Ref. Ares(2024)7899202 - 07/11/2024

D2.2 BENCHMARKING OF DIGITAL TECHNOLOGIES FOR MATERIALS MODELLING, PROCESSING AND MANUFACTURING

08/11/2024

Grant Agreement No.: 101091496
Call: HORIZON-CL4-2022-RESILIENCE-01
Topic: HORIZON-CL4-2022-RESILIENCE-01-25
Type of action: HORIZON Innovation Actions

D2.2 Benchmarking of Digital Technologies for Materials Modelling, Processing and Manufacturing

Grant agreement number	101091496	Acronym	DiMAT
Full title	Digital Modelling and Simulation for Design, Processing and Manufacturing of Advanced Materials		
Start date	01/01/2023	Duration	36 months
Project url	[HTTPS://CORDIS.EUROPA.EU/PROJECT/ID/101091496](https://cordis.europa.eu/project/id/101091496)		
Work package	WP2 - NEED: Industrial Scenarios and Requirements Analysis		
Deliverable	D2.2 – Benchmarking of Digital Technologies for Materials Modelling, Design, Processing and Manufacturing		
Task	T2.2 - Benchmarking of Digital Technologies for Materials Modelling, Design, Processing and Manufacturing.		
Due date	31/05/2023		
Submission date	08/11/2024		
Nature	Report	Dissemination level	Public
Deliverable lead	5-NTUA		
Version	1.1		

Authors	Tsitseklis Konstantinos (NTUA), Georgia Stavropoulou (NTUA)
Contributions	Stefanos Vrochidis (CERTH), Anastasios Karakostas (DRAXIS) Pablo de Andres, Kuo-I Chang (FRAUNHOFFER) Luca Canetta (SUPSI) Harrison de la Rosa Ramírez (UPV)
Reviewers	4- AITEX, Ignacio Rico, 1- CERTH, Yannis Papadimitriou
Abstract	Deliverable 2.2 aims to present the gathered knowledge about digital technologies relevant to the goals of the DiMAT project focusing on materials modelling processing and manufacturing. Information was gathered from specialized project partners. A multi-dimensional benchmarking approach was followed characterizing each technology and more specifically available and popular tools. Moreover, the state-of-the-art for each technology was surveyed from academic and professional literature.
Keywords	DIGITAL TECHNOLOGIES, BENCHMARKING, EVALUATION, STATE-OF-THE-ART, TOOLKITS, PERFORMANCE METRICS

Document Revision History

Version	Date	Description of change	List of contributor(s)
0.1	10-Feb-2023	ToC	NTUA
0.2	21-Feb-2023	Added content for Section 3	NTUA
0.3	01 -Mar-2023	Revised ToC	NTUA
0.4	10-Mar-2023	Added content for Section 4	NTUA
0.5	20-Apr-2023	Added content for Sections 5,6,7	NTUA, CERTH, DRAXIS, FRAUNHOFER, UPV, SUPSI
0.6	15-May-2023	Version submitted for internal review	NTUA
0.7	22-May-2023	Internal Review	CERTH, AITEX, UPV
0.8	24-May-2023	1st Draft addressing the comments	NTUA

		from internal reviewers	
1.0	31-May-2023	Final Quality Check and issue of Final Document	CERTH
1.1	08-Nov-2024	Update of deliverable according to PO/reviewers comments and quality check	NTUA, CERTH

DISCLAIMER

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

COPYRIGHT NOTICE

EXECUTIVE SUMMARY

Nowadays the technological landscape is constantly shifting. Novel tools are being introduced every day, making feasible the development of new applications, while, at the same time, older and previously established platforms are losing momentum. Assisting in the selection of the best tools for the design and development of DiMAT suites and toolkits, this deliverable presents the current state of various technological fields relevant to the project's scope in general but with emphasis on the materials design, manufacturing and simulation. The examined sources were both academic and professional in nature and were gathered from numerous project partners.

The technologies which were pinpointed as the most relevant to the DiMAT's objectives were material ontologies and knowledge graphs, digital twins, cost and lifecycle assessment techniques, machine learning technologies (reinforcement learning, recommender systems, etc.), big data analytics and storage, and finally, materials design and simulation platforms. Besides the state of the art for each of the above, relevant European-funded projects are discussed and relevant tools are presented focusing on open-source solutions.

This deliverable proposes an evaluation and benchmarking framework per technology that helps define its level of readiness as an overall evaluation of the respective available open-source tools. These tools are evaluated based on the computing requirements they impose, the availability of the source code, the activity of their respected community evaluated through GitHub/Gitlab metrics, YouTube available seminars, workshops, etc. and their adoption by the industry. Moreover, for every technology mentioned, specific technology-related metrics, on which the tools can be evaluated, are proposed. Finally, this deliverable also acts as an indicative recommendation guide to the partners of the DiMAT project's toolkits by matching the evaluated tools to specific toolkits.

TABLE OF CONTENTS

Co-funded by
the European Union

LIST OF TABLES

Co-funded by the European Union

ABBREVIATIONS

ACID	Atomicity, consistency, isolation, durability
AI	Artificial Intelligence
AMQP	Advanced Message Queuing Protocol
AWS	Amazon Web Services
API	Application Programmable Interface
AQL	ArangoDB Query Language
ARIMA	Autoregressive Integrated Moving Average
ASE	Atomic Simulation Environment
BDA	Big Data Analytics
BILSTM	Bidirectional Long-Short Memory
CAD	Computer Aided Design
CE	Circular Economy
CFD	Computational Fluid Dynamics
CMDB	Configuration Management DataBase
CNN	Convolutional Neural Network
CPS	Cyber-Physical Systems
CRF	Conditional Random Field
DA	Data Analytics
DB	DataBase
DBMS	DataBase Management System
DIH	Digital Innovation Hubs
DT	Digital Twin
EMMO	Elementary Multi-perspective Material Ontology
EoL	End-Of-Life

EU	European Union
FEM	Finite Element Method
GDPR	General Data Protection Regulation
GRU	Gated Recurrent Unit
HDFS	Hadoob Distributed File System
HTTP	Hypertext Transfer Protocol
IIoT	Industrial Internet-of-Things
ISO	International Standardisation Organisation
JSON	JavaScript Object Notation
KAF	Kaltura Application Framework
KG	Knowledge Graph
LCA	Life Cycle Assessment
LCI	Life Cycle Inventory
LCIA	Life Cycle Impact Assessment
LSTM	Long Short-Term Memory
MaaS	Manufacturing as a Service
MCDA	Multi Criteria Decision Analysis
MFA	Material Flow Analysis
MDO	Materials Design Ontology
ML	Machine Learning
MMKG	Metallic Materials Knowledge Graph
MPS	Material Processing Simulator
MQTT	Message Queuing Telemetry Transport
MSP	Manufacturing Simulation Platform
NER	Named Entity Recognition
NLP	Natural Language Processing

NN	Neural Network
OS	Operational System
OWL	Web Ontology Language
PCA	Principal Components Analysis
PLC	Programmable Logic Controller
PSP	Processing Structure Properties (analysis)
RDBMS	Relational DataBase Management System
RDF	Resource Description Framework
RL	Reinforcement Learning
RNN	Recurrent Neural Network
ROM	Reduced Order Model
RS	Recommender System
RVM	Resource Value Mapping
SME	Small Medium Enterprise
SPARQL	SPARQL Protocol and RDF Query Language
STOMP	Streaming Text Oriented Message Protocol
SVM	Support Vector Machine
SQL	Structured Query Language
VIMMP	Virtual Material Market Place
WoT	Web of Things
XAI	Explainable Artificial Intelligence
ZMDP	Zero Defects Manufacturing Process

1 INTRODUCTION

The aim of this deliverable is to present knowledge on various digital technologies relevant to the goals and objectives of the DiMAT project in a clear and systematized manner. The state-of-the art of these digital technologies is highlighted using as input academic sources as well as professional literature. Moreover, the available (and relevant) tools and platforms for each technology are presented alongside a brief description explaining their functionalities. In order to evaluate both the tools' overall suitability to be employed for the purposes of DiMAT and characterize each technological sector in general, a benchmarking and evaluation framework, based on qualitative metrics, that enables the comparison among the tools and the extraction of useful insights, is proposed. The results of this analysis act as guidelines for the partners that will develop the project's suites and toolkits.

Towards achieving this goal, input from numerous project beneficiaries was gathered capitalizing on each partner's expertise as it is expressed through their scientific background and their involvement in previous related European Union (EU) (co-)funded research projects that have either employed or developed relevant tools through research in the respective fields.

A number of EU (co)-funded research projects are presented for each scientific area analyzed in this deliverable. The purpose of presenting these projects is twofold. First, the number of relevant projects highlights the significance and interest assigned to the area by a funding agency and then, it allows for exploring relevant projects for a potential adoption from the DiMAT partners. It is important to note that not all the EU funded projects results took part in the evaluation and benchmarking process. This is due to the fact that some of them are still in progress, some are relevant in the area but not specifically for DiMAT and finally, the focus was mainly shifted towards the EU projects in which the partners were more actively involved.

In order to benchmark the tools for each of the examined technologies several aspects were considered. These include the level of social activity surrounding the communities built around the tools, the infrastructural constraints expressed as the level of computing resources required for the deployment of each tool, the availability of the source-code that allows for modifications and finally, as an important aspect, the level of adoption of each tool by the industry sector. It is noteworthy, that although the same benchmarking tools were applied across for the tools of all scientific areas examined, the importance assigned to them may vary. These decisions are explained in the relevant subsections of each of the following chapters. Based on how the tools score across these metrics, an indication of the status of each area can be extracted.

In addition, aiming to assist the developers of the project's toolkits and suites, for each technology, relevant metrics are pinpointed based on which they can appropriately choose between the available tools. All of the above, coupled with the partners' experience and expertise, and following a conversation among toolkit developers leads to the provision of some initial recommendations, by suggesting the use of tools that rank higher across the evaluation metrics for the development of the relevant DiMAT tools.

The rest of this document is structured as follows. Each chapter is dedicated to a relevant digital technology to DiMAT project. Initially, there is a brief presentation of the current state-of-the-art, targeted towards the purposes of materials modeling and manufacturing, followed by a presentation of relevant EU (co-)funded projects. Then, existing solutions with focus on open-source, are presented and an evaluation of these solutions across several dimensions is performed. Moreover, suitable metrics that will aid in the evaluation of each tool are discussed. Finally, a pairing between the most popular tools to tasks related to the DiMAT toolkits functions is performed. After the analysis of each of the presented technologies the document concludes with a short paragraph summarizing the work.

2 KNOWLEDGE GRAPHS AND ONTOLOGIES

2.1 STATE OF THE ART ANALYSIS

Manufacturing and Material Design Industries have shifted their interest towards digitalization to enhance quality and boost speed of production. Therefore, data related to simulations, experiments, calculations and material properties as well as prior knowledge regarding material and manufacturing data encountered in papers, documents and material databases, need to be managed and organized in a unified and coherent way. Towards this goal, material ontologies and knowledge graphs are being developed and leveraged [1]. Materials Knowledge Graphs and Ontologies have emerged to address the need of structured and unified representation of material properties and processes that could contribute to the discovery of hidden correlations between materials and the design of new ones, boosting, in this way, industrial manufacturing.

There is a close relationship between Knowledge Graphs and Ontologies. A Knowledge Graph (KG) is a graph-structure representation of data related to real-word objects, concepts and entities, constituting the nodes of the graph, while also depicting relationships and correlations among them, through edges of the graph. In this way, the semantics of the relevant field of interest are maintained and highlighted, allowing further and deeper exploration of the represented knowledge [2]. A KG is application-driven and, in most cases, stores data related to an underlying Ontology, which is a more general knowledge representation, describing rules, definitions, concepts and relationships of a specific domain in a more formal and abstract way. [4].

Material Ontologies

Elementary Multi-perspective Material Ontology (EMMO) is a project aiming to develop a standard ontology framework, basically related to information and communication technologies and physics, that can be used across different applied sciences [5]. Materials science is one of the fields that has initiated this effort to create a consistent approach to capturing and representing knowledge based on scientific principles and methodologies. Several domain-specific ontologies are being developed using EMMO as the top and middle level ontology [1][5][6][7]. EMMO consists of top-level representations and middle level ontologies that enable the extension of the material ontology to target application fields. Generally, EMMO represents real-world objects as four-dimensional entities that exist in space and time, adopting a physicalistic/nominalistic perspective, being compliant at the same time to the axioms and norms that comprise EMMO's top level [4].

The Materials Design Ontology (MDO) [8] a set of concepts and relationships that pertain to the knowledge of materials science and design, particularly to solid-state physics and condensed matter theory, incorporates the existing EMMO concept "Material", aiming to extend the ontology to sub-domains that EMMO does not cover. MDO is developed in OWL2 DL and designed to represent materials calculation, standardize the publication of materials calculation data and facilitate data exchange among heterogeneous material databases [8] by leveraging a common Application Programmable Interface (API) developed by the project OPTIMADE [9].

Another work, which uses the EMMO as a top-level ontology, is the Virtual Materials Marketplace (VIMMP) [1], a project aiming at developing a marketplace to facilitate the integration of materials modelling services and enable interoperability among them. Towards this goal, OWL2 ontologies employing EMMO as their top-level, have been modelled to coordinate and organize data regarding processes, models and services, represent materials properties in a unified and consistent way and contribute to the user-friendly data management of the platform, allowing users to retrieve data and extend existing ontologies to more targeted use-cases.

Material Knowledge Graphs

Concerning Materials Knowledge Graphs, propnet [10] is an open source extensible and simple Python framework, based on graph representation of complex material properties and their relationships, that extends existing datasets of materials knowledge. The proposed framework not only increases the available information but also facilitates the creation of multifunctional materials, by discovering correlations and similarities between their properties [10]. The software is, also, able to predict physical behaviour and generate feature vectors, assess the accuracy of property relationships and augment training sets with property descriptors that are physically relevant, thereby improving machine learning models in materials design.

In the field of Nanocomposite Materials Science, this paper [11] describes the NanoMine Knowledge Graph. The researchers curated materials data and experimental results into an extensible KG format, allowing scientists to easily explore, retrieve and visualise data concerning properties, relationships and processes of nanocomposite materials, providing at the same time customization. A consistent and structured representation was required, since nanocomposite materials are subject to properties' changes when created or processed. The KG was developed in accordance with PSP (Processing, Structure, Properties) analysis and facilitated by the employment of pre-existing materials ontologies.

This paper [12] presents an approach to construct a Metallic Materials Knowledge Graph (MMKG) based on DBpedia and Wikipedia. The population and extension of the graph is mainly performed through detection and extraction of entities related to metallic materials from DBpedia, by employing semantic distance calculation algorithms. Methods related to materials ontologies were leveraged to detect properties from HTML tables of Wikipedia, as there was inadequate knowledge in DBpedia regarding properties of materials. The evaluation of the project's results was performed by constructing a prototype, calculating machine learning metrics and conducting experiments.

Addressing production purposes, the Bosch Knowledge System for Finding Materials Science Knowledge [13] combines and integrates information from relational databases and documents of materials science into a single Knowledge Graph using the Ontology Data Access technology and Natural Language Processing (NLP) techniques. The system offers standard KG search capabilities, complex query answering facilities and multi-hop reasoning. Through reasoning steps, it is also capable of computing and justifying the origin of query answers, making them explainable.

Finally, an additional ongoing work is presented in [14], in which the development of the material graph database MatKG is described, focusing on storing and coherently representing materials data from heterogeneous sources, by leveraging graph embeddings and machine learning techniques and models, such as transformers and large language models. MatKG enables the prediction of links between materials and properties, facilitates the discovery and design of new materials and applications and allows seamless data retrieval, exploration and visualization. The database incorporates a vast amount of material knowledge, including more than 2 million relationship triads, by processing more than 80K entities derived from over 4 million publications and documents.

2.2 EU RELATED PROJECTS

Several projects, aiming at developing ontologies and Knowledge Graphs for the material and manufacturing domain, have been subsidized by the European Union. The experience gained through these projects helped shaping the knowledge and semantic management concepts that are employed in the project, most prominently in CMDB, where its background IP is linked to large extent to them. Some of them are listed below:

Acronym	Grant Agreement No.	Title	Description
EMMC	723867	European Materials Modelling Council	The European Materials Modelling Council aims at integrating the materials modelling and digitalization critical for more agile and sustainable product development. To this end, an ontology has been developed (EMMO)

			as a semantic framework for the materials design and applied sciences.
OntoCommons	958371	Ontology-driven data documentation for Industry Commons	OntoCommons focuses on standardizing data documentation for the manufacturing and materials domain, using ontologies.
SimDOME	814492	Digital Ontology-based Modelling Environment for Simulation of materials	SimDOME is an industry-ready software framework for materials modelling and industrial applications, facilitating the prediction of materials properties, and materials simulation. It develops an ontology-based simulation platform.
SimPhoNy	604005	Simulation framework for multi-scale phenomena in micro- and nanosystems	SimPhoNy develops an easy-to-use integrated multiscale modelling environment for discovery and design of nano-enabled systems and materials. The SimPhoNy Open Simulation Platform allows visualization

			and exploration of ontologies and manipulation of ontology-based data, among others.
VIMMP	760907	Virtual Materials Market Place	VIMMP is a platform that aims to connect manufacturers with materials modelling activities and resources through an open-source, user-friendly, web-based marketplace. Within VIMMP, a system of marketplace-level ontologies is developed to characterize services, models, and interactions between users.
NanoMECommons	952869	Harmonization of EU-wide nanomechanics protocols and relevant data exchange procedures, across representative cases; standardization, interoperability, data workflow	NanoMECommons intends to create a cross-border and interdisciplinary researching network focused on employing innovative solutions to address the problem of nanomechanical materials characterization across various industries. CHAMEO ontology is developed to contribute to the

			project's goals. In this project the concept of the CHADA for documenting characterization processes has been further developed. A corresponding app also has been introduced, which is planned to be applied in DiMAT.

Table 1: Digital Material Ontologies EU Funded Projects

2.3 EXISTING OPEN-SOURCE SOLUTIONS

Open-Source Solutions	Description
Neo4j	Neo4j is a highly performant, NoSQL, open-source graph database compliant to ACID standards. It is developed in Java, while graph traversals and queries are performed in Cypher Query Language in constant time. Neo4j enables highly flexible and adaptive schema representation, supporting programming in many languages, including Python, Java, JavaScript and .NET.
ArangoDB	ArangoDB is an open-source multi-model graph database, implemented in C++. It uses AQL (ArangoDB Query Language) for graph traversal and data retrieval and offers performance scaling, decreased operational complexity and support for ACID properties.
OrientDB	OrientDB is a multi-model open-source NoSQL DBMS, written in Java, that combines graph models and the flexibility of documents into a scalable database with high speed and performance. It uses Gremlin programming language and SQL to perform queries and offers schema flexibility. It also has a strong security profiling system and is ACID compliant. OrientDB supports a multi-master distributed architecture.
Protégé	Ontology editor that has many plugins allowing the visualization of ontologies. Also, it supports the latest OWL2.0 standard.

OWL API	OWL API is a Java-based framework for working with OWL (Web Ontology Language) ontologies. It provides a set of APIs for creating, modifying, and querying ontologies.
Apache Jena	Apache Jena is a framework developed in Java, aiming at building applications compliant to Link Data and Semantic Web standards.It includes a set of tools and libraries for working with RDF (Resource Description Framework) and OWL ontologies.
NeoDash	NeoDash is an open-source tool for visualizing and processing Neo4j data through interactive and customizable dashboards.
Bloom	A data visualization tool that allows users to easily explore and investigate Neo4j's graph data, without the need for any coding skills.

Table 2: Knowledge Graphs and material ontologies open-source solutions

2.4 BENCHMARKING AND EVALUATION

In order to benchmark the open-source solutions listed in subsection 2.3 we chose the following metrics that showcase the ease of use of each one as well as its current state. First, we examine whether the source code is publicly available or not. If it is, then the statistics from repositories such as GitHub or Gitlab give an indication of the frequency of maintenance of the software (Last commit) as well as the tool's popularity (number of forks and stars). Moreover, an important benchmarking dimension is the existence or not of an active community around each tool. To determine whether the community is active or inactive, the following aspects were considered:

- Existence of YouTube videos (tutorials, advertisements, conference presentations, etc.),
- Questions regarding the tools on sites like Stackoverflow,
- Existing discussions in online repositories.

Additionally, we think that an important metric of evaluating a solution is whether it is used by companies. To determine this, we searched the official websites of the available tools. Finally, the computing requirements for the execution of the software were examined.

A similar logic is followed throughout the document in the relevant subsection for each technology.

Co-funded by the European Union

Solution Name	Publicly available code	Online repository statistics			Active community	Used by industries	Minimum Requirements	
		#Forks	#stars	Last commit			RAM	Proc/or Cores
Neo4j	YES	2.2K	11.2K	March 2023	YES	YES	2GB	2
ArangoDB	YES	808	12.9K	March 2023	YES	YES	1GB	N/A
OrientDB	YES	866	4.6K	March 2023	YES	YES	N/A	N/A
Protégé	YES	223	836	Feb. 2023	N/A	N/A	N/A	N/A
OWL API	YES	309	738	April 2023	N/A	N/A	N/A	N/A
Apache Jena	YES	614	950	April 2023	YES	YES	N/A	N/A

Table 3: Benchmarking of KG and material ontology related tools

The above table indicates the existence of a diverse range of open-source tools available for developing knowledge graphs and ontologies. These tools are supported by active communities and are characterised by high levels of popularity, which is indicative of the researchers' and developers' expanding interest in leveraging such technologies.

2.4.1 Evaluation Metrics

In the following, we define metrics based on which the evaluation of the above platforms can be performed for the purposes of the DiMAT project.

- Scalability: the amount of data and/or the different instances it can handle, as well as the ability to split a logical database into smaller ones.
- Query response time: time for the graph database to process a query and return the results to the user.
- Parallel processing: Number of parallel queries that the database can process.
- Flexibility: Enabling a schema optional structure, without explicitly requiring indexes and constraints.
- Interoperability: Provision of APIs for the connection and communication with other applications, as well as the support of different programming languages.

2.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

The usage of the below mentioned tools is suggested due to their high popularity as well as their solid performance and adoption by the industry as well as the relevant experience of the project partners.

Neo4J is currently one of the most popular graph databases and the respective tools built on top of it (Neodash, Bloom) provide multiple functionalities for generating useful visualizations. Neo4j is scalable even in its open-source version and allows for deployment in resource-constraint scenarios. Moreover, it supports APIs over Python, Javascript, Go, and .NET making it one of the most interoperable database software and increases the ease of use for integrating with other scripts. Finally, its active community is essential for aiding the development effort of a toolkit that uses a knowledge graph as a central element. The Cypher query language used for querying Neo4J knowledge graphs is one of the most popular and suitable for graph format.

Finally, the use of Protégé is recommended due to its capability of allowing visualizing the ontologies, assisting in this way the user in understanding their structure and intricate elements. Finally, incorporating elements from the EMMC ontology, a result of the EMMO consortium to the developed KG, is suggested due to its applicability to represent materials and business processes.

Solution	Usage in DiMAT
Neo4j	Construct Knowledge Graph to represent and manage materials characteristics, properties and their relationships. Information retrieval, analysis and visualization techniques will be supported to reveal correlations among the available data in the KAF Toolkit.
Neodash	Dashboard builder for applications that use Neo4J graph database. Can be employed for visualization purposes and can provide overview of the developed KG via plots and figures directly from cypher queries. It is customizable and open-source.
Bloom	Visualization tools compatible with Neo4J databases, can be utilized for examining visually relationships among entities of the KG of the KAF framework.

Protégé	Visualization and handling of ontologies (reasoning, etc.)
EMMO	Ontology to act as reference point for DiMAT

Table 4: Indicative usage for KG and Ontology related tools in DiMAT

3 DIGITAL TWIN TECHNOLOGIES

3.1 STATE OF THE ART ANALYSIS

Although the notion of twinning an asset for simulation purposes dates to the days of NASA's Apollo project where a replica of a space vehicle was created, Digital Twins (DT) technology is currently in the spotlight of research. Indicative of this is the fact that it is listed among the top 10 technological trends for this decade [15], and it is expected to grow by 58% annually until 2026 [16]. In the literature there is no single definition of the DT, but the consensus among the scientific community is that this term refers to a complete virtualization of a system (i.e., the physical twin), capable of acting as a digital replica of the actual system. The ISO 23247-1:2021 tries to standardize a digital twin framework for manufacturing [17]. The development of a Digital Twin enables the efficient monitoring of a system, providing functionalities such as forecasting, simulation or emulation of processes in which the physical twin partakes. Moreover, it supports direct communication with the physical counterpart, something that differentiates it from the relative term of Digital Shadow [18]. The areas where the DT research is emerging include manufacturing, transportation, aerospace, smart cities, healthcare and many more.

In order for the DT to be developed, devices like sensors, actuators and other smart devices need to be embedded in the corresponding physical object being modelled (the Physical Twin). These devices generate data that needs to be collected, processed and fed as input to the appropriate algorithms. The researchers in [19] propose an architecture for developing a Digital Twin Network for the Industrial Internet of Things (IIoT) that employs Eclipse Hono [20] and Eclipse Ditto [21] software in order to collect data transmitted over various communication protocols (I.e., MQTT, HTTP, etc.), create and model the twins, store the data and provide it over HTTP to the respective applications. The proposed architecture supports numerous services like diagnostics, resource allocation and security management functionalities.

Focusing on Digital Twins in the manufacturing domain, the authors of [22] pinpoint the lack of consensus among experts regarding the key features of these kind of twins. The work identifies, with the aid of area experts, four key features, namely, digital modelling, analytics, timeliness and control. The digital models usually represent machines and equipment and model numerous functions (e.g., mechanical, electrical, etc.). These models can be used for machine monitoring, system design as well as diagnostics. The analytics feature concerns the ability of the system to use refined data analytics techniques to inform the system's stakeholders about optimal actions, efficient planning and provide them via forecasting techniques with predictions of the behavior of the system in the future. Timeliness refers to the ability of the Twin to be near-real time updated with the current state of the system via periodic transmissions of data from edge devices like sensors to the DT to maximize the accuracy of its functionalities. Finally, the control feature highlights the necessity of bi-directional data flows between the DT and its physical counterpart. This communication with the actual device can be used for the DT to be able to decide on its own about the best plan of action and send appropriate commands to the device.

In [23], the authors propose reusable models, stressing their benefits for the DIGITbrain project [24], a platform for DTs aiming towards the digitalization in industry 4.0 paradigm. DIGITbrain is a platform that enables Manufacturing as a Service (MaaS) for DTs. Various types of models are supported. Some of these include Co-simulation models, machine learning based models, reduced order models (ROMs) and System of Systems models. These models are decomposed in three parts: data, models, and algorithms and these are enriched with relevant metadata. It should be noted that DIGITbrain is a highly relevant project to DiMAT and it will incorporate knowledge and expertise gained from its development.

The researchers in [25] propose a three-step method for constructing a digital twin for smart manufacturing using 3D modelling, mechanism modelling and real-time synchronization. The use-case of the publication is a furniture production factory. First, the 3D model of the device is created with the help of relevant software, based on the actual dimensions of the physical device. Afterwards, a physical simulation engine (e.g., Unity3D, Unreal, etc.) is employed in order to model the physical properties of the materials (e.g., gravity, acceleration, etc.) as well as the possible motion mechanisms. Aiming to achieve real-time synchronization, data is collected from PLCs and sensors and is transferred to relevant software packages.

The authors in [26] discuss the scope and requirements for digital twin frameworks. They argue that the concept of a DT relies on the context and use-case, meaning that the DTs should be tailored for the specific applications that they will be employed. The article also proposes example use-cases of DT in manufacturing, like optimizing the planning and scheduling of activities and prediction of failures in equipment.

The work presented in [27] highlights the feasibility of open-source implementations in the design of DT architectures focused on smart manufacturing. The authors claim that closed systems are hindering the progress in Smart Manufacturing and so they suggest the development of open-source solutions to make them accessible to more specialists. The paper lists some of the existing commercial solutions, such as Microsoft Azure Digital Twin [28] and Bosch IoT [29], and notes the current limited open-source solutions for DT but goes on to provide open-source solutions that can be employed in order to create a digital twin architecture, such as open-source communication protocols and platforms for software deployment.

The scientists in [30] propose the application of DT in the injection molding process. The whole manufacturing process consists of three phases and in each one of them a Digital Twin is involved. Sensors are used for feeding the DT model with data about the current state of the machines and are given as input to appropriate simulation software. The paper mentions though, some problems in the development of DTs such as the implementation of sensors in the specific machines and the need for some manual acquisition of data.

Finally, as previously mentioned, healthcare is one area where DTs have proven to be useful. In [31], HospiT'Win is proposed, a framework to support an envisioned future hospital. The framework contains algorithms that analyze incoming data from sensor measurements and can detect emergencies in the hospital. By using stochastic simulation, it can generate scenarios that predict the near-future state as well as recommend solutions to expert decision-makers. The paper also proposes a way to connect the solution to an existing hospital.

3.2 EU RELATED PROJECTS

The European Union during the past years has (co-)funded numerous projects dedicated to developing Digital Twins. Some of these projects are listed in the following table.

Acronym	Grant Agreement No.	Title	Description
DIGITbrain	952071	Digital twins bringing agility and innovation to manufacturing SMEs, by empowering a network of DIHs with an integrated digital platform that enables Manufacturing as a Service (MaaS)	A platform that enables Manufacturing as a Service (MaaS) for DTs. Various types of models are supported. Some of these include Co-simulation models, machine learning based models, reduced order models (ROMs) and System of Systems models.
i4Q	958205	Industrial Data Services for Quality Control in Smart Manufacturing	Part of this project is the development of a DT aimed for the manufacturing domain, helping the relevant industries to achieve accurate

			production simulation.
NEPHELE	101070487	A lightweight software stack and synergetic meta-orchestration framework for the next generation compute continuum	NEPHELE project aims to develop an IoT software stack to virtualize IoT devices operating at the network edge and also provide an orchestration framework for coordinating cloud and edge computing platforms.
IoTwins	857191	Distributed Digital Twins for industrial SMEs: a big-data platform	A platform for modelling, simulating and optimizing plans in virtual environments in order to anticipate changes in the physical world. The project aimed to develop DT testbeds for the manufacturing and facility management sectors.
Change2Twin	951956	Create and Harvest Offerings to support Manufacturing SMEs to become Digital Twin Champions	The goal of this project is to make possible the deployment of DTs for all European manufacturing companies. Among others, it provides an architectural-agnostic marketplace as well as a benchmarking

			service model for making easier the development of DT solutions.
DT4GS	101056799	The Digital Twin for Green Shipping	The DT4GS project aims to make the shipping industry "greener" by providing solutions based on Digital Twins leveraging on forecasting algorithms.
STAR	956573	Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines	STAR aims at designing new technologies to enable the deployment of standard-based secure, safe, reliable and trusted human centric AI systems in manufacturing environment. STAR aims to research, develop, validate and make available to the community leading-edge AI technologies including explainable AI, active learning systems, simulated reality systems, human-centric digital twins, advanced reinforcement learning techniques

			and cyber-defense mechanisms, thus becoming a catalyst for the deployment of advanced AI systems in the manufacturing shop-floor.

Table 5: Digital Twins EU Funded Projects

3.3 EXISTING OPEN-SOURCE SOLUTIONS

Although there exist commercial solutions for deploying Digital Twin platforms like Microsoft Azure Digital Twin [28], Bosch IoT [29], XMPro [32] and numerous more, the focus lies in discovering suitable and reliable open-source platforms in order to develop the DiMAT project's software without any restrictions or charges imposed by third parties. The currently available platforms are listed in the following table. It is important to note that NEPHELE is a highly relevant to the Digital Twins concept project lead by NTUA, a DiMAT consortium partner.

Open-Source Solutions	Description
Eclipse Ditto	A platform that acts as an IoT middleware. It provides an abstraction layer for IoT solutions to interact with physical devices. A digital twin can be described and created, corresponding to an actual device. A secure connection, possible over numerous protocols (e.g., MQTT, AMQP, HTTP, etc.) can be established for their (bi-directional) communication. Moreover, Ditto enforces policies uploaded by the administrator/owner that define different user roles and their permissions regarding the actions that they can perform on the twin or the device. The platform also stores the last state of the twins by employing a MongoDB database and provides both synchronous and asynchronous APIs to access either the digital or the physical twin. It also provides the user with a graphical user interface for visual inspection of the existing twins, their states and policies. An example of a commercial application that utilizes Ditto is the Bosch IoT suite.
CPS-Twinning	CPS-Twinning is a platform designed to provide Digital Twin implementation for Cyber-Physical Systems. CPS-Twinning

	offers two modes of operation, the simulation mode where the operation of the DTs is separated from the physical counterparts as well as the replication mode in which there is communication with the devices. The platform considers possible security issues and introduces appropriate safety tools.
Eclipse Hono	Eclipse Hono's goal is to make device connectivity simplified. In more detail, it is a software that provides various interfaces for connecting many Internet-of-Things (IoT) devices to a common back-end and allowing interaction with them in a uniform way, without taking into consideration the specific device communication protocols. Hono supports many communication protocols, such as HTTP, MQTT and AMQP but also provides the capability of adding custom protocol adapters.
FIWARE	An open-source platform for developing applications. Although not specifically targeted to DTs, it provides a set of APIs enabling the creation and management of DTs. It provides real-time processing as well as security.
Digital Twin (ZDMP)	The ZDMP architecture allows the creation and management of digital twins. It supports a hierarchical way for the organization of the different components of a Twin and provides simulation and prediction tools. The Digital Twin component of ZDMP allows for visualizing trends of real-time data. It is based on Eclipse Ditto and is compatible both with it and with PI System.
CLAWDITE	Fully owned by SUPSI and developed in previous research projects (STAR H2020), Clawdite is an Industrial Internet of Things (IIoT) platform that supports the development of customized data representations of production systems. Its modular infrastructure equipped with interchangeable components facilitates the creation of digital twins [33].
NEPHELE VO	Open-source Virtual Object creation software based on the Web-of-Things standard. It allows for the creation of digital counterparts of the real objects. It is lightweight and supports communication over various channels (HTTP, MQTT, CoAP). It offers out-of-the-box capabilities such as time-series storage and basic algorithms for predictions.

Table 6: Available solutions for Digital Twins

3.4 BENCHMARKING AND ASSESSMENT

In the following table some metrics across which the available software can be compared are provided. Among them, the most important one is the openness of the software, a critical component in the development of the Digital Twin for Process Control toolkit, whose development should be based on open-source software. Then, the minimum requirements for deploying the solution should be also taken into account. The desired software should be lightweight enough so that it can be deployed both in the case of ample computing resources as well as in more resource constrained devices. Finally, the activity of the community surrounding it should be taken into account, because it provides an indication of the quality of available documentation and discussions that can ease the development of software based on top of it.

Solution Name	Publicly available code	Online repository statistics			Active community	Used by industries	Minimum Requirements	
		#Forks	#stars	Last commit			RAM	Proc/or Cores
Eclipse Ditto	YES	162	459	March 2023	YES	YES	4GB	2
CPS-Twinning	YES	18	40	2021	NO	N/A	N/A	N/A
Eclipse Hono	YES	136	398	March 2023	YES	YES	8GB	2
FIWARE	YES	41	36	Feb. 2023	NO	YES	N/A	N/A
Digital Twin (ZDMP)	NO	-	-	-	N/A	N/A	16GB	N/A
CLAWDITE	NO	0	0	Oct. 2022	N/A	N/A	8GB	2
NEPHELE VO	YES	0	2	Sep. 2024	YES	N/A	0.5GB	0.5

Table 7: Benchmarking of Digital Twin Tools

As we can see from the above evaluation it is evident that there do not exist that many open-source platforms for developing Digital Twins. This is expected as Digital Twins are a recent trend. In the future some of the above are expected to gain even more attention from researchers and also support more features.

3.4.1 Evaluation Metrics

In the following, we define metrics based on which, the evaluation of the above platforms can be performed for the purposes of the DiMAT project.

- Security: Number of available protocols for encryption of data. The existence or not of authentication mechanisms.
- Scalability: How many DTs can the platform handle? The amount of data it can store.
- Cost: Cost of deploying the platform. For all the tools mentioned in the previous subsection, it is equal to zero.
- Computing requirements: The resources needed to be dedicated for the smooth operation of the platform.
- OS Compatibility: Number of operating systems supporting the platform.
- Response time: Measured in time (ms) that is needed for the platform to respond to a user's request.
- Accuracy: How accurate are the DTs that can be developed in the platform when compared to their physical counterparts.
- Protocol interoperability: Number of different protocols (e.g., MQTT, HTTP, CoAP, etc.) for data transmission.
- Semantic interoperability*:* Information models supported (e.g., NGSI-LD, W3C WoT, etc.).

3.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

Digital Twin software is expected to be mainly employed in the development of the Digital Twin for Process Control (DTPC) toolkit. The potential usage of some of the aforementioned tools presented in subsection 3.3 and based on the benchmarking of the previous section can be seen in the following table. All three of the suggested tools are open-source and are capable of being executed in the major operating systems (i.e., Linux, Windows, MacOS) an important factor to consider when choosing software for implementing the DiMAT software components. Moreover, both Eclipse Ditto and Hono gain constantly popularity as indicated by the trend in the GitHub statistics. However, Eclipse Hono is not a Digital Twin creation software but covers essential aspects of a DT related to the connectivity of IoT devices, so its usage will be examined solely for this purpose. The NEPHELE Virtual Object stack is the result of an ongoing EU funded project. It is based on the Web-of-Things (WoT) standard and one of the DiMAT consortium's partners, NTUA, has a leading involvement in this project. It is a very lightweight and scalable software stack that can be executed even in limited resource devices deployed near the network edge and in fully dockerized environments. Both Ditto and NEPHELE offer the same number of different protocols that can support data transmission, while NEPHELE seems to be more lightweight.

Solution	Usage in DiMAT
Eclipse Ditto	Design DTs, exchange messages and store the latest measurements.
Eclipse Hono	Utilize the provided APIs to exchange messages between the devices and the twins. Hono can act as a middleware connecting the devices to the Ditto platform.
NEPHELE Virtual Object	Adopt the NEPHELE Virtual Object for the design of DTs in DiMAT.

Table 8: Indicative Usage of Digital Twin Tools for DiMAT

Co-funded by the European Union

4 ENVIRONMENTAL AND COST LIFE CYCLE ASSESSMENT

4.1 STATE OF THE ART ANALYSIS

The expanding economic activities of various industries have long raised concerns about climate change, energy security, and natural resources scarcity [34]. As a result, decoupling environmental impact from economic growth has become a critical issue among stakeholders, from policy makers and academics to professionals in various fields and practitioners. Hence, it is no surprise that "sustainability" is continuously gaining ground within industries' strategic and operational activities to increase growth and enforce their competitiveness. While manufacturing companies play a crucial role in the global production system by contributing to economic growth and employment, they also escalate energy and resource consumption, waste generation, and harmful emissions [35] As defined by the United States Environmental Protection Agency (USEPA): "Sustainable manufacturing is the creation of manufactured products through economically-sound processes that minimize negative environmental impacts while conserving energy and natural resources. Sustainable manufacturing also enhances employee, community, and product safety" (US EPA). Manufacturing companies are becoming more and more aware of the need to meet customers' requirements, legal provisions and regulations towards a more sustainable operation and the provision of more sustainable products.

To address this goal, a variety of sustainability assessment methods have been developed over the years [36]. Sustainability assessment methods may refer to ecological footprint, Life-Cycle Assessment (LCA), Resource Value Mapping (RVM), Material Flow Analysis (MFA), Integrated Assessment Models, Multi Criteria Decision Analysis (MCDA), Cost-benefit analysis etc. [37]. To unlock its full potential while reducing its environmental impacts, the manufacturing sector needs simplified, harmonized, and user-friendly sustainability assessment methods. Life cycle assessment (LCA) is a worldwide accepted methodology that allows for the quantification of the impacts arising throughout the entire life cycle of products and services. At the same time, it provides valuable information on identifying hotspots and thus allowing interventions to improve production. However, conducting a traditional LCA is not only time and resource consuming, but also requires a thorough knowledge on data handling, system boundary definition, assessment methodologies, etc. Especially for SMEs, that can invest limited budget and time on such assessments, the need for implementing simpler LCA methods to monitor the impacts of their production processes is of increased importance.

Nowadays, numerous LCA software and tools are available, primarily licensed for in-house use by large industries, consulting firms, and external experts. Among the leading LCA software, meaning the most known and widely used, are SimaPro, openLCA, and GaBi, followed by Umberto and EIME, of Bureau Veritas. Despite their wide usage and integration with various LCI databases, most of these tools are not available for free, with openLCA being an exception. These tools also provide economic assessment apart from the environmental one. The need for more simplified approaches has led to the development of web based LCA tools and streamlining approaches. Regarding web-based tools, Bilan Produit is an online LCA software which is free and simplified in terms of database and modeling. However, this is only available in French language. Finally, the European Platform on LCA provides an extensive list of LCA related tools that cover a range of topics from environmental assessment and eco-design tools to compliance management tools, which are inevitably interrelated and LCA relevant. Nevertheless, streamlined LCA can be quite challenging both in terms of simplified LCI and LCIA approaches [38].

4.2 EU RELATED PROJECTS

The European Union, during the past years, has (co-)funded numerous projects dedicated to developing, testing, and assessing materials' value chains. A great share of these projects includes LCA as a core part of their work structure. However, to the best of our knowledge, there are very few projects that are LCA-driven. Some of the most relevant projects were identified, although most of them are not LCA-driven, and are listed in the following table:

Acronym	Grant Agreement No.	Title	Description
LCA TO GO	265096	Boosting Life Cycle Assessment Use in European Small and Medium-sized Enterprises: Serving Needs of Innovative Key Sectors with Smart Methods and Tools	LCA TO GO project developed sectoral methods and tools for bio-based plastics, industrial machinery, electronics, renewable energy, sensors and smart textiles.
NanoSustain	247989	Development of sustainable solutions for nanotechnology-based products based on hazard characterization and LCA	Objective of the NanoSustain project was to develop innovative solutions for the sustainable design, use, recycling and final treatment of nanotechnology-based products

ECOGEL CRONOS	609203	High productivity manufacturing process of composite parts based on zero emissions fast curing coatings and heated moulds	The proposed project aims to the development of an ecological and innovative coating for composite parts which may be able to eliminate the styrene emissions from the workplace.
INCOM	608746	Industrial Production Processes for Nanoreinforced Composite Structures	The INCOM project aimed to develop technologically and economically viable solutions and production methods for lightweight structures based on advanced sustainable materials geared towards packaging, vehicles, sporting goods and aeronautical applications.
TREASURE	101003587	Leading the TRansion of the European Automotive SUpply chain towards a circulaR futurE	TREASURE wants to support the transition of the automotive sector towards Circular Economy (CE) trying to fill in the existing information gap among automotive actors, both at design and EoL stage.
CIRCTHREAD	958448	Circular Digital Thread project vision	CircThread's main objective is to unlock access to

			product data between stakeholders that is now in silo's and utilize it for enhanced Circular Economy decision making, across the extended product life cycle.
DENIM	958339	Digital intelligence for collaborative ENergy management in Manufacturing	DENiM wants to develop a digital intelligence platform that constitutes a suite of tools and advanced digital services used to optimize existing workflows and product manufacturing, driven by energy efficiency targets and sustainability goals. The overarching goal is to define and promote the use of best practices across industry sectors within the EU to support the realization of energy efficient manufacturing systems.

Table 9: Environmental and Cost LCA EU Funded Projects

4.3 EXISTING SOLUTIONS

Currently there are several open-source Life Cycle Assessment related solutions, which can also be found in the repository of the European Platform on LCA. Regarding the DiMAT project, the LCA-related open-source solutions that were identified as most relevant are listed below:

Open-Source Solutions	Description
"LCA to go" tool	"LCA to go" is an open-source toolbox to support Life Cycle Assessment (LCA) in SMEs. It is a web-based tool designed for sector-specific environmental assessments (i.e., bio-based plastics, industrial machines, electronics, renewable energy, sensors and smart textiles).
Bilan Produit®	The Bilan Produit Ademe tool is part of a larger set: the Impacts® database. The latter offers a set of tools for examining the contributions to environmental pollution of the processes falling within the scope of the analysis. The Product© report tool allows for an eco-design approach, for example in the context of the optimization / creation of packaging.
OpenLCA	OpenLCA is a free software for conducting sustainability and Life Cycle Assessment. It is open source and covers a wide variety of sectors.
GLAD	GLAD is a community-driven open-source project, built on open-source software.
GRETA (GREen Targets)	Fully owned by SUPSI and developed in previous research projects (DENiM H2020, CircThread H2020), GRETA is a comprehensive platform designed to assess the sustainability and circularity performances of products and processes in manufacturing contexts. GRETA offers powerful diagnostic and advisory functionality, enabling users to optimize their manufacturing practices and make data-driven decisions. The platform is modular and can be easily integrated with real production environments, AI algorithms, simulation engines and user-oriented data-entry and data visualization customizable interfaces. Supporting comparisons between different scenarios and enabling users to explore the potential impacts of different manufacturing strategies on sustainability and circularity performance, GRETA is a valuable tool for mainstreaming sustainable decision making for a low-impact manufacturing future.

Table 10: Tools for Environmental and Cost LCA

Co-funded by the European Union

4.4 BENCHMARKING AND EVALUATION

Among all software presented in Table 10:

- LCA to go is tailored for small businesses and predefined sectors, while it offers simplified assessments, lacking the detailed customization and broader applicability,
- Bilan Produit is more regionally focused, it lacks diverse database support,
- and GLAD is a more database-centric platform, lacking process modeling, offering a few options for scenario modeling and impact assessment customization.

Thus, the two most appropriate software to be applied in the context of DiMAT project are GRETA and OpenLCA.

Solution Name	**Publicly available code**	**Online repository statistics**			**Active community**	**Used by industries**	**Minimum Requirements**	
		#Forks	**#stars**	**Last commit**			**RAM**	**Proc/or Cores**
OpenLCA	Yes	26	127	May 2023	YES	YES	N/A	N/A
GRETA	NO	0	0	Mar. 2023	N/A	YES	8	4

Table 11: Benchmarking of Environmental and Cost LCA tools

Based on the information provided in the above table, it is evident that there are limited open-source tools providing environmental assessment based on the entire life cycle of the products. Furthermore, it seems that these solutions are not quite popular, although they are used by industries. In addition, although OpenLCA is an open-source software, it requires extensive knowledge on such assessments as well as increased resources.

4.4.1 Evaluation metrics

In the following, we define metrics based on which, the evaluation of the above platforms can be performed for the purposes of the DiMAT project.

- Computing requirements: The resources needed to be dedicated for the smooth operation of the LCA analysis.
- Accuracy: How accurate is the LCA that will be developed in the project.
- Data availability and flexibility: If the tool provides access to comprehensive databases for LCA, including product and region-specific data, or allows the addition of custom datasets.
- Scalability: How well the tool can handle large datasets or complex models, which is essential for projects with extensive LCA requirements, like DiMAT.

- Integration capabilities: How easily can data be inserted and exported from the tool, considering that will be further utilized in the development of DiMAT tools.
- Community support: How large and active is the user community gives indications about how often and relevant the updates and improvements of the tool are, as well as support for troubleshooting issues.

4.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

LCA software is expected to be employed for the environmental assessments of the pilot processes, which in turn produces the data used in the MEC-LCA toolkit. The potential usage of OpenLCA tool is further described and is based on evaluating the tools mentioned in section 4.4, with the criteria mentioned in section 4.4.1.

OpenLCA provides access to a variety of leading databases (such as Ecoinvent, Product Environmental Footprint (PEF), etc.), enabling users to access detailed industrial data, specific to materials and processes, related to all four pilots of the project. Finally, OpenLCA is compatible to a significant number of impact assessment methodologies, such as ReCiPe, Environmental Footprint (EF), etc., allowing users to evaluate a variety of industry-related impacts.

Solution	Usage in DiMAT
OpenLCA	Support the LCA analysis of the DiMAT solutions in all the 4 pilots

Table 12: Indicative Usage of LCA tools in DiMAT

5 DATA ANALYSIS AND MACHINE LEARNING

5.1 STATE OF THE ART ANALYSIS

5.1.1 Explainable AI

Artificial Intelligence (AI) has become a widely adopted technology in various fields, including finance, healthcare, and transportation. With the increasing use of AI, it has become crucial to ensure that AI is transparent, accountable and trustworthy. Explainable Artificial Intelligence (XAI) represents a burgeoning area within AI research, dedicated to creating AI systems capable of providing transparent and comprehendible elucidations of their decision-making processes to users.

XAI can be classified into two main types: model-agnostic and model-specific. Model-agnostic techniques aim to provide explanations that are applicable to any type of AI model, while model-specific techniques focus on explaining the decisions of specific models [39]. Examples of model-agnostic techniques include feature importance ranking, partial dependence plotting, and the creation of local surrogate models. Feature importance ranks the features used by the model based on their importance [40]. Partial dependence plots can show how the output of the model changes as a function of a particular feature [41]. Local surrogate models approximate the behaviour of the model around a specific input and output [42]. Model-specific techniques on the other hand, include decision trees, rule extraction, and attention mechanisms. Decision trees function by creating a tree structure that represents the decision-making process of the model [43]. Rule extraction techniques translate the decision-making process of the model into a set of human-readable rules [44]. Lastly, attention mechanisms highlight the input features that the model pays the most attention to [45].

XAI is becoming increasingly important in materials modelling as it provides insights into the decision-making process of complex machine learning algorithms [46]. Materials modelling involves simulating the behaviour of materials at the atomic and molecular level to understand their properties and design new materials with desired characteristics. Machine learning algorithms are often used in materials modelling to predict material properties and accelerate the discovery of new materials [47]. However, these algorithms can be difficult to interpret, which limits the ability of researchers to understand why certain predictions are made. XAI techniques can help address this issue by providing explanations for the predictions made by machine learning models [48]. Decision trees can help identify the most relevant features that influence material properties and classify materials based on their properties. Moreover, Gaussian process regression is a powerful machine learning technique that is widely used in materials modelling to predict material properties. It can provide uncertainty estimates in addition to point predictions, making it easier to interpret the

Co-funded by the European Union

results. Simple regression techniques, such as lasso regression, which is a linear regression technique, can help identify the most important features that affect material properties and is particularly useful when dealing with high-dimensional datasets.

Despite notable progress in XAI, certain challenges persist, such as balancing the trade-off between the quality of explanations and model accuracy [49]. Another challenge is to ensure that the explanations are understandable to the intended audience, including non-experts. Furthermore, there is a need to develop XAI techniques for emerging AI models, such as deep neural networks [49].

In order to address these challenges, future research in XAI should concentrate on the development of more effective and efficient techniques that can provide high-quality explanations while maintaining model accuracy. In addition, there is a need to conduct user studies to evaluate the effectiveness of XAI techniques in improving user trust and acceptance of AI systems.

In conclusion, XAI is an important area of research that seeks to address the challenges associated with AI transparency, accountability, and trustworthiness. XAI techniques can provide clear and understandable explanations of AI decisions to users. Model-agnostic and model-specific techniques have been developed to provide different types of explanations. However, challenges still exist, and future research should focus on developing more effective and efficient XAI techniques while ensuring the explanations are understandable to non-experts.

5.1.2 Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning that is increasingly being used in Industry 4.0 to optimize decision-making processes and improve automation. In RL, an agent operates within an environment and is learning to undertake actions, aiming to maximize a cumulative reward. RL can be used in various applications, including the areas of predictive maintenance and quality control in Industry 4.0 environments, scheduling of business processes and optimization of logistics processes. By enabling machines to learn from experience and improve their decision-making capabilities over time, RL can help industries reduce costs, increase productivity, and improve efficiency [50].

In the case of predictive maintenance, RL can assist by monitoring machine performance and identifying potential issues in real-time [51][52]. In this way, machines can be serviced before a breakdown occurs, reducing downtime and increasing productivity. RL can also be used for quality control, where machines can learn to identify and correct defects in real-time, reducing waste and increasing yield [53][54]. This approach is especially useful in complex manufacturing processes where manual inspection can be time-consuming and error prone. RL is also used in Industry 4.0 for scheduling and logistics optimization [55]. By learning from past data and predicting future trends, machines can optimize production schedules and logistics operations, reducing lead times and improving efficiency. RL can also be used to improve safety in manufacturing processes by automating certain tasks and providing real-time feedback on machine performance [56]. This approach can help reduce human error and increase safety in hazardous environments.

The future of RL in Industry 4.0 is likely to be characterized by increased adoption, greater integration with other technologies, and the development of more advanced autonomous systems. As the demand for smarter, more efficient, and more adaptable manufacturing processes continues to grow [57], RL is likely to become an increasingly important tool for achieving these goals.

5.1.3 Natural Language Processing (NLP)

Natural Language Processing is a subfield of computer science and artificial intelligence, involving computational techniques to interpret, analyse and generate human language either in spoken or written form. It mainly aims at facilitating human and computer interaction and boosting knowledge extraction from vast amounts of data. NLP techniques are used in a wide range of applications, including language translation, sentiment analysis, chatbots, keyword extraction, text classification and summarization and more [58]. With the increasing amount of data generated every day, NLP is becoming an essential tool to help humans interact with and make sense of the massive amounts of textual data available. Regarding material science and manufacturing processes, the use of NLP on scientific text can generate knowledge that enables data visualization, mining and analytics [59]. The derived information can be used to populate databases, reveal correlations of materials and materials properties by finding relationships between compounds, predict new materials, emerging associations between them and possible applications.

One of the major uses of NLP in material and manufacturing domain is to extract knowledge from texts in published studies. The scientists in [60] aim to compile knowledge regarding relationships between material and manufacturing concepts and entities from millions of related documents. Their goal is to develop a labelled-property data structure for query performance and materials and manufacturing knowledge retrieval. To this end, the researchers have developed and deployed a named entity recognition model (NER), an NLP technique, to associate words encountered in literature with terms and categories of the manufacturing and materials domain. Specifically, a supervised machine learning approach and a neural network using bidirectional long-short term memory (BiLSTM) and conditional random field (CRF) techniques, have been employed to extract information and classify free text from more than 500K documents related to the manufacturing and materials domain. The resulting overall accuracy (F1-score) of the NER was 88%. A similar work is presented in [61], where materials science knowledge is extracted from 500K journal papers related to the polymer's material domain. By leveraging NLP techniques, word embeddings are created for the material corpus. The training of words' vector representations is accomplished by using an unsupervised approach. In addition, time-based studies, to determine polymers' popularity for different applications, to predict new polymers and to reveal possible correlations between the materials, are performed based on the polymer data set.

Materials with useful properties have also been identified using databases based on natural language processing to train machine learning models. Examples include the discovery of the large magnetocaloric properties of HoB_2 [62] and the development of a "design to device approach" for designing dye-sensitized solar cells [63]. NLP has also been used to directly predict materials without the use of intermediary models [64]. The researchers in [65] have created word embeddings trained on documents related to materials science and by calculating dot products between word vectors and applications words (e.g., thermoelectric), conclude in materials applications predictions.

5.1.4 Classification Algorithms

Classification algorithms are an essential part of machine learning, used to categorize data into different classes or groups based on their features or attributes [66]. These algorithms have a broad range of applications, from image recognition to NLP, and are used in many industries, including healthcare, finance, and marketing.

Supervised learning is the most common approach used in classification algorithms. In this approach, a model is trained on labelled data, with the labels corresponding to the class or group that each data point belongs to [67]. Popular supervised classification algorithms include logistic regression, decision trees, random forests, and support vector machines (SVMs) [66]. Unsupervised learning algorithms are used when there is no labelled data available, and the goal is to identify patterns and structures within the data. These algorithms include techniques such as k-means clustering, hierarchical clustering, and principal component analysis (PCA) [66].

In recent years, deep learning has revolutionized the field of classification algorithms, especially for image and speech recognition tasks. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are deep learning models, widely used for classification tasks [66]. Transfer learning, involves pre-training a model on a large dataset and fine-tuning it on a smaller dataset, and has also been widely used in classification tasks to improve performance [67].

Materials modelling heavily relies on classification algorithms in machine learning to predict material properties and behaviour [68]. It involves the use of computational techniques to simulate the behaviour of materials at the atomic or molecular scale, providing insight into their properties and potential applications [69]. Supervised learning algorithms, such as decision trees, random forests, and support vector machines, are commonly used in materials modelling to predict material properties based on labelled data [70]. Unsupervised learning algorithms, such as k-means clustering and PCA, can be used in materials modelling to identify patterns and structures in large datasets of unlabelled data [71]. Deep learning algorithms, such as convolutional neural networks and recurrent neural networks, have also been used in materials modelling to predict material properties [72].

One of the primary challenges in classification algorithms is the issue of imbalanced data, where the data is not equally distributed among the different classes. This can lead to biased predictions and inaccurate results [73]. Another challenge is the interpretability of complex models, such as deep learning models, which can be difficult to understand and explain.

To summarize, classification algorithms are a crucial part of machine learning and have numerous applications in various domains. Recent advances in deep learning, transfer learning, and other techniques have improved their performance significantly [66][67]. However, challenges such as handling imbalanced data and improving model interpretability remain, and ongoing research is focused on developing new approaches and solutions to address these issues.

5.1.5 Recommender Systems

Recommender Systems (RS) are nowadays ubiquitous, permeating several domains such as Online Social Networks, streaming platforms and even manufacturing. Their main role is to advertise content and assist users in selecting the appropriate action(s) based on their tastes and preferences. Depending on the mode of operation, RS are usually divided into several broad categories. One of the most popular ones are Content-based RSs [74], where the users of a platform are recommended items similar to those they consumed beforehand. Another popular method is collaborative filtering RS [75], in which the actions of other users of the platform with similar profiles are considered. Other categories of RS include context-aware RS [76] as well as other more domain-specific RS and hybrid methods combining two or more techniques, aiming to increase their precision.

Focusing on existing Recommender Systems that assist in the materials manufacturing domain, the authors in [77] propose a context-aware RS for improving the manufacturing process modeling. The proposed system recommends the top-k similar processes that can be followed in each step, extracts independent paths for manufacturing and provides also a visualization tool. The work published in [78] focuses on the development of an RS capable of recommending materials needed for producing automobile spare parts. The system first detects the materials to suggest either by a simple filtering mechanism based on historical data or by using data analysis techniques like the k-means and PCA algorithms. Following this step, the RS calculates the number of material parts that should be recommended based on past users' ratings. Finally, the users should input the materials that they were assigned.

Moreover, a method of recommender systems that could assist in the DiMAT project goals are the Knowledge Graph-based recommender systems [79]. These systems recommend items and actions based on knowledge gathered for a specific domain (e.g., materials, manufacturing, various ontologies etc.). In general, systems of this type, employ either the embedding of the knowledge graph in a geometric space with machine learning methods such as TransE [80] or DistMult [81] or opt for a path-based approach [82] when explainable recommendations are a concern. In the PathSim algorithm presented in [82][83], the similarity between pairs of entities is extracted via aggregating the number of metapaths (which describe entity types in each step of a graph path) existing between these entities.

5.1.6 Forecasting Techniques

Forecasting is an important area of research in data science that aims to predict future values of a variable based on historical data [84]. Forecasting techniques have numerous applications in various domains, including finance, healthcare, and energy.

There are various types of forecasting techniques, including time-series methods, regression methods, and machine learning methods [85]. Time-series methods are used when the data has a natural temporal ordering and include techniques such as ARIMA, exponential smoothing, and state space models. Regression methods are used when there is a linear relationship between the dependent and independent variables and include techniques such as linear regression and logistic regression. Machine learning methods are used when there is a complex relationship between the variables and include technical methods such as decision trees, random forests, and neural networks [86].

Forecasting techniques have numerous applications in various domains. In finance, they are used for stock price prediction, portfolio optimization, and risk management [87]. In healthcare, they are used to predict patient outcomes and disease progression. In energy, they are used for load forecasting and energy demand prediction. In materials science, forecasting techniques are used to predict the properties of new materials before they are synthesized or manufactured [90]. This is important because materials properties can be difficult and expensive to measure experimentally and predicting them accurately can save time and resources. These techniques are used to predict various properties of materials, including their mechanical, thermal, and electronic properties. For example, machine learning techniques have been used to predict the elastic properties of materials based on their atomic structure [91]. Deep learning techniques have also been used to predict the band gap of materials, which is an important electronic property that determines their ability to conduct electricity [92].

In addition to predicting materials properties, forecasting techniques are also used in materials design to optimize the properties of existing materials or to design new materials with specific properties [93]. For example, machine learning techniques have been used to design new materials for energy storage applications [94]. The models predict the properties of the materials based on their chemical composition and crystal structure, and the most promising candidates are synthesized and tested experimentally. Overall, forecasting techniques have a wide range of applications in materials design, from predicting the properties of new materials to optimizing the properties of existing materials or designing new materials with specific properties.

In recent years, deep learning has shown promising results in tasks involving time-series forecasting. Recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) networks and gated recurrent units (GRUs), are typical deep learning models used for time-series forecasting [88]. Transfer learning, a technique that involves pre-training a model on a large dataset and adjusting it on a smaller dataset, has also been used in forecasting tasks to improve performance [85].

One of the primary challenges in forecasting is dealing with uncertainty and variability in the data. The accuracy of forecasts can be affected by various factors, such as data quality, outliers, and missing values [89]. Another challenge is the interpretability of complex models, such as deep learning models, which can be difficult to understand and explain.

Overall, forecasting techniques are essential for predicting future values of a variable and have numerous applications in various domains. Recent advances in deep learning and transfer learning have improved their performance significantly [85][88]. However, the aforementioned challenges concerning uncertainty and interpretability persist, with ongoing research focused on developing innovative approaches and solutions to these issues.

5.2 EU RELATED PROJECTS

The table below presents some EU funded projects that utilize or develop/research areas of the Machine Learning technological sector. In the vast majority of the listed projects, consortium partners are involved.

Acronym	Grant Agreement No.	Title	Description
ENTROPY	649849	Design of an Innovative Energy-Aware It Ecosystem for Motivating Behavioural Changes Towards the Adoption of Energy Efficient Lifestyles	Reduce energy consumption of buildings by combining deployment of IoT devices, data analysis techniques and employing recommendation techniques and gamification elements.

ARSINOE	101037424	Climate Resilient-Regions Through Systemic Solutions and Innovations	Analyze climate change related data and develop a framework in order to combine systems innovation support with the Climate Innovation Window.
AIDEAS	101057294	AI Driven Industrial Equipment Product Life Cycle Boosting Agility, Sustainability and Resilience.	AIDEAS aims to enhance the sustainability, agility, and resilience of European machinery manufacturing companies by harnessing the power of artificial intelligence (AI) throughout the entire lifecycle of industrial equipment (design, manufacturing, use, and repair/reuse/recycle stages)
I4Q	958205	Industrial Data Services for Quality Control in Smart Manufacturing	I4Q aims to enhance manufacturing processes by using data-driven predictions, ultimately leading to improvements in both process efficiency and product quality.

Table 13: Machine Learning EU Funded Projects

Co-funded by the European Union

5.3 EXISTING OPEN-SOURCE SOLUTIONS

Open-Source Solutions	Description
LIME	Local Interpretable Model Agnostic Explanations (LIME) is a software that explains what ML learning models are doing when generating predictions. It operates by perturbing the instance that needs explaining and learns a sparse linear model around it. LIME works well with algorithms that are included in the scikit-learn package.
NLTK	Platform that aids in the construction of Python applications intended to handle data related to human language. It boasts a multitude of corpora and lexical resources, as well as libraries for text processing. These libraries can be utilized for a range of functions including, but not limited to, classification, tokenization, stemming, and semantic reasoning.
SpaCy	An open-source library designed for sophisticated natural language processing tasks. It encompasses implementations of several NLP capabilities, including but not limited to, Named Entity Recognition (NER), part-of-speech-tagging, word vector representation and more. Moreover, it supports a multitude of languages with pre-trained models.
Apache OpenNLP	An open-source, Java-based, toolkit designed for handling NLP tasks. It incorporates algorithms that address prevalent NLP tasks such as Named Entity Recognition (NER), chunking, tokenization, and coreference resolution.
Stanford CoreNLP	Open-source Java implementation for NLP. It supports 8 languages, namely Arabic, Chinese, English, French, German, Hungarian, Italian, and Spanish.
Ampligraph	A Python library that includes numerous knowledge graph embedding algorithms and provides link prediction methods.
PySClump	A Python implementation of the algorithms presented in [83]. Includes the PathSim algorithm.
LightFM	A Python library that provides implementations of popular recommendation algorithms. There exist in LightFM package both algorithms that require explicit feedback (e.g., ratings) from users and also algorithms that infer the users' feedback. The library comes with a set of popular datasets on which to evaluate a recommender system.

Drools	A business rule-based management system and event processing engine that is developed in Java that can be employed for recommendation purposes by defining specific rules. This system can be used to emulate the decision-making process by experts.
Tensorflow	A well-known Python library that provides implementations of various machine learning models ranging from data embeddings, classification algorithms to recommender systems.
scikit-learn	A Python library that offers implementations of various classification algorithms like k-means, SVMs, k-Nearest Neighbors, etc., as well as dimensionality reduction (e.g., PCA).
Hugging Face	A Python library that provides APIs and access to various machine learning algorithms with pretrained models. The APIs give the opportunity to the users to quickly download and use these pretrained models. It can be used for NLP related tasks, image classification as well as audio related tasks.
Prophet	A library implemented both in R and Python by Facebook offering high quality forecasts for time-series data. This toolkit is best suited for time-series data with strong seasonal effects. For better accuracy access to historical data should be available.

Table 14: Open-source ML tools

5.4 BENCHMARKING AND EVALUATION

In Table 15 the basic benchmarking metrics used for evaluating the available solutions are presented. The software presented in the table covers various aspects of ML techniques ranging from neural networks up to natural language processing. For the case of the ML libraries that are presented below, beyond the availability of source code in which all libraries cover the requirement of being open-source, the most important one is the activity of surrounding communities as well as the rating it has received by developers indicated by the relevant repository statistics. Finally, a solid indication about the quality is whether the library is part of commercial solutions or not (Used by industries).

Solution Name	Publicly available code	Online repository statistics			Active community	Used by industries
		# Forks	# Stars	Last Commit		
LIME	YES	1.7K	10.6K	3 years ago	NO	N/A
NLTK	YES	2.7K	11.7K	March 2023	YES	YES
SpaCy	YES	4.1K	25.6K	March 2023	YES	YES
Apache OpenNLP	YES	419	1.2K	March 2023	YES	YES
Stanford CoreNLP	YES	2.7K	9K	March 2023	YES (low)	N/A
Ampligraph	YES	223	1900	March 2023	YES	YES
PySClump	YES	4	16	~ 2019	NO	N/A
LightFM	YES	662	4300	March 2023	YES	YES
Drools	YES	2.4K	5.1K	April 2023	NO	YES
Tensorflow	YES	8.2K	174K	April 2023	YES	YES
scikit-learn	YES	24.2K	54K	April 2023	YES	YES
Hugging Face	YES	20K	96K	April 2023	YES	YES
Prophet	YES	4.4K	15.8K	Feb. 2023	YES	YES

Table 15: Benchmarking of ML tools

As can be seen by examining the above table, there exist various software tools for the purpose of developing machine learning applications. Most of these tools are quite popular, having GitHub statistics in the orders of thousands and active communities surrounding them, while at the same time they are employed by businesses and academics alike. The most popular language for these tools is Python, offering implementations for the vast majority of the available algorithms. Moreover, although for the above tasks commercial solutions can be found, the open-source libraries mentioned are of high quality, so the use of commercial software that would impose both financial costs and probable limitations is avoided.

5.4.1 Evaluation metrics

In this section we mention some metrics that aid in evaluating machine learning tools. It should be noted that these metrics can be used in all the different types of machine learning techniques (e.g., forecasting, RL, NLP). These metrics include:

- Scalability: The amount of data they can handle efficiently.
- Ease of use: Measured as a factor of available documentation, tutorials, ability to handle common data types.
- Flexibility: How many of the related ML tasks can a tool support.
- Accuracy. The performance of ML techniques is usually evaluated along the following metrics:
    - Precision: The percentage of relevant instances among the actual results.
    - Recall: The percentage of relevant instances among those predicted.
    - F-score: A function of precision and recall (their harmonic mean).
    - Confusion matrix: A table that helps in visualizing the prediction results of an ML algorithm (I.e., true positive, true negative, false positive, false negative).
- Transferability: The extent to which an ML toolkit can provide models that can be trained and then reuse their knowledge in different settings.

5.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

In the following table, suggestions for possible usage of Data Analysis and Machine Learning toolkits are provided.

spaCy is suggested due to its open-source nature and its popularity compared to the other NLP toolkits presented above, as suggested by the relevant metrics highlighting its more active community. It is developed in Python, contrary to the other two tools (Apache openNLP and Standford CoreNLP) that are developed in Java, as such, it is easier integrated with other Python libraries such as Tensorflow. Moreover, it displays ease of use through a simple installation process. It is known for its speed, and It supports a variety of supported features for NLP tasks while providing fast implementations.

Ampligraph is a Python library that provides implementations of knowledge graph embedding algorithms that can be useful for the analysis of the structured KG of KAF. It is the only library that offers well documented functions that cover a variety of different techniques applied for knowledge graph embedding. Beyond the most popular algorithms supported, it also provides an implementation for weighted knowledge graph embedding, which falls in an area not so well researched. In addition, it is open source and developed by a leading industry provider, Accenture.

Scikit-learn is one the most popular and flexible Python libraries covering a huge number of applications from simple ML algorithms such as Random Forests and Linear Regression, up to image recognition. It has been starred over 54K times on GitHub and is well documented with a very active community surrounding it. Due to its flexibility, Scikit-learn can be employed in a variety of tools that employ machine learning techniques.

Drools is a rule management system, that can be applied for recommendations based on specific conditions, it is open-source and it is deemed suitable for applications that may require activation upon the satisfaction of some rules.

Solution	Usage in DiMAT
SpaCy	Perform NLP techniques to extract knowledge from materials and manufacturing related documents and populate the Knowledge Graph of Di-KAF toolkit.
Ampligraph	Embedding/analyzing purposes of the Knowledge Graph of Di-KAF toolkit
Scikit-learn	Can be used for classification purposes in order to detect materials with similar properties.
Drools	The Drools engine can be employed in order to provide recommendations on which actions the manufacturing equipment (or the respective DT) should follow, after relevant rules (by industry experts) have been defined.

Table 16: Indicative usage of ML tools in DiMAT

6 BIG DATA COMPUTING TECHNOLOGIES, DATA MANAGEMENT, AND STORAGE

6.1 STATE OF THE ART ANALYSIS

Big data technologies have emerged in recent years to tackle the challenges associated with the processing, analysis, and storage of large, complex data sets [95]. These technologies can be broadly categorized into three categories: data storage, data processing, and data analysis.

Data storage technologies have evolved significantly to keep pace with the increasing demands of big data [96]. Traditional storage systems, such as relational databases, are not well-suited for handling large and complex data sets. As a result, new data storage technologies have emerged, including Hadoop Distributed File System (HDFS) [97], NoSQL databases [98], and cloud storage [102]. The Hadoop Distributed File System (HDFS) is a distributed file system conceived to store and manage extensive data sets across clusters of computers. It offers dependable, scalable, and resilient storage for large-scale data. HDFS is based on the Google File System and is an open-source Apache project. NoSQL databases are non-relational databases that can handle large and complex data sets with ease. They provide a flexible schema and horizontal scalability, making them ideal for big data applications. Some popular NoSQL databases include MongoDB, Cassandra, and Apache HBase. Cloud storage provides on-demand access to shared resources and services over the internet. Cloud storage providers, such as Amazon Web Services (AWS) and Microsoft Azure provide a diverse array of storage choices, encompassing object storage, block storage, and file storage. Cloud storage provides scalable, durable, and secure storage for big data.

Data processing technologies are used to transform and analyze large and complex data sets [103]. Traditional data processing technologies, such as batch processing, are not well-suited for handling big data. Emerging data processing technologies include Apache Spark, Apache Storm, and Apache Flink. Apache Storm is a distributed computing system designed for real-time stream processing, based on an open-source framework. It offers fault-tolerant processing capabilities [99] of streaming data and can process millions of messages per second. Storm is used for applications such as real-time analytics, fraud detection, and internet of things (IoT) data processing. Apache Flink is a distributed computing system specifically developed for processing data streams, and it is available as an open-source solution [100]. It provides a unified API for stream processing and batch processing, making it easy to build data pipelines that can process both types of data. Flink can process data in real-time and can scale to handle large data sets.

Data analysis technologies are used to analyze and derive insights from large and complex data sets. Traditional data analysis technologies, such as statistical analysis and data mining, are not well-suited for handling big data. New data analysis technologies have emerged, including Apache Hadoop. Hadoop is a distributed computing system built for processing big data, utilizing an open-source framework [101]. It provides a scalable, fault-tolerant framework for batch processing of large data sets. Hadoop consists of two primary components: the Hadoop Distributed File System (HDFS) and MapReduce. MapReduce is a computational framework specifically designed for handling large-scale datasets, enabling their creation and manipulation.

These big data technologies have many applications, including materials design and simulation. Big data technologies can help materials scientists to accelerate the discovery of new materials by analyzing vast amounts of data. For example, machine learning algorithms can be used to predict the properties of materials based on their chemical composition.

To summarize, big data technologies have revolutionized the way we handle large and complex data sets. These technologies have many applications, including materials design and simulation. With the continuous growth of big data, it is likely that newer, more advanced big data technologies will surface to meet the demands of this rapidly evolving field.

6.2 EU RELATED PROJECTS

Acronym	Grant Agreement No.	Title	Description
MarketPlace	760173	Materials Modelling Marketplace for Increased Industrial Innovation	MarketPlace aims to integrate all scattered modelling components from materials modelling and industrial communities and provide a single point of access to all materials modelling activities in Europe. The knowledge management and app store concepts that were first introduced in this project have strongly influenced

			the design of CMDB toolkit of DiMAT.
I4Q	958205	Industrial Data Services for Quality Control in Smart Manufacturing	I4Q aspires to transform the way factories manage and analyze large amounts of data, resulting in more efficient and sustainable operations.

Table 17: Big Data EU Funded Projects

The MarketPlace project's focus is a centralized, open web-based platform to accelerate materials modeling in industry, offering comprehensive integration with a wide array of resources in materials science. It includes capabilities for collaboration on models, access to simulation tools, and training resources, with mechanisms for interoperable workflow integration across models. MarketPlace's collaborative approach aligns with DiMAT's goal to create an accessible digital ecosystem for materials sciences, facilitating real-time data sharing, interoperability, and integration of materials modeling activities across Europe. The platform's federated databases and open simulation tools support DiMAT's 'Data and Assessment Suite' by streamlining data access, management, and interoperability for a wide range of materials modeling tasks, ensuring scalability and adaptability in modeling workflows. The i4Q project provides a comprehensive suite of IoT-based tools, designed to manage large volumes of industrial data with high reliability and real-time performance. This suite is particularly suited for zero-defect manufacturing, offering capabilities such as continuous process qualification and quality diagnostics, which are essential for modern digital manufacturing. DiMAT's 'Simulation and Optimization Suite' can leverage i4Q's real-time data analysis tools for optimizing materials manufacturing processes, contributing to DiMAT's goals of enhancing quality, sustainability, and efficiency in materials production. i4Q's reliable data management and process reconfiguration capabilities support the integration of advanced digital technologies, especially for SMEs aiming to adopt sustainable, data-driven manufacturing practices.

6.3 EXISTING SOLUTIONS

Open-Source Solutions	Description
i4QDA (On Edge)	The i4Q Data Analytics (i4QDA) solution is designed to optimize manufacturing processes and improve product quality through

	data-driven predictions. Operating on edge devices, it offers seamless integration with existing hardware infrastructure. i4QDA uses Docker containers for configuration, enabling the scaling of container numbers in the service upon execution.
i4QBDA (Cloud)	The i4Q Big Data Analytics Suite (i4QBDA) enhances the capabilities of the i4Q Data Analytics (i4QDA) solution by providing a ready-to-implement software package designed for cloud instances. The i4QBDA solution is accessible through a web-based user interface, enabling users to select and customize supported technologies prior to deployment, and it adapts to the cloud infrastructure, ensuring optimal configuration for the available resources.
Apache Cassandra	An open-source NoSQL database platform, known for its capacity to handle vast quantities of data across numerous standard servers. It is engineered to ensure continuous availability, eliminating any single point of failure by utilizing a distributed wide column store architecture.
Apache Solr	Solr is a popular and fast open-source search platform, suitable for covering e-commerce and data analytics needs. It is a tool built on Apache Lucene.
MongoDB	A source-available cross-platform document-oriented database program. Essentially a NoSQL database program, MongoDB uses JSON-like documents with optional schemas.
Apache NiFi	An open-source, real-time data ingestion platform that enables users to automate and manage the flow of data between systems.
Elasticsearch	A search engine based on the Lucene library, providing a distributed, multitenant-capable, full-text search engine with an HTTP web interface and schema-free JSON documents.
RabbitMQ	An open-source message-broker software that originally implemented the Advanced Message Queuing Protocol (AMQP) and has since been extended with a plug-in architecture to support Streaming Text Oriented Messaging Protocol (STOMP), Message Queuing Telemetry Transport (MQTT), and other protocols.
Redis	A project that provides a distributed, in-memory key-value store with the possibility of persistence. It functions primarily in memory and offers a variety of data structures.

PostgreSQL	An open-source Relational Database Management System (RDBMS) known for its focus on scalability and adherence to SQL standards.
ASE (Atomic Simulation Environment)	ASE consists of Python modules and tools designed to facilitate the execution, manipulation, visualization, and analysis of atomistic simulation tasks.
pymatgen (Python Materials Genomics)	A powerful and open-source Python library for materials analysis. It's used for creating, analyzing and manipulating materials data.
AiiDA (Automated Interactive Infrastructure and Database for Computational Science)	An adaptable and expandable platform that facilitates the organization, safeguarding, and sharing of data, simulations, and workflows in the realm of current computational science, while maintaining reproducibility.
Open Babel	A chemical toolbox designed to be compatible with the many languages of chemical data. It facilitates searching, storing and analysis of data from areas such as molecular modeling, chemistry, biochemistry.
RDKit	An open-source cheminformatics and machine learning software that makes it easy to build and manipulate molecules, calculate properties, build predictive models, and more.

Table 18: Existing Big Data Tools (open-source & commercial)

6.4 BENCHMARKING AND EVALUATION

In our benchmarking of the solutions listed above, we employed several metrics: (1) the public availability of the tool, (2) online repository statistics (including the number of forks, stars, and the date of the last commit) to evaluate whether the tool is actively maintained and developed, (3) evidence of an active community, inferred from the repository statistics, and (4) the tool's usage in industrial applications, determined through an analysis of relevant webpages. Please note that MongoDB and Elasticsearch were added to the table later, which accounts for the difference in their last commit dates compared to the other tools.

Solution Name	Publicly available code	Online repository statistics			Active community	Used by industries
		#Forks	#stars	Last commit		
PostgreSQL	YES	3,887	12,350	May 2023	Yes	Yes
RabbitMQ	YES	3,885	10,671	May 2023	Yes	Yes
MongoDB	YES	5.574	26.285	October 2024	Yes	Yes
Redis	YES	22,537	59,812	May 2023	Yes	Yes
ASE	YES	72	144	May 2023	Yes	N/A
AiiDA	YES	162	358	May 2023	Yes	Yes
pymatgen	YES	736	1,096	May 2023	Yes	N/A
Apache Cassandra	YES	3,378	7,962	May 2023	Yes	Yes
Apache Solr	YES	447	733	May 2023	Yes	N/A
Apache nifi	YES	2,464	3,769	May 2023	Yes	N/A
Elasticsearch	YES	24,800	70,000	October 2024	Yes	Yes
Open Babel	YES	390	809	May 2023	Yes	N/A
RDKit	YES	765	2,029	May 2023	Yes	N/A
i4QBDA	NO	N/A	N/A	February 2024	Yes	Yes
i4QDA	NO	N/A	N/A	April 2024	Yes	Yes

Table 19: Benchmarking of Big Data tools

As can be seen from the previous subsections there exists a variety of solutions for handling Big Data related tasks such as storing and data analysis with most of them being open-source. Tools were assessed based on key metrics such as scalability, reliability, interoperability, and alignment with DiMAT's core digitalization needs in materials sciences. For example, MarketPlace's federated approach to data and model interoperability closely aligns with DiMAT's vision of creating integrated materials data and simulation workflows. Similarly, i4Q's suite was evaluated for its suitability in real-time data handling and process optimization, critical for DiMAT's aim of achieving high efficiency in digital manufacturing. Specific tools were included in the 'Potential Usage in DiMAT' list based on their performance under these evaluation metrics. decisions were based on factors like community support, ease of integration, and specific performance under DiMAT's targeted use cases in materials design and simulation.

6.4.1 Evaluation Metrics

In this section we mention some metrics that aid in evaluating the available big data and data management tools. Some of these tools are dependent on the probability that big data will be used in the project:

- Storage Capacity: The total volume of data that can be stored.
- Scalability: The capacity to manage increasing workloads in an efficient and seamless manner. This includes both vertical and horizontal scalability.
- Processing Speed: The speed at which data can be processed and analysed. This can be measured in terms of latency (time to process a single item) and throughput (items processed per unit of time).
- Data Transfer Rate: The speed at which data can be transferred to and from the storage system.
- Concurrency: The number of tasks that can be processed simultaneously without degrading system performance.
- Fault Tolerance: The ability of a system to continue functioning when there are failures in some of its components.
- Data Durability: The likelihood that data will not be lost in the event of a system failure.
- Data Consistency: Ensuring that all copies of the data show the same values, even after updates. This is important in distributed systems.
- Data Availability: The ability to access the data whenever needed. This is usually measured as a percentage of uptime.
- Data Recovery: The speed and ease with which data can be recovered after a disaster.
- Security: The ability to prevent unauthorized access and protect sensitive data. This includes both data at rest and data in transit.
- Compliance: Adherence to regulations governing data handling, storage, and processing. This can include regulations such as GDPR, HIPAA, etc.

6.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

The use of some of the following tools is dependent on the use of Big Data within the project. Each tool included has been evaluated for its compatibility and potential impact on DiMAT's objectives. Selection criteria focused on the tools' scalability, data integration capabilities, and specific strengths in managing materials data, supporting digital modelling workflows, and optimizing manufacturing processes. Tools from various domains—data storage, processing, analysis, and simulation—were chosen based on their demonstrated performance and suitability for DiMAT's needs. By leveraging these tools, DiMAT aims to provide European SMEs with cost-effective, flexible, and advanced digital solutions to improve materials design, production efficiency, and overall sustainability across the materials value chain.

From the below mentioned tools, PostgreSQL and Elasticsearch can find potential usage in the Cloud Materials Database toolkit (CMDB). PostgreSQL can be employed in CMDB for handling large datasets (a common occurrence when working with materials datasets) and high workloads. The tools are open source and have a strong community. In addition, Elasticsearch could play a role in CMDB due to its capabilities of analyzing large volumes of data. Elasticsearch also fulfills the benchmark prerequisites also.

ASE (Atomic Simulation Environment): Open-source Python framework that provides flexibility for handling atomic structures and automating simulations with various calculators, ideal for custom workflows and high-throughput calculations.

OpenBabel: Open-source tool with extensive support for chemical file conversion and structure manipulation, enabling seamless interoperability between different file formats and molecular editors.

pymatgen (Python Materials Genomics): Open-source library with powerful tools for analyzing and manipulating crystal structures and computing properties, facilitating integration with data from the Materials Project database and DFT calculations.

Solution	Usage in DiMAT
MongoDB	A document-oriented database program that is available as a source code, compatible across different platforms. It is categorized as a NoSQL database program and utilizes JSON-like documents, which adhere to a prevalent schema throughout the project.
RabbitMQ	Message queuing technology that can be used to handle computational jobs.
Redis	Redis provides lightning-fast data access by storing data in memory, making it ideal for caching, real-time analytics, and session management.
Apache nifi	Apache NiFi simplifies data flow automation by enabling the seamless movement, transformation, and

	management of data between systems. Its intuitive, drag-and-drop interface and real-time monitoring make it ideal for integrating diverse data sources with minimal coding.
PostgreSQL	A tool for managing relational databases that toolkits (such as CMDB) will need.
Apache Cassandra	Its main advantage lies in its high availability, ensuring that there is no single point of failure.
ASE (Atomic Simulation Environment)	ASE can enable manipulating, running, visualizing and analyzing atomistic simulations.
pymatgen (Python Materials Genomics)	pymatgen is used for creating, analyzing and manipulating materials data.
AiiDA (Automated Interactive Infrastructure and Database for Computational Science)	AiiDA facilitates the management, preservation, and dissemination of simulations, data, and workflows within the realm of modern computational science.
Elasticsearch	Elasticsearch enables fast searching, analyzing, and visualizing of large data volumes in real-time, ideal for log monitoring, full-text search, and data analytics with easy integration for visualization and data ingestion.
Open Babel	Open Babel provides a platform for individuals to search, convert, analyze, and store data originating from molecular modeling, chemistry, biochemistry, and other related fields.
RDKit	RDKit simplifies the process of constructing and modifying molecular structures, evaluating their characteristics, creating predictive models, among other functionalities.

Table 20: Indicative usage of Big Data tools in DiMAT

Co-funded by the European Union

7 MANUFACTURING SIMULATION PLATFORMS

7.1 STATE OF THE ART ANALYSIS

The concept of manufacturing simulation and simulation modeling for manufacturing systems [104][105] first emerged between the 1950s and the 1960s. Over the last 60 years, this concept has been updated and adapted as digital technology advances. Since then, the trend has been towards developing digital tools for integrating networked manufacturing, visualization, and analysis of manufacturing procedures before being carried out on real scenarios, which involve all stages of a product cycle, from conceptualization and design to delivery to the end users. About 15 years ago, a research paper described the importance of simulation in manufacturing. The authors stated that the importance of simulation in manufacturing is due to the possibilities offered by simulation programs to predict the behavior of a previously designed production system, in addition to adding benefits to decision-making in terms of reduced planning and analysis time, reduced investment costs, and the security provided [106]. So far, this importance has remained the same since more and more applications of digital technologies are developed and employed in different industry sectors [107].

Numerous Manufacturing Simulation Platforms (MSP) have been developed with a focus on integrating multimethod-simulation modeling tools. These tools enable the simulation and evaluation of system changes' impact on manufacturing processes, the layout of physical facilities, logistics, transportation, and supply chain management [108]. The modeling of these platforms, such as Anylogic® [109][110], Goma, OpenFOAM® [111], Simul8, COMSOL Multiphysics®, etc., are based on either statistical data, finite element analysis, or drawing schemes; it all depends on the architecture of the platform and the concern of its creation.

Besides that, other MSPs (e.g., Abaqus©, Moldex3D®, and Autodesk Netfabb) [112] are developed to simulate specific scenarios of manufacturing processes or allow the analysis of different manufacturing scenarios and processes regardless of their nature, with just a few functional adjustments, like building new solvers based on the needs. A 2020 research case illustrates this. The authors studied the three-dimensional (3D) simulation of the filling stage of thermoplastic materials injection molding with the help of a solver built for an open-source simulation platform [113].

The prediction of residual stresses and shape deformation in printed geometries is another study showing the effectiveness of developing and implementing specific simulation tools within manufacturing simulation platforms. In this study, the authors used Abaqus© software (propriety) to create the simulation tool to model thermoplastic composites' Extrusion Deposition Additive Manufacturing. As a result, they obtained a strong correlation between modeled measure and the validation experiment [114].

According to the literature review and the many publications concerning this field, deposition-based Additive Manufacturing is one of the processes that has gained relevant attraction to implementing modeling software for its optimization [115][116][117]. In addition, other processes -including the design of extrusion die for thermoplastic materials and the cooling stage in the extrusion process [118][119][120], have also been studied by simulation modeling.

In light of the above, simulation modeling in manufacturing processes is considered an efficient and practical tool that allows for optimizing the performance of any process prior to its execution in a building facility with real tooling.

7.2 EU RELATED PROJECTS

The European Union, during the past years, has (co-)funded numerous projects dedicated to developing Manufacturing Simulation platforms. Some of these projects are listed in the following table. Taking into consideration that in the framework of these projects, the simulation platforms have been developed for specific purposes, the objective of gathering them into a summarized table is to have a brief description of their functionality and evaluate the possibility of whether any of them can be used within the DiMAT project.

Acronym	Grant Agreement No.	Title	Description
SIMPHONY	604005	Simulation framework for multi-scale phenomena in micro- and nanosystems	Developing an extendable, easy-to-use, freely available, and open platform for integrating various simulation and pre- and post-processing software. For advances in multiscale nano-enabled materials modeling by significantly reducing the gap between materials modeling and European nanotechnology industries.
MODENA	604271	Modeling of morphology Development of micro- and nanostructures	MoDeNa took the task to construct a generic computational environment that combines an arbitrary number of computational methods solving problems local to the applicable scale and material and letting different software modules interact so that a product's

			behavior and properties are simulated and predicted.
COMPOSELECTOR	721105	Multi-scale Composite Material Selection Platform with a Seamless Integration of Material Models and Multidisciplinary Design	The objective was to develop a Business Decision Support System (BDSS), which integrates materials modeling, business tools, and databases into a single workflow to support the complex decision process in selecting and designing polymer-matrix composites. This will be achieved by an open integration platform which enables interoperability and information management of materials models and rich materials modeling layer with industry-standard business process models.
Neosaware	887621	Modeling Software platform for ceramic body optimization	Enhancing ceramic body composition is instrumental to developing new products, increasing profitability, and reducing environmental impact. Traditional methods could be more efficient. NEOSAWARE project proposes a software solution based on artificial intelligence, computer simulation, and mathematical optimization to hasten the learning process. The solution will allow the industry to carry out exact forecasts during the manufacturing process and on the final product. The solution reduces production costs, guarantees quality standards, and decreases environmental impact.
ReaxPro	814416	Software Platform for Multiscale Modelling of Reactive	Reactive process design has primarily been based on trial-and-error experimentation and similarly. On the other hand, physics-based modeling approaches are emerging as highly promising in developing

		Materials and Processes	new catalytic materials and reactive processes. ReaxPro has identified a set of academic software tools which will be upscaled into easy-to-learn, user-friendly.
NanoSim	604656	A Multiscale Simulation-Based Design Platform for Cost-Effective CO2 Capture Process using Nano-Structured Materials	NanoSim project created an efficient and cost-effective multi-scale simulation platform based on free and open-source codes in the context of reactive fluid-particle flows. This platform connected models of various types and spanned an extensive range of scales. Specifically, the project achieved to connect atom-scale phenomena and mesoscale phenomena in fluid-particle systems with process-scale aspects such as economic feasibility or optimal process conditions. Therefore, the NanoSim project developed one offline linking and one online coupling software tool.

Table 21: Materials Manufacuring and Simulation EU Funded Projects

7.3 EXISTING SOLUTIONS

Simulation can provide a helpful way to identify, understand, and address improvement opportunities related to the operations that make up the manufacturing network. For such cases, the use of open codes is recommended, allowing reducing accessibility and implementation costs. In addition, open code software allows the modification or incorporation of new functionality to be used and adapted according to the needs by recoding the main source of the program, which can be an essential feature for the development of the Manufacturing Processing simulator toolkit within the DiMAT project.

Several open-source software solutions are available for Computational Fluid Dynamics (CFD) simulation. Some of the most popular ones are listed in the following table, which briefly describes their functionality.

Open-Source Solutions	Description
OpenFOAM®	This widely used open-source CFD software supports a wide range of solvers for different types of flows -including incompressible, compressible, multiphase, and turbulent flows. It also includes a pre-processing tool for mesh generation and a post-processing tool for visualizing simulation results [121].
SU2	Besides the various types of flows that SU2 supports, like compressible, incompressible, and multiphase flows, it also includes an optimization module that can be used to optimize the design of aerodynamic systems [122].
SALOME – MECA software	SALOME is a numerical simulation platform that involves different pre-processing steps including code development. SALOME software can also be used for post-processing, such as visualization. This multi-module software integrates individual interfaces for each module using two languages (C++ and Python) [123].
Gmsh	Gmsh is a mesh generator that can generate high-quality meshes for CFD simulations. It supports various mesh formats and can create both structured and unstructured meshes [124].
FEniCS	FEniCS is a popular open-source finite element software that can solve various partial differential equations, including those involved in CFD simulations. It includes a range of solvers and can simulate steady-state and transient flows [125].
Code_Saturne	Code_Saturne is an open-source CFD software used as a pre-processing tool for mesh generation and a post-processing tool for visualizing simulation results. As SU2 software, Code_Saturne can simulate different types of flows (compressible, incompressible, and multiphase flows) [126].

Table 22: Materials Manufacturing and Simulation open-source solutions

These are just a few examples of today's open-source CFD software packages. Each software package has its strengths and weaknesses, and the choice of software depends on the user's specific needs.

For Finite Element Method (FEM) simulation, commercial software packages will be the first choice due to the capacity and mature development, but open-source tools are also in consideration. Some examples are listed in the following table, including open-source and commercial ones.

Solutions	Description	Open-source
Abaqus	Abaqus is a popular commercial software suite that uses mathematical modeling throughout the finite element method. It is used for a wide range of engineering problems in both industry and academia.	No
Digimat	Digimat is a commercial software able to simulate in detail the behaviour of non-homogeneous materials during, but not only, manufacturing processes.	No
Calculix	Calculix is a free finite element analysis software suite that is designed to solve field problems using the finite element method. It can be used for linear and non-linear calculations, static, dynamic and thermal analysis.	Yes
FreeFEM	FreeFEM is described as a partial differential equation solver, specifically used in non-linear multi-physics frameworks (1D, 2D, 3D, and 3D border domains). It is a popular PDE solver used by thousands of researchers across the world.	Yes
FEniCS	As described in the previous table.	Yes
Elmer	Elmer composes an open-source finite element software for Multiphysics processes, co-developed by CSC – IT Center for Science.	Yes
Code-Aster	Code Aster is a free finite element software, developed by the French company EDF and the French Atomic Energy Commission (CEA). It solves mechanical, thermal, and multiphysics problems.	Yes
TexGen	TexGen is a free software to generate 3D models of composite materials for use in finite element analysis. It is developed by the University of Sheffield.	Yes

Table 23: Finite Element Method (FEM) simulation tools

Multiscale modelling will involve molecular dynamics simulations to determine material properties. Some available free software solutions are presented below:

Solutions	Description	Open-source
GROMACS	GROMACS (GROningen MAchine for Chemical Simulations) is a molecular dynamics simulation software developed by the University of Groningen and the Max Planck Institute.	Yes
LAMMPS	LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator). This software allows analyzing molecule and atom physical displacements, enabling, thus, the behavior in a lapse of time.	Yes
Prepomax	PrePoMax is an open source pre and post-processor using Calculix FEM solver and an interface to analyses the results of the simulations.	Yes
FreeCAD	FreeCAD is an open-source 3D CAD computer-aided engineering and design for mechanical engineering assistance and mechanical element design. This software allows the use of different digital tools to create an accurate environment for recreating some industrial conditions.	Yes

Table 24: Multiscale Modelling tools

7.4 BENCHMARKING AND EVALUATION

Typical tasks of different numerical simulation platforms are the pre-and post-processing steps that can be carried out depending on the platform architecture. However, complex numerical simulation requires the interaction of varying simulation codes that sometimes are not integrated into a single software. Therefore, finding a simulation platform that addresses various issues and unifies all the basic functionalities, besides allowing the creation of additional modules, represents a competitive solution against other simulation platforms that offer limited functionalities. Consequently, as the Manufacturing Simulation platforms described in the EU Related Projects subsection 7.2 are developed for specific solutions and do not meet the specifications of a simulation platform that integrates all the basic functionalities for recording and creating additional modules, they were not included in this benchmarking and evaluation section. Instead, we included in this section, those solutions listed in the Existing Solutions subsection 7.3, which allows code development and wide integration of simulation codes. To assess the performance, handling complexity, and critical aspects of those open-source and commercial software solutions available for Computational Fluid Dynamics (CFD) and Finite Element (FE) simulation, we employed various metrics.

The benchmarking process began with a comprehensive evaluation of generic aspects such as the programming language used and publicly available code. We then delved into the realm of community participation and access to different platforms from the programmer's website, including tutorials, use of forums, and file-sharing libraries. This exploration extended to the YouTube platform, where tutorials and solutions to frequent problems are especially valued. Additionally, we searched repositories like GitHub to understand the involvement of a more specialized community.

To encapsulate these aspects, we adopted the terms; Higher (H) for high community participation, Medium (M), and Lower (L) for reduction of participation on websites. Additionally, we conducted a search on software websites to determine their usage by companies. Lastly, we examined the computing requirements necessary for executing the software.

Solution Name	Language Code	Publicly available code	Online Community			Specialty	Used by industries	Minimum Requirements	
			website	You Tube	Git Hub			RAM	Proc/or Cores
Open Foam	C++/Python	YES	H	H	H	Computational fluids dynamics/Allows recoding	YES	16GB	4
SU2	C++	YES	H	H	H	Multiphysics analysis and aerodynamic design optimization	YES	16GB	4
Salome-MECA Software	C++/Python	YES	M	M	M	Multiplatform of industrial studies of physics simulations	YES	16GB	4
Gmsh	C++	YES	L	M	H	A meshing tool with parametric input and flexible visualization capabilities	YES	N/A	N/A
Fenics	C++/Python	YES	M	L	M	Platform solving partial	N/A	N/A	N/A

Co-funded by the European Union

						differential equations			
Code_Saturne	C++/Python	YES	M	H	M	Computational fluid dynamics	YES	16GB	4
Digimat	N/A	NO	N/A	N/A	N/A	Multiphase material simulation	YES	4GB	1
Abaqus	Python	NO	M	M	M	FEA software with wide materials simulation	YES	16GB	4
Calculix	Python	YES	L	H	H	FEA software with wide materials simulation	YES	N/A	2
FreeFEM	C++	YES	M	M	M	Multiphysics materials analysis	YES	N/A	N/A
Elmer	C++	YES	M	M	L	Multiphysics materials analysis	YES	8	2
Code-Aster	Python	YES	L	L	L	Multiphysics materials analysis	YES	8	2
TexGen	Python	YES	L	L	L	Modelling textile geometries and variety of properties	YES	N/A	N/A
GROMACS	Python	YES	L	L	M	Multiphase molecular dynamic simulation	YES	16	4
LAMMPS	C++	YES	L	L	L	Multiphase molecular dynamic simulation	YES	N/A	N/A
FreeCAD	C++/Python	YES	M	H	H	Multiplatform modelling and extensible simulation	YES	8	4
PrePoMax	Python	YES	M	M	M	Multiphysics materials analysis	YES	16	4

Table 25: Benchmarking of Materials Modelling and Simulation Tools

Co-funded by the European Union

7.4.1 Evaluation Metrics

Herein are defined several metrics by which both the open-source and commercial software solutions available for Computational Fluid Dynamics (CFD) and Finite Element (FE) simulation, listed in the above section, can be evaluated for the DiMAT project.

- *Handling complexity*: How intuitive the platform is for being operated.
- *Reliability, Software stability/software release life cycle*: Background and today's information on the continuous updating and improvement of the software technology.
- *Efficiency*: The level of values or ratio of inputs and outputs. The number of inputs operating needed for coming across desirable outputs.
- *Maintainability*: The accessibility of keeping the system debugged and managed in troubleshooting. As well as extending new features or functionality.
- *Capabilities:* The set of physical models that the tool implements and can simulate.

7.5 POTENTIAL USAGE IN DIMAT (RELEVANT TOOLS)

After considering the aspects that characterized each software package from the list previously presented in the benchmarking subsection 7.4, and the evaluation metrics subsection 7.4.1, several of these platforms were excluded as solutions for potential usage in the development of the Material Processing Simulator (MPS) or the Material Designer (MD) toolkits within the framework of the DiMAT project.

For the MPS toolkit, the first criterion was to consider whether the software package to be used consisted of free and open-source or commercial ones. Consequently, the following software package, which implies a paid commercial license, was excluded:

- Abaqus

Continuing with the evaluation, the next criterion was the *Specialty* of each software package and its alignment with DiMAT aims. In this sense, the following software packages were excluded since they are not aligned with DiMAT aims:

- SU2: its specialty is Multiphysics analysis and aerodynamic design optimization.
- TexGen: this software is more related to the modeling of textile geometries.
- GROMACS: this software package is more related to the molecular dynamic method, for simulating proteins and lipids.
- LAMMPS: this software package is more related to the molecular dynamic method, which uses spatial decomposition techniques.

Judging by the evaluation metrics described in subsection 7.4.1, the following software packages were excluded:

- Gmsh, which is a meshing tool with parametric input. However, in terms of capacities, the same function can be operated with more complete software that includes other functionalities, such as PrePomax.
- Fenics, despite being an open-source software for solving partial differential equations (PDEs) with the finite element method (FEM), in terms of *maintainability*, the online community refers to the disadvantages and issues when running in different environments, such as the Windows Subsystem for Linux (WSL).
- Code_Saturne: Compared with more popular and more efficient software for the resolution of computational fluid dynamics problems, such as OpenFoam, the Code_Saturne package possesses lower community support and less *reliability*, which means that the background information, continuous updates, and software improvement are behind from other CFD software.
- Calculix: although this open-source package can be used as pycalculix in the Python environment, it can also be used integrated into another software package, such as the case of PerPomax and OpenFoam, as explained ahead.
- Elmer: As a finite element software, it can be useful for different domains, such as acoustic, fluid mechanics, heat transfer, and electromagnetics. However, its best performance is reached when calculating electromagnetism problems.
- Code_aster: Compared with other software for finite element analysis, such as Calculix, the code_aster is described as more powerful and efficient. However, when it comes to *Handling complexity,* it is also the hardest to operate.

Finally, after applying the evaluation metrics, together with the above-mentioned criterion for selecting the relevant tools, the following software packages were identified as solutions for potential usage within the DiMAT project and specifically for implementing the Material Processing Simulator (MPS) or Material Designer (MD) toolkit.

- The first one is OpenFOAM®, described as a free and open-source toolbox that allows the development of new numerical solvers for specific applications.

- The second one is SALOME-MECA software, a generic platform of free license for pre-/and post-processing.

- The third one is FreeCad, a free and open source that, in terms of *handling complexity*, is accessible, multiplatform, and intuitive. In terms of *Reliability*, this software is highly customizable and extensible, capable of reading and writing in vast numbers of file formats, Such as TEP, IGES, STL, SVG, DXF, OBJ, etc. In terms of *efficiency*, it has a wide range of features, including the ability to import and export various file formats, create and edit 2D and 3D sketches, generate meshes, and perform finite element analysis. In terms of *maintainability*, the support of a great number of users is available and it has a geometry modeler.

- PrePomax, is s an open source pre and post-processor for the Calculix FEM solver. In terms of *handling complexity, it is* simple and intuitive to use. It has become one of the most popular software of its kind, which makes it a *Reliability* option. In terms of *capabilities,* it is one of the best and most user-friendly interfaces with a powerful efficiency of performance. Also, this program has a geometry importer and a Mesher capable of creationg mesh in tetrahedral form.

- Finally, Digimat is a valid candidate for usage in toolkits where the behavior of complex and non-homogeneous materials needs to be modeled and investigated. Despite being a commercial tool, its capabilities make it a potential candidate for usage in the Material Designer toolkit.

The following table briefly describes the role of the selected software to be implemented in developing the MPS or the MD toolkits.

Solution	Usage in DiMAT
FreeCAD	FreeCAD can be employed to design the elements that integrate the material processing simulator.
OpeanFOAM®	It can be used as a multifunctional platform and solver for material processing after indicating its main conditions and parameters.
PrePoMax	It is an interactive pre and post-processor for the Calculix FEM solver wiht capacity for Linear and nonlinear static analyses
SALOME – MECA software	Salome can be used as an interface to visualize the behavior of models previously processed in FreeCAD and OpenFoam.
Digimat ®	It can be used to investigate the mechanical behavior of non-homogeneous and multiphase materials

Table 26: Indicative usage of tools in DiMAT

8 CONCLUSIONS

In this document, we identify technologies relevant to the DiMAT project's scope based on input from numerous beneficiaries. For each such technology, first, the state-of-the-art was characterized with emphasis on the progress marked in the materials design, simulation and manufacturing sectors. The sources for this encompassed both academic literature and professional reports. We also highlighted related EU projects, many of which involve project partners, to underline the significance that the EU places on these fields. In addition, we provided available tools, focusing mainly on open-source solutions. For these tools, we proposed a benchmarking framework based on qualitative metrics. The purpose of this framework is twofold. On the one hand, it aids in highlighting the popularity, the availability and the processing requirements of the tool, indicating how appropriate its selection could be for accomplishing DiMAT goals. On the other hand, it provides an overview of the current status of each technology sector. Moreover, we proposed certain evaluation metrics per technology, intending to guide the developers who will implement DiMAT's suites in selecting the right tool for component development. Finally, we presented indicative tools for some of the DiMAT's components.

REFERENCE

[1] Horsch, M. T., Chiacchiera, S., Seaton, M. A., Todorov, I. T., Šindelka, K., Lísal, M., & Cavalcanti, W. L. (2020). Ontologies for the virtual materials marketplace. *KI-KünstlicheIntelligenz*, *34*, 423-428.

[2] Ontotext. Online: https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

[3] Ontotext. Online: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

[4] EMMO. Online: https://github.com/emmo-repo/EMMO/blob/master/top/top.ttl

[5] Morgado, J. F., Ghedini, E., Goldbeck, G., Hashibon, A., Schmitz, G. J., Friis, J., & Anne, F. (2020). Mechanical Testing Ontology for Digital-Twins: a Roadmap Based on EMMO. In *SeDiT@ ESWC*.

[6] Horsch, M. T., Toti, D., Chiacchiera, S., Seaton, M. A., Goldbeck, G., & Todorov, I. T. (2021). OSMO: Ontology for simulation, modelling, and optimization.

[7] Horsch, M. T., Chiacchiera, S., Schembera, B., Seaton, M. A., & Todorov, I. T. (2020). Semantic interoperability based on the European Materials and Modelling Ontology and its ontological paradigm: Mereosemiotics. *arXiv preprint arXiv:2003.11370*.

[8] Li, H., Armiento, R., &Lambrix, P. (2020). An ontology for the materials design domain. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19* (pp. 212-227). Springer International Publishing.

[9] Andersen, C. W., Armiento, R., Blokhin, E., Conduit, G. J., Dwaraknath, S., Evans, M. L., & Yang, X. (2021). OPTIMADE, an API for exchanging materials data. *Scientific data*, *8*(1), 217.

[10] Mrdjenovich, D., Horton, M. K., Montoya, J. H., Legaspi, C. M., Dwaraknath, S., Tshitoyan, V., & Persson, K. A. (2020). Propnet: a knowledge graph for materials science. *Matter*, *2*(2), 464-480.

[11] McCusker, J. P., Keshan, N., Rashid, S., Deagen, M., Brinson, C., & McGuinness, D. L. (2020, November). NanoMine: A knowledge graph for nanocomposite materials science. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II* (pp. 144-159). Cham: Springer International Publishing.

[12] Zhang, X., Liu, X., Li, X., & Pan, D. (2017). MMKG: An approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia. *Computer Physics Communications*, *211*, 98-112.

[13] Strötgen, J., Tran, T. K., Friedrich, A., Milchevski, D., Tomazic, F., Marusczyk, A., & Kharlamov, E. (2019, October). Towards the Bosch Materials Science Knowledge Base. In *ISWC (Satellites)* (pp. 323-324).

[14] Venugopal, V., Pai, S., & Olivetti, E. (2022). MatKG: The Largest Knowledge Graph in Materials Science--Entities, Relations, and Link Prediction through Graph Representation Learning. *arXiv preprint arXiv:2210.17340*.

[15] Cearley, D., Burke, B., Smith, D., Jones, N., Chandrasekaran, A., and Lu, C. (2019). Top 10 strategic technology trends for 2020. Technical report, Gartner, Stamford, CT, USA.

[16] Digital Twin Market Growth Analysis, Industry Report, Forecast, 2031. Online: https://www.marketsandmarkets.com/Market-Reports/digital-twin-market-225269522.html/. Accessed 20/2/2023.

[17] ISO 23247-1:2021 Automation systems and integration — Digital twin framework for manufacturing — Part 1: Overview and general principles. Online: https://www.iso.org/standard/75066.html/. Accessed: 20/2/2023.

[18] Sepasgozar, S. M. (2021). Differentiating digital twin from digital shadow: Elucidating a paradigm shift to expedite a smart, sustainable built environment. Buildings, 11(4), 151.

[19] Kherbache, M., Maimour, M., & Rondeau, E. (2022). Digital Twin Network for the IIoT using Eclipse Ditto and Hono. IFAC-PapersOnLine, 55(8), 37-42.

[20] Connect, Command & Control IoT Devices: Eclipse Hono. Online: https://www.eclipse.org/hono/. Accessed: 20/2/2023.

[21] Eclipse Ditto – open source framework for digital twins in IoT. Online: https://www.eclipse.org/ditto/. Accessed: 20/2/2023.

[22] Neto, A. A., da Silva, E. R., Deschamps, F., & de Lima, E. P. (2021). Digital twins in manufacturing: An assessment of key features. Procedia CIRP, 97, 178-183.

[23] Zambrano, V., Mueller-Roemer, J., Sandberg, M., Talasila, P., Zanin, D., Larsen, P. G., & Stork, A. (2022). Industrial digitalization in the industry 4.0 era: Classification, reuse and authoring of digital models on Digital Twin platforms. Array, 14, 100176.

[24] DIGITbrain. Online: https://digitbrain.eu/. Accessed: 20/2/2023.

[25] Qin, H., Wang, H., Zhang, Y., & Lin, L. (2021, May). Constructing digital twin for smart manufacturing. In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 638-642). IEEE.

[26] Shao, G., &Helu, M. (2020). Framework for a digital twin in manufacturing: Scope and requirements. Manufacturing Letters, 24, 105-107.

[27] Damjanovic-Behrendt, V., & Behrendt, W. (2019). An open source approach to the design and implementation of Digital Twins for Smart Manufacturing. International Journal of Computer Integrated Manufacturing, 32(4-5), 366-384.

[28] Azure Digital Twins Documentation – Tutorials, API reference. Online: https://learn.microsoft.com/en-gb/azure/digital-twins/ . Accessed: 20/2/2023.

[29] Bosch IoT Suite – A toolbox in the cloud for IoT developers. Online: https://bosch-iot-

suite.com/ . Accessed: 20/2/2023.

[30] Liau, Y., Lee, H., & Ryu, K. (2018, March). Digital Twin concept for smart injection molding. In IOP Conference Series: Materials Science and Engineering (Vol. 324, No. 1, p. 012077). IOP Publishing.

[31] Karakra, A., Fontanili, F., Lamine, E., & Lamothe, J. (2019, May). HospiT'Win: a predictive simulation-based digital twin for patients pathways in hospital. In 2019 IEEE EMBS international conference on biomedical & health informatics (BHI) (pp. 1-4). IEEE.

[32] The Only No-Code Digital Twin Composition Platform – XMPro. Online:https://xmpro.com/. Accessed: 17/3/2023.

[33] Montini, E., Cutrona, V., Bonomi, N., Landolfi, G., Bettoni, A., Rocco, P., & Carpanzano, E. (2022). An iiot platform for human-aware factory digital twins. *Procedia CIRP*, *107*, 661-667.

[34] OECD (2009). Sustainable Manufacturing and Eco-innovation: Towards a Green Economy, OECD Directorate for Science, Technology and Industry. https://www.oecd.org/env/consumption-innovation/42957785.pdf

[35] Marconi, M., Menghi, R. A sustainable manufacturing tool for the analysis and management of resource consumption within production processes. Int J Interact Des Manuf 15, 65–68 (2021). https://doi.org/10.1007/s12008-020-00720-6

[36] Schneider, D., Paul, M., Vernim, S., Zaeh, M.F. (2022). Sustainability Assessment of Manufacturing Systems – A Review-Based Systematisation. In: , et al. Towards Sustainable Customization: Bridging Smart Products and Manufacturing Systems. CARV MCPC 2021 2021. Lecture Notes in Mechanical Engineering. Springer, Cham. https://doi.org/10.1007/978-3-030-90700-6_117

[37] National Research Council. 2011. Sustainability and the U.S. EPA. Washington, DC: The National Academies Press. https://doi.org/10.17226/13152.

[38] Arzoumanidis, I.; Raggi, A.; Petti, L. Considerations When Applying Simplified LCA Approaches in the Wine Sector. Sustainability 2014, 6, 5018-5028. https://doi.org/10.3390/su6085018

[39] Adadi, A., &Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160.

[40] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863)

[41] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[42] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

Co-funded by the European Union

[43] Breiman, L., Friedman, J., Stone, C. J., &Olshen, R. A. (1984). Classification and regression trees. CRC press.

[44] Lakkaraju, H., Bach, S. H., &Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1675-1684).

[45] Bahdanau, D., Cho, K., &Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[46] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

[47] Rajan, K. (2015). Materials informatics: The materials "gene" and big data. Annual Review of Materials Research, 45, 153-169.

[48] Ghorbani, A., Abid, A., & Zou, J. (2019, July). Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3681-3688).

[49] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[50] R. Nian, J. Liu, B. Huang, A review on reinforcement learning: Introduction and applications in industrial process control, Computers & Chemical Engineering, Volume 139, 2020, 106886, ISSN 0098-1354, https://doi.org/10.1016/j.compchemeng.2020.106886.

[51] M. Feng and Y. Li, "Predictive Maintenance Decision Making Based on Reinforcement Learning in Multistage Production Systems," in IEEE Access, vol. 10, pp. 18910-18921, 2022, doi: 10.1109/ACCESS.2022.3151170.

[52] M.L. Rodríguez, S. Kubler, A. de Giorgio, M. Cordy, J. Robert, Y. Le Traon, Multi-agent deep reinforcement learning based Predictive Maintenance on parallel machines, Robotics and Computer-Integrated Manufacturing, Volume 78, 2022, 102406, ISSN 0736-5845, https://doi.org/10.1016/j.rcim.2022.102406.\

[53] P. D. Paraschos, G. K. Koulinas, D. E. Koulouriotis, Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures, Journal of Manufacturing Systems, Volume 56, 2020, Pages 470-483, ISSN 0278-6125, https://doi.org/10.1016/j.jmsy.2020.07.004.

[54] J. Shin, T. A. Badgwell, K. Liu, Jay H. Lee, Reinforcement Learning – Overview of recent progress and implications for process control, Computers & Chemical Engineering, Volume 127, 2019, Pages 282-294, ISSN 0098-1354, https://doi.org/10.1016/j.compchemeng.2019.05.029.

[55] M. Panzer, B. Bender, N. Gronau, Deep Reinforcement Learning In Production Planning And Control: A Systematic Literature Review. In: Herberger, D.; Hübner, M. (Eds.): Proceedings of the Conference on Production Systems and Logistics : CPSL 2021.

Hannover

[56] Q. Liu, Z. Liu, B. Xiong, W. Xu, Y. Liu, Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function, Advanced Engineering Informatics, Volume 49, 2021, 101360, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2021.101360.

[57] M. Hillebrand, M. Lakhani, R. Dumitrescu, A design methodology for deep reinforcement learning in autonomous systems, Procedia Manufacturing, Volume 52, 2020, Pages 266-271, ISSN 2351-9789, https://doi.org/10.1016/j.promfg.2020.11.044.

[58] Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

[59] Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y. J., &Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews, 7(4), 041317

[60] Kumar, A., &Starly, B. (2022). "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. Journal of Intelligent Manufacturing, 33(8), 2393-2407.

[61] Shetty, P., & Ramprasad, R. (2021). Automated knowledge extraction from polymer literature using natural language processing. Iscience, 24(1), 101922.

[62] Castro, P. B. D., Terashima, K., Yamamoto, T. D., Hou, Z., Iwasaki, S., Matsumoto, R.,& Takano, Y. (2020). Machine-learning-guided discovery of the gigantic magnetocaloric effect in HoB2 near the hydrogen liquefaction temperature. NPG Asia Materials, 12(1), 35.

[63] Cooper, C. B., Beard, E. J., Vázquez-Mayagoitia, Á., Stan, L., Stenning, G. B., Nye, D. W., & Cole, J. M. (2019). Design-to-device approach affords panchromatic co-sensitized solar cells. Advanced Energy Materials, 9(5), 1802820.

[64] Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., ... & Wolverton, C. (2022). Recent advances and applications of deep learning methods in materials science. npj Computational Materials, 8(1), 59.

[65] Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571(7763), 95-98.

[66] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[67] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. Journal of Big data, 3(1), 1-40.

[68] Rajan, K. (2015). Materials informatics: The materials "gene" and big data. Annual Review of Materials Research, 45, 153-169.

[69] Regli, W., Rossignac, J., Shapiro, V., & Srinivasan, V. (2016). The new frontiers in computational modeling of material structures. Computer-Aided Design, 77, 73-85.

[70] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. npj Computational Materials, 3(1), 54.

[71] Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. Apl Materials, 4(5), 053208.

[72] Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., & Scheffler, M. (2015). Big data of materials science: critical role of the descriptor. Physical review letters, 114(10), 105503.

[73] Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245-251). IEEE.

[74] Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., & Luo, S. (2021). A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(3), 274-306.

[75] Ekstrand, M. D., Riedl, J. T., &Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction, 4(2), 81-173.

[76] Casillo, M., Colace, F., Conte, D., Lombardi, M., Santaniello, D., & Valentino, C. (2021). Context-aware recommender systems and cultural heritage: a survey. Journal of Ambient Intelligence and Humanized Computing, 1-19.

[77] Wang, J., Gao, S., Tang, Z., Tan, D., Cao, B., & Fan, J. (2021). A context-aware recommendation system for improving manufacturing process modeling. Journal of Intelligent Manufacturing, 1-22.

[78] Hernández-Rodríguez, S., Hernández-Lira, C. A., Aguilar-Rico, A., Flores-Becerra, G., Flores-Méndez, J., & Vázquez-Cuchillo, O. (2016). A recommender system applied to the indirect materials selection process (RS-IMSP) for producing automobile spare parts. Computers in Industry, 82, 233-244.

[79] Chicaiza, J., &Valdiviezo-Diaz, P. (2021). A comprehensive survey of knowledge graph-based recommender systems: Technologies, development, and contributions. *Information*, *12*(6), 232.

[80] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in Advances in neural information processing systems, 2013, pp. 2787–2795

[81] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," arXiv preprint arXiv:1412.6575, 2014.

Co-funded by the European Union

[82] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “Pathsim: Metapath-based top-k similarity search in heterogeneous information networks,” Proceedings of the VLDB Endowment, vol. 4, no. 11, pp. 992–1003, 2011.

[83] Li, X., Kao, B., Ren, Z., & Yin, D. (2019, July). Spectral clustering in heterogeneous information networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4221-4228).

[84] Hassani, H., & Silva, E. S. (2015). Forecasting with big data: A review. Annals of Data Science, 2, 5-19.

[85] Brownlee, J. (2018). How to develop LSTM models for time series forecasting. Machine Learning Mastery, 14.

[86] Mitchell, T. M. (2007). Machine learning (Vol. 1). New York: McGraw-hill.

[87] Qiu, J., Wang, B., & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. PloS one, 15(1), e0227222.

[88] Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., &Troncoso, A. (2021). Deep learning for time series forecasting: a survey. Big Data, 9(1), 3-21.

[89] Hyndman, R. J., &Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[90] Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., ... & Morgan, D. (2012). AFLOW: An automatic framework for high-throughput materials discovery. Computational Materials Science, 58, 218-226.

[91] Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials, 2(1), 1-7.

[92] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. npj Computational Materials, 3(1), 54.

[93] Rupp, M., Tkatchenko, A., Müller, K. R., & Von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters, 108(5), 058301.

[94] Honrao, S. J., Yang, X., Radhakrishnan, B., Kuwata, S., Komatsu, H., Ohma, A., ... & Lawson, J. W. (2021). Discovery of novel Li SSE and anode coatings using interpretable machine learning and high-throughput multi-property screening. Scientific Reports, 11(1), 16484.

[95] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[96] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Co-funded by the European Union

[97] Shvachko, K., Kuang, H., Radia, S., &Chansler, R. (2010). The Hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). IEEE.

[98] Stonebraker, M. (2010). SQL databases v. NoSQL databases. Communications of the ACM, 53(4), 10-11.

[99] Iqbal, Muhammad Hussain, and Tariq Rahim Soomro. "Big data analysis: Apache storm perspective." International journal of computer trends and technology 19.1 (2015): 9-14.

[100] Apache flink: Stream and batch processing in a single engine." The Bulletin of the Technical Committee on Data Engineering 38.4 (2015).

[101] Lam, Chuck. Hadoop in action. Simon and Schuster, 2010.

[102] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... &Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[103] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... &Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12) (pp. 15-28).

[104] D. Mourtzis, M. Doukas, and D. Bernidaki, "Simulation in Manufacturing : Review and Challenges," Procedia CIRP, vol. 25, pp. 213–229, 2014, doi: 10.1016/j.procir.2014.10.032.

[105] L. Zhang, L. Zhou, L. Ren, and Y. Laili, "Modeling and simulation in intelligent manufacturing," Computers in Industry, vol. 112. 2019. doi: 10.1016/j.compind.2019.08.004.

[106] F. Hosseinpour and H. Hajihosseini, "Architecture_based_on_dynamic_graphs_for," pp. 285–288, 2009.

[107] A. Negahban and J. S. Smith, "Simulation for manufacturing system design and operation: Literature review and analysis," J Manuf Syst, vol. 33, no. 2, pp. 241–261, Apr. 2014, doi: 10.1016/j.jmsy.2013.12.007.

[108] M. Jahangirian, T. Eldabi, A. Naseer, L. K. Stergioulas, and T. Young, "Simulation in manufacturing and business: A review," Eur J Oper Res, vol. 203, no. 1, pp. 1–13, May 2010, doi: 10.1016/j.ejor.2009.06.004.

[109] I. Daniyan, K. Mpofu, B. Ramatsetse, E. Zeferino, G. Monzambe, and E. Sekano, "Design and simulation of a flexible manufacturing system for manufacturing operations of railcar subassemblies," in Procedia Manufacturing, Jan. 2020, vol. 54, pp. 112–117. doi: 10.1016/j.promfg.2021.07.018.

[110] D. Muravev, H. Hu, A. Rakhmangulov, and P. Mishkurov, "Multi-agent optimization of the intermodal terminal main parameters by using AnyLogic simulation platform: Case study on the Ningbo-Zhoushan Port," Int J Inf Manage, vol. 57, p. 102133, Apr. 2021, doi: 10.1016/j.ijinfomgt.2020.102133.

[111] P. Higuera, J. L. Lara, and I. J. Losada, “Simulating coastal engineering processes with OpenFOAM®,” Coastal Engineering, vol. 71, pp. 119–134, Jan. 2013, doi: 10.1016/j.coastaleng.2012.06.002.

[112] M. H. Kobir et al., “Prediction of recoater crash in laser powder bed fusion additive manufacturing using graph theory thermomechanical modeling,” Progress in Additive Manufacturing, pp. 1–26, Aug. 2022, doi: 10.1007/s40964-022-00331-5.

[113] J. Pedro, B. Ramôa, J. M. Nóbrega, and C. Fernandes, “Verification and validation of openInjMoldSim, an open-source solver to model the filling stage of thermoplastic injection molding,” Fluids, vol. 5, no. 2, pp. 1–24, 2020, doi: 10.3390/fluids5020084.

[114] B. Brenken, E. Barocio, A. Favaloro, V. Kunc, and R. B. Pipes, “Development and validation of extrusion deposition additive manufacturing process simulations,” AdditManuf, vol. 25, pp. 218–226, Jan. 2019, doi: 10.1016/j.addma.2018.10.041.

[115] A. Chouhan, A. Aggarwal, and A. Kumar, “Model Development in OpenFOAM for Laser Metal Deposition-based Additive Manufacturing Process,” Transactions of the Indian Institute of Metals, vol. 71, no. 11, pp. 2833–2838, Nov. 2018, doi: 10.1007/s12666-018-1440-7.

[116] B. Behdani, M. Senter, L. Mason, M. Leu, and J. Park, “Numerical study on the temperature-dependent viscosity effect on the strand shape in extrusion-based additive manufacturing,” Journal of Manufacturing and Materials Processing, vol. 4, no. 2, p. 46, May 2020, doi: 10.3390/jmmp4020046.

[117] R. Schoof, G. F. Castelli, and W. D. ¨ Orfler, “The 8th European Congress on,” 2022.

[118] J. M. Nóbrega et al., “Using OpenFOAM to Aid the Design of Extrusion Dies for Thermoplastics Profiles,” no. June, pp. 23–26, 2014, Accessed: Mar. 25, 2023. [Online]. Available: https://www.researchgate.net/publication/263463212

[119] J. Kattinger, J. Kettemann, and C. Bonten, “Numerical 3d simulation of a co-kneader in openfoam,” in AIP Conference Proceedings, 2020, vol. 2289, p. 20060. doi: 10.1063/5.0028367.

[120] C. Fernandes, F. Habla, O. S. Carneiro, O. Hinrichsen, and J. M. Nóbrega, “Development of a new 3D OpenFOAM® solver to model the cooling stage in profile extrusion,” in AIP Conference Proceedings, 2016, vol. 1713, p. 120008. doi: 10.1063/1.4942260.

[121] H. Jasak, “OpenFOAM: Open source CFD in research and industry,” International Journal of Naval Architecture and Ocean Engineering, vol. 1, no. 2, pp. 89–94, Dec. 2009, doi: 10.2478/ijnaoe-2013-0011.

[122] T. D. Economon, F. Palacios, S. R. Copeland, T. W. Lukaczyk, and J. J. Alonso, “SU2: An open-source suite for multiphysics simulation and design,” AIAA Journal, vol. 54, no. 3, pp. 828–846, Dec. 2016, doi: 10.2514/1.J053813.

[123] A. Ribes and C. Caremoli, “Salomé platform component model for numerical

simulation," Proceedings - International Computer Software and Applications Conference, vol. 2, pp. 553–564, 2007, doi: 10.1109/COMPSAC.2007.185.

[124] C. Geuzaine and J. F. Remacle, "Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities," Int J Numer Methods Eng, vol. 79, no. 11, pp. 1309–1331, Sep. 2009, doi: 10.1002/nme.2579.

[125] A. Logg, K.-A. Mardal, and G. N. Wells, "Finite element assembly," 2012, pp. 141–146. doi: 10.1007/978-3-642-23099-8_6.

[126] F. Archambeau, N. Mechitoua, and M. Sakiz, "Code Saturne: A Finite Volume Code for Turbulent flows - Industrial Applications," International Journal on Finite Volumes, vol. 1, no. 1, 2004, Accessed: Mar. 27, 2023. [Online]. Available: http://www.latp.univ-mrs.fr/IJFV/spip.php?article3.ï¿¿hal-01115371ï¿¿