

Semantic Enrichment of Additive Manufacturing Technology Data for Machine Learning-based Optimization of Laser Beam Powder Bed Fusion Process

Katsiaryna Kosarava^{1,*}, Bartłomiej Adam Wysocki²

¹*Faculty of Mathematics and Natural Sciences, Cardinal Stefan Wyszyński University in Warsaw, 1/3 Kazimierza Wóycickiego str., Warsaw, 01-938, Poland*

²*Multidisciplinary Research Centre, Cardinal Stefan Wyszyński University in Warsaw, Dziekanów Lesny, 05-092, Poland*

Abstract

This paper presents novel methodology integrating ontological knowledge representation, knowledge graph construction, and agent-based automated feature engineering for predictive modeling in additive manufacturing (AM). The proposed framework addresses the critical challenge of embedding domain-specific physical knowledge into machine learning (ML) pipelines by leveraging structured information extracted from scientific literature and material property databases. In the context of Laser Beam Powder Bed Fusion (PBF-LB), this challenge is of growing industrial importance, as this process is widely used for high-value applications in aerospace, electronics, and medicine, where complex geometries, expensive powder materials, and efficient use of scarce elements are critical. When a new powder material is introduced into PBF-LB production, the first practical objective is typically to identify process parameters that minimize critical defects such as porosity and cracking; only once defect formation is controlled does optimization usually proceed toward microstructure tailoring, post-process heat treatment, surface quality, or build-rate improvement. For newly developed alloys, conventional parameter optimization based on extensive experimental campaigns or Design of Experiments is costly and time-consuming, which makes data-efficient ML and transfer learning approaches particularly attractive for accelerating the search for defect-minimized and time-efficient process windows. At the core of the contribution is a Feature Engineering Agent capable of autonomously generating physics-informed descriptors based on an enriched ontology and a populated knowledge graph, effectively bridging the gap between data-driven modeling and first-principles process understanding. Experimental validation was conducted on an H13 tool steel dataset created in PBF-LB processes, encompassing key process parameters such as laser power, laser scanning speed, and distance between scanning vectors ("hatch spacing"), with eight ML models evaluated on the task of porosity prediction. Model evaluation on the updated dataset incorporating the newly derived descriptors revealed heterogeneous algorithmic responses. The SGD model demonstrated substantial improvements across all performance metrics ($\Delta R^2 = +103.00\%$, with corresponding reductions in error metrics), whereas the remaining algorithms exhibited only moderate or negligible gains. Feature importance analysis confirmed that the new descriptor preserved the predictive information of the original linear energy density while introducing additional physical context, accompanied by observable redistribution of feature importance in select models.

Keywords

additive manufacturing, laser beam powder bed fusion (PBF-LB), ontology, knowledge graph, feature engineering, machine learning, physics-informed learning

1. Introduction

The recent industrial expansion of metal additive manufacturing technologies has further increased the importance of Laser Beam Powder Bed Fusion (PBF-LB) process, which is widely regarded as one of the leading for producing geometrically complex metallic components with high precision and high material efficiency. Its relevance is particularly pronounced in aerospace, medical, and advanced electronic applications, where PBF-LB is increasingly used for high-value parts made from expensive or scarce materials, and minimization of powder consumption and scrap is therefore an important

SeMatS 2026: The 3rd International Workshop on Semantic Materials Science co-located with the 25th International Semantic Web Conference (ISWC 2026), October 25th, 2026, Bari, Italy

*Corresponding author.

✉ k.kosarava@uksw.edu.pl (K. Kosarava); b.wysocki@uksw.edu.pl (B. A. Wysocki)

ORCID 0000-0001-7326-5307 (K. Kosarava); 0000-0002-4119-621X (B. A. Wysocki)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

practical objective.

From an industrial perspective, introducing a new powder material in PBF-LB typically begins with identifying a stable process window that minimizes fundamental defects, especially pores and cracks, since these govern mechanical performance and reliability. Only after a defect-lean range is established does optimization usually shift to advanced objectives such as microstructure control, surface quality, or build productivity.

Recent advances in laser-based powder bed fusion (LPBF), particularly PBF-LB, increasingly rely on machine learning (ML) for predicting material properties and optimizing process parameters. Key approaches, such as Gaussian Process Regression, Support Vector Machines, ensemble tree algorithms and Artificial Neural Networks have proven effective for parameter optimization, density and porosity prediction, defect detection, and melt pool geometry characterization [1, 2, 3, 4, 5]. In industrial PBF-LB workflows, such models are valuable because they reduce the number of trial builds needed to identify parameter sets that achieve both acceptable part quality and efficient production times.

However, despite these successes, a significant gap persists in PBF-LB parameter optimization. Most existing ML studies focus on tuning laser parameters as isolated variables, without adequately accounting for the process ontology, the material's physical and mechanical properties, or leveraging previously accumulated empirical knowledge in the form of knowledge graphs (KG) for other materials. This gap is particularly pronounced when new alloys are introduced into production, because the amount of available PBF-LB data for such materials is usually too limited to support robust model development from scratch, even though rapid process qualification is critical in industrial contexts. Overcoming these limitations requires a paradigm shift toward novel methodologies that integrate structured process knowledge. Recent efforts have focused on formalizing PBF-LB ontologies that capture defect knowledge and causal relationships [6, 7], as well as constructing KGs to represent interconnections among process parameters, mechanisms, and defects [8]. Furthermore, ontologies for process planning and knowledge bases utilizing semantic web rules language (SWRL) have been proposed to support decision-making in laser scanning strategy, part orientation on the building platform, support generation and model slicing [9]. These foundational studies pave the way for intelligent systems capable not only of prediction but also of explaining decisions by integrating heterogeneous knowledge, thereby opening new avenues for more efficient and interpretable LPBF optimization.

2. Ontology-Driven Framework for PBF-LB Process Optimization

To address the limitations of conventional ML approaches that treat laser parameters as isolated variables without leveraging structured process knowledge, we propose an ontology-driven framework that integrates semantic representations of the PBF-LB process with ML optimization. The architecture of the proposed framework is illustrated in the system diagram (Figure 1) and comprises four interconnected layers with bidirectional feedback loops enabling continuous refinement.

The foundation of the framework is a **Knowledge Base** that formalizes domain expertise through three interconnected components:

- Papers Database (folder with JavaScript Object Notation - JSON) – a structured repository of experimental data extracted from scientific literature, where each paper is stored as a JSON file containing standardized information about process parameters, material properties, experimental conditions, and reported outcomes.
- Ontology. The ontology organizes PBF-LB knowledge by hierarchically linking process parameters (laser power, scan speed, hatch spacing, etc.), material properties, and defect types through explicit causal relationships, providing a semantic backbone for consistent data interpretation and automated reasoning.
- Knowledge Graph – an instantiated graph structure that links entities from the ontology with experimental evidence extracted from papers. The knowledge graph enables querying of prior experiences across different materials and process conditions, effectively representing accumulated empirical knowledge.

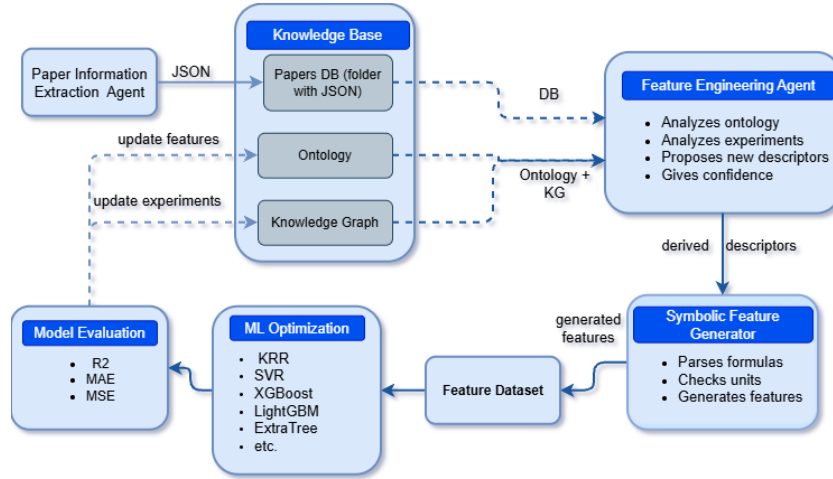


Figure 1: Ontology-driven PBF-LB process optimization framework

The **Paper Information Extraction Agent** automatically populates the Papers Database by parsing scientific publications and converting relevant information into structured JSON format, ensuring systematic accumulation of experimental data. Iterative Prompt Refinement (IPR) was applied to extract structured data from papers using a closed-loop workflow: a base prompt was used for extraction, expert analysis identified errors and omissions, and the prompt was refined accordingly until satisfactory performance was achieved. Two models were evaluated: DeepSeek V4.1 Flash and GPT-5.6 Luna, both achieving comparable extraction accuracy with no significant performance difference observed between them. Notably, both models demonstrated substantial improvements in extraction quality when plain text extracted from articles via the textract library (version 2.1.0) was provided as input, rather than the original PDF files [10].

The **Feature Engineering Agent** serves as the core reasoning component that bridges domain knowledge with ML-ready features. This agent performs four key functions:

- Interprets semantic relationships to identify relevant process parameters and material descriptors.
- Examines historical experimental data to understand parameter distributions and their effects on quality metrics.
- Generates candidate derived features (dimensionless numbers, combined parameters, or physics-informed metrics) based on domain knowledge encoded in the ontology.
- Assigns confidence scores to proposed descriptors based on their theoretical grounding and empirical support from the knowledge base.

The **Symbolic Feature Generator** operationalizes these new descriptors by parsing their underlying mathematical formulas, performing rigorous dimensional and unit consistency checks, and calculating the resultant features. These computed features are aggregated into a dataset, which is then fed into the ML optimization pipeline.

The **ML Optimization** module applies a diverse ensemble of supervised learning algorithms to model the relationship between augmented dataset features and target properties (density, porosity, melt pool geometry). The framework supports multiple state-of-the-art regressors (eg., Extreme Gradient Boosting, LightGBM, etc.). This diversity allows comparative evaluation and selection of the most appropriate model for specific prediction tasks.

The **Model Evaluation** module assesses prediction performance using standard regression metrics: R^2 (coefficient of determination), **MAE** (Mean Absolute Error), **MSE** (Mean Squared Error) and **MAPE** (Mean Absolute Percentage Error).

A distinctive feature of the framework is its **closed-loop architecture**, which enables continuous knowledge accumulation and model improvement. The feedback process is activated after the completion of model training and testing, as well as after the assessment of feature importance used for

prediction. At this stage, two main actions are performed: *updating the KG* with training results and *extending the ontology* with validated new features. Model metrics and feature importance are added to the KG, linking fresh data to existing entities and enriching the understanding of process relationships. If newly proposed features improve predictive performance, they are incorporated into the ontology with proper semantic relationships to capture their physical influence on product properties. In this way, each optimization cycle not only identifies the optimal combination of PBF-LB process parameters for the current material but also expands the systematized knowledge about process physics, including melt pool behaviour, defect formation mechanisms, and other intermediate phenomena. As the KG becomes increasingly comprehensive and the ontology is enriched with new descriptors of proven predictive power, subsequent iterations can rely on more informative features and make better-informed decisions when optimizing PBF-LB process for different materials.

3. Semantic representation of PBF-LB knowledge

To enable systematic knowledge representation and reasoning within the proposed framework, we developed a formal ontology for PBF-LB domain. The ontology is implemented in OWL 2 (Web Ontology Language [11]) and captures the fundamental entities, their attributes, and the causal relationships governing the PBF-LB process. The ontology structure is designed to support both human interpretation and machine reasoning, facilitating the integration of domain knowledge with machine learning workflows.

The ontology defines seven primary classes that represent the key conceptual categories within the PBF-LB process:

1. **MachineParameter** – Represents the controllable machine and process parameters of the PBF-LB equipment. These include operational settings that can be adjusted during manufacturing, such as laser power [W], scan speed [mm/s], spot diameter [μm], hatch distance [μm], layer thickness [μm], platform size [cm^2] and preheating temperature [$^{\circ}\text{C}$], wiper type and mechanism, etc.
2. **MaterialParameter** – Encapsulates the intrinsic physical properties of the feedstock material. These parameters are material-specific and influence the process outcome independently of machine settings. Examples include theoretical density [g/cm^3], powder bulk density [g/cm^3], specific heat capacity [$\text{J}/(\text{kg K})$], thermal conductivity [$\text{W}/(\text{m K})$], melting point [$^{\circ}\text{C}$ or K], thermal diffusivity [mm^2/s], optical absorptivity (at the laser wavelength) [–], coefficient of thermal expansion [$1/\text{K}$], electrical conductivity [S/m], etc.
3. **DerivedDescriptor** – Denotes composite or physically informed features that are calculated from combinations of machine and material parameters (Figure 2). These descriptors often capture fundamental physical phenomena and serve as more informative predictors for ML models. Examples include Volumetric Energy Density (VED) [J/mm^3], Linear Energy Density (LED) [J/mm], laser–powder-bed interaction time [s], and normalized melt pool or cooling descriptors [12, 13, 14, 15].
4. **Formula** – Represents the mathematical expressions used to compute derived descriptors from base parameters. Each formula includes its algebraic expression, associated variables, units, and source attribution. For example, $\text{VED} = P/(v \cdot h \cdot t)$, where P is laser power [W], v is scan speed [mm/s], h is hatch spacing [mm], and t is layer thickness [mm], giving VED in units of energy per unit volume [J/mm^3] or $\text{ED} = P/f(v, d, H, \lambda)$, where P is laser power [W], v is scan speed [mm/s], d is laser spot diameter [mm], h is hatch spacing [mm], and λ is laser wavelength [mm], and $f(\cdot)$ is a function that combines these parameters to yield an effective energy input per unit area or volume [J/mm^2 or J/mm^3].
5. **PhysicalPhenomenon** – Captures intermediate physical processes and mechanisms that occur during PBF-LB and mediate the relationship between process parameters and final material properties. Examples include melt pool dynamics, cooling rate, formation of keyhole and lack of fusion porosity, etc.

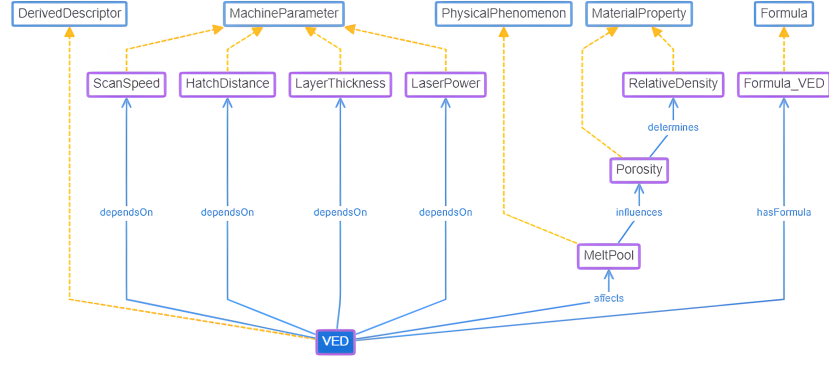


Figure 2: DerivedDescriptor example for Volumetric Energy Density (VED)

6. **MaterialProperty** – Denotes the final quality attributes of the fabricated part that are of primary interest for optimization. These include relative density, porosity, microstructure features, surface roughness, mechanical properties (hardness, tensile strength), etc. .
7. **Process** – Represents the overall manufacturing process itself, specifically the PBF-LB or PBF-EB process, which serves as the contextual container that relates all other entities.

The ontology defines a rich set of object properties that establish semantic relationships between classes, enabling causal and dependency modeling: *affects*, *dependsOn*, *determines*, *influences*, *hasFormula*, *usesParameter*, and *processedBy*. These properties link **MachineParameters** and **MaterialParameters** to **DerivedDescriptors** (e.g., VED, LED, Interaction Time) via *dependsOn*, while *hasFormula* connects each descriptor to its mathematical **Formula**. The *affects* property then associates derived descriptors with intermediate **PhysicalPhenomena** such as MeltPool, Keyhole, LackOfFusion, and CoolingRate. Subsequently, the *influences* property links these physical phenomena to final **MaterialProperties** (Hardness, Porosity, RelativeDensity), and *determines* captures quantitative dependencies between properties, as in Porosity → RelativeDensity. Concrete instantiations materialize these causal chains, for example, LED → MeltPool → Porosity → RelativeDensity, while OverlapRatio → LackOfFusion → Porosity, and InteractionTime → CoolingRate → Microstructure → Hardness. Finally, the ontology is populated with material-specific data, including density, thermal conductivity, and melting point, which are linked to the PBF-LB process via *usesParameter* and *processedBy*, enabling quantitative reasoning and physics-informed feature engineering.

4. Knowledge Graph for ML-Driven PBF-LB Optimization

We developed a KG that extends the PBF-LB ontology with experimental workflows and model evaluation data to operationalize closed-loop feedback. The KG is formalized in OWL and comprises five core classes that structure the experimental lifecycle: **Dataset**, **Experiment**, **Feature**, **Model**, and **Metrics**, along with specialized subclasses for feature types and importance measures.

The **Dataset** class captures the data used for model training and evaluation, including the specific features employed and the target variable.

The **Feature** hierarchy distinguishes between **InitialFeature** (raw process parameters) and **GeneratedFeature** (derived descriptors). This distinction enables the system to track which features originate from direct measurements and which was generated.

The **Experiment** class serves as the central hub, linking all components of each ML optimization run. Each experiment is instantiated as an individual that connects a specific **Model** (e.g., XGBoost, Extra Tree, etc.) with the **Dataset** used, the **Material** under investigation, and the resulting **Metrics**.

Performance metrics are captured through a dedicated **Metrics** class with datatype properties for R^2 , MAE, MAPE, MSE and training CV R^2 . The KG stores both training and test performance, enabling

direct comparison across models. The KG further captures optimal hyperparameters and optimal feature values identified during optimization, linking optimal configurations to specific experiments and enabling systematic tracking of promising parameter regimes for future reference.

FeatureImportance has two subclasses (**PermutationImportance** and **SHAPImportance**), each storing per-feature values. The properties *hasOptimalFeatureValues*, *hasPermutationImportance*, and *hasSHAPImportance* enable closed-loop feedback, feeding new importance patterns back into the ontology for future decisions.

By formalizing experimental provenance, model performance, feature importance, and optimal parameter configurations in a machine-readable KG, this representation enables systematic knowledge accumulation across experiments, facilitates meta-analysis across different ML models, and provides a structured foundation for the iterative refinement of both feature sets and optimization strategies within the proposed ontology-driven framework.

5. Semantic Feature Engineering

We developed a semantic **Feature Engineering Agent**, powered by an LLM, that systematically generates physically informed descriptors for PBF-LB process optimization. It operates under a constrained reasoning framework integrating three knowledge sources: structured experimental data (JSON), the PBF-LB ontology (OWL), and a KG of causal chains and prior results (OWL/TTL). Unlike conventional approaches, the agent leverages the ontology’s semantic structure and the KG’s relational topology to identify physically meaningful candidates, guided by a prompt encoding domain expertise in PBF, materials science, physics-informed ML, and knowledge engineering. The agent targets fundamental physical mechanisms (e.g., heat transfer, melt pool stability, solidification) while avoiding duplication of existing features or meaningless combinations. For each candidate, it provides a structured evaluation: the mathematical formula, required variables, relevant ontology concepts, explicit KG paths substantiating the causal rationale, the underlying physical principle, expected relationship direction (positive, negative, or nonlinear), novelty assessment, confidence score (0–1), and computability status. This systematic evaluation ensures that each proposed descriptor is traceable to both physical theory and empirical evidence encoded in the knowledge graph. The ablation results demonstrate that incorporating the KG primarily enhances the novelty and causal grounding of the generated features, raising the reasoning-quality score, whereas incorporating the ontology primarily improves practical viability, increasing the fully computable feature count and raising the mean confidence from 0.71 to 0.82. The full configuration (JSON + KG + ontology) achieves the best balance, combining high productivity and high reasoning quality with the broadest coverage of target properties (porosity, relative density, hardness, lack of fusion, and keyhole). The JSON-only baseline is consistently dominated, yielding no fully computable features and no source traceability. These findings indicate a clear synergy between the knowledge graph and the ontology, with the former supplying mechanistic relationships and the latter supplying standardized, computable descriptors.

For illustration, consider the *ThermalDiffusivityAdjustedLED* descriptor proposed by the agent:

$$\Phi = (P \cdot \alpha) / (v \cdot d),$$

where P is laser power, v is scan speed, d is spot diameter, and α denotes thermal diffusivity. This feature is a phenomenological descriptor that scales line energy density (P/v) by the thermal diffusivity and the beam diameter, approximating the combined effect of heat input and heat spreading in the powder bed and substrate. The underlying causal chain, extracted from the knowledge graph, is $P \xrightarrow{\text{dependsOn}} \text{LED} \xrightarrow{\text{affects}} \text{MeltPool} \xrightarrow{\text{influences}} \text{Porosity}$, with an additional branch $\alpha \xrightarrow{\text{affects}} \text{HeatDiffusion} \xrightarrow{\text{influences}} \text{CoolingRate} \xrightarrow{\text{influences}} \text{Microstructure} \xrightarrow{\text{influences}} \text{Hardness}$.

The descriptor is grounded in the ontology via the **DerivedDescriptor** class, exhibits a high novelty, and is estimated with a confidence of 0.85, though its computability is limited by the availability of material-specific thermal diffusivity data.

After the agent generates new features, the resulting JSON file is passed to the **Symbolic Feature Generator** for parsing and verification of the applicability of these features for training ML models. The Symbolic Feature Generator is an automated pipeline that translates semantically proposed descriptors into computable features and integrates them into the dataset. It operates through three stages: (1) semantic mapping—using a sentence-transformer [16] to map abstract variables to dataset columns via similarity thresholds; (2) symbolic computation—parsing formulas with SymPy [17] and applying them element-wise to datasets for features with confidence > 0.7 ; and (3) correlation-based filtering—removing redundant features with inter-correlation > 0.9 while retaining those most correlated with the target. The pipeline outputs augmented datasets and a JSON log of eliminated features, and is designed for modular integration with the evolving ontology and KG through closed-loop feedback.

6. ML for PBF-LB process parameters optimization

The ML optimization stage covers data preparation and preprocessing and model training. The input to the ML optimization module consists of the augmented dataset containing calculated values of novel features proposed by the Feature Engineering Agent. Following preliminary data preprocessing, which includes encoding of categorical features and normalization of numerical variables, the preprocessed data are fed into the set of models, which are trained using 5-fold cross-validation (CV) [18] over a predefined hyperparameter search space. Since the incoming data with base features are assumed to have been already cleaned of duplicates and outliers prior to ingestion into the framework, these steps are omitted from the preprocessing pipeline. The model with the best-performing hyperparameters is then evaluated on the test dataset using a comprehensive suite of metrics within the Model Evaluation module, where feature importance is additionally assessed via both permutation importance [19] and SHAP (SHapley Additive exPlanations) methods [20]. To identify the optimal PBF-LB process parameters, a parameter grid is generated from predefined ranges, and the selected target property (e.g., density, porosity, or other quality metrics) is evaluated for each parameter combination within this grid. Such grid-based evaluation is particularly valuable for industrial PBF-LB applications in aerospace, electronics, and medical sectors, where exhaustive experimental exploration of all parameter combinations is impractical due to high machine time and powder costs. Finally, the obtained results, including optimal parameter configurations, model performance metrics, and feature importance scores, are fed back to update both the KG and the process ontology, thereby enabling continuous knowledge accumulation and iterative refinement of the optimization framework. Crucially, only those new features that provided a statistically significant improvement in the metrics of at least one model are added to the ontology, preventing the accumulation of non-informative or redundant features.

7. Semantic Interpretation

The proposed approach extends classical ML-based optimization by embedding predictions within a semantically meaningful context built upon a KG. For each prediction, the system retrieves a causal subgraph that links the input process parameters to the target material property through intermediate physical phenomena encoded in the ontology. This allows the user to trace how a specific combination of laser parameters influences melt pool behavior, cooling rates, and ultimately density or porosity — rather than receiving an opaque numerical value. In addition, SHAP and permutation importance values are projected onto the ontological nodes, revealing which physical mechanisms contribute most to the prediction. As a result, every model prediction becomes traceable and scientifically grounded, enhancing transparency, interpretability, and trust in the optimization outcomes. This type of semantically grounded explanation is particularly useful in regulated industries (e.g., aerospace and medical devices), where process engineers must justify parameter choices and defect mitigation strategies in terms of underlying physical mechanisms, rather than solely relying on opaque ML predictions.

8. Experiment results

The experiment was conducted on example H13 steel PBF-LB parameter optimization dataset described in [15], in which the authors fabricated 189 cylindrical specimens (fabrication process was conducted using Aconity3D GmbH, Herzogenrath, Germany) for training and testing ML models using variable process parameters: laser power (250–350 W), scanning speed (1050–1300 mm/s), and hatch spacing (65–90 μm). Following the complete data cleaning and preprocessing pipeline described in the original study, the resulting training and testing sets comprised 166 and 43 samples, respectively, and included the following features: laser power P [W], scanning speed v [mm/s], hatch spacing h [μm], volumetric energy density VED [J/mm^3], linear energy density LED [J/mm], sample position on the build platform (marked on each sample by a number on the top surface) "Placement", and sample marking depth "Mark depth" (0.5 mm—A or 0.1 mm—B), along with the target property, sample density D [g/cm^3]. The only difference in our preprocessing pipeline was the replacement of the simple encoding of the "Mark depth" feature with One-Hot Encoding (the feature values were converted into binary vectors: (1, 0) for A and (0, 1) for B), which introduced an additional column into the dataset.

The ontology was augmented with material property information for H13 steel, including theoretical density, specific heat capacity, thermal conductivity, melting point, and thermal diffusivity, while the knowledge graph was populated with the results of ML model training experiments conducted on the base features. For model training, we adopted the same set of models as in the original study: (1) Support Vector Regression (SVR), (2) Kernel Ridge Regression (KRR), (3) Random Forest Regression (RFR), (4) Stochastic Gradient Descent (SGD), (5) Extreme Gradient Boosting (XGBoost), (6) Extreme Gradient Boosting with limited depth (XGBoost LD), (7) Extra-Trees Regressor (ETR), and (8) Light Gradient Boosting Machine (LightGBM). To identify the best hyperparameters for models (3) - (8), we also utilized the Fast Library for Automated Machine Learning & Tuning (FLAML) framework, which autonomously optimizes hyperparameters for ML models beyond predetermined ranges [21]. The best hyperparameters for KRR and SVR models were obtained using grid search method on the predefined ranges of hyperparameters values with 5-fold CV. A thorough evaluation of the trained models for porosity prediction was conducted using R^2 , MAPE, MAE and MSE metrics.

To populate the knowledge base, we selected several articles dedicated to PBF-LB parameter optimization for various materials [22, 23, 24, 25] and, using the Paper Information Extraction Agent, generated a folder of structured JSON files. Given the structured nature of these files, an alternative storage format could be a database enabling filtering by material type, target property for optimization, and other relevant criteria. Subsequently, the Feature Engineering Agent (based on OpenAI chatgpt 3.5 turbo model), taking as input the JSON files, the ontology file, and the KG file, generated novel candidate features. In our experiment, the agent proposed seven new features (Table 1 include only features with confidence > 0.85). However, most of these features exhibited high correlation with each other and with the base features, primarily because the underlying variables required for their computation - such as spot diameter, layer thickness, and material properties (e.g., thermal diffusivity) - were not varied in the laser parameter experiments and thus were absent from the dataset. Consequently, after correlation-based filtering, the Symbolic Feature Generator retained only one novel feature, *ThermalDiffusivityAdjustedLED*, while discarding the base feature LED. The results of model training on the new feature set and the percentage changes in evaluation metrics relative to the baseline are illustrated in Table 2.

9. Discussion

The introduction of the *ThermalDiffusivityAdjustedLED* feature yielded a distinctly heterogeneous response across the evaluated ML models, revealing important insights into both the nature of the descriptor and the sensitivity of different algorithms to domain-informed feature engineering.

The most pronounced positive effect was observed for the SGD model, which demonstrated substantial improvements across all metrics: R^2 increased by +103.00%, while MAE, MSE, and MAPE decreased

Table 1

Features proposed by the Feature Engineering Agent.

Feature Name	Formula	Rationale
Thermal Diffusivity Adjusted LED	$\frac{P \cdot \alpha}{v \cdot d}$	Adjusts linear energy density by thermal diffusivity, accounting for heat spreading through powder bed and substrate, affecting melt pool dimensions, cooling rates, and porosity.
Diffusion- Normalized Line Energy	$\frac{P}{v \cdot \sqrt{\alpha \cdot t}}$	Relates line energy (P/v) to thermal diffusion length and layer thickness, approximating how much energy per unit track is available relative to the material's ability to diffuse heat and the required fusion depth.
Scan Speed Over Thermal Diffusivity	$\frac{v}{\alpha}$	Direct measure of cooling rate relative to thermal diffusivity; higher values lead to faster cooling, affecting hardness and residual stress.
Combined Heat Input Power Density	$\frac{P}{v \cdot h \cdot t} \cdot \frac{P}{\pi(d/2)^2}$	Combines VED and power density, accounting for total heat input and spatial concentration, determining melt pool stability and defect propensity.
Laser Power To Hatch Times Layer Thickness Ratio	$\frac{P}{h \cdot t}$	Effective energy input per unit cross-sectional area of powder bed, influencing melting efficiency and porosity.

Note: Only features with a confidence score above 0.8 are shown; t denotes layer thickness, P - laser power, v - scan speed, d is spot diameter, α - thermal diffusivity.

Table 2

Model performance metrics on the test set.

Model	R^2	MAE	MSE	MAPE	ΔR^2 (%)	Δ MAE (%)	Δ MSE (%)	Δ MAPE (%)
ETR	0.974	0.018	0.001	0.002	-0.10%	0.000%	0.000%	0.000%
KRR	0.971	0.022	0.001	0.003	0.000%	0.000%	0.000%	0.000%
LightGBM	0.974	0.019	0.001	0.002	0.000%	0.000%	0.000%	-33.33%
RFR	0.973	0.018	0.001	0.002	0.000%	0.000%	0.000%	0.000%
SGD	0.228	0.115	0.021	0.015	+103.00%	-68.92%	-90.99%	-68.75%
SVR	0.868	0.051	0.004	0.007	0.000%	0.000%	0.000%	0.000%
XGBoost LD	0.972	0.019	0.001	0.003	-0.10%	0.000%	0.000%	0.000%
XGBoost	0.972	0.018	0.001	0.002	+0.21%	-10.00%	0.000%	-33.33%

Green cells indicate the best values of test metrics and their improvement (positive change for R^2 , negative change for error metrics). Red cells indicate the worst values of test metrics and their degradation (negative change for R^2 , positive change for error metrics).

by 68.92%, 90.99%, and 68.75%, respectively. This dramatic enhancement suggests that SGD, being a linear model sensitive to feature scaling and relevance, benefited significantly from the physically meaningful normalization introduced by the new descriptor. Similarly, XGBoost showed moderate improvements ($\Delta R^2 = +0.21\%$, Δ MAE = -10.00%, Δ MAPE = -33.33%), indicating that even tree-based ensemble methods, which are generally robust to irrelevant features, can leverage well-constructed physics-informed descriptors to refine their predictions. Interestingly, LightGBM exhibited a notable reduction in MAPE (-33.33%) with no change in other metrics, implying that the new feature primarily improved prediction stability on specific data subsets rather than overall accuracy.

Several models - KRR, SVR, RFR, and, to a large extent, ET and XGBoost LD showed negligible or zero change across all metrics. This invariance can be attributed to several factors. First, KRR and SVR are kernel-based methods that operate in high-dimensional feature spaces where the marginal contribution of a single feature may be diluted, especially if the kernel function implicitly captures similar nonlinear interactions. Second, RFR and ET are ensemble methods that rely on random subspace

selection; if the new feature was not frequently selected during tree splitting due to high correlation with existing features or low information gain, its impact would remain minimal. Third, the hyperparameter configurations used for these models — particularly the fixed depth constraints in XGBoost LD may have restricted the model’s capacity to exploit additional feature information, resulting in unchanged performance. A slight degradation observed for ET and XGBoost LD suggests that in some cases, the new feature introduced noise or collinearity that slightly impaired model generalization. This marginal decrease, however, is within the range of statistical variance and does not constitute a significant deterioration.

Feature importance analysis confirms the performance results obtained across all evaluated models. For the majority of models (KRR, SVR, LightGBM, RF and ET) the importance of the newly introduced *ThermalDiffusivityAdjustedLED* feature remained virtually unchanged compared to the importance of the removed base feature LED, as assessed by both permutation importance and SHAP methods. This indicates that the new descriptor effectively preserves the predictive information contained in the original LED while introducing additional physical context. However, a notable exception was observed for LightGBM, where the introduction of the new feature was accompanied by a substantial increase in the importance of the VED feature. This suggests that *ThermalDiffusivityAdjustedLED* may act synergistically with VED in the LightGBM model, enhancing the model’s ability to capture complementary aspects of energy input and thermal diffusion.

For the XGBoost and XGBoost LD models, the importance of the new feature decreased slightly relative to the original LED (by approximately 1%), while the importance of VED increased by a roughly equivalent margin. This trade-off implies that the two features share overlapping predictive information and that the model redistributes attention between them without substantial net gain or loss in overall performance.

The most pronounced effect was observed for SGD, where the importance of the new feature not only increased substantially compared to the removed LED (by over 100%) but was also accompanied by a several-fold increase in the importance of all other features. This suggests that the introduction of a physically normalized descriptor fundamentally improved the feature space for the linear SGD model, enabling it to better exploit the entire set of input variables and leading to the dramatic performance improvements reported in Table 2.

10. Conclusions and Results

This experimental study proposed and validated a methodology that integrates an ontological approach, KG construction, and agent-oriented feature generation for predicting material properties in additive manufacturing. The developed framework encompasses stages of extracting structured information from scientific literature, enriching the ontology with physico-mechanical material characteristics, and automatically generating new descriptive features based on physical process models. The effectiveness of the proposed approach was demonstrated on the task of predicting steel sample density in PBF-LB process, employing eight machine learning models: SVR, KRR, RFR, SGD, XGBoost, XGBoost LD, ETR, and LightGBM. A hybrid hyperparameter optimization strategy, combining automated tuning via the FLAML library with deterministic grid search for kernel-based methods, ensured robust configuration of all models. For high-value PBF-LB applications that use expensive or scarce alloy powders, such as metastable titanium alloys, nickel-based superalloys, or steels with specialized functional properties, the ability to reuse accumulated process knowledge via the ontology and KG has the potential to reduce the number of trial builds required for process qualification, thereby lowering development costs and material consumption.

The conducted experiment revealed that while the agent can propose novel physically grounded features, their utilization in ML models may be constrained due to high mutual correlation. This limitation arises because the generated features may require information about parameters that were not varied in the experimental dataset, such as spot diameter and layer thickness. This finding highlights a key limitation of the proposed methodology - its effectiveness is directly depending on the completeness

of the input data and the diversity of varied process parameters.

The feature importance analysis, embedded within the proposed framework, can be utilized to evaluate the derived descriptors and inform decisions regarding the inclusion of new descriptors in the process ontology, thereby providing additional physical context.

The proposed framework demonstrates the potential for automated construction of physically grounded features for materials science, yet its practical applicability faces several constraints. The effectiveness of descriptor generation depends on the completeness of the ontological description, and correlation-based filtering may discard useful features lacking an immediate linear relationship with the target.

Future work should expand the ontological base with multi-material databases, thermophysical models, and melt pool dynamics parameters, while also developing a confidence assessment methodology, such as a composite Hypothesis Score integrating Physics, Literature, and Data utility metrics, to objectively quantify the reliability of LLM-generated descriptors. Additionally, implementing a dynamic database with filtering by material type and target property could enhance the system's utility for real-time process optimization.

Acknowledgments

Dr Bartłomiej Wysocki would like to acknowledge the Polish Ministry of Science and Higher Education, which financially supported this work with the grant “New Polish Metallic Materials for Additive Manufacturing Technology (3DPrinting)” (Grant No. MEiN/2022/DPI/1064) and the project “Towards Innovations” within the “Excellent Science” Program (DNK/SN/550991/2022).

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly v.1.173.2.0 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] G. Tapia, A. Elwany, H. Sang, Prediction of porosity in metal-based additive manufacturing using spatial gaussian process models, *Additive Manufacturing* 12 (2016) 282–290. doi:10.1016/j.addma.2016.05.009, special Issue on Modeling & Simulation for Additive Manufacturing.
- [2] G. Barrionuevo, J. Ramos-Grez, M. Walczak, et al., Comparative evaluation of supervised machine learning algorithms in the prediction of the relative density of 316l stainless steel fabricated by selective laser melting, *International Journal of Advanced Manufacturing Technology* 113 (2021) 419–433. doi:10.1007/s00170-021-06596-4.
- [3] M. Gor, A. Dobriyal, V. Wankhede, P. Sahlot, K. Grzelak, J. Kluczyński, J. Łuszczek, Density prediction in powder bed fusion additive manufacturing: Machine learning-based techniques, *Applied Sciences* 12 (2022) 7271. doi:10.3390/app12147271.
- [4] M. Kuehne, K. Bartsch, B. Bossen, et al., Predicting melt track geometry and part density in laser powder bed fusion of metals using machine learning, *Progress in Additive Manufacturing* 8 (2023) 47–54. doi:10.1007/s40964-022-00387-3.
- [5] S. Eshkabilov, I. Ara, F. Azarmi, A comprehensive investigation on application of machine learning for optimization of process parameters of laser powder bed fusion-processed 316l stainless steel, *International Journal of Advanced Manufacturing Technology* 123 (2022) 2733–2756. doi:10.1007/s00170-022-10331-y.
- [6] N. Hasan, M. H. Rahman, A. Wessman, T. Smith, M. Shafae, Process defects knowledge modeling in laser powder bed fusion additive manufacturing: An ontological framework, *Manufacturing Letters* 35 (2023) 822–833. doi:10.1016/j.mfglet.2023.08.132, supplement.

- [7] B. M. Shahriar, M. H. Rahman, A knowledge-driven llm-based decision-support system for explainable defect analysis and mitigation guidance in laser powder bed fusion, arXiv (2026). URL: <https://arxiv.org/abs/2605.01100>, arXiv:2605.01100.
- [8] A. P. Madathil, O. Bylya, A. Sherlock, A physics informed machine learning framework towards density prediction of additively manufactured components, *Key Engineering Materials* 1047 (2026) 73–82. doi:10.4028/p-vvwf5p.
- [9] Z. Li, M. Huang, Y. Zhong, Y. Qin, A description logic based ontology for knowledge representation in process planning for laser powder bed fusion, *Applied Sciences* 12 (2022) 4612. doi:10.3390/app12094612.
- [10] textract developers, textract, <https://textract.readthedocs.io/en/stable/>, 2026. Accessed: 2026-09-10.
- [11] W3C OWL Working Group, Owl 2 web ontology language: Document overview, W3C Recommendation, 2012. URL: <https://www.w3.org/OWL/>.
- [12] A. Dong, D. Du, M. tour sangsaraki, H. She, L. Lei, B. Sun, Overlap ratio and scanning strategy effects on laser powder bed fusion ti6al4v: numerical thermal modeling and experiments, *The International Journal of Advanced Manufacturing Technology* 125 (2023) 1–15. doi:10.1007/s00170-023-10900-9.
- [13] J. Ye, S. A. Khairallah, A. M. Rubenchik, M. F. Crumb, G. Guss, J. Belak, M. J. Matthews, Energy coupling mechanisms and scaling behavior associated with laser powder bed fusion additive manufacturing, *Advanced Engineering Materials* 21 (2019) 1900185. doi:10.1002/adem.201900185.
- [14] Review on modeling techniques for powder bed fusion processes based on physical principles, *Additive Manufacturing* 47 (2021) 102336. doi:<https://doi.org/10.1016/j.addma.2021.102336>.
- [15] K. Kosarava, P. Widomski, M. Ziętała, D. Dobras, M. Muzyk, B. Wysocki, Machine learning-assisted optimisation of the laser beam powder bed fusion (pbf-lb) process parameters of h13 tool steel fabricated on a preheated to 350 °c building platform, *Materials* 19 (2026) 210. doi:10.3390/ma19010210.
- [16] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597, model available at: <https://huggingface.co/BAAI/bge-small-en-v1.5>.
- [17] SymPy Development Team, Sympy: Python library for symbolic mathematics, 2025. URL: <https://docs.sympy.org/latest/index.html>.
- [18] M. Stone, Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (1974) 111–147.
- [19] L. Breiman, Random forests, *Machine Learning* 45 (2001) 5–32. doi:10.1023/A:1010933404324.
- [20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
- [21] Microsoft, Flaml: Fast library for automated machine learning & tuning, 2025. URL: <https://microsoft.github.io/FLAML/>, accessed on 31 December 2025.
- [22] Y. Liu, J. Li, T. Cheng, Z. Fan, W. Li, W. Xia, Q. Wei, Parameter automatic optimization strategy for laser powder bed fusion using neural network infrared radiation intensity prediction model, *Additive Manufacturing* 92 (2024) 104373. doi:10.1016/j.addma.2024.104373.
- [23] M. H. Mosallanejad, H. Gashmard, M. Javanbakht, B. Niroumand, A. Saboori, A novel feature engineering approach for predicting melt pool depth during lpbf by machine learning models, *Additive Manufacturing Letters* 10 (2024) 100214. doi:10.1016/j.addlet.2024.100214.
- [24] A. Paraschiv, G. Matache, M. Condruz, T. Frigioescu, L. Pambaguian, Laser powder bed fusion process parameters' optimization for fabrication of dense in 625, *Materials* 15 (2022) 5777. doi:10.3390/ma15165777.
- [25] D. Inam, I. Ali, A. A. S. Syed, I. U. Ahad, A review of machine learning methods for process parameter optimization in laser powder bed fusion, in: *Proceedings of the 4th International Conference on Key Enabling Technologies (KEYTECH 2024)*, Atlantis Press, 2024, pp. 302–311. doi:10.2991/978-94-6463-602-4_40.