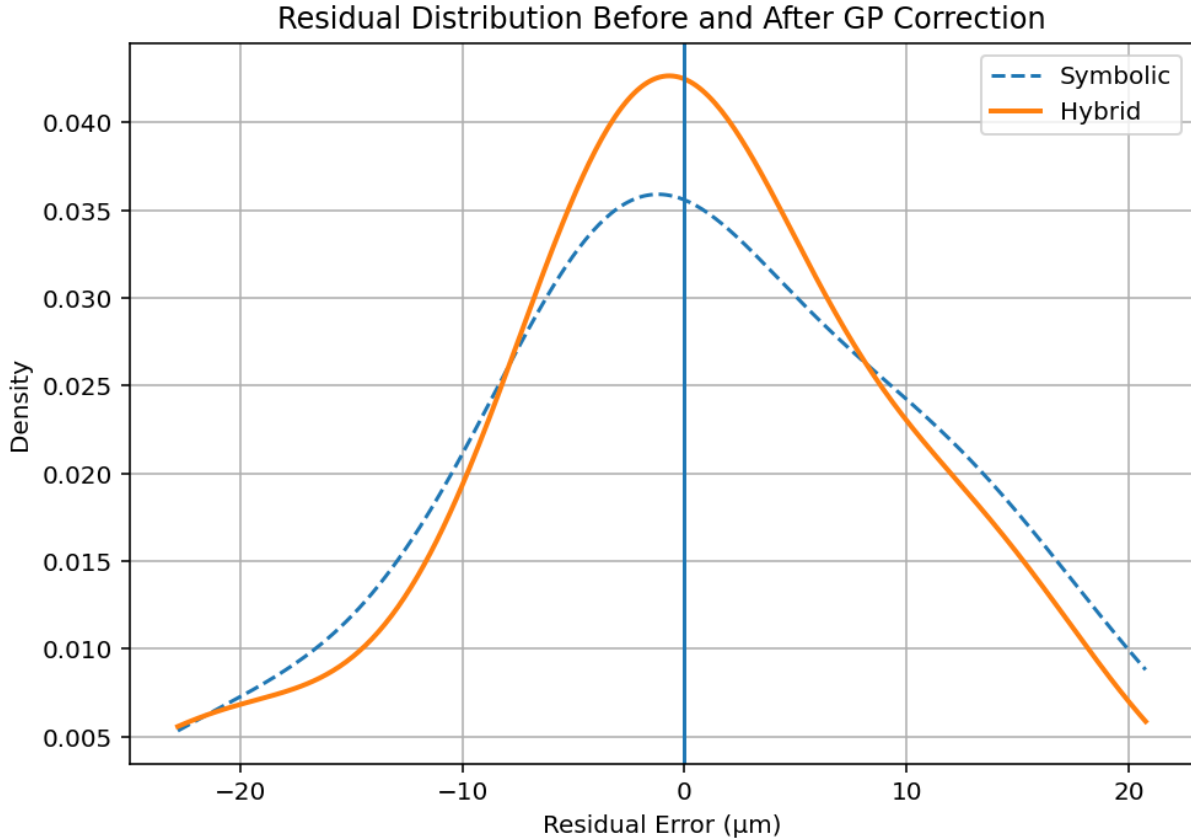


Figure 4. Distribution of residuals before (symbolic) and after (hybrid) Gaussian Process correction.

The residual distributions in Figure 4 illustrate the intended role of the GP. The symbolic equation absorbs the dominant trend and leaves an error component that can be modeled separately. After GP correction, the residuals are more concentrated around zero. The hybrid architecture is easier to interpret than a more complex symbolic search would be, because the correction has a clearly defined target: the residual error of the selected baseline.

A useful residual decomposition should not be mistaken for a causal explanation. The GP residual is not a new physical mechanism; it represents structure that the symbolic baseline does not explain under the chosen representation. Some of it may come from process variables that were not modeled, from metrology noise, or from local nonlinearities, and without further experiments or direct process measurements these sources cannot be separated.

The feature-engineering stage also shows up here. The final expression relies on two engineered quantities, $\varphi_{\Delta P}$ and φ_{CVT} , in addition to the raw holding pressure, so the designed representation allowed the search to express interaction structure compactly. The study does not include an ablation in which the same PySR search is repeated on raw variables only. We can therefore show that engineered features were used and that the selected equation depends on them, but not that any feature group on its own caused the observed performance.

4.5 Optimization results

The symbolic and hybrid surrogates were optimized separately over the same explored ranges. The symbolic surrogate gave a minimum predicted diameter deviation of about 33.25 μm , and the hybrid surrogate about 23.73 μm . The difference of 9.52 μm is about 28.6% of the symbolic-model optimum. These values are surrogate predictions; the process was not rerun at the optimized settings. Both also lie below the smallest deviation among the 27 conditions (about 50 μm ; Figures 2 and 3), and the gap between them is about the size

of the LOOCV RMSE (9.32–10.19 μm). They should therefore be read as indications of direction and not as expected outcomes.

Surrogate	Predicted minimum	Difference vs. symbolic baseline	Experimental confirmation
Symbolic	33.25 μm	Baseline	Not performed
Hybrid SR + GP	23.73 μm	–9.52 μm (predicted change)	Not performed

Table 3. Model-based optimization results. The differences are surrogate predictions and were not validated experimentally.

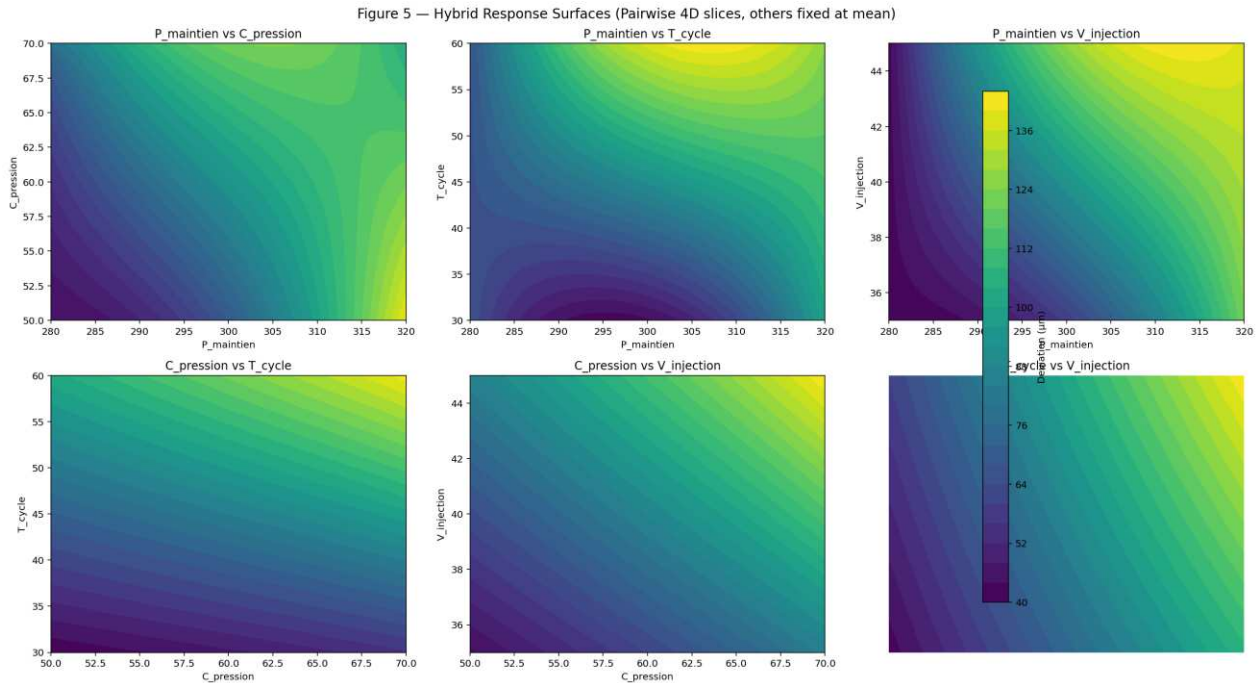


Figure 5. Pairwise response surfaces of the hybrid surrogate over the optimization domain. Variables not shown on the axes are held at their mean values.

Figure 5 gives a qualitative picture of the predicted response surface. Each panel is a conditional slice of a four-dimensional surrogate with the remaining variables held at their means, so the panels are not independent two-variable experiments. The figure shows the geometry of the model and not separately measured response surfaces.

The main point of the optimization is not that the hybrid model has found the true process optimum. It is that adding a residual correction can noticeably change the shape of the predicted objective landscape, and with it the parameter recommendation the optimizer returns. That is exactly why the recommendation needs experimental validation: a lower surrogate value is useful only if the model remains credible in that region of the design space.

4.6 Predictive uncertainty

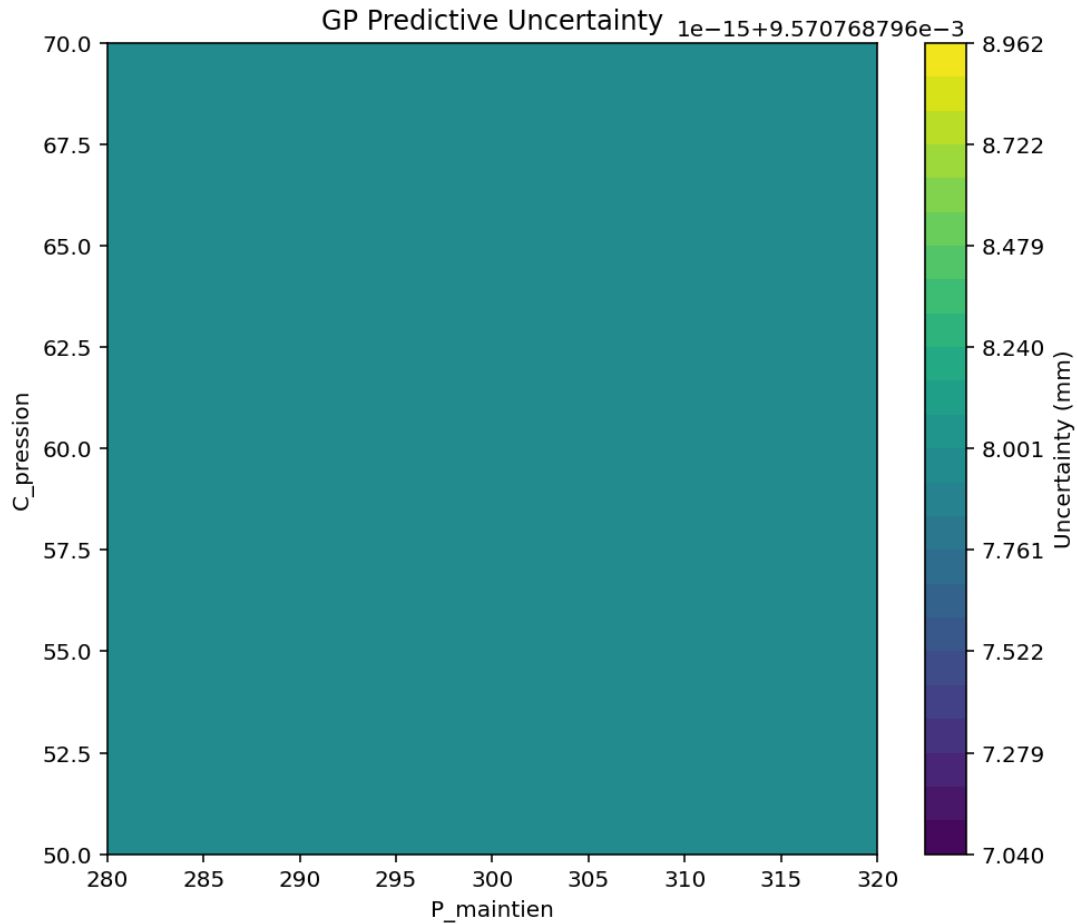


Figure 6. GP predictive uncertainty over a representative two-dimensional slice (P , C) of the optimization domain. The map is essentially uniform and should be interpreted cautiously.

A main advantage of the GP layer is that it gives a predictive variance alongside the mean residual correction. This lets the analyst tell a low predicted response apart from a low predicted response that comes with high uncertainty. That matters in surrogate optimization because model behavior can differ between densely sampled conditions and the sparsely sampled edges of the domain.

In this study the uncertainty map is essentially uniform across the plotted slice (Figure 6). That should not be taken to mean that uncertainty is uniform over the whole four-dimensional process space. With 27 representative conditions and four inputs, the residual GP has little spatial information from which to estimate how uncertainty varies. The flat map may reflect the actual residual structure, the behavior of the kernel, or simply the sparse sampling, and we have no independent calibration experiment that could tell these explanations apart.

Predictive variance is therefore presented here as a decision-support quantity and not as a validated probability of error. A stronger claim would need additional test points, interval-coverage analysis, or repeated holdout experiments.

4.7 Overall interpretation of the results

First, the symbolic model gives a useful deterministic representation despite the small number of representative operating conditions. The LOOCV result shows that the selected equation keeps substantial predictive capability when every condition is treated as unseen. The evidence is strongest for interpolation within the studied operating region and does not establish transfer to another machine, polymer, mold, or geometry.

Second, GP residual learning refines the baseline in a disciplined way. The hybrid model neither discards the symbolic relation nor buries it inside a neural architecture: the equation stays visible and the GP has a specific correction task. That makes the hybrid easier to communicate and potentially easier to maintain as new observations arrive.

Third, the predictive gain needs context. An increase of 0.0258 in R^2 and a reduction of 0.87 μm in RMSE are useful but not dramatic. The main benefit of the GP may therefore lie in the change it makes to the local surrogate landscape and in the availability of a predictive variance, more than in a large global improvement in the metrics.

Fourth, feature engineering is a central methodological element, but this study does not isolate its effect with a controlled ablation. The strongest defensible statement is that the final symbolic equation was obtained from a process-informed representation and that it uses two engineered features, $\varphi_{\Delta P}$ and φ_{CVT} , in addition to the raw holding pressure. Quantifying the contribution of each feature group is a direction for future work.

Finally, the optimization result shows the framework working as a decision-support pipeline, but it does not show the manufacturing performance of the recommended settings. Experimental confirmation at the candidate operating point is still needed before the predicted reduction in diameter deviation can be counted as a realized improvement.

5. Limitations and Reproducibility Considerations

Sample size and scope. The final model rests on 27 representative operating conditions. That suits the small-data motivation of the study but limits the statistical resolution of complex effects. The model should be regarded as a local surrogate over the investigated domain and not as a universal injection-molding law.

External validation. Performance was assessed by LOOCV within the available dataset, and no independent external campaign is reported. Validation on a different batch, machine configuration, material condition, or mold geometry would be needed to establish transferability.

Experimental validation of the optimum. The predicted minima of 33.25 μm and 23.73 μm were obtained computationally, and the optimized settings were not retested experimentally. The 9.52 μm difference is therefore a model-based prediction, not an experimentally demonstrated quality improvement. Both minima also fall below the smallest measured deviation, so they lie outside the range of responses the models were trained on.

Nested model selection. LOOCV is strictly out-of-sample for the whole modeling pipeline only if expression selection and hyperparameter tuning are carried out inside the training portion of each fold. Here scaling was refitted within each fold, while the symbolic expression and the kernel were chosen by comparing LOOCV scores on all 27 conditions. The reported metrics should therefore be read as conditional on the selected model structure, and a fully nested analysis would give a stricter estimate.

Equation reproducibility. The selected PySR expression is defined in normalized feature space. Using it in raw process units requires the exact min–max scaling constants together with the feature definitions given in Table 1.

Measurement and process metadata. The conclusions depend on the experimental context. Reproducing the experiments and judging transferability require the machine configuration, the polymer, the mold and part geometry, the measurement method, the number of repeated measurements per condition, and the exact definition of diameter deviation. These details are not needed to follow the modeling logic, but they are needed to reproduce the experimental work.

Uncertainty calibration. The GP provides a predictive variance, but its interval coverage has not been checked against independent observations. The uncertainty maps should be read as model-derived indicators and not as validated confidence statements.

Feature ablation. The results establish that process-informed features were used but do not quantify the marginal value of each feature family. An ablation comparing raw-only, ratio-only, interaction-only, and full representations would strengthen the interpretation of the feature-engineering contribution.

6. Conclusion

This study developed an interpretable hybrid surrogate for diameter deviation in injection molding under a small-data experimental regime. Two DoE campaigns contributed 127 runs, which were consolidated into 27 representative operating conditions. Four process variables were retained, and their representation was expanded through process-informed ratios, contrasts, and interaction terms before symbolic regression.

The selected PySR equation gave an explicit analytical baseline with full-data $R^2 = 0.8665$ and $RMSE = 9.37 \mu m$, and LOOCV values of $R^2 = 0.8423$ and $RMSE = 10.19 \mu m$. A Gaussian Process residual model with an RBF + White kernel lowered the LOOCV error to $9.32 \mu m$ and raised R^2 to 0.8681 . The result supports GP residual learning as a refinement layer, although the global gain is modest and should be read that way.

When the two surrogates were embedded in Differential Evolution, the predicted minimum moved from about $33.25 \mu m$ for the symbolic baseline to about $23.73 \mu m$ for the hybrid surrogate. The shift shows that residual correction can noticeably influence the optimization landscape. It remains a model-based result and not an experimentally confirmed process improvement.

The broader methodological point is that interpretability and flexible correction need not be mutually exclusive. A symbolic model can provide the dominant analytical structure while a GP learns the residual behavior and supplies a predictive variance. For small-data industrial problems, this division of roles is a practical alternative to choosing between a rigid analytical model and a fully opaque surrogate.

The next step is rigorous validation rather than added complexity: repeat the modeling with fully nested cross-validation, archive the exact preprocessing and PySR configuration, test the predicted optimum experimentally, and evaluate the workflow across other materials, geometries, machines, and process windows. These steps will show how far the present local surrogate can be generalized.

Declarations

Funding

Not applicable. This research received no external funding.

Conflicts of interest/Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Amine ABOUTAIB: methodology, coding/modeling, analysis, and manuscript writing. Hanaa OUAOMAR: methodology and analysis. Souad Touairi: manuscript writing and analysis. Mourad KADIRI: manuscript writing and analysis.

References

- [1] Heinisch, J.; Lockner, Y. P.; Hopmann, C. Comparison of design of experiment methods for modeling injection molding experiments using artificial neural networks. *Journal of Manufacturing Processes* 2021, 61, 357–368. <https://doi.org/10.1016/j.jmapro.2020.11.011>.

- [2] Selvaraj, S. K.; Raj, A.; Mahadevan, R. R.; Chadha, U.; Paramasivam, V. A review on machine learning models in injection molding machines. *Advances in Materials Science and Engineering* 2022, 2022, 1949061. <https://doi.org/10.1155/2022/1949061>.
- [3] Párizs, R. D.; Török, D.; Ageyeva, T.; Kovács, J. G. Machine learning in injection molding: An Industry 4.0 method of quality prediction. *Sensors* 2022, 22(7), 2704. <https://doi.org/10.3390/s22072704>.
- [4] Schulze Struchtrup, A.; Kvaktun, D.; Schiffrers, R. A holistic approach to part quality prediction in injection molding based on machine learning. In: *Advances in Polymer Processing 2020*; Springer, 2020; pp. 137–149. https://doi.org/10.1007/978-3-662-60809-8_12.
- [5] Chen, W.-C.; Nguyen, M.-H.; Chiu, W.-H.; Chen, T.-N.; Tai, P.-H. Optimization of the plastic injection molding process using the Taguchi method, RSM, and hybrid GA-PSO. *The International Journal of Advanced Manufacturing Technology* 2016, 83, 1873–1886. <https://doi.org/10.1007/s00170-015-7683-0>.
- [6] Fernández, A.; Clavería, I.; Pina, C.; Elduque, D. Predictive methodology for quality assessment in injection molding comparing linear regression and neural networks. *Polymers* 2023, 15(19), 3915. <https://doi.org/10.3390/polym15193915>.
- [7] Muaz, M.; Yu, H.; Sung, W. L.; Liu, C.; Drescher, B. A multitask encoder-decoder model for quality prediction in injection moulding. *Journal of Manufacturing Processes* 2023, 103, 238–247. <https://doi.org/10.1016/j.jmapro.2023.08.030>.
- [8] Qian, Q.; Yu, J.; Zhan, H.; Wang, R. A novel DL-BiGRU multi-feature fusion and deep transfer learning based modeling approach for quality prediction of injection molded products using small-sample datasets. *Journal of Manufacturing Processes* 2024, 120, 272–285. <https://doi.org/10.1016/j.jmapro.2024.04.030>.
- [9] Kim, I.; Na, J.; Yun, J. P.; Lee, S. Deep feature selection framework for quality prediction in injection molding process. *IEEE Transactions on Industrial Informatics* 2024, 20(1), 503–512. <https://doi.org/10.1109/TII.2023.3268421>.
- [10] Liparoti, S.; Di Pasquale, V.; Volpe, V.; Adinolfi, F.; Pantani, R. Artificial neural network-based flow length prediction system under different machine settings for micro-injection moulding. *The International Journal of Advanced Manufacturing Technology* 2025, 136, 1281–1295. <https://doi.org/10.1007/s00170-024-14928-3>.
- [11] Ko, J. U.; Lee, J.; Kim, T.; Kim, Y. C.; Youn, B. D. Deep-learning-based fault detection and recipe optimization for a plastic injection molding process under the class-imbalance problem. *Journal of Computational Design and Engineering* 2023, 10(2), 694–710. <https://doi.org/10.1093/jcde/qwad011>.
- [12] Lockner, Y. P.; Hopmann, C.; Zhao, W. Transfer learning with artificial neural networks between injection molding processes and different polymer materials. *Journal of Manufacturing Processes* 2022, 73, 395–408. <https://doi.org/10.1016/j.jmapro.2021.11.014>.
- [13] Luo, L.; Yao, Y.; Gao, F.; Zhao, C. Mixed-effects Gaussian process modeling approach with application in injection molding processes. *Journal of Process Control* 2018, 62, 37–43. <https://doi.org/10.1016/j.jprocont.2017.12.003>.
- [14] Jung, J.; Park, K.; Cho, B.; Park, J.; Ryu, S. Optimization of injection molding process using multi-objective Bayesian optimization and constrained generative inverse design networks. *Journal of Intelligent Manufacturing* 2023, 34(8), 3623–3636. <https://doi.org/10.1007/s10845-022-02018-8>.
- [15] Ren, Z.; Cai, J.; Zhang, B.; et al. A learning-based model predictive control scheme for injection speed tracking in injection molding process. *Complex & Intelligent Systems* 2024, 10, 7845–7861. <https://doi.org/10.1007/s40747-024-01588-9>.
- [16] Cheng, Y.; Zhan, H.; Yu, J.; Wang, R. Data-driven online prediction and control method for injection molding product quality. *Journal of Manufacturing Processes* 2025, 145(2), 252–273. <https://doi.org/10.1016/j.jmapro.2025.04.054>.
- [17] Muaz, M.; Sajid, S.; Schulze, T.; Liu, C.; Klasen, N.; Drescher, B. Explainable AI for correct root cause analysis of product quality in injection moulding. *Journal of Manufacturing Processes* 2025, 145, 371–380. <https://doi.org/10.1016/j.jmapro.2025.03.114>.
- [18] Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*; MIT Press: Cambridge, MA, USA, 1992.
- [19] Schmidt, M.; Lipson, H. Distilling free-form natural laws from experimental data. *Science* 2009, 324(5923), 81–85. <https://doi.org/10.1126/science.1165893>.
- [20] Udrescu, S.-M.; Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances* 2020, 6(16), eaay2631. <https://doi.org/10.1126/sciadv.aay2631>.

- [21] Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. arXiv:2305.01582, 2023.
- [22] Rasmussen, C. E.; Williams, C. K. I. Gaussian Processes for Machine Learning; MIT Press: Cambridge, MA, USA, 2006.
- [23] Queipo, N. V.; Haftka, R. T.; Shyy, W.; Goel, T.; Vaidyanathan, R.; Tucker, P. K. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences* 2005, 41(1), 1–28. <https://doi.org/10.1016/j.paerosci.2005.02.001>.
- [24] Forrester, A. I. J.; Sobester, A.; Keane, A. J. *Engineering Design via Surrogate Modelling: A Practical Guide*; Wiley, 2008. <https://doi.org/10.1002/9780470770801>.
- [25] Cross, E. J.; Rogers, T. J.; Pitchforth, D. J.; Gibson, S. J.; Zhang, S.; Jones, M. R. A spectrum of physics-informed Gaussian processes for regression in engineering. *Data-Centric Engineering* 2024, 5, e8. <https://doi.org/10.1017/dce.2024.2>.