msamDB: Towards addressing data-scarcity challenges in PBF-LM additive manufacturing.

Jigar Patel[1], Chris Vuong[1], Mihaela Vlasea[1,*], Tamer Özsu[2]

[1] Multi-Scale Additive Manufacturing Lab, University of Waterloo, Waterloo, Canada N2L 6V3

[2] Data Systems Research Group, University of Waterloo, Waterloo, Canada N2L 6V3

* mihaela.vlasea@uwaterloo.ca

Abstract: Data science techniques, particularly machine learning (ML), have proven to be valuable tools in PBF-LM research. While ML can rapidly model the large process parameter space of PBF-LM, their efficacy is dependent on large, informative and diverse training datasets. However, scarcity in the development and availability of such datasets is an on-going challenge. This work outlines the on-going progress to address this challenge through the development of a database platform, tentatively named msamDB (multi-scale additive manufacturing database). This platform, specifically created to manage PBF-LM academic research data, is a modular, extensible and scalable database that can promote data-sharing among researchers. The initial architecture of msamDB focuses on surface roughness data generated throughout the PBF-LM lifecycle. This work highlights the findings and challenges encountered in the design, implementation and pilot data population stages of msamDB. In its current stage, msamDB data spans data from approximately 30 builds, multiple research and industry studies, 3 different powder materials and a broad range of process parameters. Data has been collected from various stages such as powder characterization, build planning, process parameter selection, surface characterization, etc. In reference to surface roughness measurements, the database currently has more than 1000 data points across various surface orientations. This work represents first known effort to curate research PBF-LM data at scale for PBF-LM. The potential impact of such a database is to promote federated data for PBF-LM researchers, which allows for data-driven model development to have increased usability.

Keywords: data management, data scarcity, PBF-LM database, relational data

1. Introduction

Product variability remains a significant concern in additive manufacturing (AM) and in laser powder bed fusion (PBF-LM), specifically [1]. Machine learning (ML) approaches have gained significant traction in PBF-LM research for fast exploration of the process space. ML models complement the high-fidelity but costly experimentation and simulation approaches. However, the effectiveness of ML and other data-driven models depends on data availability, and there is a scarcity of large, diverse and usable data. This is illustrated in Figure 1, where selected datasets [2–16] are compared for their size (i.e. rows), dimensionality (i.e. columns) and variety (number of builds, materials, machines, etc. used). The dataset composition shown in Figure 1 (a) is used to create a data variety score for Figure 1 (b) (normalized between 0-6, where 6 indicates highest variety). The red line depicts the rule-of-thumb which proposes 100 datapoints (rows) for each dimension (columns) [17]. It is clear that most dataset sizes have high dimensionality but smaller size and variety.

While there is precedence in data management efforts for AM data [1], there are few working implementations in the research domain. This work is a scalable, extensible implementation specifically for heterogenous research data.

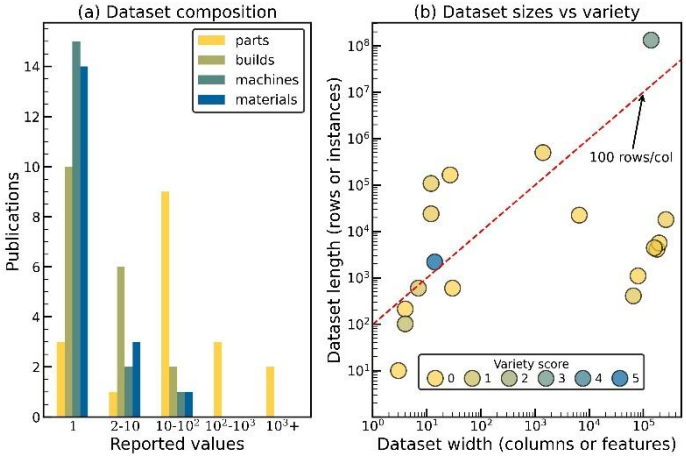

Figure 1: Visualized data scarcity in datasets reported in literature. (a) Shows the composition of datasets in terms of parts, builds, machines and materials. (b) Scatterplot indicating relationship between dataset length (rows – Y axis) and width (columns - X axis)

2. Materials and methods

As illustrated in Figure 2, this work demonstrates a framework which facilitates data aggregation from different stages of PBF-LM lifecycle. The second stage of the framework systematically indexing data at relevant scales. To ensure a high level of trust and reduce data cleaning burden in downstream analytics efforts, msamDB also includes a comprehensive validation layer, which conducts rigorous rule-based and statistical testing of incoming datapoints.

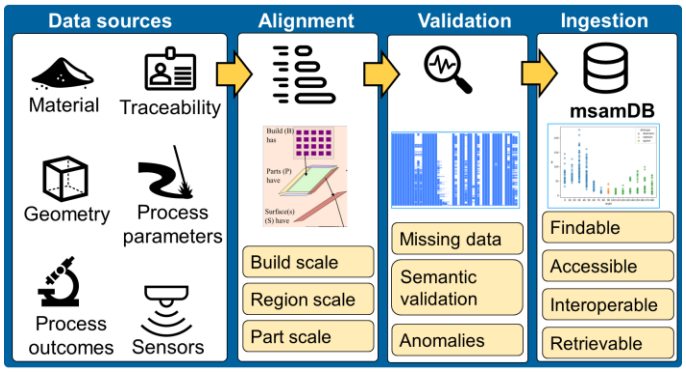

Figure 2: Graphical illustration of data ingestion workflow for the database.

2.1 Modelling multi-scale PBF-LM data

An entity relationship (ER) diagram was constructed for a preliminary understanding of the multi-scale nature of PBF-LM data. ER diagrams allow for a high degree of abstraction to the database development process to conceptualize relationships between different data entities. Figure 3 shows an excerpt from the ER diagram (full diagram excluded for brevity) that shows the creation of entities such as "parts", "measurements" and "measurement conditions", which are linked by specific relationships. The developed database is postgres instance on a Linux server (Ubuntu OS). Postgres was chosen as the database platform due to significant community support and open-source nature.

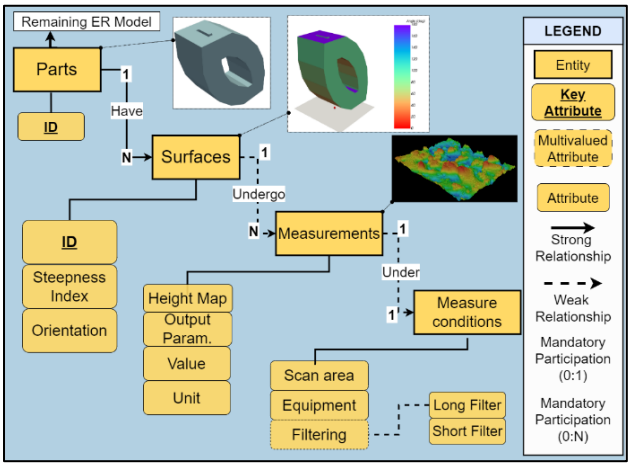

Figure 3: An excerpt from the ER diagram of the database, showing PBF-LM entities such as parts and surfaces

2.2 Data indexing and validation layers

By leveraging the relationships between entities as shown in Figure 3, data from different sources is first linked to a unique scaled based identifier. For example, data pertaining to a specific surface is indexed to a unique "region_id" ("regions" formulated as a generalization of "surfaces" in the database.). Next, data is passed through a validation layer. Selected examples of data validation scenarios are given in Table 1, along with potential outcomes:

Table 1: Examples of data validation scenarios implemented for ensuring baseline levels of data quality

Test scenario	Success	Failure
Candidate datapoint has high probability of anomaly	Safe to insert	Warning
Value (Power) with erroneous/missing unit (Watts)	Safe to insert	Error – reject insertion
"Region" data inserted with a missing link to a part	Safe to insert	Error – reject insertion
Part data inserted without any process outcomes	Safe to insert	Warning

3. Results and discussion

For demonstrating the efficacy of the msamDB, surface roughness data was extracted from 10 different PBF-LM builds printed across two machines, a continuous laser system (EOS M290) and a modulated laser system (Renishaw AM400). The data encompasses different part geometries, three different ferrous alloys and surfaces with different orientation angles. Roughness measurements in the form of height maps and areal roughness parameters were extracted. Surface roughness measurements were collected using a confocal laser profilometer. Process parameters data was extracted from templated process parameter sheets and powder properties (thermal and morphological) was extracted from material sheets. These data were extracted from builds not intended for this work, hence simulating conditions where heterogenous data with variation in structure and quality is encountered. For reference, Table 2 includes baseline data retrieval metrics for two sample queries.

Table 2: Baseline retrieval information for a simple query (single table) and joined query (two related tables) from the database

Query type	Query description	Data rows	Planning time [ms]	Execution time [ms]
Simple	Fetch all parts from builds = [X,Y]	528	0.032	0.058
Joined	Fetch all parts and their linked (roughness) data	1043	0.087	0.336

We also demonstrate the data validation layer where incoming data is assessed baseline quality. For brevity, we illustrate through one example wherein two candidate values of laser power are evaluated against the existing distribution. This is visualized in Figure 4. Given the distribution of the datapoints already in the database, the probability of the candidates being an outlier is computed. Once again, this evaluation will lead to an ingestion with warning, depending on the threshold for "probability of an outlier" chosen. Coupling the database with such built-in statistical evaluation processes can help maintain baseline quality of data added to the database, contributing to the need for improved data quality ML [18].

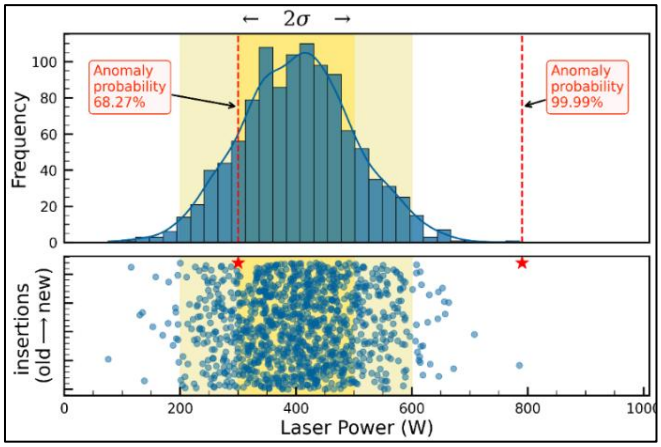

Figure 4: Visualization of the statistical monitoring of data quality. Two candidate datapoints (star markers) with the distribution of the current datapoints (blue), and the probability of candidate data point being an anomaly is annotated. (σ = standard deviation)

4. Conclusions

This work shows prelimnary development of a novel data management solution that can help address a longstanding challenge with ML based approaches: data scarcity. This work shows that part-scale dataset size and variety can be achieved with research data, with access to a shared data platform which automatically evaluates incoming data to assure baseline data quality. msamDB can contribute to several data-driven efforts such as those illustrated in Figure 5. By integrating a data quality evaluation module with the database, we propose that it can facilitate scaling of data ingestion from 30 builds to hundreds of builds collected by different researchers. The authors welcome collaboration opportunities to demonstrate benefits of data sharing. The current limitations our work are: (a) only surface roughness data ingested as process outcome, (b) only prelimnary architecture to ingest high volume temporal sensor data (e.g. photodiode streams) and (c) development of a user-interface (UI) to facilitate smoother data sharing. The authors hope to address (c) as future work. While (a) was intentionally chosen to manage scope of current work, the relational model proposed was designed to be easily adapted for other process outputs. (e.g. density related to part data). Finally, the authors hope to address (c) as future work by developing a prelimnary UI for data interaction.

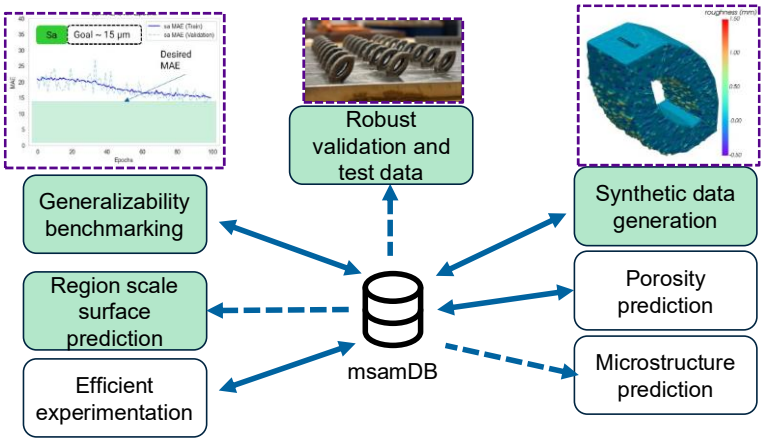

Figure 5: Examples of data science applications possible with current (highlighted in green) and future extensions of msamDB.

5. Acknowledgements

Authors acknowledge funding support from NSERC (CGS-D), Federal Economic Development Agency for Southern Ontario, grant #814654, Mohsen Keshvarz, Jerry Ratthapakdee and other members of MSAM lab for their support with data curation.

6. References

[1] Prater T. Database development for additive manufacturing. Prog Addit Manuf. 2017 Jun;2(1–2):11–8.
[2] Lapointe S, Guss G, Reese Z, Strantza M, Matthews MJ, Druzgalski CL. Photodiode-based machine learning for optimization of laser powder bed fusion parameters in complex geometries. Addit Manuf [Internet]. 2022;53.
[3] Bao H, Wu S, Wu Z, Kang G, Peng X, Withers PJ. A machine-learning fatigue life prediction approach of additively manufactured metals. Eng Fract Mech [Internet]. 2021;242.
[4] Kappes B, Moorthy S, Drake D, Geerlings H, Stebner A. Machine Learning to Optimize Additive Manufacturing Parameters for Laser Powder Bed Fusion of Inconel 718. In: Ott E, Liu X, Andersson J, Bi Z, Bockenstedt K, Dempster I, et al., editors. Proceedings of the 9th International Symposium on Superalloy 718 & Derivatives: Energy, Aerospace, and Industrial Applications. Cham: Springer International Publishing; 2018. p. 595–610. (The Minerals, Metals & Materials Series).
[5] Gobert C, Reutzel EW, Petrich J, Nassar AR, Phoha S. Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. Addit Manuf. 2018 May 1;21:517–28.
[6] Gobert C, Kudzal A, Sietins J, Mock C, Sun J, McWilliams B. Porosity segmentation in X-ray computed tomography scans of metal additively manufactured specimens with machine learning. Addit Manuf. 2020 Dec 1;36:101460.
[7] Scime L, Siddel D, Baird S, Paquit V. Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: A machine-agnostic algorithm for real-time pixel-wise semantic segmentation. Addit Manuf. 2020 Dec;36:101453.
[8] Gaikwad A, Williams RJ, de Winton H, Bevans BD, Smoqi Z, Rao P, et al. Multi phenomena melt pool sensor data fusion for enhanced process monitoring of laser powder bed fusion additive manufacturing. Mater Des. 2022 Sep 1;221:110919.
[9] Koo J, Park E, Baek AMC, Kim N. The Research of Surface Roughness Prediction with Machine Learning According to Process Parameters in Laser Powder Bed Fusion. In 2022. p. 65. (Lecture Notes in Mechanical Engineering).
[10] 1Akbari P, Ogoke F, Kao NY, Meidani K, Yeh CY, Lee W, et al. MeltpoolNet: Melt pool characteristic prediction in Metal Additive Manufacturing using machine learning. Addit Manuf. 2022 Jul;55:102817.
[11] Baumgartl H, Tomas J, Buettner R, Merkel M. A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring. Prog Addit Manuf. 2020;5(3):277–85.
[12] Schwerz C, Nyborg L. A neural network for identification and classification of systematic internal flaws in laser powder bed fusion. CIRP J Manuf Sci Technol. 2022 May;37:312–8.
[13] Schmid S, Krabusch J, Schromm T, Jieqing S, Ziegelmeier S, Grosse CU, et al. A new approach for automated measuring of the melt pool geometry in laser-powder bed fusion. Prog Addit Manuf. 2021 May 1;6(2):269–79.
[14] Shevchik SA, Kenel C, Leinenbach C, Wasmer K. Acoustic emission for in situ quality monitoring in additive manufacturing using spectral convolutional neural networks. Addit Manuf. 2018;21:598–604.
[15] Estalaki SM, Lough CS, Landers RG, Kinzel EC, Luo T. Predicting Defects in Laser Powder Bed Fusion using in-situ Thermal Imaging Data and Machine Learning. Addit Manuf. 2022 Jul 1;103008.
[16] Gao J, Zhu ,Chenyang, Gao ,Shubo, Ji ,Weiming, Xue ,Ming, and Zhou K. Generative adversarial network–enabled microstructural mapping from surface profiles for laser powder bed fusion. Virtual Phys Prototyp. 2025 Dec 31;20(1):e2499925.
[17] Jain AK, Chandrasekaran B. 39 Dimensionality and sample size considerations in pattern recognition practice. In: Handbook of Statistics [Internet]. Elsevier; 1982 [cited 2025 Jul 3]. p. 835–55. (Classification Pattern Recognition and Reduction of Dimensionality; vol. 2).
[18] Xie J, Sun L, Zhao YF. On the Data Quality and Imbalance in Machine Learning-based Design and Manufacturing—A Systematic Review. Engineering. 2025 Feb 1;45:105–31.