Key Engineering Materials
ISSN: 1662-9795, Vol. 1050, pp 103-111
doi:10.4028/p-wWX8Z3
© 2026 The Author(s). Published by Trans Tech Publications Ltd, Switzerland.
Submitted: 2026-01-19
Revised: 2026-02-25
Accepted: 2026-02-25
Online: 2026-04-16

Residual Learning-Based Synthetic Data for Hybrid Metamodeling in the Stamping of an Automotive Door Panel

Laura Muñiz[1,a*], Javier Trinidad[1,b], Julian Ramirez de Okariz[2,3,c], Eduardo Garcia[4,d] and Lander Galdos[1,e]

[1]Faculty of Engineering, Mechanics and Industrial Production, Mondragon Unibertsitatea, Loramendi 4, 20500 Mondragon, Spain

[2]Fagor Arrasate Koop. Elk., San Andres 20, 20500 Mondragon, Spain

[3]Koniker Koop. Elk. Zentro Teknologikoa, San Andres Auzoa 20, 20500 Mondragon, Spain

[4]FORD Motor Company, 46440 Almussafes, Spain

[a]lmuniz@mondragon.edu, [b]jtrinidad@mondragon.edu, [c]j.rokariz@fagorarrasate.com, [d]egarci75@ford.com [e]lgaldos@mondragon.edu (*corresponding author)

Keywords:. stamping, hybrid metamodel, synthetic data, discrepancy model.

Abstract. Hybrid twins and residual-learning strategies are increasingly used to reconcile the broad coverage of physics-based simulations with the fidelity of production measurements. In industrial stamping, however, truly matched simulation-experiment cases are scarce, while large-scale monitoring data are clustered around a narrow operating window. This work proposes an industrially practical hybrid metamodel in which the discrepancy (ignorance) model is trained exclusively from approximate residuals computed at in-domain production inputs, considering a surrogate of a thermal-enabled AutoForm Sigma Design of Experiments (DoE). To prevent uncontrolled extrapolation, learning and evaluation are restricted to an explicitly validated domain defined through multivariate kNN support in standardized space derived from the DoE cloud. The residual model is selected through five-fold cross-validation on 10,446 in-domain production samples and then retrained on the full approximate-residual set. A small set of seven matched cases is kept as an external check based on true residuals. The resulting hybrid predictor enables the generation of synthetic, experiment-informed corrected data while retaining the DoE coverage required for downstream modelling tasks.

Introduction

The maturation of machine learning (ML) and the availability of high-resolution production monitoring have renewed interest in data-driven support tools for metal forming. In practice, ML surrogates are often introduced to speed up design and robustness studies, cut down on try-out iterations and enable near-real-time decision-making. However, in industry, data comes from a variety of sources, each with its own limitations: numerical simulations are flexible and can explore a wide parameter space, while production data accurately reflect the plant, but are concentrated around a limited operating window [1-3].

Finite element (FE) models remain essential for sheet forming because they encode geometry, contact, and constitutive behaviour, but they are also affected by systematic bias [3]. In stamping, discrepancies between simulation and reality arise from unavoidable modelling assumptions (e.g., friction, heat transfer, tool compliance, and simplified material descriptions), from parameter uncertainty, and from the time evolution of the process. Increasing model complexity (for instance by including thermal effects) can reduce error, but it rarely eliminates it and often increases computational cost [1,4].

Hybrid modelling addresses this tension by combining a physics-based predictor with a data-driven component that learns the remaining ‘ignorance’-the part of the response that the physical model does not capture. Within the broader ‘hybrid twin’ paradigm, the correction is typically formulated as a residual model that adds a learned discrepancy to the numerical prediction. This residual-learning

This article is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0)

view is particularly attractive in manufacturing because the discrepancy is often smoother and lower-dimensional than the full response, which reduces the data burden of the ML component [5-7].

A practical obstacle in industrial settings is the limited availability of strictly matched simulation–experiment pairs. Matched cases allow one to compute 'true' discrepancies by comparing measurements with simulation under identical inputs, but generating them requires additional effort and is not always feasible, as plants often rely on a pre-defined design of experiments rather than an iterative matching campaign. In contrast, large-scale monitoring provides thousands of measured parts, but typically only within the plant's operating window. A methodology that can leverage this statistical richness without relying on many matched cases is therefore of high practical value [5,6].

This paper proposes a hybrid metamodel for an industrial stamping process in which the discrepancy model is trained exclusively from approximate residuals computed from production measurements and a numerical surrogate. A thermal effects-enabled DoE is generated with AutoForm Sigma and used to train a fast simulator surrogate $\hat{y}_{\mathrm{FEM}}(x)$. For each retained production sample, an approximate residual is defined as $r_{\mathrm{approx}}(x_j) = y_{\mathrm{EXP}}(x_j) - \hat{y}_{\mathrm{FEM}}(x_j)$, and a residual metamodel $\hat{r}(x)$ is learned from these targets. The key precaution is to restrict residual construction and learning to an explicitly validated domain defined through a multivariate support criterion derived from the numerical DoE cloud (kNN distance in standardized space) to prevent extrapolation. The necessity of this restriction is quantified in the Results and Discussion section. Model selection is carried out by five-fold cross-validation on the in-domain residual samples, and the final residual model is retrained on the full approximate set. A small set of seven FEM-experimental matched cases is reserved for an external check based on true residuals. The contributions are therefore: (i) an industrially transparent domain-validity statement that separates modelling error from extrapolation error; (ii) an approximate-residual formulation that allows discrepancy learning at scale without requiring many matched cases; (iii) a cross-validated training-and-freeze procedure followed by an external check on scarce matched points; and (iv) a hybrid predictor that can generate synthetic corrected data with numerical coverage and experimental consistency.

Methodology

The target response is the draw-in measured at the defect-critical location of an automotive door inner panel under industrial conditions. The case study is based on production monitoring collected during the manufacturing campaign comprising on the order of 70,000 inner door panels; in this study, a representative subset of about 10,446 monitored parts includes synchronized inputs and draw-in measurements suitable for model development. Three sources of information are considered: (a) a numerical Design of Experiments (DoE) of 67 simulations generated in AutoForm Sigma [8] (see input setup in Figure 1) with thermal effects enabled, providing simulated draw-in values y_{FEM} at designed input combinations; (b) a production dataset with monitored inputs and measured draw-in y_{EXP} for a large number of stamped parts; and (c) a small set of matched simulation-experiment configurations that can be used exclusively for external validation. Press speed is pre-filtered to a narrow range and treated as fixed, so that the modelling focuses on the remaining process and material variability.

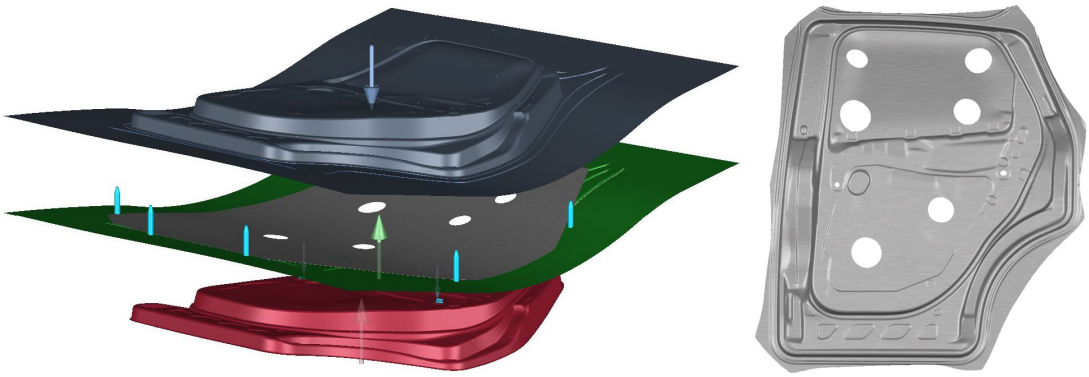

Fig. 1. AutoForm FE-model set-up (left) and inner door panel analyzed (right). Arrows indicate the prescribed tool motions and blankholder force directions used in the AutoForm setup. Reproduced from [9] under the Creative Commons CC BY 4.0 license.

The predictor vector x contains 12 variables available both in simulation and production monitoring: thickness (t), ultimate tensile strength (R_m), yield strength ($R_{p0.2\%}$), plastic anisotropy (r_n), die temperature, ambient temperature, two shim-height signals (Shim 20 and Shim 21), and four cushion forces (C1-C4). To improve numerical conditioning and to support distance-based validity checks, the inputs are standardized using statistics computed on the numerical DoE only: $x_k^{std} = x_k - \mu_k^{DoE} / \sigma_k^{DoE}$. Using DoE statistics makes the validity statement explicit-predictions are intended to be trustworthy only within the region where the numerical model has been sampled.

Because approximate residuals require evaluating the numerical surrogate at production inputs, extrapolation must be handled explicitly. In this work, an admissible operating domain Ω is defined through a multivariate support criterion derived from the numerical DoE cloud: the k-nearest-neighbour (kNN) distance of each standardized production sample to the DoE samples is compared to a threshold obtained from the internal DoE spacing (e.g., a high percentile of DoE kNN distances). Samples exceeding the threshold are flagged as out-of-support and excluded from residual construction. Of the 70,000 production measurements at 16 spm, 10,446 fall within the kNN-validated domain, forming the subset used for residual construction and model training. The remaining outside Ω, either the residual correction is gated down to zero—thereby reverting to the numerical surrogate—or the condition is marked for future DoE enrichment.

A surrogate of the numerical simulator, $\hat{y}_{FEM}(x)$, is trained using only the AutoForm Sigma DoE, considering the DoE inputs mentioned previously as inputs, and considering draw-in in the critical zone as output. Gaussian Process Regression (GPR) [10] is used here as a strong baseline for moderate DoE sizes and smooth interpolation. Five-fold cross validation was used for model validation. The surrogate is not calibrated to experiments; its role is to emulate the simulator with negligible runtime so that residual targets can be computed at scale within Ω.

For every retained production sample, an approximate residual target is then constructed as $r_{approx}(x_j) = y_{EXP}(x_j) - \hat{y}_{FEM}(x_j)$. This target captures the dominant simulation-to-plant discrepancy structure under real operating conditions, while also inheriting surrogate error and measurement noise. Accordingly, all claims are restricted to Ω and the residual model is validated with a conservative protocol.

The discrepancy model $\hat{r}(x)$ is learned by supervised regression on the approximate residual targets r_{approx}. Extreme Gradient Boosting regression trees (XGBoost regressor [11]) are adopted due to their robustness on mixed industrial signals and their favourable bias-variance trade-off when combined with cross-validation. Hyperparameters were selected via five-fold cross-validation on the in-domain approximate-residual dataset (10,446 samples after multivariate support gating). After selecting the configuration, the residual model is retrained on the full approximate-residual dataset to obtain a frozen model for deployment and for synthetic data generation. A set of seven matched

simulation-experiment cases, not used in training, was reserved for an external check based on true residuals. Although small, this check provides a direct sanity test of whether the approximate-residual training leads to a correction that remains physically consistent when compared against matched conditions.
Finally, the hybrid predictor is defined as $\hat{y}_{hyb}(x) = \hat{y}_{FEM}(x) + \hat{r}(x)$. Synthetic corrected data can be generated by evaluating $\hat{y}_{hyb}$ on the numerical DoE points (and, if required, on additional samples drawn within Ω), yielding a corrected dataset that retains the DoE coverage while embedding an experimentally informed bias correction

To evaluate the predictive accuracy of the metamodels and the hybrid strategy, four primary statistical indicators are used. Two main metamodels are evaluated in this work, each with a different reference quantity y_i. For the GPR numerical surrogate, y_i stands for the draw-in output from AutoForm; performance is reported via five-fold cross-validation on the 67 DoE points. For the XGBoost residual model, the reference is the approximate residual. The Root Mean Square Error (RMSE) provides a measure of the average error between the residual or numerical draw-in (depending on the metamodel) measurements (y_i) and the model predictions ($\hat{y}_i$), giving higher weight to larger discrepancies:

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

Additionally, the Mean Absolute Error (MAE) is utilized to represent the average absolute difference magnitude, offering a more linear representation of the error:

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

And to identify any systematic over-prediction or under-prediction by the models, the Bias is calculated as:

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (3)$$

The overall proportion of variance captured by the model is evaluated using the Coefficient of determination, or R-squared (R^2). This metric indicates the goodness-of-fit by comparing the model residuals to the total variance of the experimental data:

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

where $\bar{y}$ represents the mean of the numerical or measurements. An R^2 value of 1 indicates a perfect match between the model and the physical process.

Finally, to ensure that the reported performance is robust and not dependent on a specific data split, a k-fold Cross-Validation (CV) procedure is implemented. In this study, the mean and standard deviation of these metrics are calculated across five folds, ensuring that the model generalizes properly to unseen data.

Results and Discussion

Applying the multivariate kNN support screening to the production monitoring data yields an in-domain subset suitable for residual construction. In the present case, 10,446 samples remain after screening. Table 1 summarizes the datasets and splits used in this work.

Table 1. Summary of datasets and their roles in the hybrid modelling pipeline.

Dataset	Role	Size	Notes
Numerical DoE (AutoForm Sigma)	Train $\hat{y}_{FEM}$	67	Thermal-enabled; covers broad design space
Production in-domain subset	Train/validate $\hat{r}$	10,446	After multivariate kNN support screening
Matched cases (reserved)	External check	7	True/anchored residuals; not used for training

The numerical surrogate $\hat{y}_{FEM}$ is evaluated only on inputs deemed in-domain, so that surrogate quality and residual structure are not confounded with extrapolation. This separation is critical in industrial reporting: if an out-of-support condition occurs, the methodology does not attribute the resulting error to the hybrid model but rather flags the condition for additional simulation or for a dedicated local model. This also enables a clean operational guideline: within Ω the hybrid predictor can be used for rapid what-if analysis, while outside Ω the system reverts to the numerical surrogate or triggers a DoE update cycle. The physics surrogate is a Gaussian Process Regression model with Matérn 5/2 kernel, constant basis function, and noise σ = 1.19 mm. Error metrics of the numerical surrogate model are shown in Table 2. Five-fold cross-validation was used for validation.

Table 2. GPR Error Metrics.

CV R-squared [-]	CV RMSE [mm]	CV MAE [mm]
0.9800	1.9280	1.4937

Regarding the discrepancy model $\hat{r}(x)$, XGBoost model was trained and validated with approximated residuals. Five-fold cross-validation on the 10,446 approximate residual samples provides an internal estimate of generalization under typical operating variability. The mean and standard deviation of the error metrics across folds for the approximate residual dataset is reported in Table 3. Hyperparameters were selected via cross-validation on the approximate residual dataset, and the final configuration used in all reported results is summarized in Table 4. In addition to global metrics, fold-wise diagnostics are useful to verify that performance is stable and not driven by a single favourable split.

Table 3. Five-fold cross-validation results on the approximate residual dataset.

Model	CV RMSE (mean±std)	CV MAE (mean±std)	CV Bias (mean±std)	CV R-squared (mean±std)
Residual model $\hat{r}$ XGBoost	1.5416± 0.0361 mm	1.0535 ± 0.0139 mm	-0.0027 ± 0.0377 mm	0.9412 ± 0.0032

Table 4. XGBoost optimized hyperparameters used for the residual model.

Parameter	Value
Number of boosting rounds	264
Maximum tree depth	7
Shrinkage factor	0.0242
Row sampling ratio	0.7192
Feature sampling ratio	0.8139
Minimum child weight	3
Minimum split loss	0.3885
tL1 regularization	0.1381
tL2 regularization	6.7886

After model selection, the residual regressor is retrained on the full approximate-residual dataset. The residual model $\hat{r}(x)$ is then evaluated on the seven matched cases that provide true residuals, being these cases simulations that exactly match experimental conditions in seven different configurations. Despite the limited sample size of seven points, this external validation is significant because it

evaluates the model's correction against true residuals derived from matched simulation-experiment pairs, rather than surrogate-based approximations. Figure 2 illustrates the relationship between real and predicted residuals for these matched cases, alongside the approximate residuals used during training.

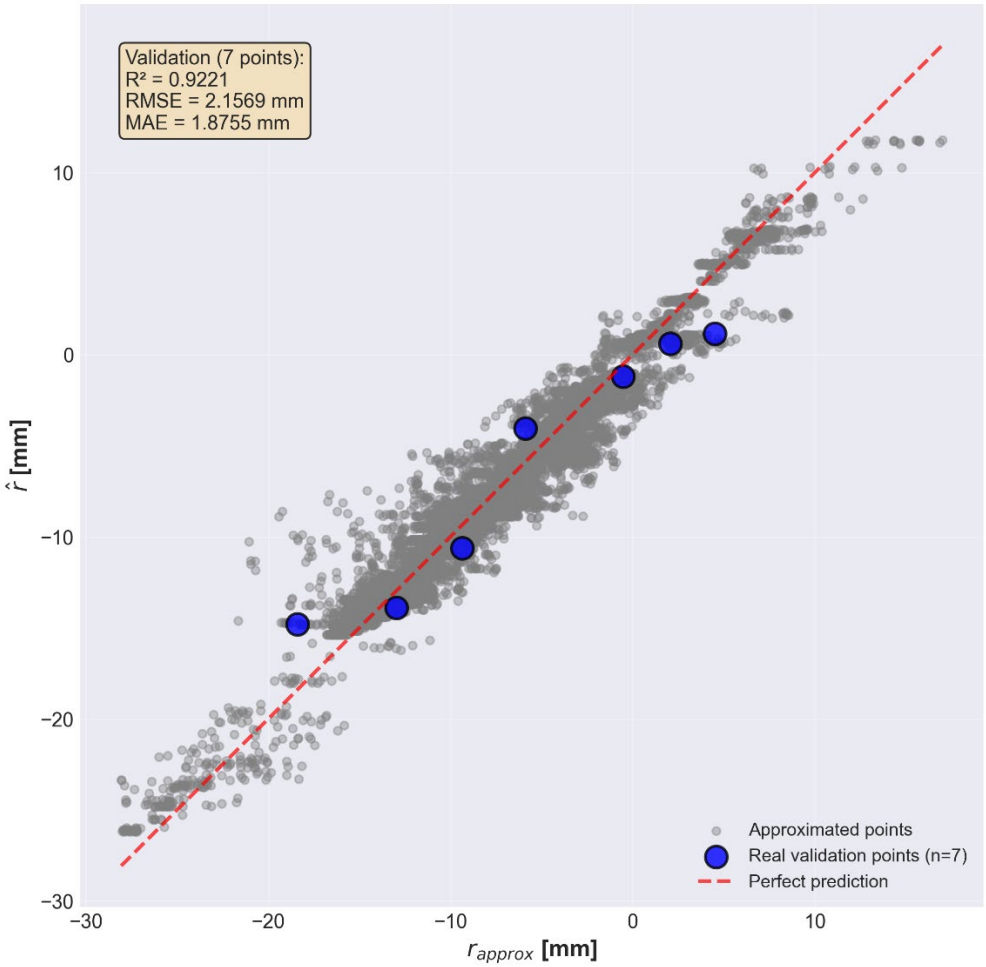

Fig. 2. Predicted residual $\hat{r}$ versus residual targets: production (surrogate-based) and anchored validation.

Figure 2 provides a compact sanity check of whether a residual model trained on surrogate-based production residuals can remain consistent when confronted with matched (anchored) conditions. The grey cloud shows that the XGBoost model learns a near one-to-one mapping between predicted residuals $\hat{r}$ and the approximate targets r_{approx} over the in-domain production dataset, with limited systematic deviation from the ideal 1:1 line. Importantly, the seven reserved anchored validation points (blue markers), which were not used during training or hyperparameter tuning, lie close to the same trend and yield strong agreement ($R^2 \approx 0.92$, RMSE $\approx$ 2.16mm, MAE $\approx$ 1.88mm). Although the anchored set is small and therefore not intended as a statistically exhaustive validation, its alignment with the learned mapping supports the physical plausibility of the correction and suggests that learning on approximate residuals can transfer to matched conditions within the validated domain.To further see the hybrid model use improvement, the hybrid predictor $\hat{y}_{hyb}$ is then evaluated on the reserved matched cases that provide true residuals.

Figure 3 contrasts the error distributions of the baseline numerical surrogate and the hybrid predictor within the validated domain Ω. On the full in-domain dataset ($n = 10{,}453$), the baseline shows a clear systematic offset ($\mu \approx 7.60$ mm) and broad dispersion ($\sigma \approx 6.35$ mm), whereas the hybrid error is effectively centered ($\mu \approx 0.00$ mm) and substantially narrower ($\sigma \approx 1.38$ mm), indicating that the learned discrepancy term mainly acts as a bias compensator and variance reducer under the operating regime covered by the data. Importantly, the same trend is observed on the reserved matched

validation set (n = 7), where the mean error shifts from μ ≈ 5.79 mm (baseline) to μ ≈ −0.33 mm (hybrid) and the dispersion decreases from σ ≈ 8.35 mm to σ ≈ 2.30 mm. While the "all-data" distribution can be optimistic if it includes samples used during residual training, the matched-case statistics provide an external, physically grounded sanity check that the correction does not collapse when evaluated under matched conditions. Residual tails and any remaining multimodality suggest that multiple operating regimes and/or unmodelled or bad-modelled factors may still be present, reinforcing the need for domain gating and continued expansion/curation of the matched-case set. Residual tails suggest that the process may still switch between operating regimes and that some effects are not fully captured (temperature modeling being a possible contributor). This supports a practical deployment view in which temperature could be handled as a disturbance.

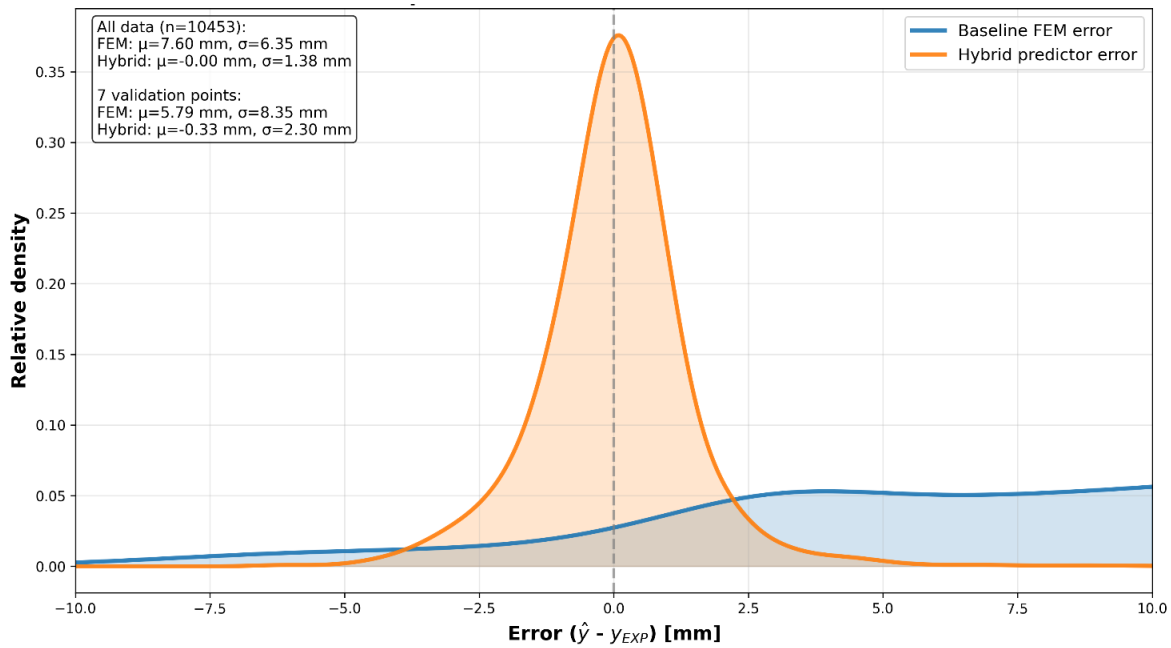

Fig. 3. Qualitative effect of residual correction on prediction error.

A key consideration is that the discrepancy model is trained on surrogate-based residuals $r_{approx} = y_{EXP} - \hat{y}_{FEM}$, and therefore inherits part of the uncertainty of the numerical surrogate (here, the GPR emulator of the FEM model). Consequently, the learned correction reflects both simulation–production mismatch and any residual surrogate error within the validated domain. A further limitation of the present study is the small number of matched simulation–experiment cases available for the external anchored check. While the observed agreement is encouraging, a larger matched set would enable a more reliable quantification of out-of-sample performance and more systematic sensitivity analyses. To address this, Table 5 reports a sensitivity study comparing three prediction strategies across the in-support (10,446 samples) and out-of-support (59,554 samples) regions. The kNN-filtered model achieves an 85.3% RMSE reduction in-support (from 10.68 mm to 1.54 mm). Training on all 70,000 points without the kNN filter yields only 79.4% improvement (from 10.68 mm to 2.20 mm), observing that out-of-support approximate residuals can contaminate the model.

Table 5. Sensitivity analysis of kNN support filtering: Root Mean Squared Error (RMSE) in mm of draw-in prediction for in-support and out-of-support production measurements under four prediction strategies.

Region	Samples	RMSE Surrogate-only [mm]	RMSE Filtered-train (no gate) [mm]	RMSE Unfiltered-train (no gate) [mm]
In-Support	10,446	10.68	1.54	2.20
Out-of-Support	59,554	10.87	6.29	2.22

The proposed pipeline remains attractive in practice because it provides an explicit and conservative validity statement, and it offers a scalable path to progressively strengthen the hybrid predictor as additional matched cases become available.

Conclusions

An approximate-residual hybrid metamodel was formulated for an industrial stamping process to reconcile numerical coverage with production fidelity under limited matched-case availability. The approach introduces an explicit, reproducible validity statement by restricting residual construction and learning to the numerical DoE support, optionally refined through multivariate support checks in standardized space. Within this validated domain, a fast numerical surrogate $\hat{y}_{\mathrm{FEM}}$ enables the computation of large-scale approximate residual targets from production data, and a residual regressor $\hat{r}$ is selected by five-fold cross-validation on 10,446 in-domain samples and then retrained on the full set for deployment.

The resulting hybrid predictor $\hat{y}_{\mathrm{hyb}}(x) = \hat{y}_{\mathrm{FEM}}(x) + \hat{r}(x)$ provides an experiment-informed correction while retaining the broad sampling of the numerical DoE. This enables the generation of synthetic corrected datasets that can support downstream modelling tasks that require coverage beyond the narrow production operating window. An external check based on seven matched cases provides a conservative sanity test of physical consistency; expanding the matched-case set remains a priority to strengthen external validation and to refine domain-gating strategies in out-of-support regions.

Acknowledgements

The authors acknowledge that the data used in this study were generated within the framework of the iStamp project (Grant Agreement IDI-20220061) under SMART Eureka program, funded by the Centre for the Development of Industrial Technology (CDTI) of the Spanish Ministry of Science and Innovation. The analyses, interpretations, and conclusions presented in this paper were developed after the completion of the project.

References

[1] P. de Souza and B. Rolfe, “Characterising material and process variation effects on springback in sheet metal forming,” Int. J. Mech. Sci., 2010.

[2] M. Golmohammadi et al., “A machine learning-based model to predict residual stress distribution considering FE modelling and experimental uncertainties,” Int. J. Mech. Sci., 2025.

[3] L. Muñiz, L. Galdos, and J. Trinidad, “Metamodel-based control algorithms for the correction of bending angle after springback in an industrial U-Bending process,” Int. J. Mater. Form., vol. 18, art. no. 44, pp. 1–20, May 2025, doi: 10.1007/s12289-025-01906-7.

[4] S. B. Choi and co-authors, “Replacing FEA for sheet metal forming by surrogate modeling,” Cogent Eng., 2014.

[5] F. Chinesta, A. Huerta, G. Rozza, and K. Willcox, “Virtual, Digital and Hybrid Twins: A New Paradigm in Data-Based Engineering and Engineered Data,” Arch. Comput. Methods Eng., 2019 (online 2018).

[6] S. Torregrosa, F. Chinesta, and co-authors, “Hybrid twins based on optimal transport,” Appl. Math. Comput., 2022.

[7] A. Forrester, A. Sóbester, and A. Keane, “Engineering Design via Surrogate Modelling,” Wiley, 2008.

[8] AutoForm Engineering GmbH, “AutoForm-Sigma: Robust Processes,” product brochure, 2024.

[9] L. Muñiz, J. Trinidad, E. Garcia, I. Peinado, N. Montes, and L. Galdos, “On the Use of Advanced Friction Models for the Simulation of an Industrial Stamping Process including the Analysis of Material and Lubricant Fluctuations,” Lubricants, vol. 11, art. no. 193, Apr. 2023, doi: 10.3390/lubricants11050193.

[10] C. Rasmussen and C. Williams, “Gaussian Processes for Machine Learning,” MIT Press, 2006.

[11] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD ’16), 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

The publication of this article was funded by the Mondragon Goi Eskola Politeknikoa, J.M.A. S.Coop