Mannheimer Institut für Intelligente Systeme in der Medizin (MIISM)
Abteilung für Automatisierung in der Medizin und Biotechnologie
(Direktor: Prof. Dr. Jan Stallkamp)

A standardization concept for machine actionable and reusable scientific data

Inauguraldissertation
zur Erlangung des

Doctor scientiarum humanarum (Dr. sc. hum.)

der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von:
Axel Wilbertz

aus Eberbach
2024

Dekan: Prof. Dr. med. Sergij Goerdt
Referent: Prof. Dr.-Ing. Jan Stallkamp

Table of contents

1. Introduction 1

1.1 Current challenges in drug development 1

1.2 Therapeutic mAbs and formulation development 1

1.3 In silico-supported liquid formulation development 3

1.4 Data standardization for advanced analytics in drug discovery 5

1.5 Aim and Outline of the thesis 5

1.6 Thesis structure 6

2 Analysis of requirements 8

2.1 Background of protein modeling and descriptor predictions 8

2.1.1 Protein structure 8

2.1.2 Structure modeling 9

2.1.3 Structure-derived protein descriptors 11

2.1.4 Supervised learning algorithms 11

2.1.5 Decision trees and random forest 12

2.2 Assessment of data standardization for ML 13

2.2.1 Data set 1: HTS formulation data 14

2.2.2 Data set 2: company-internal biologics dashboard data 15

2.2.3 Data set 3: external pharmaceutical data provider 16

2.2.4 Impact of data standardization on ML results 19

2.3 Data complexity 20

2.3.1 Simple and complex data 20

2.3.2 Data storage requirements 20

2.3.3 Data collection strategies 21

2.4 Requirements for a biologics data standardization concept 23

3 State of the art 25

3.1 Lab automation and high-throughput screening concepts 25

3.1.1 System Suitability Tests 27

3.1.2 Analytical methods standardization 28

3.1.3 Workflow standardization using positive and negative controls 28

3.1.4 Errors and sample replication 29

3.2 Scientific data standardization 30

3.2.1 Standard Data management 30

3.2.2 Data quality 30
3.2.3 The FAIR guiding principles 31
3.2.4 FAIR assessment options 32
3.2.5 Current FAIRification approaches and FAIR implementations 33
3.2.6 Method and device-focused data standardization 35
3.3 Standardization concept for biologics data 36

4 Methodology 39
4.1 Data standardization process using FAIR 39
4.1.1 Standardization objective 40
4.1.2 Reassessment of standardization of the three data sets using FAIR 41
4.1.3 Optimization of FAIR assessment score 44
4.2 Redefinition of the standardization objective: machine actionability 49
4.2.1 Automated data comparability 51
4.2.2 Challenges and concept proposition 52
4.3 Result-based standardization concept 55
4.3.1 Data semantification using a self-developed semantic model 57
4.3.2 Qualified aggregated metadata – decisions, threshold, and ranges 64
4.3.3 Inclusion of subject matter expert (SME) comparability logic 74

5 Evaluation 76
5.1 Experimental raw data 79
5.2 Semantification of raw data 80
5.3 Querying qualified aggregated metadata 81
5.4 Application of the SME comparability logic 83
5.5 Overall comparability assessment results 83
5.6 Error influence within and across screenings 86

6 Discussion 88
6.1 Data reuse challenges 88
6.2 Verification of the requirements 89
6.3 Interpretation of FAIR 90
6.4 FAIR implementations 90
6.5 Implications 91
6.6 Benefits of semantics 93
6.7 Generalization 94

6.8 Limitations and future work....95

7 Summary....96

8 Supplementary information....97

9 References....102

10 Curriculum vitae....112

11 Acknowledgement....114

List of figures

Figure 1 ML contribution throughout the drug development process *4*
Figure 2 Envisioned solution for utilizing biologics data for Advanced Analytics *6*
Figure 3 Schematic depiction of a monoclonal antibody (mAb) *9*
Figure 4 Homology Modeling steps *10*
Figure 5 Decision tree schema *13*
Figure 6 Example protein descriptors of different mAbs *18*
Figure 7 Inverse data strategy *22*
Figure 8 Overview of implications from the poor ML results *23*
Figure 9 Fully automated laboratory system *26*
Figure 10 Modular high-throughput formulation screening design *27*
Figure 11 Overview of the 15 FAIR guiding principles. *32*
Figure 12 RDF example. *34*
Figure 13 Different classification categories *35*
Figure 14 Overview of the Allotrope Foundation framework *36*
Figure 15 Desired standardization concept. *37*
Figure 16 The FAIRification process *40*
Figure 17 Relation of FAIRification score and machine actionability *41*
Figure 18 Sub-score calculation for the findability FAIR attribute *42*
Figure 19 Allotrope HPLC ontology model *47*
Figure 20 Different concepts and levels to increase the semantic expressivity *48*
Figure 21 Adaptation of the FAIRplus data maturity levels (DSM) *52*
Figure 22 Challenges for the method-based approach and potential solutions *54*
Figure 23 The Result-based standardization concept *55*
Figure 24 Semantification process of experimental raw data *58*
Figure 25 The adapted OBI schema *62*
Figure 26 Diagram of a sample measurement semantic representation *63*
Figure 27 FAIRplus data maturity levels (DSM) and metadata hierarchy *66*
Figure 28 Top-level tree of the qualified aggregated metadata *67*
Figure 29 System Suitability Test (SST) qualified aggregated metadata *68*
Figure 30 Workflow/process qualified aggregated metadata subtree *70*
Figure 31 Analytical method qualified aggregated metadata subtree *71*
Figure 32 Replicates qualified aggregated metadata subtree *73*
Figure 33 Example table for the sample replicate calculation *74*
Figure 34 Flow chart of the SME comparability logic *75*
Figure 35 High-level workflow of data set comparability assessment. *76*
Figure 36 Detailed workflow of the result-based standardization concept *78*
Figure 37 Screening and stress-specific error comparison *87*

List of tables

Table 1 The ten most predictive selected protein descriptors 19
Table 2 The two data collection strategies for simple and complex data. 22
Table 3 Multiple data quality definitions from literature 31
Table 4 Summary table of the FAIR assessment results 43
Table 5 Example RDF structure and components 46
Table 6 Ontology reference and description for each class. 60
Table 7 Different metadata types and their characteristics. 65
Table 8 Example calculation for the expected range of the positive and negative controls 69
Table 9 SPARQL query results for the analytical method part 82
Table 10 Detailed comparability results for the different stress conditions 85
Table 11 Overall example screenings comparability results 86

List of listings

Listing 1 Owl class definitions *61*
Listing 2 Owl numerical value specification for a sample *80*
Listing 3 SPARQL quality classification examples *81*
Listing 4 Reasoning example *82*

Acronyms

ADC	Antibody–drug Conjugates
ADF	Allotrope Data Format
ADM	Allotrope Data Models
AFO	Allotrope Foundation Ontologies
AI	Artificial Intelligence
API	Active Pharmaceutical Ingredient
ASM	Allotrope Simple Model
BFO	Basic Formal Ontology
BMI	Body Mass Index
CDE	Common Data Element
CDR	Complementary Determining Region
CDS	Chromatography Data System
CMDO	Clinical MetaData Ontology
CQA	Critical Quality Attributes
DL	Deep Learning
DMM	Data Maturity Model
DOE	Design of Experiments
DOI	Digital Object Identifier
DS	Data Science
ELN	Electronic Lab Notebooks
ETL	Extract, Transform, Load
Fab	Antigen-Binding Fragments
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
FDMF	FAIR Data Maturity Framework
FDOs	FAIR Digital Objects
HMW	High Molecular Weight
HPLC	High-Performance Liquid Chromatography
HTS	High-throughput Screening
IAO	Information Artifact Ontology
ICE	Information Content Entity

ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
JSON	JavaScript Object Notation (JSON)
KPI	Key Performance Indicator
LIMS	Laboratory information management system
LMW	Low Molecular Weight
mAb	Monoclonal Antibody
MD	Measurement Datum
ML	Machine Learning
MOE	Molecular Operating Environment
NCE	New Chemical Entities
OBI	Ontology for Biomedical Investigations
OWL	Web Ontology Language
PDB	Protein Data Bank
RDA	Research Data Alliance
RDF	Resource Description Framework
RFC	Random Forest Classifier
SDM	Semantic Data Model
SEC	Size-Exclusion Chromatography
SM	Supplementary Material
SME	Subject Matter Expert
SPARQL	SPARQL Protocol And RDF Query Language
SQL	Structured Query Language
SST	System Suitability Test
Turtle	Terse RDF Triple
URI	Uniform Resource Identifier

Glossary and definitions

Data comparability	Capability to relate data sets to one another based on appropriate criteria with the goal of selecting only data relevant to answering a specific question. Is the prerequisite before data can become machine actionable.
Data complexity	A continuum demanding different standardization and normalization efforts to reach machine actionability. **Basic data**: Require minimal semantic context, e.g., currency exchange rates. **Complex data**: Require standardization and data value normalization utilizing a semantic model.
Data FAIRness	**Non-FAIR**: Explicit and relevant metadata are missing. Data not fit for purpose. Data not represented in a semantic model. Solely manual data integration by SME is possible. **FAIR**: Explicit metadata via semantic representation, e.g., ontologies. Quality aspects need to be included. Sufficient to achieve machine actionable for non-complex data. **FAIR + comparability logic**: Required for complex data. Adds a semantic layer that includes quality and SME logic aspects to determine data set comparability automatically.
Data integration	Differently processed data sets of different formats and meanings are merged to form a larger data set. Examples: technical interoperability, semantic interoperability.
Data quality	Measuring of the suitability of a data set to be fit for its purpose, i.e., to be machine actionable.
Interoperability	Capability to combine sub-data sets from different sources into a uniformed data set for further reuse, e.g., literature vs. experimental data. **Semantic interoperability:** Achieving unambiguous semantic meaning of data through metadata and concepts like ontologies. **Technical interoperability:** The exchange of differently formatted digital entities e.g., CSV transformation into a standard JSON format **Process interoperability (workflow):** Depending on the data's complexity, this can be achieved through metadata or workflow standardization. Iterative approaches may impact standardization/normalization and interoperability, especially for complex data.
In silico	Computer-based calculations, simulations, or modeling.
Metadata	**General metadata:** Used to enrich data with additional context.

	Method-based metadata: Enable machine and human readability of data. Focused on the applied method and the replication of an analytical method, e.g., the Allotrope data model. It is not sufficient to achieve machine actionability. **Qualified aggregated metadata:** Used to set analytical results into context and include additional SME knowledge. Serve a specific reuse purpose when included in a semantic layer. Help to qualify a sample, which is required to determine manual or automatic data comparison.
Machine readability	Enables the machine access to the data, e.g., through metadata or a semantic representation through ontologies.
Machine actionability	The capability of the machine to act on the data in the same manner as a human subject matter expert (SME) would. Machine readability is a prerequisite. Machine actionable data are Artificial Intelligence (AI)-ready and can be directly used through AI technologies, e.g., Machine Learning (ML) or Deep Learning (DL).
Automatic comparability	Automation of machine actionability decision
Normalization	Adjusting data to a common scale or an agreed-upon standard **Implicit normalization:** Alignment of data values to a common scale, e.g., between 0 and 1, without a formal standard. **Explicit normalization:** Normalization against a standard.
Knowledge triples/ Resource Description Framework (RDF)	Semantic knowledge representation from the semantic web. A triple consists of subject, predicate, and object. The predicate connects the subject and object via a relation, e.g. Subject: "Scientist A" Predicate: "does research in" Object: "formulation development". The triples can be linked to ontology terms for each part of the triple to enable a machine readable format.

Standardization	A domain-specific agreed-upon reference that enables data normalization for achieving interoperability of said data. **Process standardization:** Consists of identifying all critical parameters of a process and harmonizing them, e.g., predefined stress conditions (40°C for 7 days). **Data standardization:** Converting data into an agreed-upon format or against a specific reference as a requirement for normalization, e.g., unit conversion, scale calibration.
Subject matter expert (SME)	Expert in a given domain, such as, high-performance liquid chromatography (HPLC) with specialization for size-exclusion chromatography (SEC). The SME can define and decide on decision criteria, threshold and ranges for data set comparability.

Abstract

The reuse of scientific data through sophisticated algorithms has the potential to advance drug development (Mak & Pichika, 2019) (Narayanan et al., 2021) (Paul et al., 2021). For this, data standardization is required (Kush et al., 2020). Data standardization is the conversion of data into an agreed-upon format or against a reference. For biologics such as therapeutic monoclonal antibodies (mAbs), this is challenging because data is diverse, analytical methods are complex, and the data inherently suffers from ground noise (Taylor, 2021).

Laboratory automation and high-throughput concepts enable process standardization to cope with the challenges on the lowest level, which allows for fast and reliable generation of standardized data sets. Unfortunately, manual data integration is still required to determine which data sets are of sufficient quality to be reused through advanced analytics. Therefore, human scientists rely on their scientific expertise to make a decision on data set comparability, which is the capability to relate data sets to one another based on appropriate criteria with the goal of selecting only those data relevant for answering a specific question. However, machines are incapable of substituting the human factor because standardization concepts for biologics, which allow for data standardization, are either missing or unsuited. The limited existing concepts do not offer relevant metadata for assessing the quality of biologics data sets, which involve sophisticated analytical techniques like liquid chromatography - a critical method in drug development.

The Findable, Accessible, Interoperable and Reusable (FAIR) principles are such a data standardization concept, making scientific data machine actionable for machine and human reuse is the goal of these guiding principles (Wilkinson et al., 2016). Semantic web technologies are established in the FAIR community. In many cases, these are used to enable FAIR data. The question is whether these concepts are sufficient to render biologics analytical data machine actionable enough so that a machine has the same capability to act on the data as a human does. To achieve this, the machine requires human-like knowledge to determine the comparability of biologics data sets. FAIR and existing semantic concepts only partially address the problem because the comparability of data sets is not explicitly covered by FAIR. Additional steps are required to enable automated data set comparability for the machine.

In this thesis, the current level of biologics standardization is reviewed. For this, the suitability of standardization concepts like Allotrope and FAIR to standardize biologics data is elaborated. Furthermore, it has been identified that, for biologics data, these concepts are only partially usable to enable comparable data. As a result, a new standardization concept using semantic technologies that enables the automatic decision on biologics data set comparability, similar to a human scientist, is developed. The concept can be applied to other domains that face similar challenges with complex data integration and lack of standardization

1. Introduction

1.1 Current challenges in drug development

Enhancing drug development efficiency through increased data reuse is important since drug development is increasingly cost and resource intensive. Saving costs and improving efficiency by increasing data reuse is the essential first step in preparing the data for advanced analytics.

Therapeutic mAbs have a probability below 15 % to be clinically successful and show high attrition rates (Hutchinson & Kirk, 2011). These compounds are especially challenging to develop due to their multidimensionality, complex structure, and function. They are difficult to analyze and stabilize due to their natural origin in living organisms.

Furthermore, they are often formulated as highly concentrated solutions, which makes drug development and drug manufacturing difficult due to stability and manufacturability challenges (Goswami et al., 2013). Nevertheless, they have become essential to treating inflammatory and autoimmune diseases and cancer. Their unique advantage lies in the ability to bind specifically to their targets, showing fewer side effects.

Given their sensitive nature, most therapeutic mAbs are administered through injection. Anti-inflammatory and autoimmune therapeutics require regular dosing and are applied mainly by the patient at home as a liquid via a self-injection syringe. In contrast, mAbs for cancer treatment are predominately lyophilized. Compared to intravenous injection in the clinic, self-injection pens can be administered at home by the patient. Thus, they are 50 percent less cost-intensive and show improved benefits for the patients (Heald et al., 2021). Furthermore, this helps to decrease health care system costs.

Administration of large doses of up to 2 g with a high protein concentration of up to 150 mg/ml pose challenges that need to be overcome during development, which adds to the importance of thorough liquid formulation development (Hendrikx et al., 2017). Liquid formulations are composed of a specific drug in combination with drug-stabilizing additives.

1.2 Therapeutic mAbs and formulation development

MAbs are technically engineered proteins with a size of around 150.000 Dalton. Compared to Aspirin, with about 100 Dalton, they are three magnitudes bigger and have a very complex three-dimensional structure which determines their function. Their specificity and high affinity to binding targets make them increasingly important for diagnostic and therapeutic applications. Therapeutic mAbs have become the predominant drug form in the last few decades and are increasingly important in treating a wide range of diseases (Lu et al., 2020).

Formulation development is the start of the drug development process, which involves developing an active pharmaceutical ingredient (API) into a safe-to-use drug and drug format for the patient (Chan & Carter, 2010). Formulation development is the creation and optimization of the drug in a specific form that can be safely administered. Moreover, it is the process of systematically researching drug-stabilizing additives (excipients) to develop a stable therapeutic mAb.

The formulation composition is the main factor assuring protein stability during drug development, manufacturing, transportation, storage, and injection. Suppose the drug is developed for application through a syringe or an autoinjector. In that case, the process is referred to as liquid formulation development since the drug and the additives are contained and developed in a watery solution. The drug is combined with multiple excipients that ensure protein stability over time and aid during drug application. Possible excipients include buffers, antioxidants, isotonizers, stabilizers, surfactants, solubilizers, preservatives, and other potential additives. These excipients are combined with the drug to form multiple varying formulation compositions. From a formulation development perspective, these combinations of drugs and excipients are also referred to as samples.

Multiple stress conditions are applied, such as accelerated temperature stress over days and weeks, shaking stress, freeze-thaw stress, and light stress. The stress-induced stability changes between the different compositions are analyzed through a wide range of techniques. Determining the most stable formulation that performs best throughout all analytics is challenging. Physical and chemical properties, referred to as Critical Quality Attributes (CQAs), are monitored to reduce risk and ensure the quality, safety, and efficacy of the product throughout its development (Alt et al., 2016).

High-performance liquid chromatography (HPLC) is used during drug development to ensure the CQAs are fulfilled. It is a crucial technique for quality control, used to separate, identify, and quantify the API and its components (formulation) in a drug product. HPLC aids in formulating, optimizing, and establishing specifications for the drug. It also monitors stability during short- or long-term storage, ensuring the drug's quality and shelf-life.

Size-exclusion chromatography (SEC) is an analytical method in HPLC to determine the size and shape of macromolecules like proteins. SEC separates the molecules based on their size, allowing for the differentiation of components into monomers, fragments, and aggregates This helps evaluate the purity and molecular weight distribution of the drug substance and product. It is crucial to determine the purity of the drug as impurities and degradation products can have a significant impact on its safety and efficacy (Fekete et al., 2014).

Pharmaceutical companies rely on lab automation to efficiently achieve drug development. High-throughput screening (HTS) concepts are utilized to identify the most stabilizing formulation to develop a safe liquid formulation for a drug using minimal volume. Micro titer plates are used with minimal volume analytics that require sample volumes of a few grams. This additionally increases the number of possible excipient combinations since less drug substance is required. A key factor in determining stability is a liquid formulation's pH and the drug substance concentration.

Pharmaceutical companies must meet regulatory requirements and expand the knowledge regarding the target formulation. Additionally, they are requested to perform Design of Experiment (DOE) approaches to better understand the formulation design space and follow the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidelines. To ensure a product's quality and safety, a quality by design approach is used (Rathore & Winkle, 2009).

The costs for developing mAbs have increased over the recent years, including dosage and annual therapy costs (DiMasi et al., 2003). Pharmaceutical companies' development pipelines shift away from classic therapeutic mAbs towards a diverse landscape of novel molecule formats, such as highly engineered compounds like bispecific antibodies, ADC, or fusion proteins. These formats further increase development costs due to their complex production process and limited drug yield.

1.3 In silico-supported liquid formulation development

In silico-supported liquid formulation, which uses computer models to support or replace drug development experiments in the laboratory, can enable faster and more cost-effective research on new therapeutic treatments (Kolluri et al., 2022). Figure 1 shows the different stages of the drug development process (A) and where an ML contribution can be useful (B). Existing experimental data can be used to train predictive algorithms to uncover hidden knowledge or generate new knowledge for new drugs. In drug discovery, ML can aid in reducing the potential compounds to a smaller subset of candidates with desired affinity or a specific function. This subset is furtherly reduced based on physical and chemical stability properties in the developability assessment a drug.

For formulation development, predictions can range from suggesting parts of the formulation, e.g., single stabilizing excipients, to the complete formulation prediction in silico. The goal of all pharmaceutical companies is to enable computer-based drug development without extensive laboratory experiments. In silico characterization of unknown molecules by predicting stability can shorten development timelines and efforts and evolve formulation development to the next level. Ideally, data from different organizations and sources can be integrated and used for this matter.

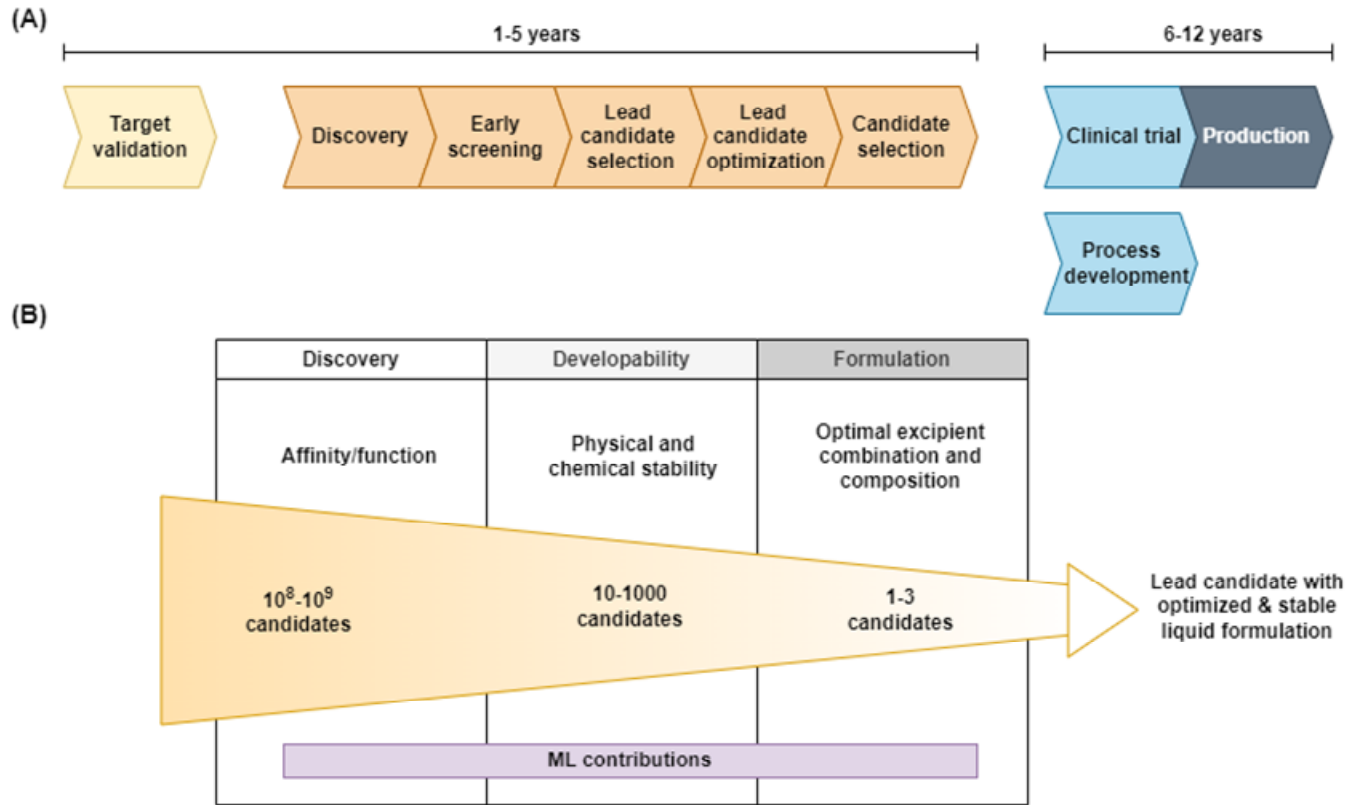


Figure 1 ML contribution throughout the drug development process . The process is a multiple-year effort that includes multiple steps (A). ML can potentially be established at multiple stages (B). Figure adapted from (Narayanan et al., 2021).

Unfortunately, pharmaceutical companies face difficulties effectively establishing ML due to the complexity of drug-related development data (Narayanan et al., 2021). Especially biologics data is diverse and multidimensional. Experimental data from analytical methods inherit noise, and adapting these methods to new molecules is challenging.

Furthermore, the availability of biologics data sets for ML is limited, even within companies. This data limitation leads to inaccuracies in the ML models. Scientists within the departments and groups know and own the few available data sets. They manually try to gather and prepare more analytical data from different departments and groups within the company. Processes and analytical methods differ between these. This makes it challenging to compare and analyze data from different sources or time periods in a meaningful and accurate way.

Moreover, the capability to relate data sets to one another based on appropriate criteria with the goal of selecting only those data relevant for answering a certain question is required (data comparability). Furthermore, data standardization ensures that data is cross-functionally collected, recorded, and presented in a consistent format or against a certain reference so that it can be compared. This is the first step before it can be reused through sophisticated algorithms such as ML.

1.4 Data standardization for advanced analytics in drug discovery

A recent example demonstrated how well-standardized protein data can be reused to enable advanced analytics. Jumper et al. demonstrated how they leveraged existing protein structural data to solve the 50-year grand challenge, a longstanding problem in biochemistry and molecular biology. It involved predicting the three-dimensional structure of a protein solely on its amino acid sequence. The goal was to reach high enough accuracy comparable to experimental techniques such as X-ray crystallography and cryo-electron microscopy. Jumper et al. solved this structure prediction of unknown proteins based on their amino acid sequence with a DL algorithm (Jumper et al., 2021). Well-standardized protein data from the Protein Data Bank (PDB), including x-ray crystallography data and data from unknown proteins from the UniProt database, were used for their approach. Their technology helps understand protein interactions with each other and their environment, leading to new medical treatments and more efficient drug discoveries (Ren et al., 2023).

Although they revolutionized general science, the impact on formulation development is neglectable. As displayed in Figure 2, Drug Discovery is the first step in a three-stage drug development process. The protein structure is predetermined during drug development and the formulation stage and rarely changes. The goal of formulation development is to explore the drug design space. Therefore, a deeper understanding of the drug's physico-chemical properties through stability experiments with multiple excipients is required. Nevertheless, their technology demonstrates that data standardization is necessary to enable data reuse through advanced analytics.

1.5 Aim and Outline of the thesis

The general aim of this thesis was the development of a strategy to enable a biologics data standardization concept. Biologics data refer to data related to biological drugs such as mAbs and related analytical results obtained from analytical techniques such as the HPLC. While the emphasis is on HPLC experiments, the concept should apply to other domains and analytical methods to enable the reuse of scientific data in preparation for reuse through AI. Figure 2 depicts the aim and the envisioned solution. Biologics data, e.g., internal biologics formulation data, biologics data from company-internal dashboards, and publicly available protein data (green), are complex due to its natural noisiness and diverse data formats, making it unsuitable for direct use in advanced analytics, such as ML (dotted arrow).

Currently, manual data integration is required to prepare data for AI approaches, which is time-consuming and unfeasible when automated decision-making is desired. A data standardization concept (blue) is crucial to address these challenges. The goal of the concept is to make biologics data comparable for humans and machines. Therefore, it must incorporate all relevant criteria to derive a biologics data comparability decision, such as experiments, workflows and processes, lab automation, analytical methods, analytical results, and data quality (purple). Concepts such as the FAIR principles can be used to achieve this. This thesis focuses on developing a biologics standardization concept but did not focus on the application and data reuse for advanced analytics itself (red). The problem is that the currently available data standardization concepts only partially addressed the problem and required adaptation.

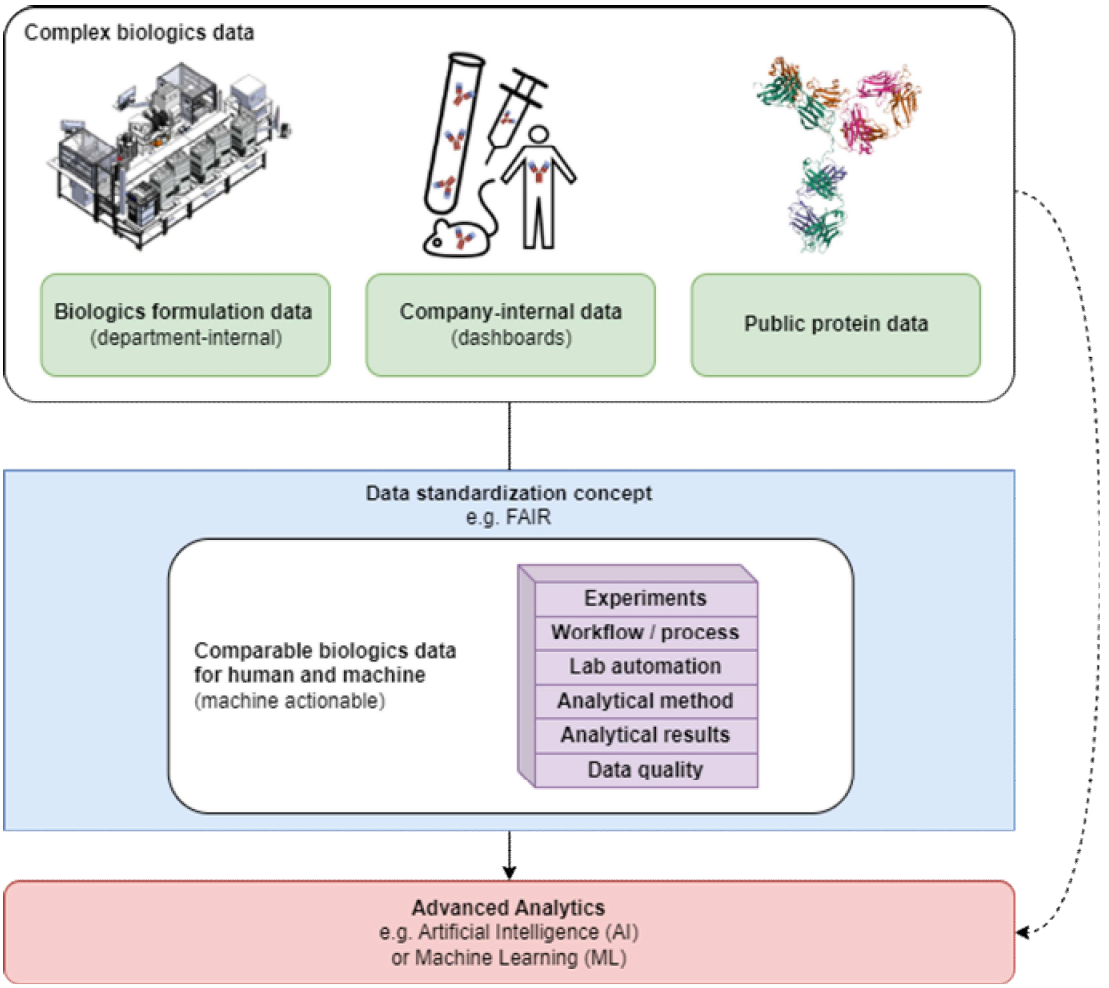


Figure 2 Envisioned solution for utilizing biologics data for Advanced Analytics (red) such as Artificial Intelligence (AI). Data from different sources (green): internal biologics formulation data, company-internal data from dashboards, and publicly-available protein data should be directly used for this purpose (indicated by the dotted arrow). However, due to the complexity of biologics data, a data standardization concept (blue) is required to achieve comparable biologics data. This allows humans and machines to interpret the data and determine which data sets are suitable for this task (white).

1.6 Thesis structure

Chapter 2 starts with the assessment of different data complexity levels and their implications. Moreover, the data collection requirements are explored. Furthermore, the current standardization level of biologics data and thus their standardization is reviewed. For this, three biologics data sets from different sources were manually prepared for reuse through advanced analytics. A lack of data quality and an insufficient metadata context indicated low data standardization. Therefore, poor data comparability and insufficient predictive accuracy during advanced analytics are observed. Overall, data reuse is identified as a challenge for complex data. Consequently, the requirements for a standardization concept that enables data reuse for both humans and machines, allowing for automated

determination of data comparability and ensuring the generation of machine-actionable data, are defined.

In Chapter 3, data standardization concepts are introduced. Standardization is separated into lab automation and scientific data standards. The scientific data standardization concepts FAIR, Allotrope, and ontologies are introduced. These were used in Chapter 4 to develop a standardization concept for biologics data. Furthermore, in Chapter 3 different FAIR assessment methods that are well-suited to evaluate data standardization levels are presented.

Chapter 4 starts with a detailed standardization assessment of the biologics data sets from Chapter 2. An existing FAIR assessment method was adapted. The assessment was conducted, and the results indicate an overall low standardization. The current state-of-the-art standardization concepts for biologics data and whether they enable machine comparable data were closer examined. Furthermore, current standardization concepts, which do not enable the machine to automatically decide on biologics data set comparability, are outlined. Moreover, the requirements for a new concept are described and novel concept based on a semantic model for biologics data is proposed. The concept comprises three parts: 1. Qualified aggregated metadata describing the essential criteria for biologics data comparability decisions. 2. A novel semantic model based on existing ontologies to represent the qualified aggregated metadata. 3. The SME comparability logic that is transformed into a flow chart. The three parts enabled the machine to automatically decide on data set comparability.

In Chapter 5, the developed concept was validated through three example biologics data sets. For each data set, the comparability was automatically assessed through the machine. Afterward, the results and the comparability-influencing components of the concept were examined. Finally, the concept was compared to the requirements from the analytics chapter.

The results from Chapter 5 are set in a broader scientific context in Chapter 6. The interpretation of FAIR in this thesis is set in contrast to others throughout the data standardization community. Furthermore, the alignment of the proposed concept with other practical FAIR implementations is evaluated. Next, the limitations of the developed concept are highlighted. At last, the generalization of the approach and how it can aid other fields that are challenged with the standardization of complex data, e.g., the health care sector, is reviewed.

2 Analysis of requirements

The following chapter contains an assessment of the standardization of biologics data. For this, data sets from different sources are prepared and reused through advanced analytics. Additionally, protein structure information is calculated through antibody modeling to achieve a better characterization of the mAbs. ML algorithms, such as random forest classifier models, are trained to predict the pH of different mAbs. Furthermore, the results from the trained models are assessed to conclude each data set's standardization level. Next, different data complexity levels and the implications for data collection are introduced. The importance of data availability for advanced analytics is highlighted. Finally, requirements for an improved standardization concept that enables an increased reuse of biologics data are defined.

The performance of data analytics depends on the data quality. More sophisticated approaches like ML or DL require an increased amount of data compared to classical statistics. Compared to statistical models, the accuracy of ML models increases with data availability (volume). Data scarcity is an overall challenge. Concepts that artificially generate data (synthetic data) can be used to increase the data volume but have the disadvantage of not reflecting real-world data. Moreover, synthetic data has the potential to bias models and not capture outliers and events. This is especially true for biological data due to its high complexity. Too few or no data makes it impossible to establish the relationship between data inputs and outputs, which leads to poor results or no results at all. However, scientific techniques such as structure modeling can be used to derive additional information and data from the 3D structure of a protein.

2.1 Background of protein modeling and descriptor predictions

The following section describes protein structures and the methods and techniques used to calculate protein structure models. Structure antibody modeling was used to increase the amount of data from 3D models of different molecules and to characterize the targeted mAbs. Additionally, the calculation of protein descriptors based on these models is explained. Next, advanced analytic technologies such as predictive algorithms are introduced, which were used to train the predictive models. In a later step, the results from these were evaluated to draw conclusions on biological data standardization.

2.1.1 Protein structure

The 3D structure of mAbs and proteins in general determines their function. It can be separated into three parts.

- Primary structure: it is a sequence of amino acids that form a chain. It contains no structural information.
- Secondary structure: it is characterized by common repeating protein motifs such as -helix, -sheet and turns. They are stabilized by hydrogen bonds. Each amino acid has a specific angle towards the protein’s backbone. The angle located in front of the c-atom is referred to as (phi) and the one after is called (psi) angle.

- Tertiary structure: the proteins 3D structure, also known as the fold of the protein, is determined by the forces and bonds between the proteins side chains and amino acids.
- Quaternary structure: the association of two or more protein chains into one greater assemble.

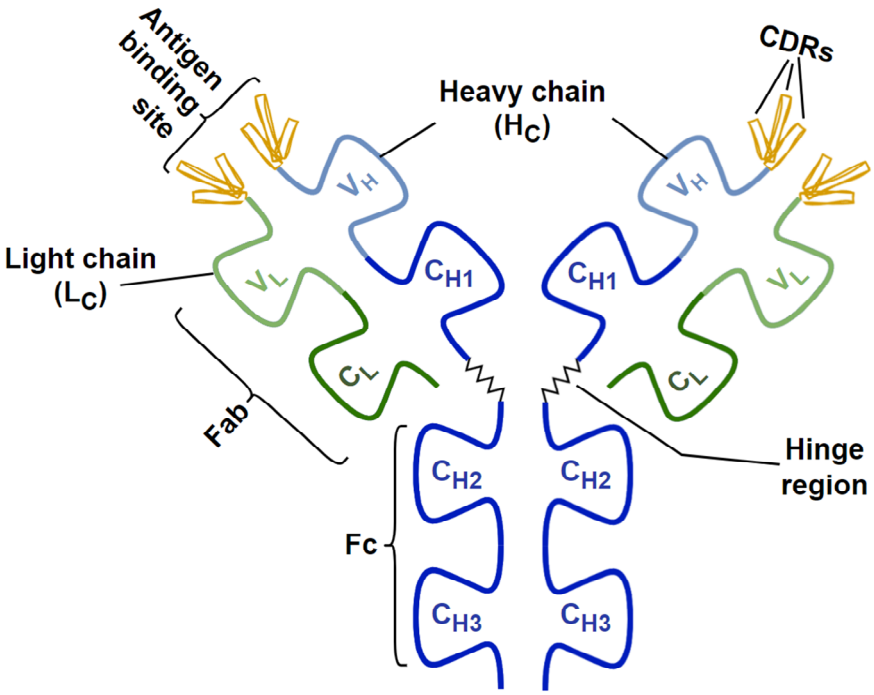


Figure 3 Schematic depiction of a monoclonal antibody (mAb) including all essential parts. Figure adapted from (Moradi-Sardareh et al., 2016).

Figure 3 shows the structure of a mAb that contains four chains: two identical large, heavy chains (blue) and two identical small light chains (green), which are connected by disulfide bonds to form the typical y-shaped form. The light chains consist of a constant domain C_L and a variable domain V_L. The heavy chains consist of one variable V_H and three constant domains, C_H1, C_H2, and C_H3. The constant and variable domains of one heavy chain and light chain are called the Fab region. The variable domain of one heavy and one light chain forms the antigen binding site and is known as the F_v region. The F_v region contains three loops of the CDR. They are essential for binding to an antigen and determining the antibodies specificity.

2.1.2 Structure modeling

Structure modeling is also referred to as homology modeling or template-based modeling. It is the prediction of an unknown protein structure (target) based on its amino acid sequence. The sequence of the unknown structure is compared (similarity) to sequences with known structures (template). In contrast, de novo modeling calculates the protein structure from sequence without a similar template, making the calculation highly computationally expensive. This and the fact that sequence differences between mAbs mainly occur within the binding region led to the decision that only homology modeling were to be used in this thesis. Due to evolutionary principles, protein structures from the same protein family stay

conserved in their core. Minor changes in the sequence lead to minor changes in structure. Once one member's structure is experimentally proven, the rest can be predicted based on sequence similarity since they are related (Fiser, 2010). Homology modeling can be performed for the whole antibody or different parts, e.g., Fab or F_v part. The calculation costs for Fab and F_v are significantly lower compared to the entire antibody. If, for example, the binding affinity at the antigen binding site is investigated, only the F_v part of the molecule must be modeled. Nevertheless, pharmaceutical development requires exploring the overall characteristics of the whole antibody. Due to this, only full antibody models were created and used for further calculations in this thesis. To characterize a molecule, a 3D structural model of the amino acid sequence is required, which can be calculated through Homology modeling. The homology modeling process comprises multiple steps/methods, as shown in Figure 4.

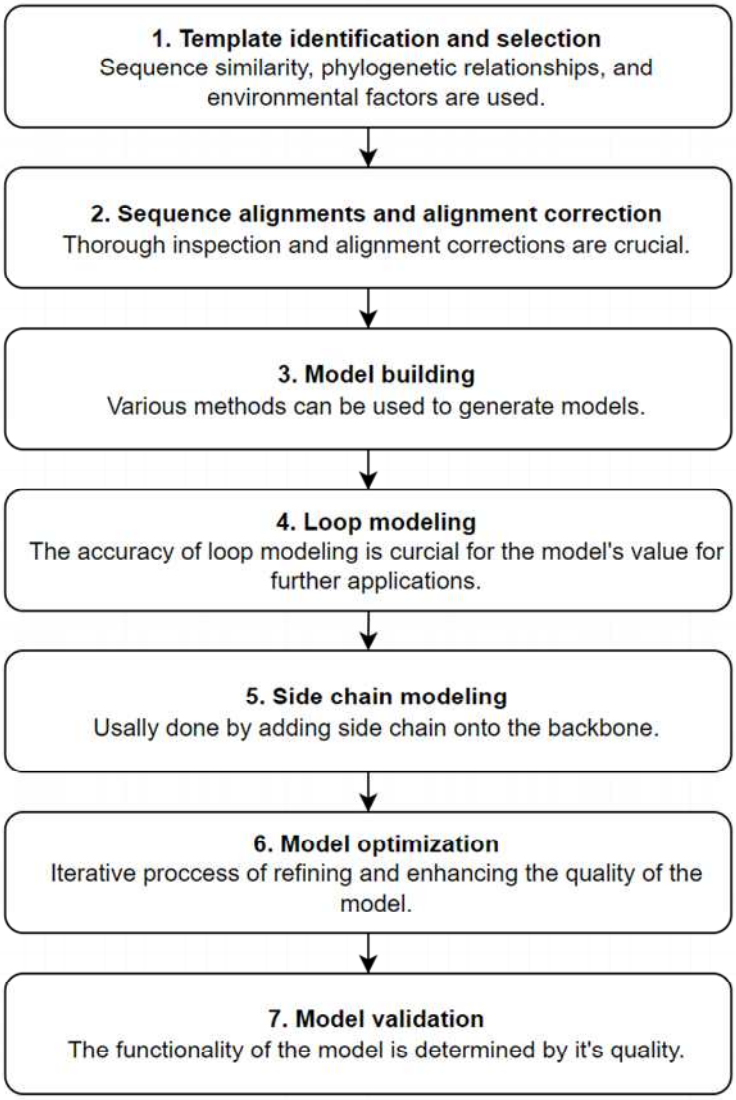


Figure 4 Homology Modeling steps. Figure adapted from (Muhammed & Aki-Yalcin, 2019).

2.1.3 Structure-derived protein descriptors

In order to gain protein insight and to better characterize a molecule and its biomolecular chemistry, theoretical protein descriptors or descriptors can be calculated based on the 3D structural model. The idea is that unfavorable compositions can be identified based on favorable or unfavorable descriptors before performing a stability screening. Basic descriptors can be calculated on a single static 3D structural model. These basic descriptors do not consider that protein structures change through external influences and interactions. More sophisticated and precise descriptors are calculated through complex molecular dynamics simulations that derive descriptors from multiple structural conformations over time. These simulations are based on calculating atom and molecule movements over a fraction of seconds. Compared to static descriptors, the computational effort is higher multiple folds since atomic interactions are being calculated for each time step repeatedly (Klepeis et al., 2009).

Furthermore, pH-dependent calculation of protein descriptors can focus on protein protonation and structural flexibility through multiple conformations. A stepwise pH change that influences a protein's binding affinity is introduced due to different protonation states of ionizable groups. For each pH step, protein descriptors can be calculated. The pH-dependent descriptors can be used to improve protein characterization and drug development since favorable and unfavorable descriptors influence protein stability (Onufriev & Alexov, 2013). The descriptors were stepwise calculated for each protein between pH 4 and 7 since this is the most common range within which therapeutic mAbs are formulated.

2.1.4 Supervised learning algorithms

Supervised learning algorithms were used in the context of this thesis to check if biologics data were sufficiently standardized to derive accurate predictions based on protein descriptors. Supervised learning uses inputs (features) to predict a certain output (labels). The algorithms are called "supervised" since they require the assignment of labels to the data so that the machine can learn to map a function or model during training. The goal is to apply the function to new data as input and to generate an output. An important part of the quality of the model is the label's correctness. Incorrectly labeled data will affect the accuracy and effectiveness of the model. Typical supervised learning applications are:

- Regression is used to map a function to predict a numerical and continuous output to given inputs. One example could be predicting a protein's column retention time (liquid chromatography) based on previous experimental values and protein descriptors.
- Classification categorizes (categorical classification) data based on the input. One example would be classifying predicted protein descriptors from homology modeling at a given pH into a stable or unstable formulation group.

Unsupervised learning, on the other hand, uses unlabeled data to find patterns and rules for clustering. In this thesis, only supervised learning was used to deduce data standardization requirements, whereas unsupervised learning is not further explained.

2.1.5 Decision trees and random forest

The decision tree is a supervised learning method primarily used to solve classification or regression problems. The goal is to train a model with training and test data including attributes (features). The algorithm determines which attributes are predictive and lead to a conclusion or classification. It is often visualized in a tree- or flowchart-like structure. This representation helps to better understand the model's classification logic. Decision trees belong to the white box models (contrasting decisions from artificial neural network are difficult to interpreted and thus belong to black box model). Decision trees are read from top to bottom. Each node represents a decision based on specific feature criteria from the data set. Except for the leaves at the bottom of the tree, since these are referred to as leaf nodes and represent a final classification. Decision trees are simple to implement, easy to understand and allow insight into the decision-making process.

The disadvantage of decision trees is their narrow fitness to the training data set. They are less robust to new data with a slightly different context and yield less accurate results. Decision trees have a chance of becoming overfitted. Overfitting occurs when a model is too tightly adapted to the training data. In this case it does not capture general patterns in the data, which lets it perform well on the training data, but inaccurate on new data.

Random forests, on the other hand, are more accurate and robust when presented with new data. They use random sampling of the data input features to build a pre-defined number of decision trees. They can effectively handle noisy and outlier data, while they are less prone to overfitting. Furthermore, only a subset of the features is used when splitting a node of the tree.

Figure 5 depicts the training and prediction process of a random forest model using an example data set describing dogs and cats based on specific features. The figure shows three decision trees and their respective classification. During the training process a data set is used that describes dogs and cats based on specific features. Multiple decision trees are trained to learn which feature indicate a particular class (cat or dog). For examples, the presence of whiskers or the ability to bark could be distinguishing features.

When presented with new data including features, but without classification, the random forest model can determine whether the given features lead to a classification as dog or cat. Each decision tree decides on the features and the classification individually. During majority voting the most frequent class is selected as the final classification.

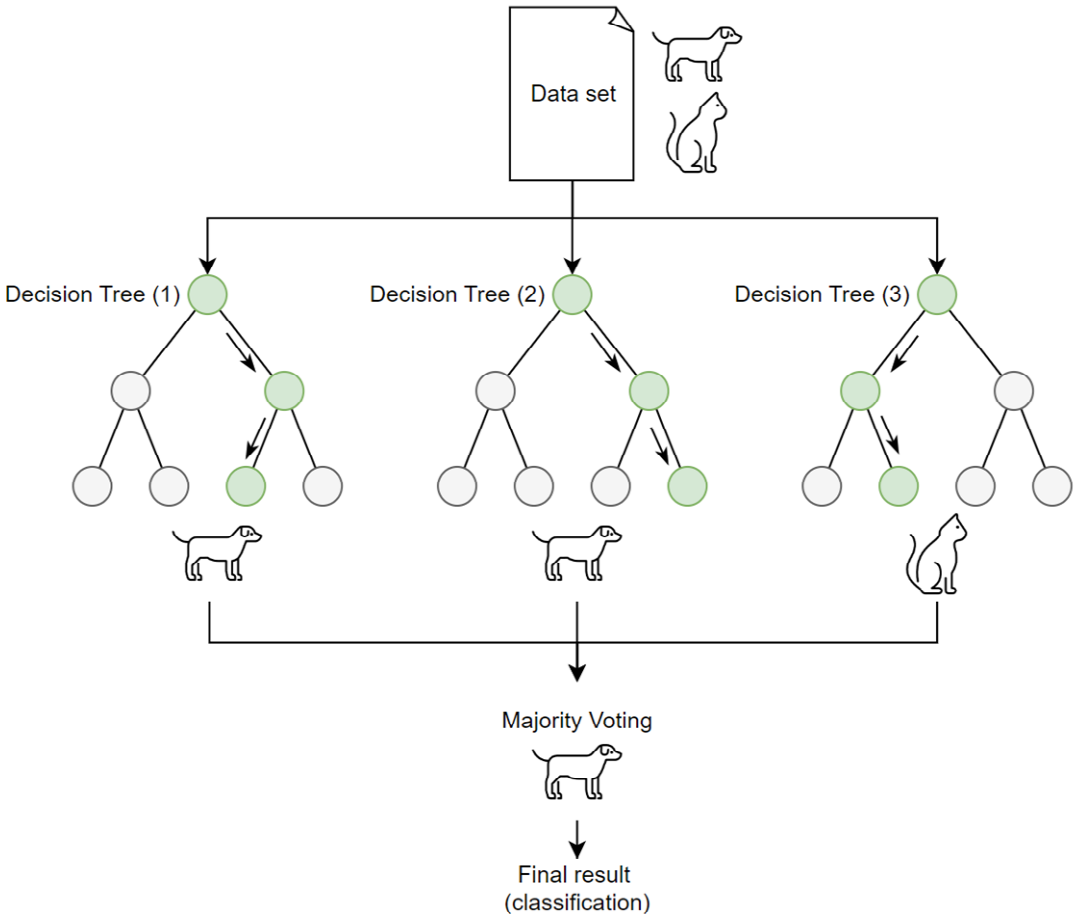


Figure 5 Decision tree schema. Multiple decision trees are generated during a random forest training process. The green nodes represent a decision that led to the next node. The trees are traversed from top to bottom. A classification is achieved by reaching a bottom leaf of a decision tree. Majority voting over all trees and their classifications is performed to assigned the result class (Machado et al., 2015).

2.2 Assessment of data standardization for ML

Three biologics data sets from different sources were chosen and used for predictive approaches such as ML algorithms. The three data sets were reviewed in terms of their level of data standardization. The initial idea was to find multiple data sets that include High-performance liquid chromatography (HPLC) data since it is the most important analytical method in biologics drug development. However, it was impossible to find HPLC data for all use cases.

The first data set focused on predicting the ideal pH from structure modeling for three mAbs. The ideal pH for these three mAbs was previously assessed experimentally through stability screenings using HPLC. However, during the model training process, the analytical readouts were not utilized as input features since they were already indirectly considered in selecting the optimal pH during the screening. PH-dependent protein descriptors were calculated from the structures. The protein descriptors were used to train a Random Forest Classifier (RFC)

model. The second data set originated from a company-internal, globally available dashboard that contains biologics data for different mAbs. The dashboard was accessed to retrieve a data set and test whether the data would be suitable for training predictive methods. However, the HPLC results were limited in the dashboard and only partially available for the intended purpose.

The third data set contained pH-related filing data from a commercial data provider. The optimal pH values for the mAbs in the third data set may have been experimentally determined through HPLC measurements. Like the first data set, structure modeling data was created for a dozen mAbs. PH-dependent protein descriptors were derived. An RFC model was trained with the protein descriptors. The model was later tested with the filed ideal pH of each mAb.

2.2.1 Data set 1: HTS formulation data

The first biologics data set contained HTS data (formulation data) combined with protein descriptors from homology modeling. The aim was to predict the size-exclusion chromatography (SEC) Monomer based on previously calculated protein descriptors from the sequence using molecular dynamics simulations for an unknown molecule at a given pH. Furthermore, if data with a sufficient amount and quality is available to achieve this goal. Identifying the ideal pH for a new molecule is the most important stabilizing factor in formulation development, achieved through pH stability screenings. During these, the quality can be identified by a high Monomer content at high temperature stresses, indicating a stable formulation.

Out of 12 available screenings for different mAbs, only 3 contained step-wise pH data, including HPLC analytics for each step. The other nine available screenings were conducted to determine stabilizing excipient concentrations instead of a systematical pH determination. From these three pH screenings, a relevant subset was extracted for training and testing. During these screenings, the following pH ranges were tested: 4.0, 4.6, 5.2, 5.8, 6.4, 7.0, and 7.6. SEC was measured after the application of temperature stress conditions for 40°C after 7 and 21 days. This resulted in 21 SEC Monomer values for the three mAbs and the seven pH steps, ranging from a minimal Monomer of 94.56 to a maximum value of 97.66.

The protein sequences were used to generate mAb structure models through homology models with the antibody modeling software Molecular Operating Environment (MOE)[1]. Based on the structure, 45 protein descriptors were calculated through molecular dynamics simulation and pH-dependent protonation for the same pH steps used during the screening. The protein descriptors include:

[1] MOE, 2020 Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2024

- Surface area and patch characteristics for the whole structure and near CDR, e.g., hydrophobic protein patches
- Electrostatic and charge-related descriptors, e.g., protein net charge
- Structural and physical properties, e.g., protein mass
- PI prediction, e.g., structure-based pI

These calculated pH-dependent protein descriptors were combined with the experimental screening results, including Monomer values for each pH step after temperature stress. All 21 Monomer values were evenly labeled as good or bad based on the Monomer value, with an equal proportion between the number of good and bad classifications. The protein descriptors were used as features. A decision tree classifier from the Python-based Scikit-learn library (Pedregosa et al., 2011)[2] was used to find the most significant features that could predict if specific protein descriptors at a given pH lead to a good or bad classification. The decision tree classifier was trained with default parameter values. From the overall 21 data points with 45 features.

Cross-validation and resampling of the data were used to determine the robustness and accuracy of the model. Therefore, multiple decision tree models were trained to map a function between the protein descriptors and the resulting pH classification into a good and bad category. Each model is learned by randomly selecting training and testing data (resampling). Two random thirds of the joined data were used for each iteration to train the model. The remaining third was held back from the training and separately used to test the model after each iteration. After each model training iteration, testing for each iteration was conducted to evaluate the predictive robustness and accuracy of the model. The evaluation of the model's robustness showed that the model's accuracy varies between 0.28 and 0.86, whereas 0.28 indicates poor performance, and 0.86 indicates a well-trained model, indicating unstable accuracy. These variations may result from an unstable or not well-fitted model, which could be caused by various reasons such as over- or underfitting the data or a too low amount of data.

2.2.2 Data set 2: company-internal biologics dashboard data

The second data set was received from a company-internal globally available web-based dashboard. The dashboard was created through the Business Intelligence and visualization software TIBCO Spotfire Analyst 10.10.4, TIBCO Software Inc., Somerville, Massachusetts., USA, 2024. The dashboard retrieves the data from multiple relational databases. It collects data from different departments and groups (data owner-groups or departments), from early discovery and late-stage development to clinical data. It is a potential data source for retrieving a sufficiently standardized data set with an adequate amount of data. The data format ranges from descriptive Key Performance Indicators (KPIs) to experimental raw data. Each data owner decides what type of data and metadata is uploaded to the dashboard. The

[2] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

dashboard also includes mAbs data and additional data regarding engineered mAbs variants. Therefore, the following data set criteria were defined. The goal was to use these to filter the data in the dashboard to retrieve a well-suited data set:

- Only classical mAbs species and no other variants should be included.
- Protein concentration between 80 to 150 mg/ml.
- Analytical data from only SEC experiments.
- Different formulation compositions (pH range).

The dashboard was accessed through the web front end. A first approach to limit the data according to the criteria in the dashboard was unsuccessful because not all required filtering fields (metadata) were available. The protein concentration range could not be selected because the corresponding field contained text instead of numerical values required for filtering. Therefore, the data was downloaded and manually filtered through Python. Data for 41 classical mAbs were available in the desired concentration range. From these, only three mAbs contained chromatographic data. Data for further processing could not be retrieved because the amount of available data was insufficient. The limited data availability could be caused by various reasons, such as: 1. The dashboard is a historically grown internal storage solution. 2. The filter criteria were too narrow, and the purpose of reuse was too unique. 3. Only a fraction of the data from different groups was shared through the dashboard. 4. A lack of data governance, e.g., missing metadata, missing values (NA), duplicate entries, and incorrectly spelled manual entries, results in less usable data. In summary, no data could be reused for predictive models due to a lack of data volume, governance, and standardization.

2.2.3 Data set 3: external pharmaceutical data provider

The goal was to train a model with calculated pH-dependent protein descriptors to predict which descriptors lead to a favorable (stable) or unfavorable (unstable) pH. PharmaCircle, an external commercial pharmaceutical data provider, was chosen to retrieve a third data set. The provider collects authoritative information, global insight, and expert analysis regarding biotech and pharmaceutical data throughout all stages of drug development. The provider extracts information and data from drug approval documents globally (Drug-Dev.com, 2021). During the drug approval process, pharmaceutical companies must provide extensive data on the drug's critical quality attributes (CQAs), including the analytical methods used to measure these attributes. The details of the information vary depending on the region where the therapeutic drug is approved. Some countries require the publication of the drug sequence together with a detailed description of the drug's formulation composition, including the targeted ideal pH.

The received data set was similar to the HTS formulation data (data set 1) because structure models were built to generate protein descriptors through a molecular dynamics simulation at various pH steps. These were used to train a random forest classifier from the Scikit-learn

Python library (Pedregosa et al., 2011)[3] to predict if protein descriptors lead to a stable or unstable pH. The difference was that the pH was not experimentally measured as in data set 1, but that the filed pH from the approval documents was assumed to be ideal - most stabilizing for the given mAb.

The following criteria were used to filter and retrieve a suitable data set:

- Data includes drug approval information such as mAb product name, mAbs species, drug product dosage form, protein concentration, sequence, and target pH.
- Only approved mAbs with liquid formulation (lyophilized products were not in scope).
- The protein concentration was between 80 and 150 mg/ml.

The data were retrieved from the PharmaCircle web front end as a data dump in table format. Unfortunately, the sequences for some mAbs were either missing or provided in a picture format. Some of the missing and picture-based sequences could be found through a manual sequence search in the SabDab (Dunbar et al., 2014) and KEGG databases (Kanehisa & Subramaniam, 2002). The remaining sequence pictures that could not be found in the databases were scanned through an OCR algorithm. The algorithm transformed the picture information into a text-based amino acid sequence. Two scientists manually compared picture-based sequences with results from the OCR algorithm to ensure sequence correctness, resulting in 26 antibody sequences that were successfully retrieved.

These sequences were used to build full antibody models through homology modeling in MOE. In the following step, 45 pH-dependent protein descriptors were calculated through molecular dynamics simulations and a protein protonation for each mAb.

The selection of calculated protein descriptors was equal to Data set 1 (HTS formulation data), but the values changed due to the different mAb sequences. The pH ranged from 4.0 to 7.6 in 18 evenly distributed steps. This resulted in 468 data points for the 26 mAbs. For each mAb, a target pH range of +-0.5 was defined based on the approved pH from the approval documents. All pHs within this range and all corresponding calculated protein descriptors were assumed to favor the mAbs stability and classified (labeled) accordingly. 119 of 468 pH steps were classified as stable. The pHs and corresponding calculated protein descriptors outside the range were classified (labeled) as unfavorable for the mAbs stability. As a result, 349 of 468 pH steps were classified as unstable. Figure 6 shows a schematic overview of the pH-dependent classification of the calculated protein descriptors based on the distance to the ideal pH.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

molecule	calculated pH	approved pH	calculated protein property 1 (feature)	calculated protein property 2 (feature)	calculated protein property n (feature)	label / classification
external mAb 1	4.0	5.5	10	-5	...	unstable
external mAb 1	5.0	5.5	14	0	...	stable
external mAb 1	6.0	5.5	18	30	...	stable
external mAb 1	7.0	5.5	12	-10	...	unstable
external mAb 2	4.0	4.5	8	-4	...	stable
external mAb 2	5.0	4.5	16	20	...	stable
external mAb 2	6.0	4.5	12	5	...	unstable
external mAb 2	7.0	4.5	7	-5	...	unstable

Figure 6 Example protein descriptors of different mAbs used for the RFC model training. The calculated protein descriptor columns were used as features to train the model. If the pH-dependent calculated protein descriptors were within a pre-defined range around the approved pH, they were labeled (classified) as stable (dark green color). If they exceed the range of 0.5 between the calculated and approved pH, they are flagged as unstable (dark red color)

Prior to the RFC model training, the ten most predictive features were selected using the SelectKBest module from the Scikit-learn library (Pedregosa et al., 2011). The RFC model was trained with the pH-dependent protein descriptors as input features and called with default parameter values[4]. The classification for each pH step was used as labels during training based on the approved pH. The data set was randomly split into three-thirds. Two-thirds of the mAbs were used to train the RFC model. One-third was excluded and used for the following testing. The model showed an accuracy after a predictive test with test data of 0.74. The ten selected features and their relative importance are displayed in Table 1. Based on the relative importance scores, the dipole moment has the highest impact, followed by positive, negative, and ionic-charged patches.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Table 1 The ten most predictive selected protein descriptors(features). The features were calculated from mAb sequences through a full antibody molecular dynamics simulation. They are a subset of 45 available descriptors from the homology modeling software MOE.

Protein descriptor	**Relative importance** (higher is better)	**Description**
ens_dipole	0.17	Measure of the separation and distribution of electric charge
patch_ion_n	0.13	Amount of ion patches
patch_neg	0.12	Overall, negatively charged patches
patch_ion	0.10	Overall ion patches
patch_neg_n	0.09	Number of negatively charged patches
patch_cdr_ion_1	0.09	Biggest ion patch near binding region (CDR)
patch_pos_n	0.09	Number of positively charged patch
patch_cdr_pos_1	0.08	Biggest positive charged patch near binding region (CDR)
patch_ion_1	0.07	Biggest ion patch
patch_pos_1	0.06	Biggest positive charged patch

2.2.4 Impact of data standardization on ML results

The three data sets indicate an overall lack of data standardization. While preparing every data set, data and metadata were missing or did not exist. A limited predictive accuracy for the first data set was observed. The scientific reasons, such as the low number of only three available internal pH screenings, resulted from a missing company-internal data strategy to consistently conduct standardized formulation screenings with an equal scope, such as pH determination. Conducting more equal screenings, meaning that the same type of screening is performed repeatedly, can help improve the accuracy and reliability of ML algorithms.

Furthermore, more data can be collected and used for advanced analytics. Both reasons indicated a misleading data strategy that failed because the data could not be comparable. Biologics data are challenging to integrate and make comparable. For this type of data, a larger data context is required. Therefore, it was challenging to reuse complex data for predictive approaches. The different data collection strategies will be explained in detail in Chapter 2.3.

The second data set serves as a reminder of the consequences of a lack of metadata annotation, leading to missing or unfindable data. It also underlines the necessity of adopting a different data collection strategy. The current approach, which involves collecting data first in a dashboard and attempting to integrate it later, is challenging. These include the absence of context for integration, missing metadata, poor data quality, data variability, data noisiness, and the heterogeneity of biologics data. A more effective strategy is needed to address these issues and ensure the availability, accessibility, and, ultimately, data reuse.

The third data set showed that well-structured, well-prepared, and well-standardized data can improve data reuse. This was achieved through manual data preparation and collection

from multiple sources, resulting in improved predictive model accuracy. However, the data set only contained 468 data points for 26 mAbs, which is considered insufficient for sophisticated algorithms such as a random forest classifier. Even if more data were available, the manual amount needed to review the data and metadata would increase significantly and become a bottleneck for human preprocess. It is assumed that the lack of data availability and quality is due to a lack of data standardization and standardization strategies for complex data. Therefore, a closer examination of the significance of thorough data collection for different data complexity levels and the overall data standardization and implications for data collection strategies was performed.

2.3 Data complexity

2.3.1 Simple and complex data

Data complexity is crucial as it determines the level of effort required to integrate, analyze, and interpret data. It is influenced by factors such as data size, structure, quality, and the tools needed for data access. Different levels of data complexity can be distinguished:

- Simple data is easily understood and interpreted. No additional context is required to compare it, as it is precise and exact. An example is currency data from the financial sector. This data is milliseconds exact. Furthermore, financial metrics are presented in a consistent numerical format, which aids comparability.
- Complex data is relative, multivariate, and depends on multiple variables. Achieving data comparability for this type is challenging due to multiple reasons: the data originates from different sources, data is represented in different formats, the data shows differing accuracy, and the data depends on multiple factors that influence variety; a larger context is required to interpret the data. Biologics data, e.g., mAbs data, are a good example of complex data since they are very large molecules that consist of various species with a complex structure and function. Compared to small molecules, more possible sequences and solution conditions exist, which increases the potential screening space (Narayanan et al., 2021). Biologics are particularly difficult to characterize due to their sensitivity, fragility, and tendency to degrade. Due to this, analytical readouts inherit a higher level of variation because the characterization of biologics is a particular challenge. Consequently, analytical techniques are descriptive, and the corresponding results must be set in relation to a well-known reference standard. The analytical results depend on the analytical method and how it is used. Even simple analytical methods such as pH measurement depend on multiple parameters, such as sample preparation, when data comparison between molecules is desired. More complex analytical methods like the HPLC require a larger context, including multivariate parameters, to integrate and understand results. This includes measurement uncertainties (analytical method errors), process variables, and workflow execution. These factors play an essential role during data comparison.

2.3.2 Data storage requirements

Complex data requires additional effort during data preparation for further reuse and processing, e.g., through training in predictive models. Therefore, on the one hand, data complexity directly impacts how data is collected and increases the requirements for

thorough data storage. For simple data, a small metadata collection is sufficient to interpret the data. Data comparability can be achieved with minimal effort. Relational databases are often sufficient to achieve simple data comparison.

On the other hand, complex data requires a broader information context to integrate and make sense of the data – an increased number of metadata is required. Completeness of metadata has a significant impact on comparability. Lack of metadata makes data interpretation and comparison challenging. Complex data is expressed in diverse formats and data types, such as structured, semi-structured, and unstructured data. Moreover, complex data can come from various sources. The source systems rely on different technologies to store data, e.g., relational and non-relational databases. This raises the requirement for sophisticated data storage solutions or frameworks such as the Hadoop framework, which allows storage requirements for both simple and complex data to be satisfied.

2.3.3 Data collection strategies

Data collection is the process of gathering data, which enables researchers to answer questions or to derive knowledge. Accurate data capture is necessary to ensure valid, reliable, and integer research. In turn, data collection strategies define the previous planning step before the actual data collection. They are not to be confused with the different ways to collect data, such as surveys, questionnaires, interviews, experiments, or documents. Often, data collection is initiated without a proper strategy, which leads to poorly organized data that is difficult to navigate and manage.

Different data collection strategies exist: 1. collect data first and ask questions later 2. think about the questions first and start the data collection later. The first strategy provides a faster start because no planning is required. This is often the preferred choice when using the data to better understand a problem at hand, similar to exploratory research. This strategy can be a good fit for simple data that does not require additional efforts to achieve comparability. Often, pharmaceutical R&D organizations rely on this strategy during their digitalization efforts. They choose to collect data in a data lake across different groups and departments, which fails because, over time, the data collection transforms into a chaotic state. Data standardization and governance rules that enable adequate data organization and management are missing, leading to messy data.

Consequently, this results in overall poor data quality and failing digitalization efforts. The second strategy brings on a delay in data collection efforts because a planning phase is previously performed to draft which data is required to answer certain questions about an existing problem. Furthermore, the delay is increased due to the specification of data governance rules, collection guidelines, and quality criteria. This strategy is necessary for complex data since it requires a thorough and adequate data collection strategy. For simple data, this strategy exceeds the requirements. The general disadvantage of this strategy is that it can lead to overplanning. The cause is that the preparation phase is too long and complex. For example, uncertainty about the questions that should be answered leads to months-long preparation. Over this time, the requirements change, which leads to wasted efforts and, finally, the cancellation of the overall initiative. Table 2 compares different data complexity levels and the corresponding data collection strategy.

Table 2 The two data collection strategies for simple and complex data.

	Simple data	**Complex data**
Collect data first, ask questions later	applies	does not apply
Think about questions first, start data collection later	partially applies	applies

The data sets indicate that biologics data necessitates an adjusted data collection strategy due to its complexity. The strategy should prioritize the questions that must be addressed before initiating data collection. The practice of collecting data first and formulating questions later is not viable. The reversal of this approach is essential. Ideally, the AI approach's objective, including the formulation of questions and potential answers, should be roughly determined in advance. Considering these, a biologics standardization strategy can be defined as gathering the necessary and essential data and metadata appropriately (regarding data standardization concepts). Subsequently, the data can be collected as required. This alternative approach, which results in an optimized data collection strategy, is shown in Figure 7.

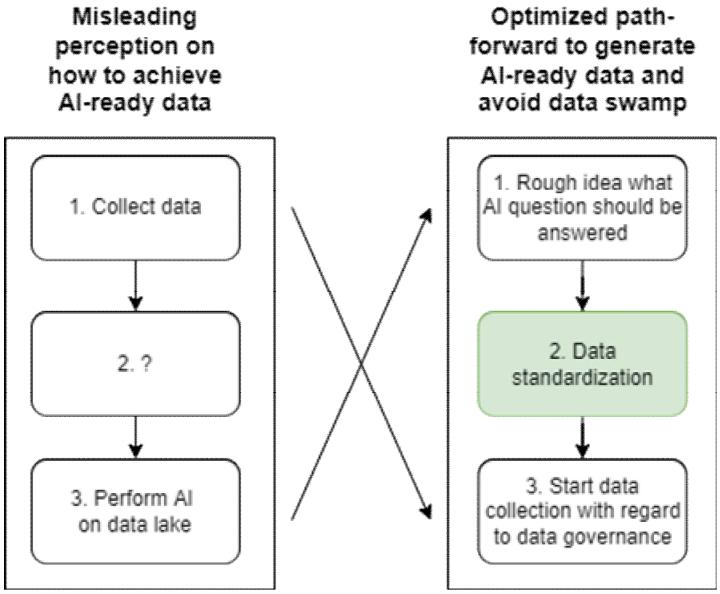


Figure 7 Inverse data strategy with the end in mind. A data standardization concept is required to enable the successful collection of data.

Independently of the data complexity level, data preparation is the required first step before any application of data analytics. Data Scientists spent 70 percent of their time upcycling data through manual data wrangling: collecting, preprocessing, normalizing, and cleaning the data to achieve sufficient data quality before the application of AI (Zhang et al., 2003) (Ahuja et al., 2016). At the latest, they notice an insufficient quality when the trained models show poor accuracy during testing or cannot be trained successfully. Increasing digitalization and automation raises the need for automated solutions performed by the machine.

2.4 Requirements for a biologics data standardization concept

The demonstrated data sets highlight the importance of data standardization for advanced analytics, such as ML. They also emphasize the challenge of utilizing various data sources, such as internal biologics formulation data, company-internal data from dashboards, and publicly available protein from PharmaCircle, for advanced analytics.

The poor ML results indicate an overall low data quality and lack of data availability, which impacts the effectiveness and accuracy of ML models. Standardization concepts are required to ensure the comparability of biologics data and to identify which data can be used for ML. Standardization assessment methods must examine this lack of standardization more closely.

The FAIR principles (Findable, Accessible, Interoperable, and Reusable) promote reusable scientific data and can thus aid as a standardization framework. Furthermore, FAIR allows one to assess whether data aligns with FAIR requirements, indicating the level of standardization. Figure 8 shows an overview of the challenges and the theory that the level of standardization must be further explored through FAIR.

The complexity inherent in biologics data makes it necessary to establish a standardized concept that enables consistent and machine actionable data interpretation for humans and machine. This facilitates the identification of data set quality differences in data sets, allowing both humans and machines to assess which datasets are most suitable for the intended tasks. The question is how the FAIR principles may aid in the creation of a biologics standardization concept.

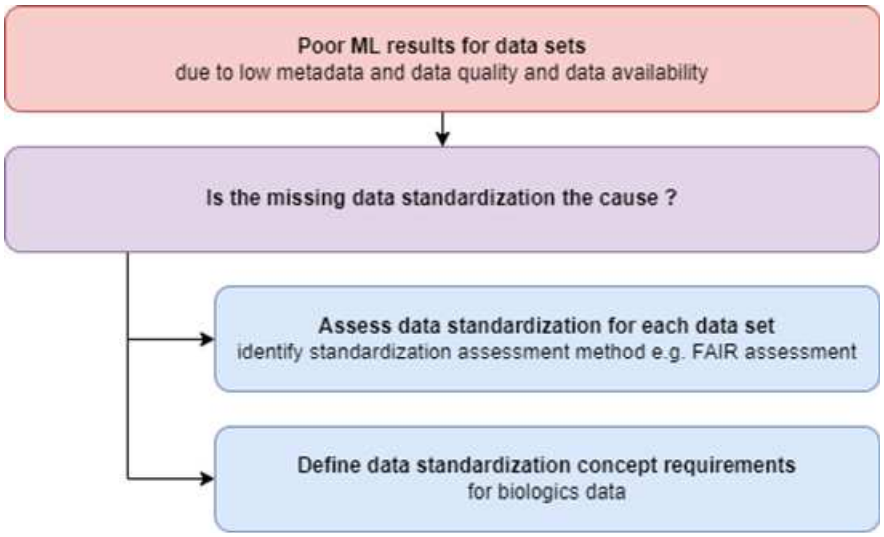


Figure 8 Overview of implications from the poor ML results. Missing data standardization is identified as a potential cause for the results. Therefore, a data standardization assessment for the three biologics data sets and the definition of requirements for data standardization concepts are required

However, a data standardization concept for biologics data must fulfill the following requirements:

1. Machine actionability, so that the machine has the capability to reuse biologics data similar to a human.
2. The standardization concept should provide a sufficient context for the machine to set biologics data results from different data sets in relation to being able to compare different data sets.
3. A sufficient collection of relevant metadata that enables manageable data integration.
4. The metadata should reflect data quality, identifying whether a data set is useful for applying predictive concepts and determining its comparability with other data sets.
5. Data representation should be in an ideal format to store complex data so that the machine can automatically access it.
6. Reusing existing concepts is a desired goal instead of developing new concepts.
7. The concept should enable the automation of the scientist's logic for data set comparability so that the machine can automatically act on the data without human intervention.

The SEC method is the most important analytical technique, used in many industries to accurately and precisely analyze and characterize drugs and other compounds. The standardization concept should, therefore, concentrate on SEC data, similar to the three biologics data sets.

3 State of the art

This chapter explores how lab automation and process standardization can aid in the generation of systematic and reproducible data sets. These data sets are the foundation to facilitate data standardization through standardization concepts such as FAIR and semantic concepts such as Allotrope foundation models. Additionally in this chapter, different methods to assess data standardization are introduced and the term data quality is defined. The chapter highlights that there is currently no biologics data standardization concept. Therefore, the development of a biologics data standardization concept is being proposed.

3.1 Lab automation and high-throughput screening concepts

Lab automation enables pharmaceutical companies to standardize processes and workflows and systematically generate standardized data sets. By standardizing workflows and processes, organizations can reduce errors and variations that may arise from manual handling. It is used to increase sample throughput, run repeatable experiments, and free up manual capacities (Chapman, 2003).

In the context of this thesis, the term “lab automation” refers to the integration of different lab devices from different vendors combined in a fully automated laboratory system. This fully automated laboratory system, as depicted in Figure 9, is able to handle 96 and 384 well plates for high-throughput liquid formulation drug development. Moreover, the term lab automation includes related aspects such as workflow, processes, analytical methods, samples, controls, errors, and analytical results.

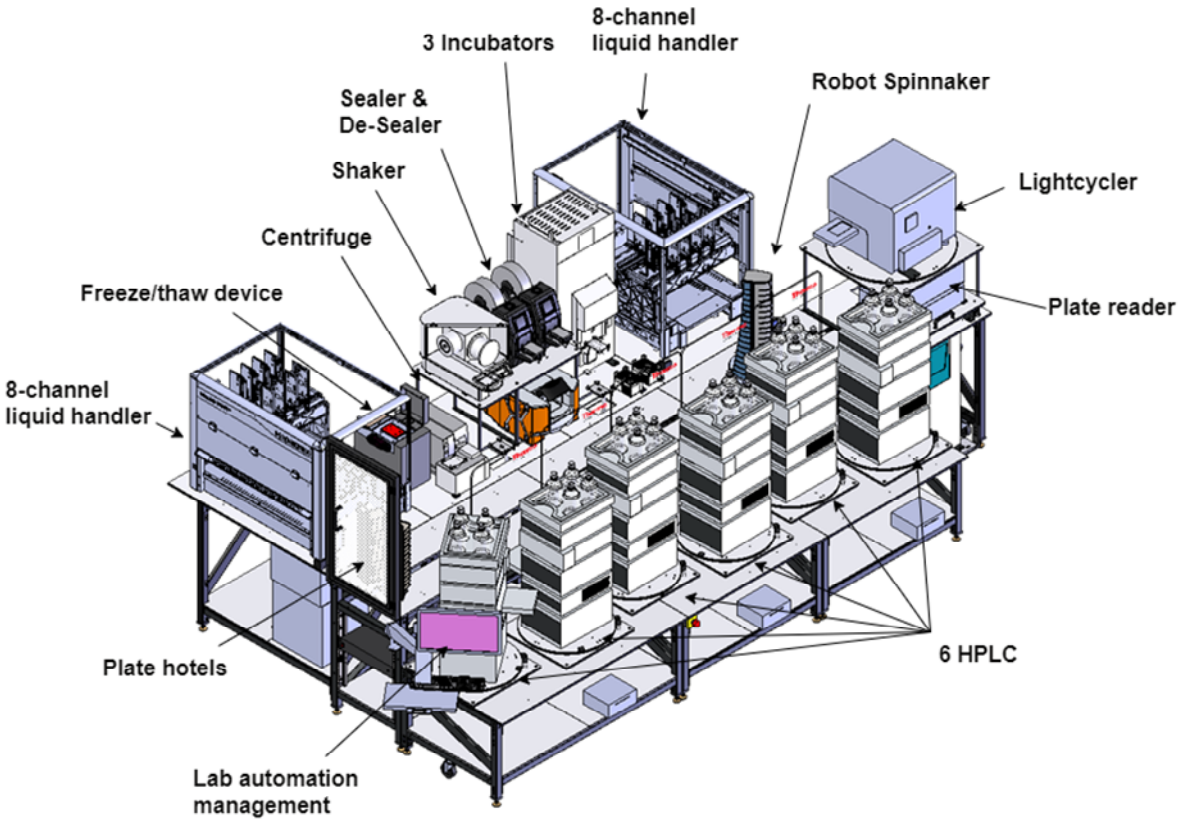


Figure 9 Fully automated laboratory system for high-throughput liquid formulation development. The system includes several devices from different vendors, such as plate handle robots, plate readers, liquid handling systems, High-Performance Liquid Chromatography (HPLC), shakers, sealers, de-sealers, freeze/thaw devices, and incubators. It is controlled and programmed by lab automation management software (Siedler et al., 2020).

A lab automation-based high-throughput screening (HTS) approach enables standardization on multiple levels: workflows, experiments, plate handling, analytical methods, and data management. The aim is to quickly collect standardized data sets required to generate a sufficient product understanding and stable product (Chan & Carter, 2010). The use of microtiter-based plates enables a miniaturized lab automation approach. This results in an increased throughput while generating more data points; only a minimal sample volume is required.

Additionally, standardization of the screening designs enables the systematic derive standardized scientific conclusions and, thus, standardized data sets. This is achieved through standardized screenings (modules). For example, the formulation's pH and protein concentration significantly influence product stability and crystallization (Sjuts et al., 2020). Figure 10 depicts different stability screening scopes. Therefore, in the first screening, the pH optimum is explored by varying the pH and protein concentrations adding a buffer and surfactant (module 1). The pH-dependent drug stabilizing excipients are explored in a second and consecutive screening to determine the optimal composition (module 2). A third screening combines the insight of the first two screenings to fine-adjust the formulation components and concentrations to explore the formulation design space (module 3). The

flexibility of the plate layout increases with each screening. The plate layout refers to the degree of variation of different formulation compositions. Module 1 has no flexibility since the pH ranges are predefined and fixed. Module 2 depends on Module 1, allowing more freedom to vary the pH-dependent excipients. Module 3 has the most flexibility and allows for greater change of the formulation parameters. A standardized set of different stress conditions is used for all screenings to examine stress-induced stability differences of each composition. The result can be a systematic and normalized data collection. These are well-formatted, standardized, and can be reused without intensive manual data preparation.

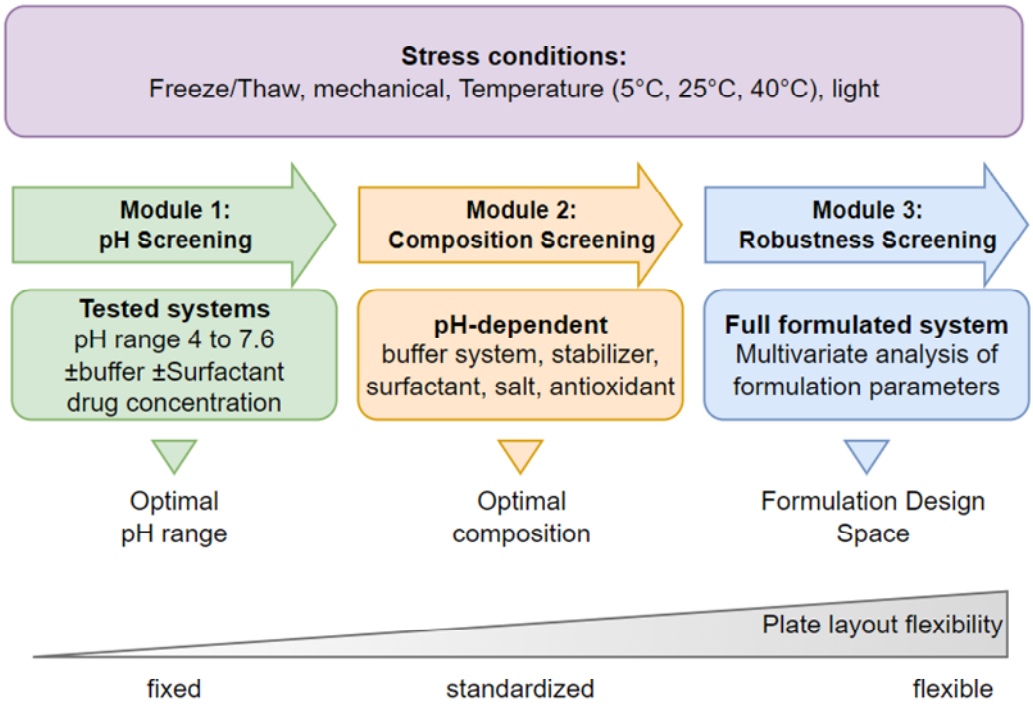


Figure 10 Modular high-throughput formulation screening design. The concept enables the generation of standardized data sets comprising three consecutive screenings. Each screening has a different scientific scope (Siedler et al., 2020).

3.1.1 System Suitability Tests

System Suitability Tests (SST) play a crucial role in drug development to ensure the reliability and precision of analytical methods (Jenke, 1996).These tests are performed before the actual experiment to assess the performance and suitability of the analytical system. They ensure that the systems are working as expected, enabling the generation of valid analytical results. This provides confidence in the equipment and ensures the product's quality. SSTs include the instruments, equipment, and analytical methods. They consist of predefined acceptance criteria tested with a reference molecule with well-known characteristics. The SST is considered successful if the reference molecule behaves as expected during the test runs and the predefined criteria are matched. This indicates that the equipment and methods are well suited for the actual analytics to generate accurate and reliable data. SSTs are the first important step in ensuring a basic level of data standardization and enabling data quality.

3.1.2 Analytical methods standardization

Plate-based assays and, thus, plate-based analytics are important in the life-science industry. They allow for quick and efficient data generation on a bigger scale compared to the use of single vials. The use of multi-well plates is ideal for conducting stability screenings because the well plate, which contains all samples, can be subjected to different stress conditions. The following analytics step allows the detection of stress-induced stability differences (Aucamp et al., 2005). The analytics – also referred to as analytical methods – are grouped into molecule-specific and unspecific analytical methods.

Prior to a screening, the method verification ensures that the analytical method correctly characterizes the target molecule. Often, small adjustments to the analytical methods are required, which could potentially render the analytical readouts incomparable to the results of previous molecules. This is prevented by result-based method optimization. Although some method parameters or settings change, the readouts stay comparable. HPLC is one of the most important techniques in modern research and development departments for physicochemical analysis (Snyder et al., 2010). Due to its speed, accuracy, and reproducibility, the SEC is the predominant analytical chromatographic method that separates molecules according to differences in size as they pass through an SEC medium packed in a column (Hong et al., 2012).

The size-exclusion chromatography (SEC) method is primarily used for impurity detection. Depending on the molecule's characteristics, quality, and liquid formulation, the molecule elutes in three distinct parts over time. The larger parts, such as aggregated protein structures, are the first to elute from the column and are known as High Molecular Weight (HMW) components. The main intact part of the molecule, referred to as the Main or Monomer, elutes after the HMWs. This part defines the majority of the analytical readout. If the molecule fragments into smaller pieces, these are the last to elute from the column and are called Low Molecular Weight (LMW) components. Unlike other chromatographic methods, such as ionic exchange or affinity chromatography, molecules do not bind to the medium during SEC.

3.1.3 Workflow standardization using positive and negative controls

The use of controls in experiments is good scientific practice. The controls function as a reference against an expected analytical outcome or to detect the unexpected performance of the machine. During formulation screening, well-characterized molecules are used, one as a positive control with good stability and one as a negative control with poor stability. These controls are subjected to the same stress conditions as all formulation samples during the experimental workflow. Equally to the samples, the stress influence is measured and analyzed for each control after each stress pull point, e.g., 40°C temperature stress for 21 days duration. The controls serve two important standardization purposes:

- As a benchmark to measure the stability performance of the current molecules. In case the analytical results of the current molecule are similar or closer to the performance of the positive control, the molecule tends to be more stable. If the results are similar to the negative control, this indicates poor stability.

- As an indication of the successful execution of the workflow, processes, and stresses. All positive and negative controls analyzed in previous stability screenings, including some variations due to external (environment, method performance) and internal influences (control batch), span an expected range for each stress. If the controls in the current screening perform within this expected historical control range, the specific stress was successfully applied to the well plate, including the controls and the samples.

In contrast to the well-characterized company-internal controls, recent research is focused on developing industry-wide and company-independent reference antibodies. These potentially improve cross-industry comparability and standardize liquid chromatography results.

Controls are important in ensuring that different workflows or parts of the workflow (stress conditions) can be compared between different molecules. They are also essential in enabling data standardization. If a control is invalidated due to technical or scientific issues, the data quality and validity of the experiment cannot be ensured. This may lead to the invalidation of all sample results of the related stress condition.

3.1.4 Errors and sample replication

Molecule-specific analytical methods require adaption for new molecules. With each adaption, an estimate of the analytical method error is required. This error is influenced by various factors such as device parameters, the chromatography equipment such as the column, sample preparation, the analytical method, and the molecule's characteristics. In an automated laboratory setting with high-throughput screening, the analytical method error can be estimated at around 2-3% of the monomer. The error estimation is crucial in differentiating between various samples and the influence of the formulation composition on their stability.

Furthermore, the automated workflow may increase the overall error and add to the analytical method error. Differences in the workflow (processing) of samples and controls, such as sample preparation, stress conditions, well plate handling, and incubation times, may significantly impact analytical readouts and increase the overall error through variations. Considering and controlling these factors is crucial since they must be considered when estimating data quality. Therefore, they play an important role in data standardization concepts. One way to better determine or mitigate these errors is using sample and control replication. By replicating the same sample (with the same formulation composition) on the same well plate, the overall error can be more accurately determined by calculating the mean stability of the replicated samples or controls (Borman, 2021). This enables the improvement of the data quality. However, replication comes with the cost of a lower number of available wells on the well plate, resulting in a smaller number of different formulation conditions that can be screened.

3.2 Scientific data standardization

Scientific data standardization concepts aid data set comparability and reproducibility. They also allow for system interoperability of analytical results and enable faster and more efficient data analysis.

3.2.1 Standard Data management

Increased data generation through automation exceeds manual data evaluation and demands more efficient automated data processing and evaluation (Paton, 2007). Often, a heterogeneous system landscape like Electronic Lab Notebooks (ELNs), Laboratory information management systems (LIMS), and Chromatography Data Systems (CDS) from different vendors is used to collect and manage automation data (Machina & Wild, 2013). Integrating these systems and their multiple vendor-specific unstandardized analytical export formats is challenging and complex.

To avoid these obstacles, some departments self-develop data management solutions within the function. An example is a self-developed Python data management software at AbbVie, referred to as HTS-Studio, which is the critical element for data management and governance (Siedler et al., 2020). It is the HTS starting point for plate-based screening planning of the formulation compositions and the screening itself. Additionally, it functions as an upload and download gateway connected to a distributed Hadoop infrastructure. Uploaded analytical results are automatically processed and evaluated by Python scripts, and visualizations are created in case of scientific or technical issues that render single data points unsuited for evaluation. If the results do not meet specific quality criteria based on thresholds, they are automatically invalidated or manually flagged as dropped, leading to their exclusion from further data analysis. In both cases, every intervention is logged. If data are excluded from evaluation, all calculations are reprocessed, and the visualizations are automatically updated. This ensures regulatory expectations on data lineage and integrity and increases data quality. Composing all standardized and normalized HTS stability data for all analyzed molecules in a data pool makes the data suited for applying AI algorithms (Siedler et al., 2020).

3.2.2 Data quality

High quality information and knowledge can only be derived from high quality data (Redman, 1998). In principle, a model can be trained successfully with low quality data and even show high predictability. However, it will not be able to represent the real world and remain a fantasy (Nisbet et al., 2017). Quality can be described in different contexts:

- Quality in the context of a semantic model as an attribute qualifier (descriptor), e.g., color, size, amount.
- A single or multiple attributes during CQAs identification during drug development.
- Data quality as a measure for the intrinsic data correctness (Liaw et al., 2013): are the data fit for their intended purpose? It has multiple dimensions like completeness, accuracy, consistency, uniqueness, validity, and timeliness. An additional necessity (often forgotten) is that it must be the right data to answer a certain question. If both requirements are met, the data reaches a machine actionable state.

The latter definition is used in this thesis to describe data quality. Table 3 lists multiple definitions for data quality as an intrinsic data correctness measure.

Table 3 Multiple data quality definitions from literature(Fürber, 2016)

Year	Data Quality Definition
1996	Data is described as fit for use by data consumers (Wang, 1996).
2001	Data is of high quality when it is fit for the intended uses in operations, decision-making, and planning. It is classified as fit for use if it is free of defects and possess the desired features (Redman, 2001).
2002	Data conformant to specifications and meeting or exceeding consumer expectations (Kahn et al., 2002).
2003	Data has quality if it satisfies the requirements for its intended use (Olson, 2003).

In this thesis, data quality is defined as the suitability of a data set to be fit for its purpose, i.e., for predictive concepts. Only if it is of high quality is comparability with other data sets ensured.

3.2.3 The FAIR guiding principles

The pharmaceutical industry aims to effectively develop drugs to treat and cure medical conditions while generating shareholder value. Academics seek to create new knowledge to advance existing knowledge and develop new theories and ideas through scientific research. They invest time and resources in projects and publications but miss out on the fact that they gather valuable data. This data is not considered an asset beyond the scope of the projects and publications and is forgotten after a specific question has been answered. Furthermore, some have identified the value of data but are reluctant to share their data with others. Both are a cultural problem.

The FAIR guiding principles emerged from these issues and proposed to provide an answer. The recently formulated FAIR guiding principles act as guidelines to improve the reuse of scholarly data (Wilkinson et al., 2016). Finding and accessing data is the necessary first step in preparing data for further reuse. The FAIR principles render the data in a machine actionable state in which the machine can automatically find, access, and understand the data. When the machine is subjected to a digital object or resource, it can do the following: "1. Identify the type of object (concerning both structure and intent), 2. Determine if it is useful within the context of the agent's current task by interrogating metadata and data elements, 3. Determine if it is usable, concerning license, consent, or other accessibility or use constraints, and 4. Take appropriate action, in much the same manner that a human would." (Wilkinson et al., 2016).

Figure 11 shows an overview of the principles. Improving data reuse can be a game changer for pharmaceutical companies, while not doing so could be a competitive disadvantage for companies and, thus, the patient (Wise et al., 2019) (Alharbi et al., 2021). The costs for not having FAIR research data were estimated to be up to €10.2bn per year for the EU, which was issued in a report by the European Union (European Commission, 2019). An adverse

impact on innovation could add further €16bn in costs, resulting in a total of €26.2bn per year.

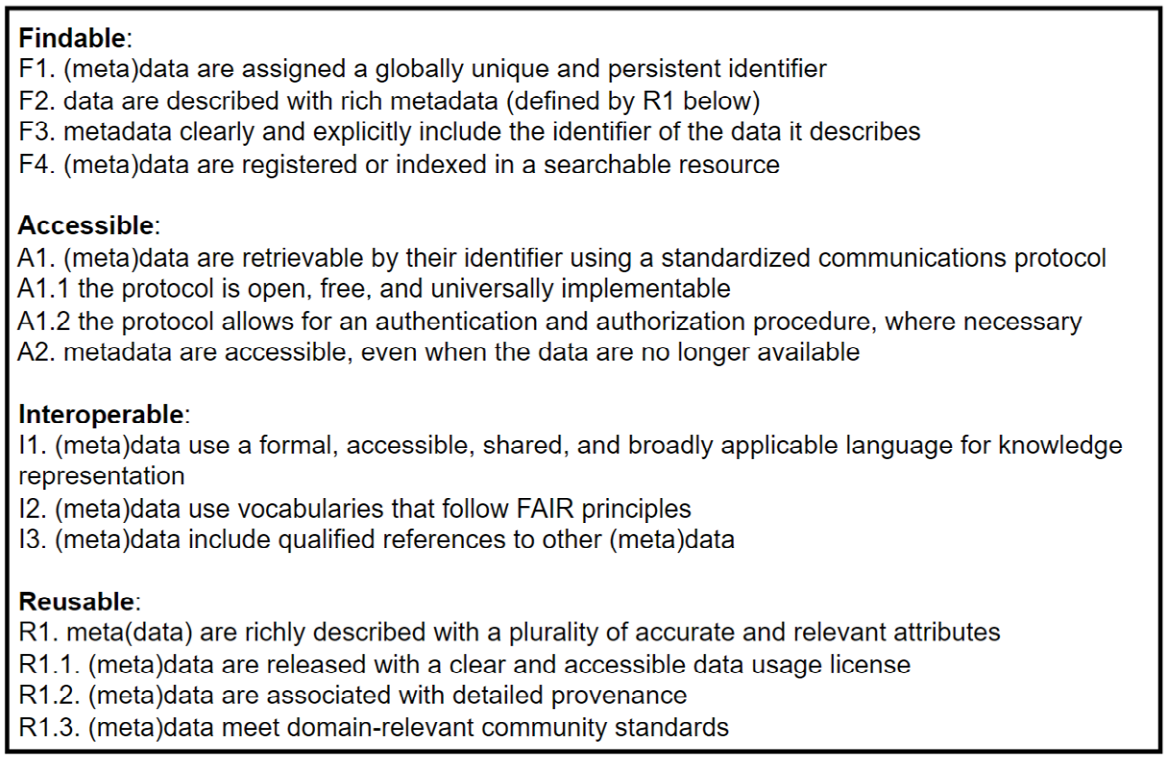


Figure 11 Overview of the 15 FAIR guiding principles. Figure adapted from (Wilkinson et al., 2016)

3.2.4 FAIR assessment options

An essential part of the reversed data strategy approach is data standardization. The lack of data standardization must be explored to improve the AI capability of the data sets. One concept to improve the reusability of scientific data and enable accessibility for the machine (machine actionability) are the FAIR guiding principles. The data FAIRness can be measured through a FAIR assessment. In this chapter, the FAIRness of the data sets is assessed. Moreover, steps to increase the FAIRness are proposed. Three options were available to assess the FAIRness of the three data sets:

1. The Research Data Alliance (RDA) published a FAIR Data Maturity Model (DMM), which includes a collection of multiple measurable indicators (referred to as RDA indicators) for each of the four FAIR principles (findable, accessible, interoperable, and reusable). These help to assess the FAIRness of a digital object or resource. Each of the 41 indicators has a priority classification (essential, important, and useful) that reflects its importance for the related FAIR principle. The RDA indicators can be compared to a checklist, where each indicator can be passed or failed. The evaluation is performed by a data owner or a person who is familiar with the data. Each indicator is described by a text. During assessment, they are manually reviewed and checked subjectively by an assessor. Two methods for indicator evaluation exist: 1. To rate the

progress of each indicator based on five levels. This method is used to get an impression of the improvement potential for each indicator. 2. To rate the indicators in a binary pass-or-fail decision. After the assessment, an evaluation methodology is provided, and the results can be categorized into different maturity levels (RDA FAIR Data Maturity Model Working Group, 2020).

2. A web-based automated framework to evaluate the FAIRness of digital resources (Wilkinson et al., 2019). The framework evaluates the FAIRness through three components: automatically measurable FAIR behavior indicators, small web apps that test the digital resources against maturity indicators similar to the RDA indicators, and an evaluator that displays and reports results. Prior to evaluation, the data set must be registered by the evaluator. After the assessment, the results of the compliance tests can be published. An example of a FAIRness evaluation for a digital resource is provided (Wilkinson et al., 2019) (Neal et al., 2020).

3. The FAIRplus FAIR Data Maturity Framework (FDMF) builds on top of the RDA indicators and extends them. The RDA indicators describe each of the four FAIR principles but do not allow a comparable FAIRness score since different combinations of the indicator results can yield an equal score for two different data sets. The FAIRplus FDMF provides a more comparable score. Furthermore, a FAIR assessment through the RDA indicators is subjective to the executing assessor. Some of the RDA indicators descriptions are unclear and not explicit, e.g., “Rich metadata ..." (RDA-F2-01M)(RDA FAIR Data Maturity Model Working Group 2020, 11) or “... other data” (RDA-I3-04M)(RDA FAIR Data Maturity Model Working Group 2020, 12). The FAIRplus FDMF improves this through a more concrete and precise language and alignment to the ISA framework[5]. Moreover, they propose the definition of a business-dependent FAIRification goal to guide the FAIRification process through data usage scenarios. Unfortunately, the ongoing work of FAIRplus FDMF is still in progress and has not yet fully been published (Beyan et al., 2021).

3.2.5 Current FAIRification approaches and FAIR implementations.

Current attempts to render data FAIR are focused on the creation of a Semantic Data Model (SDM) (Jacobsen, Azevedo, et al., 2020) (Vesteghem et al., 2020) (Guizzardi, 2020). SDMs are used to bring data and metadata in a semantic form to make them more explicit. Digital objects or entities are rendered understandable for humans and machines by providing a formal machine-understandable description of data and metadata. They are classified and grouped, and their relations are described.

Public and domain-specific ontologies make the metadata, data, and knowledge relations of SDMs understandable. An ontology is a formal knowledge representation. The simplest form of an ontology can be compared to a vocabulary that describes terms to avoid ambiguities. A more complex form includes dependencies and relations between the terms. Figure 13 shows different knowledge representation possibilities from a controlled vocabulary to an ontology.

[5] https://www.isacommons.org/

A controlled vocabulary is the simplest solution, whereas an ontology represents the highest level of knowledge representation. They are a formal knowledge representation with concepts and classes within a domain. Concepts and classes are connected through relations. Ontologies can be serialized and stored in the OWL format. Single digital data or metadata resources in an ontology are assigned a unique identifier, such as a URI or a DOI. The logic and relations between resources are expressed in triples, referred to as RDFs. An RDF consists of the connection of subject, predicate, and object. In the example "Drug substance is part of the formulation", as depicted in Figure 12. Turtle is used to serialize the RDFs in a machine readable format, e.g., a text-based file.

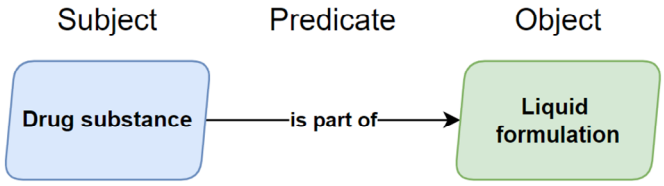


Figure 12 RDF example. An RDF consist of three parts: Subject (blue), Predicate(black arrow) and Object (green). The subject and object can be represented by ontology classes. The predicate connects the subject with the object.

A suitable example for the "Drug substance" class could be the "drug product" class[6] from the Drug Ontology (DrOn) (Hanna et al., 2013). However, other classes from formal ontologies could also be used to represent the "Drug substance" class, e.g., the National Cancer Institute Thesaurus (NCIT) or the BioAssay Ontology (BAO).

The RDFs can be stored in triple stores. The SPARQL Protocol and RDF Query Language (SPARQL) (Pedro Manuel Díaz Ortuño, 2005) (DuCharme & Beijing, 2013), that can be compared to Structured Query Language (SQL) is used to retrieve the RDFs from the triple store. SPARQL is utilized for multiple purposes:

1. Primarily to Extract, Transform, Load (ETL) data or information from RDFs similar to SQL operations on databases and to perform arithmetic calculations.

2. Reasoning: To verify the consistency and correctness of class assignments, e.g., a sample's correct classification as invalid when it exceeded the analytic threshold. Furthermore, reasoning can derive implicit information by inferring rule-based logic to assign new classes and properties. An example demonstrates a transitive relationship and the assignment of an inferred property: all humans are mortal; Spartan warriors can be defined as a subclass of humans. Therefore, reasoning can infer that all Spartans are mortal and assign their respective property.

6 https://ontobee.org/ontology/DRON?iri=http://purl.obolibrary.org/obo/DRON_00000005

Furthermore, RDF, ontologies, and SPARQL are essential parts of semantic web concepts, which aim to make information and data more useful in the World Wide Web and the context of FAIR.

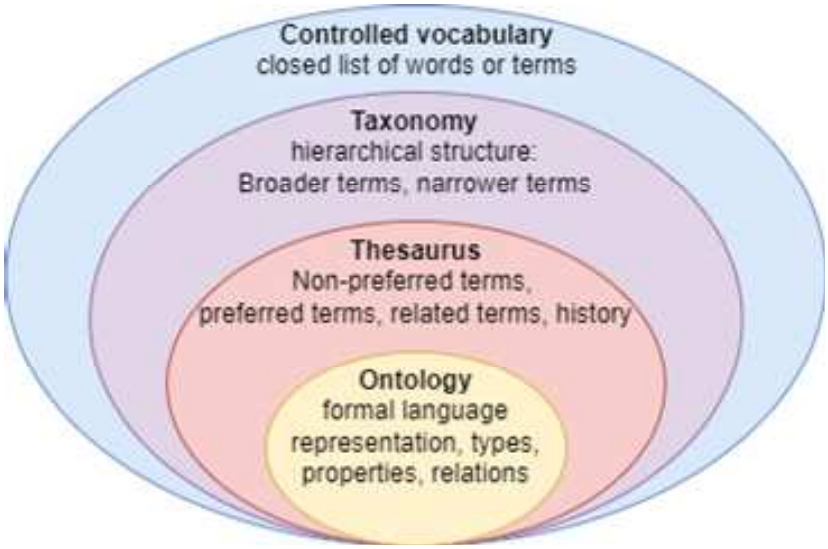


Figure 13 Different classification categories with increasing knowledge representation from outer to inner circle. Figure modified from (Kopácsi et al., 2017).

3.2.6 Method and device-focused data standardization

One attempt that aimed to develop and establish a data standard in the scientific community based on semantic technologies was developed by the Allotrope Foundation[7]. The foundation is a global consortium founded in 2012 by key players from the scientific research industry and device vendors. This enabled the scientific companies to highlight the requirements for standardized output formats to the vendors, which was a promising undertaking. Unfortunately, the many involved members slowed the momentum and development, so more members left the consortium over time. Nevertheless, the Allotrope Foundation proposed practical approaches to improve scientific data management.

A framework of technologies and tools was developed to enable scientific data integration and laboratory data exchange (Millecam et al., 2021). The framework originally consisted of three main parts: Allotrope Data Format (ADF), Allotrope Foundation Ontologies (AFO) (Aloulen et al., 2019), and Allotrope Data Models (ADM). Recently, the framework was extended with the Allotrope Simple Models (ASM) (Haynie et al., 2024). Figure 14 displays all four framework parts. The AFO are well-established ontologies that aid in the contextualization of experimental data, as well as the description thorough terms and their relations within a domain. The ADF provides data standards that rely on the platform-independent file format HDF5. The ADM provides semantic relationship of terms through

[7] https://www.allotrope.org/

tabular or semantic data models. The ASM is a lightweight JavaScript Object Notation (JSON) exchange format.

Some of the Allotrope Foundation's latest projects focus on exporting Chromatography Data Systems (CDS) into the standardized ADF format. Some vendors included the ADF format as a standardized output in their CDSs but later excluded it due to slow development progress and failed acceptance. Nevertheless, ADF could be a potential solution since it combines the semantic description and the data in a Java-based system-independent container format. This option was not further explored due to the currently halted development in this area. However, an Allotrope data model for HPLC exists. Whether the model is suited to store chromatography data will be explored in this thesis at a later point.

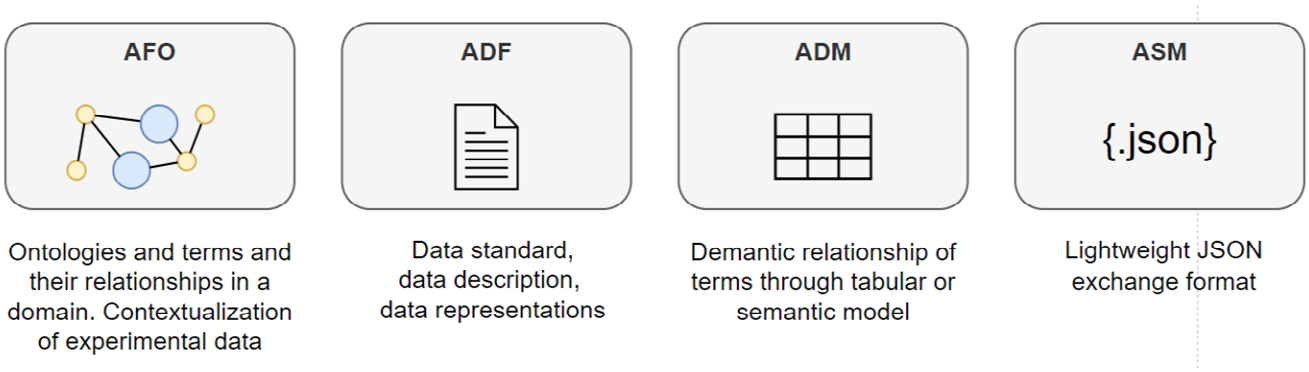


Figure 14 Overview of the Allotrope Foundation framework. The framework consists of four parts: AFO, ADF, ADM, and ASM. ontologies and data models, which are the principal building blocks of data descriptions, data cubes, and data packages (*Driving Innovation with the Allotrope Framework - Astrix*, n.d.).

3.3 Standardization concept for biologics data

Lab automation uses technology and equipment to automate the development of laboratory processes for biologics drug formulation. This automation generates large amounts of data using a high-throughput approach. An SME plans experiments and workflows executed in an automated laboratory setting, as depicted in Figure 15. The data quality is influenced by the workflow/process, the analytical method and derived results, samples and controls, errors, and sample replication. The lab automation generates reproducible experiments and outputs complex biologics data. The SME manually pre-processes this data to answer a specific question or to be reused for ML.

Data standardization concepts such as FAIR or Allotrope use semantic concepts and representations for data and metadata, promoting the reuse of scientific data. These approaches may be used to standardize lab automation data, which may benefit ML and advanced analytics approaches.

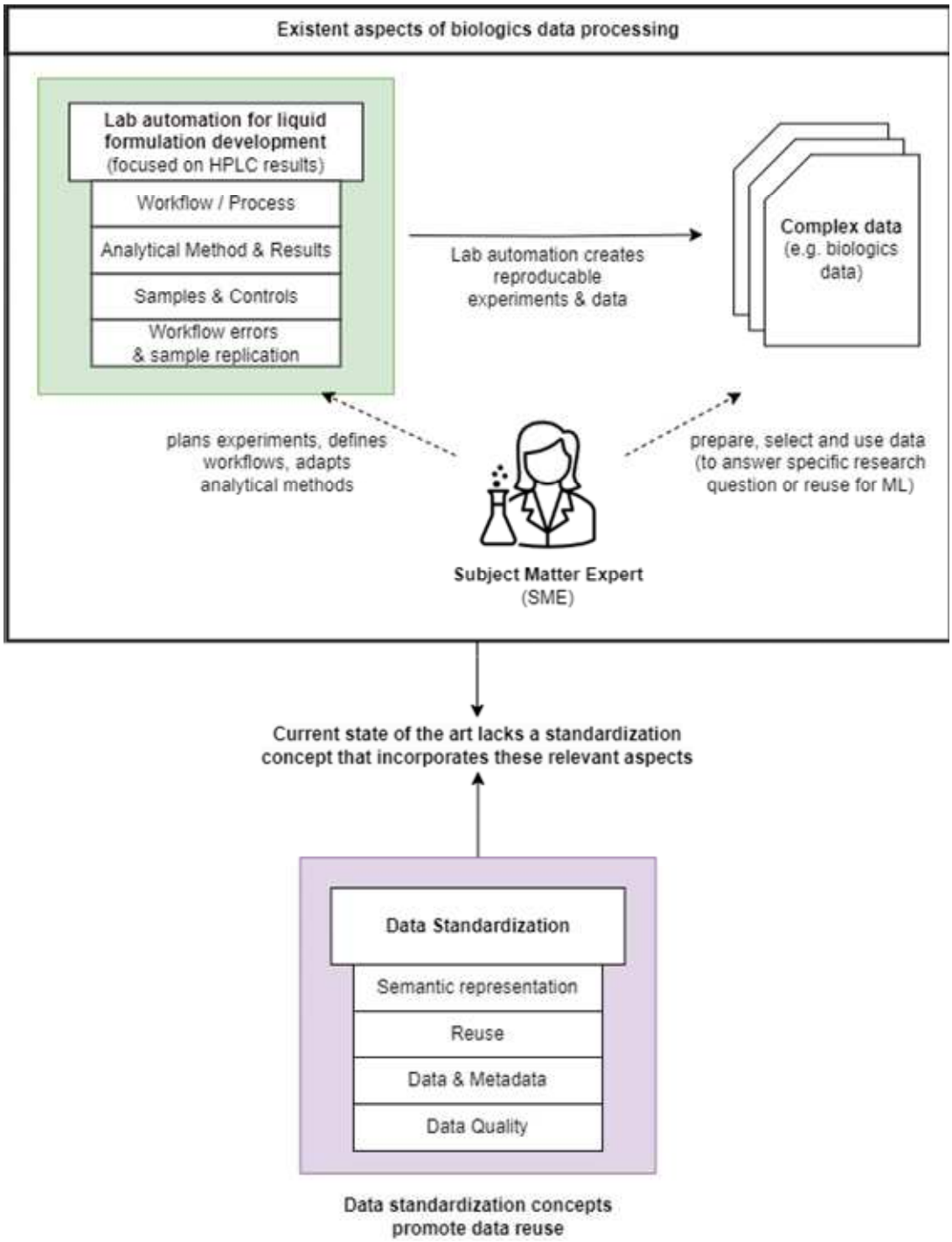


Figure 15 Desired standardization concept. The SME is well-educated in the relevant fields required to understand the data context, such as lab automation, experiments, and analytical methods. Manual data preparation for further reuse is required by the scientist. A data standardization concept that includes all relevant aspects of biologics formulation data is missing but can promote data reuse.

Currently, no standardization concept for biologics data includes all relevant aspects required to make biologics data from lab automation reusable for further automatic

processing through ML. In Chapter 3, several potential solutions were introduced that may aid in developing a biologics standardization concept. FAIR can potentially enable data reuse for humans and machines but has the disadvantage that interpretation is ambiguous and a guide to practical implementation is missing. Allotrope provides a narrow scope that focuses on device and analytical methods. Ontologies and semantics are ideal to represent knowledge, but identifying an ideal ontology is challenging. A choice is to self-build an ontology based on existing concepts from other ontologies.

Some recent publications demonstrate how the introduced concept can enable data reuse for complex data. One example showed how Next Generation Sequencing genomics data was reusable through FAIR and semantic schemas, focusing on essential data based on existing ontologies (van der Velde et al., 2022). Others identified how missing marine image standards increased heterogeneity and prevented objective comparison. They used FAIR Digital Objects to make the marine images reusable (Schoening et al., 2022). Garabedian et al. demonstrated how research tribology has become FAIR using ontologies. This is similar to the scope of this thesis because they included a bigger context, like the processes, equipment, and experiment results, which were required to make complex data reusable and comparable (Garabedian et al., 2022). However, no solution for biologics data includes experimental workflows and analytical results. The question is how can data standardization aid in promoting biologics data reuse so that the machine can automatically reuse the data as defined by the requirements in Chapter 2

The following chapter demonstrates how these standardization concepts can be used to build a standardization concept for biologics data that fulfills the requirements.

4 Methodology

This chapter used FAIR to guide standardization efforts through the FAIRification process. First, a FAIR objective was identified. Second, an assessment of the biologics data sets from Chapter 2 was conducted using an adapted FAIR assessment method. Third, a biologics standardization concept that fulfilled the requirements from Chapter 2.4 was developed. It consists of three parts: (1) a novel and self-developed semantic model based on existing ontologies, (2) qualified aggregated metadata that describe the essential criteria for biologics data set comparability, and (3) the SME comparability logic that was transformed into a flow chart. These three parts enabled the machine to automatically decide on data set comparability.

4.1 Data standardization process using FAIR

The use of FAIR to assess the data standardization level and to enable an increased data reuse is well documented (Garcia et al., 2019) (Chen et al., 2022). Semantic technologies are ideal for making data and metadata more explicit to the machine and thus enabling increased data standardization according to FAIR (Touré et al., 2023).

The FAIRification process, a collaborative effort to transition from raw to FAIR data a FAIRification process was employed (Jacobsen, Kaliyaperumal, et al., 2020) (FAIRplus, n.d.). The process contains multiple steps, starting with the definition of the FAIRification objective (step 1). An example objective could be to increase the semantic of data and metadata with the goal to enable an improved data reuse. The objective can be revisited at any time to assess the current state of FAIRification.

The next step is to analyze the status quo of the data and metadata standardization level (step 2). Here, the data representations (formats) and the data meaning (semantics) are closely examined. A FAIR assessment can be used to derive a numeric score that expresses the current level of standardization according to FAIR metrics (Lin et al., 2022). Possible FAIR assessment options were outlined in Chapter 3.2.4.

The next step involves increasing the data and metadata standardization level, including data and metadata representations, e.g., data semantics. In this step, the semantic and metadata model is defined (step 3). Finding a well-suited model is time-consuming; in some cases, a new model must be created. Ideally, existing semantic concepts from existing ontologies are used (Garabedian et al., 2022) (Matentzoglu et al., 2022).

Next, the data, metadata values, and classes are linked and semantically represented through the semantic model (step 4). They are instanced through knowledge triples in the Resource Description Frameworks (RDFs). For this, the Web Ontology Language (OWL) is used.

In the next step, the semantic data and metadata are hosted in a semantic data store (step 5), enabling access through a web frontend, which allows querying the data and metadata through the SPARQL Protocol and RDF Query Language (SPARQL).

Finally, the standardization efforts are reviewed to determine if the FAIRification objective was successfully achieved (step 6).

The process can be iterated multiple times. Figure 16 displays the FAIRification process, including all six steps. In the following section, the FAIRification process was applied to increase the standardization of biologics data sets. Furthermore, a concept was developed that enabled automated biologics data reuse.

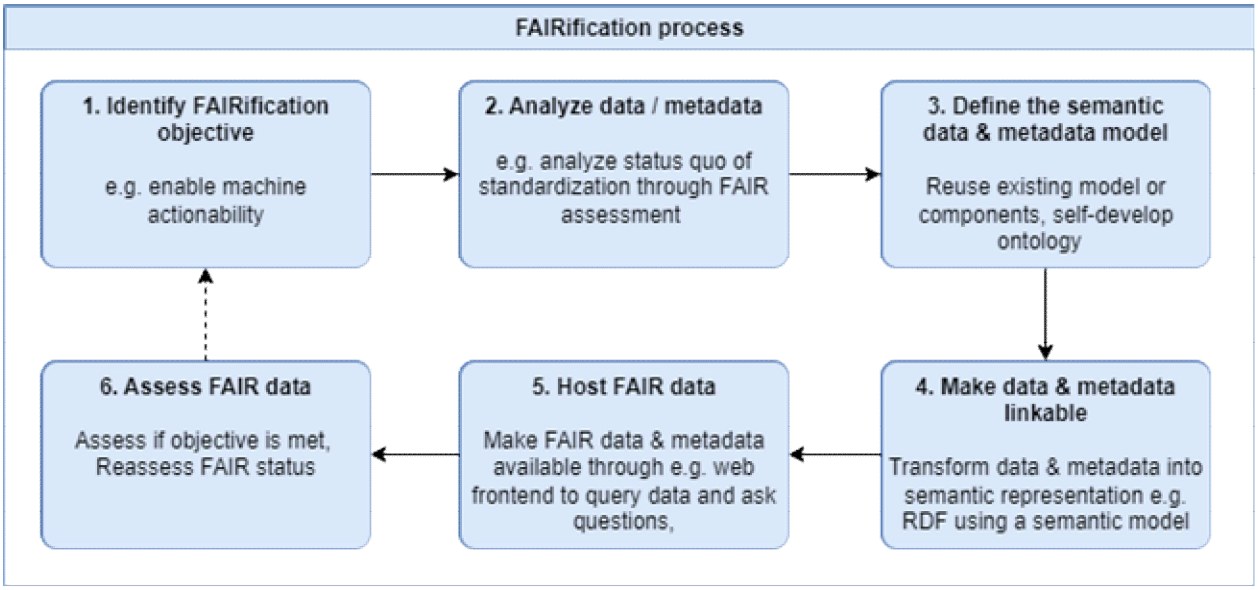


Figure 16 The FAIRification process. The process consists of 6 steps: 1. identification of a FAIRification objective, 2. status quo FAIR data assessment, 3. definition of a semantic data and metadata model, 4. linking of data and metadata, 5. to host FAIR data, and 6. Assessment if the objective is achieved. The figure from Jacobsen, et. al. was adapted and simplified (Jacobsen, Kaliyaperumal, et al., 2020).

4.1.1 Standardization objective

The FAIR principles provide guiding steps to enhance the reuse of scholarly data through good data stewardship. Data and metadata standards are used to make the data more explicit through a semantic expression. The goal is to render the data machine actionable, increasing the data reuse for humans and the machine. The term "machine actionable" is defined as

> "a continuum of possible states wherein a digital object provides increasingly more detailed information to an autonomously acting, computational data explorer" (Wilkinson, 2016,p.3).

When the agent is subjected to an unknown digital object, it has the ability to

> "a) identify the type of object (with respect to both structure and intent), b) determine if it is useful within the context of the agent's current task by interrogating metadata and/or data elements, c) determine if it is usable, with respect to license, consent, or other accessibility or use constraints, and d) take appropriate action, in much the same manner that a human would." (Wilkinson, 2016,p.3).

To guide the standardization efforts of biologics data, machine actionability and thus the machine's capability to automatically reuse biologics data sets was set as the FAIRification objective for the following steps. The hypothesis is that increased data FAIRification is not just a goal but a practical means to achieve machine actionable data, enabling the machine to reuse it automatically. As demonstrated by Wilkinson et al., machine actionability can be improved by semantic technologies (Wilkinson et al., 2017). The idea is to increase the data FAIRness to a point at which the objective is sufficiently completed. Figure 17 displays the relation between the FAIRification score from FAIR assessments that increases by semantics and the accomplishment of the FAIRification objective as machine actionability. Nevertheless, a numeric threshold at which the FAIRification efforts are sufficient to reach machine actionable data is not documented.

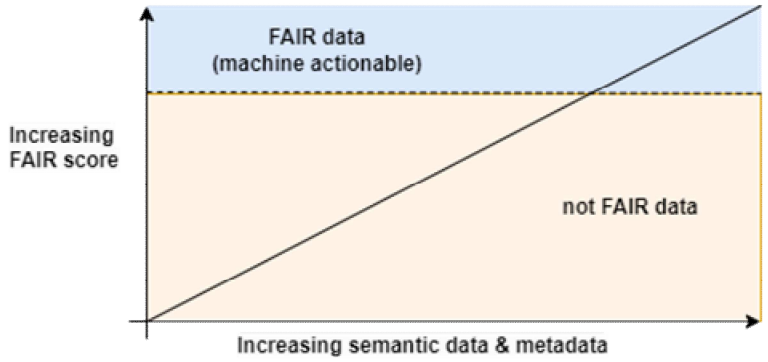


Figure 17 Relation of FAIRification score and machine actionability. The data and metadata semantic improvement increases the overall FAIRification and, thus, the score from FAIR assessments. The FAIR objective of machine actionability may be sufficiently reached at a certain threshold.

4.1.2 Reassessment of standardization of the three data sets using FAIR

4.1.2.1 Selection of a FAIR assessment option

The first option, the assessment through the RDA indicators(section 3.2.4), was the best option for a FAIR assessment since these are well-known and well-established throughout the FAIR community. The second option, the assessment through a web-based automated framework, does not comply with internal policies that prohibit the publication or registration of company-internal biologics pipeline data to an external service provider. Although the third option, the assessment through the FAIRplus FAIR Data Maturity Framework (FDMF), was a promising option, the method was not released when this thesis was conducted. This precluded its utilization, but it may develop into a potential option for a future reassessment after the FDMFs publication.

4.1.2.2 Optimization of the FAIR RDA indicators assessment

The RDA indicators were developed to support the FAIR assessment of a digital object or resource. The authors did not specify the type, format, digital object, or resource required for the assessment method (RDA FAIR Data Maturity Model Working Group, 2020). Data from different sources were used during three biologics data sets. Some of the RDA indicators are

related to the data itself and the data storage system. Therefore, the FAIR assessment, using the RDA indicators, was conducted for the data together with the data storage system. Both were reflected in the pass or fail decision for each indicator. For this thesis, the RDA indicators will be adapted to yield a more detailed FAIR assessment. A sub-score was calculated for each of the four FAIR principles. These sub-scores were calculated by grouping and summing up only the indicator results that belong to the related FAIR principle. Figure 18 shows the sub-score calculation of the findable FAIR attribute for the HTS formulation data use case. Two of the seven possible indicators passed the criteria, indicated by a 1 in the "Assessment Essential" column.

The "non-essential" column did not apply to the findable attribute, indicated through a dash. The essential and non-essential values were summed up in the "Assessment overall" column. The two out of seven possible passes result in a 28.57 % fulfillment of the findable attribute. The FAIR attribute-specific sums were a custom adaption for assessment in this thesis. Before, only a total FAIR score incorporating all FAIR attributes was available. This was achieved by grouping up all the questionnaire items related to one of the FAIR attributes. This enabled a more detailed assertion about each of the FAIR attributes.

The single principle sub-scores also enabled a result comparability for different data sets. After all, 41 indicators were reviewed by an assessor, and an overall sum was calculated by counting the passed (1), failed (0), or not applicable (NA) indicators in the table. The resulting sum was split into a sum for the essential and the non-essential (important and useful) priority groups. Furthermore, the percentage proportion was calculated for each of the three sums (total, essential, and non-essential) by dividing the number of fulfilled indicators (passed) by the total number of possible indicators. Two additional percent proportions were calculated for each of the three sums. These included the indicators that were scored as NA. Since no NA scores occurred during the FAIR assessment of the three biologics data sets, they were irrelevant and were not discussed further. The described procedure was equally adapted for the sub-scores for each FAIR principle. The difference was that only the indicators related to one FAIR principle were counted in each sub-sum. The same applies to the percentage proportion calculation.

					HTS	HTS	HTS
	Sub-principle	ID	Indicator	Priority	Assessment overall	Assessment Essential	Assessment non-essential
Findable	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	0	0	-
Findable	F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	0	0	-
Findable	F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1	1	-
Findable	F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	0	0	-
Findable	F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	1	1	-
				Findable sub sum	2	2	0
Findable subscore					28.57%	28.57%	0.00%
Findable subscore applicable					28.57%	28.57%	0.00%
Findable subscore %NA					0.00%	0.00%	0.00%

Figure 18 Sub-score calculation for the findability FAIR attribute. The FAIR attribute-specific calculation was adapted from the original RDA questionnaire for a more detailed assessment. The remaining three FAIR attributes (accessible, interoperable, and reusable), including their indicators (rows), were hidden to provide a simplistic example.

4.1.2.3 Results of the FAIR RDA indicators assessment

A summary table of the results of each FAIR principle and the average of the four principles (total) is displayed as rows in Table 4. The table includes only the percentage proportion of the essential and the non-essential priority group results for each of the three biologics data sets. Each FAIR attribute score results from multiple questionnaire questions from the RDA assessment. The score for each FAIR attribute is calculated based on the number of passed questionnaire questions. One example for the findability attribute: if four out of ten questions for this attribute are successfully passed, the numeric score is 40%. If, for accessibility, two out of five different questions are passed, the score is equally high, at 40%. Due to this, some scores may be equal between different data sets and sub-scores. The extended tables that included all results were attached in the supplementary material chapter[8].

Table 4 Summary table of the FAIR assessment results for the three biologics data sets in percent. The maximum reachable score for each cell was 100 percent. Each column represents one of the three different data sets. Each row apart from the last describes the sub-scores for the corresponding FAIR attribute. The last row is the average of the four FAIR attributes (rows above).

FAIR principle	**HTS formulation data (data set 1 - greater is better)**	**Company-internal biologics dashboard (data set 2 – greater is better)**	**External pharmaceutical data provider (data set 3 – greater is better)**
Findable	28.57	14.29	14.29
Accessible	41.67	16.67	25.00
Interoperable	0.00	25.00	25.00
Reusable	40.00	30.00	20.00
Total (average of the above)	26.83	21.95	21.95

The HTS formulation data set showed the highest total FAIR assessment score, indicated by the percent proportion. A closer examination of the FAIR principle's sub-scores and the related RDA indicators allowed for a better understanding of the results. Figures 39, 40, and 41 display a detailed overview of the assessment results for the HTS formulation (data set 1), Biologics dashboard (data set 2), and PharmaCircle (data set 3) data sets.

The HTS formulation data set (data set 1) data was stored **findable** in the Hadoop file system. The data could be filtered down and received through SQL queries in the Hadoop web front end or a self-developed Python GUI (HTS-Studio). Identifiers such as the compound or conducted screening name could be used to find data. The multiple options to access the data benefited the **findability**. The indicator *RDA-F4-01M Metadata is offered in such a way that it can be harvested and indexed*: here, the HTS formulation data set benefited from storage in Hadoop since all metadata were indexed and searchable through the system. The

[8] **Fehler! Verweisquelle konnte nicht gefunden werden.**, **Fehler! Verweisquelle ko nnte nicht gefunden werden.**, **Fehler! Verweisquelle konnte nicht gefunden werden.**

BiologicsDashboard data set and PharmaCircle data set were presented on a web front end that was not as flexible in searching and finding data as the HTS formulation data set.

The assessment of the **accessibility** showed a similar result. The BiologicsDashboard data set (data set 2) profited from *RDA-A1-01M Metadata contains information to enable the user to get access to the data* because the dashboard contained metadata that related to the scientist that conducted the experiment and additional cross-references to other data storage systems that lead to the experimental raw data. The PharmaCircle data (data set 3) set did not fulfill this criterion since it was unclear where the data originated. Occasionally, the originating filing document was stated, but no direct link was provided. Therefore, it was unclear if the user could find and access the data through a filing authority like the FDA. The HTS formulation data set scored higher for the indicators *RDA-A1-02M Metadata can be accessed manually (i.e., with human intervention)* and *RDA-A1-02D Data can be accessed manually (i.e., with human intervention)* because metadata and data could be downloaded through the HTS Studio. In the Biologics Dashboard, data cannot be accessed easily through Spotfire. PharmaCircle data could only be accessed by storing the entire web front-end page that contains the data. The indicators, *RDA-A1-04M Metadata, is accessed through standardized protocol*, and *RDA-A1-04D Data is accessible through standardized protocol,* were fulfilled since the HTS formulation data set and the PharmaCircle data set web front ends both relied on HTML. The BiologicsDashboard, in contrast, was stored in a Spotfire library from which data or metadata could not be easily accessed using a standardized protocol.

In the case of an HTML presentation, the data could at least be accessed by web scraping the page and retrieving it. Interestingly, the HTS formulation data set scored the lowest for **interoperability**, while the other two scored better. *RDA-I3-01M Metadata includes references to other metadata*, *RDA-I3-01D Data includes references to other data*, and *RDA-I3-02M Metadata includes references to other data*, revealing that the BiologicsDashboard data set and PharmaCircle data set both fulfilled this indicator because the metadata and data included cross-references to other systems or documents and the BiologicsDashboard data set often included the name of a contact person. The indicators were fulfilled since references were available (without covering data access for the user), although RDA-A1-01M was not fulfilled.

The HTS formulation data set scored highest in **reusability**. *RDA-R1-01M Plurality of accurate and relevant attributes provided to allow reuse* was only fulfilled by the HTS formulation data set because the data included enough metadata and data that enabled further reuse. Control data and intermediate calculation steps were included in the data. The other data sets did not include such data. *RDA-R1.3-01D Data complies with a community standard* and is fulfilled by the HTS formulation data set and the BiologicsDashboard data set, but not from the PharmaCircle data set, since some protein sequences were merely available in a picture format instead of a regular sequence of amino acids as text.

4.1.3 Optimization of FAIR assessment score

The question is: How can the data set FAIRness be improved to generate a higher FAIR score and thus yield better results for advanced analytics training? What actions are required to improve the score according to the FAIR principles? The RDA indicators and the 14 FAIR

guiding principles (Figure 11) help to guide an improvement and to identify actionable means to achieve this. Following the hypothesis from the FAIRification objective (4.1.1), increasing the semantic representation will increase the standardization and, thus, the machine actionability of the data (Wilkinson et al., 2017).

The overall FAIRness can be improved by rendering the metadata and data more explicit through an SDM. More precisely, semantic web technology tools render the data and metadata understandable and accessible to humans and machines. The HTS formulation data set was selected to demonstrate how the overall FAIRness could be improved by enhancing each FAIR principle.

The findability could be improved using registered identifiers that allow metadata and data discovery (Wilkinson et al., 2017). The OBI ontology could be used to describe the well plate class that contains the liquid formulations. The class code OBI_0400076 directly links to the description of the term in the OBI ontology and is defined as "...A multi-well plate is a vessel that can deliver multiple samples... ". An additional class example could be the subjection of the well plate to temperature stress to induce protein degradation. The "temperature" class from the AFO AFR_0001584" is well-suited for this example. The object property to relate both classes may be "participates in" (RO_0000056).

All in all, the example describes how a multi-well plate participates in temperature stress. The connection of the two classes with the object property forms an RDF. Serializing digital objects as RDFs and storing them in a database would improve accessibility since these databases use web-based standardized communication protocols to enable data access. An additional way to access the data could be through SPARQL. The digital objects in the form of RDF instances could be stored in RDFlib in memory. A more persistent solution would be stored in a graphical database, e.g., Blazegraph. Table 5 displays the potential RDF example. A second RDF (table) would be required to specify the "time unit" value as 21.

Table 5 Example RDF structure and components for a multi-well plate subjected to temperature stress conditions. The stress duration is specified as storage unit as time unit in days.

RDF element	Ontology class or property	Example value	Ontology reference code
Subject	Multi-well plate	96wellplate1	OBI_0400076 [9]
Predicate	participates in	participates in	RO_0000056 [10]
Object	Temperature	40°C temperature stress	AFR_0001584 [11]

The potential implementation of these steps would increase the standardization of the HTS formulation data set to a high machine-understandable level (machine actionable) according to FAIR, resulting in a new estimated FAIRness score of approximately 90%, compared to the previous assessment. While a semantic representation would significantly increase the FAIRness, the comparability of data would still pose a challenge for machines as it is not within the scope of FAIR at this point. Therefore, there is no direct relationship between an increased FAIR score and data comparability. Furthermore, machine actionability was understood in terms of data findability and not in the sense that the machine should be able to determine data set comparability. For the latter, essential factors for comparability were missing. These are explained in the following section.

4.1.3.1 Machine actionability of the method-based Allotrope HPLC model

The Allotrope Foundation is an international organization composed of different pharmaceutical companies and hardware and software vendors. They develop scientific data standardization solutions, e.g., frameworks, data formats, data models, and semantic concepts to support scientific data standardization efforts. One of their solutions is a high-performance liquid chromatography (HPLC) model (Millecam et al., 2021). It is a representative model representing a metadata-based expression of a semantic domain model, more precisely, of an analytical method. Figure 19 shows a simplified version of the original Allotrope HPLC ontology model. The original model contains a complete description of the UPLC system, including all the main device groups: device, material (sample), result, and process (HPLC run). More subordinate concepts, classes or properties are autosampler, detector, column, pump, sequence of injections, temperature control unit, mobile phase, measurement process (sequence), and analytical method. The recently introduced ASM provide a sample JSON file (Allotrope-HPLC-SampleJSON, n.d.) and a JSON schema of the model (Allotrope-HPLC-JSON-Schema, n.d.). However, due to its recent introduction, the ASM JSONs were not available during the conception of this thesis and were not taken into

[9] http://purl.obolibrary.org/obo/OBI_0400076 (visited on 16/12/2023)

[10] http://purl.obolibrary.org/obo/RO_0000056 (visited on 16/12/2023)

[11] http://purl.allotrope.org/ontologies/result#AFR_0001584 (visited on 16/12/2023)

consideration. Therefore, the term Allotrope HPLC model in this thesis refers the model presented by (Oberkampf, 2018).

In the model, the process describes the sequence (the order) in which the samples are processed during an HPLC run. It does not refer to the automation workflow/process (sample history) of an HTS screening, which is one important aspect of data set comparison.

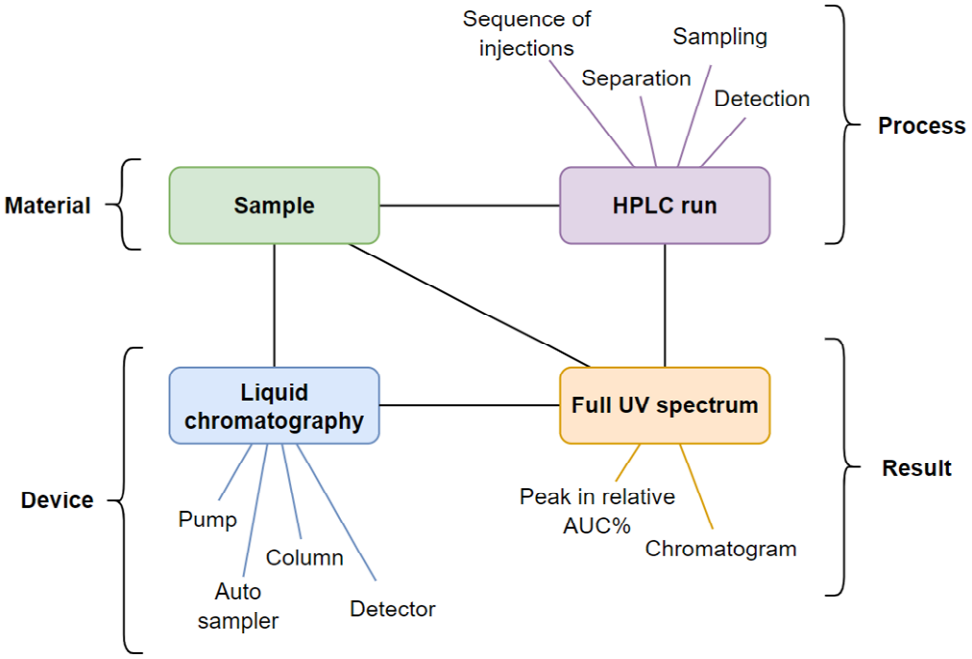


Figure 19 Allotrope HPLC ontology model. The model is split into the categories: material, device, process, and result. Each category has specific classes such as: sample, liquid chromatography, HPLC run and Full UV spectrum. Other classes or properties can be attached to the classes. The figure is a simplification of the original model and was adapted from (Oberkampf, 2018) [12].

The Allotrope HPLC model semantically describes the device (hardware and all its components) and represents the analytical results in peak form (areas under the curve in percent). Its primary focus is on all the information necessary to describe the method and all parts of the device in detail, making it a method-based model. Leveraging established ontologies and semantic web concepts based on metadata, the model provides a comprehensive understanding of the device and the analytical results it produces.

Figure 20 indicates the additional semantic benefit of sophisticated knowledge representations such as ontologies. The use of such concepts leads to an increased FAIRification, as indicated by the increased FAIR score from the HTS formulation data set. Following the FAIR principle, a high FAIR score ultimately leads to machine actionable data,

[12] ADM catalog: https://allotrope.gitlab.io/adm-patterns/

which means that the machine can understand and further process the data, e.g., for advanced analytics. This can be the case for simple analytical methods such as pH measurement, where only information (metadata) is required for the machine to understand the data and invoke comparable analytical results and thus reach a machine actionable state. This is not the case for biological analytical results derived from a complex analytical method such as chromatography.

As described in 2.3.1, data comparability for complex biologics data is challenging due to multiple reasons: data comes from different sources, data is differently formatted, differing analytical method accuracy (errors), integration requires a great semantic context, e.g., sample preparation, workflow, and the compound processing are important. Biologics data is especially complex due to mAbs complex structure and function. These compounds are challenging to characterize, and the analytical methods are prone to noise. Here, an increased FAIR score does not automatically lead to machine actionable data in the context that the machine understands the data and can compare different results.

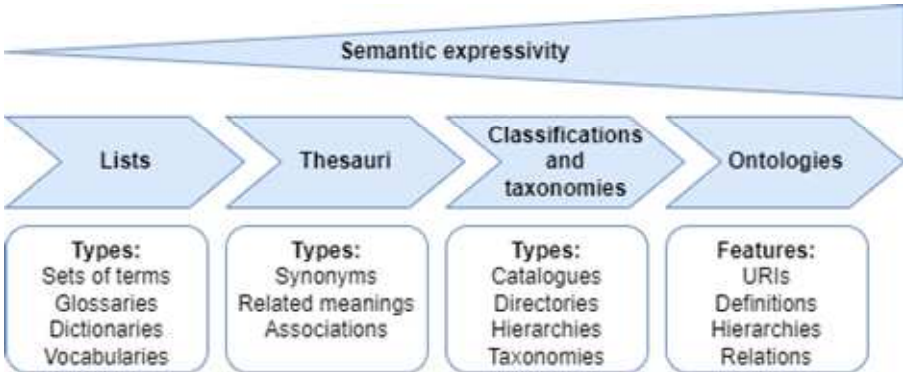


Figure 20 Different concepts and levels to increase the semantic expressivity. The semantic expressivity increases from lists to ontologies. Figure adapted from (Harrow et al., 2019).

4.1.3.2 Data comparability enabling factors

Multiple factors influence data set comparability from liquid chromatography results. One factor is the increasing artificiality of proteins and, thus, the requirement to adapt analytical methods. Therapeutic mAbs evolve in complexity over time (i.e., differing antibody species or highly engineered molecules), and the analytical methods must be adapted to fit the protein's characteristics (Chirino & Mire-Sluis, 2004). In the case of liquid chromatography, standard platform methods are developed that support a broad range of different proteins while ensuring the generation of comparable results for the differing compounds, as described in 3.1.1.2. These methods benefit analytical standardization between different mAbs and should, therefore, be included in the comparability of data sets.

Not only is the method to be adapted, but it also depends on the protein's characteristics and project constraints, so the laboratory workflow requires adaptation. In drug development, the preparation, processing, application of stress conditions, and workflow that a sample is subjected to influence significantly the analytic readouts. For example, the liquid formulation's viscosity depends on the protein's characteristics. The viscosity of increased protein concentrations during formulation development can go from a watery solution to a

viscous gel. A highly-viscous solution may require the adaption of the analytical methods or the workflow, e.g., the adaptation of automated pipetting during well plate handling. Uncarefully selected pipetting ejection forces can subject the solution to additional undesired sheer stress. In the continuing laboratory processing these viscous solutions can clog and damage a chromatography tubes and column. As a result, the workflow must be adapted to exclude these samples from further processing. The workflow and the order of orchestration of different stresses play an important role in determining analytical result comparability. If the workflow is substantially changed, e.g., new stresses are introduced, only the data from stresses that were part of previous screenings can be compared. If errors occur during the workflow, the stress can be incomparable to other screenings.

In conclusion, as long as the workflow remains unchanged and is precisely repeated, the results are comparable between different stability screenings. As described in 3.1.1.3, well-characterized positive and negative controls can be used to ensure that stress conditions are correctly applied. The controls are equally subjected to all stresses as the samples. If the analytical readout of the controls, i.e., at accelerated temperatures, is outside of the expected performance, the stress conditions were imprecise. This means that also the samples were falsely stressed.

Additional external conditions in the laboratory, like humidity, light, and room temperature, influence the molecule, workflow, and thus analytical results over time. Stability screenings during drug development may take up to several months. Analytical method performance varies over time. Increased throughput and, therefore, increased device wear down may add to the problem. System suitability tests ensure proper performance of the analytical devices and thus must be included in comparability decisions.

Overall, the analytical error during stability screenings with platform-based methods must be considered. This includes the variation of the analytical method and the influence from the workflow. The use of sample replication may allow for a more precise estimation of analytical errors and is thus an important factor for result comparability.

The question is whether these factors that influence analytical biologics performance and comparability are sufficiently covered in a metadata-based approach like the Allotrope HPLC model. Moreover, is the current FAIR approach to semantically expressing a domain (or complex analytical method such as liquid chromatography) with metadata sufficient to fulfill the machine actionable requirement?

4.2 Redefinition of the standardization objective: machine actionability

The following section contains four tasks that demonstrate the ambiguity of the term machine actionability. The definition and the four ability examples (a, b, c, and d – see 4.1.1) result in no clear understanding of "machine actionability" and when it is achieved. Although the definition describes that a digital object should provide increased information to the data explorer, it is not defined to which degree this is required. The four abilities concertize that the machine must be able to access and understand the object (a, c). The access is closely linked to the use intent (current task) of the agent (a, b).

Furthermore, the agent should be able to augment human behavior and take actions according to (d). This closely connects and conditionalizes the requirements to achieve machine actionability to the fulfillment of a human task (intent). For further illustration, three possible tasks based on different data use intent are provided:

Task 1:
A data crawler that scans for new digital objects to index them for a search engine.

Task 2:
An agent that scans experimental data in a dashboard to check the data availability and integrity.

Task 3:
An agent that gathers analytical results from a simple analytical method such as pH measurements to prepare them for a statistical evaluation to answer multiple scientific questions.

Task 4:
An agent that collects suitable and comparable biologics data sets (e.g., SEC data) as a preparational step to reuse data for advanced analytics.

Each of the four tasks is a valid human-like action and, therefore, should theoretically be achievable through FAIR data, which should lead to machine actionability. Each task puts a different emphasis on the FAIR attributes and demonstrates a slightly different meaning of machine actionability.

In Task 1, a data crawler looks for new objects for indexing purposes. Therefore, the file name and/or the file content is scanned to provide content-based rich metadata for the search engine. Machine actionability in this task is understood as data findability.

Task 2 is similar to Task 1, focusing on accessibility. The agent must be authorized to access the data and check the data integrity by calculating and comparing hash values.

Task 3 extends the requirements from Task 2 because the data needs to be interoperable and ready for further statistical evaluation through the machine. It is assumed that each pH device is calibrated, and each analytical result per se is comparable. In this case, the machine does not need to understand the scientific context of how the pH measurements were conducted. Therefore, the data complexity can be categorized as simple data, which makes it possible to compare the data without additional effort.

Task 4 further extends the requirements to reach machine actionable data. Compared to the pH measurement from Task 3, the agent in Task 4 needs to find, access, and reuse data from a complex analytical method, such as results from liquid chromatography data. In this case, data reuse is impossible for the machine because simple calibration information is insufficient for such a method. Additional data and metadata are required to integrate different data sets and determine the comparability. Similar to a human scientist (Chromatography SME), the agent must understand which analytical chromatography results are comparable.

Task 4 is similar to the increased FAIR score of the HTS formulation data set, which achieves reusable data for advanced analytics. Both Task 4 and the HTS formulation resemble complex biologics data sets, which require a machine actionable state to fulfill the task. Furthermore, this adds an automation layer to the comparability process, resulting in automated data comparability. For the agent to achieve this, a certain functional capability is required to distinguish and actively decide on the data set quality and what data sets may ideally be reused.

The four tasks demonstrate the ambiguity of the machine actionable definition originating from the FAIR guiding principles. Machine actionability is not a single state but rather a range of states that depend on the task. Task complexity can range from simplistic findability to comparability of complex analytical methods.

The current perception of FAIR to reach a machine actionable state is to achieve a detailed semantic expression of the data, including their domain, in this case, of the chromatography method. It is presumed that a high FAIR score is equivalent to machine actionable data as defined by the FAIR principles. This raises two questions: The first question is about the "machine actionable" definition. The second question is if the current steps to increase the FAIRness are sufficient to reach the goal of machine actionable data, which includes comparability.

4.2.1 Automated data comparability

The specification of the actual purpose of reuse through the machine (automatic comparability) is beyond the initial definition and understanding of FAIR. FAIR does not specify the purpose of reuse. In the case of comparability, the machine is required to augment the capability of a scientist to decide on different data set aspects. The specification of the reuse and to set it as the FAIRification objective exceeds current requirements for a solely semantic expression and the capability of the machine for complex biologics data. A good example is the FAIRplus maturity levels, which were introduced to distinguish between different semantic data set maturities. Figure 21 depicts the initial five maturity levels from 0 to 5, where 5 is the highest level. Potential data reuse improves with every maturity level because the data increases semantically, thus making it increasingly interpretable for the machine. Therefore, the machine should be able to reuse the data. However, the actual reuse is not specified. The comparability of data sets is not explicitly covered.

For simple data, e.g., finance data (see 2.3), a reuse specification is unnecessary because FAIR renders the data intrinsically comparable. All semantic available metadata is sufficient to compare different data sets. Therefore, no additional effort is required to determine how different data sets relate to one another.

Complex data from analytical methods such as liquid chromatography, on the other hand, require a larger semantically available context to achieve data set comparability. A sixth maturity level (comparable data) was added. This level includes not only semantic data and metadata but also the required factors that influence biologics data set comparability, including metadata that describes data set quality factors. These influencing factors also referred to as qualified aggregated metadata, can be classified as level 3 (standardized data) since they fulfill community standards for liquid chromatography. In addition, some sort of

agent is required to actively query the relevant metadata to decide on the data set quality and the suitability for reuse. An additional seventh level was introduced, enabling the machine to automatically compare data sets by augmenting the scientist's decision. By fulfilling levels 0 to 5, the data reaches a FAIR state that enables intrinsic data reuse, which is sufficient for simple data. Levels 6 and 7 actively specify the purpose of data reuse by automatically comparing data sets through the machine, which is beyond the scope of FAIR.

As a reference, the biologics use cases in their initial state and their low FAIR score might be ranked between levels 0 and 2 (see 4.1.2.3). Since they are semantically well-described through community standards, increasing their semantics could potentially improve their maturity to level 3 or 4. Nevertheless, as pointed out, real comparability was not achievable through a method-based Allotrope chromatography model.

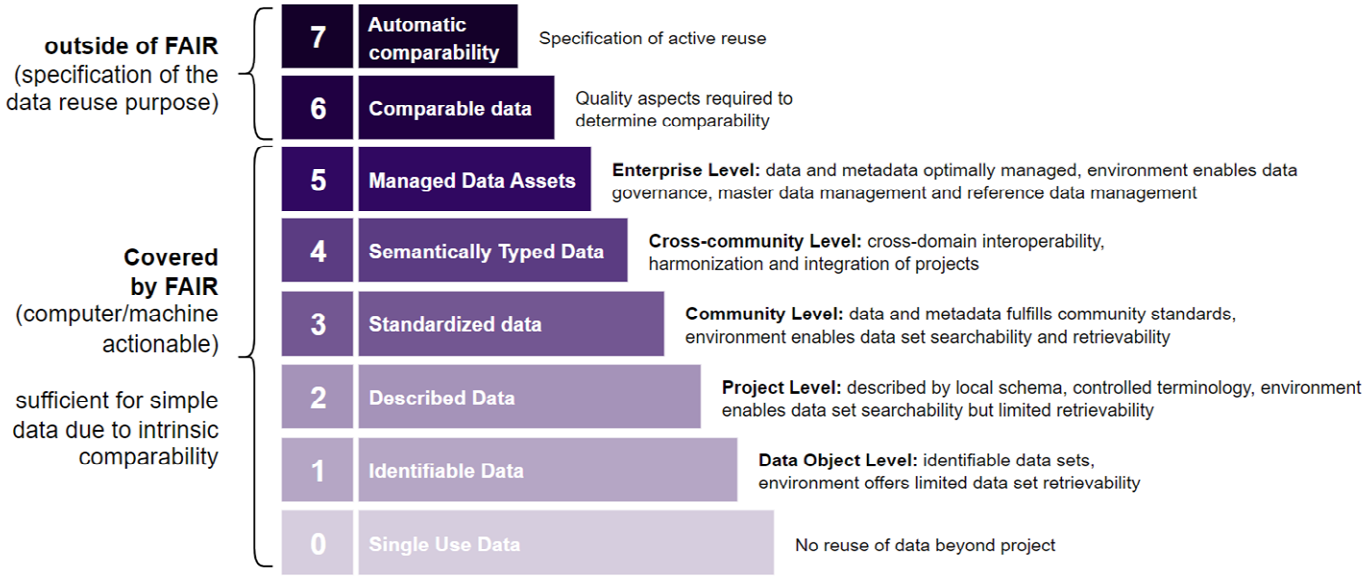


Figure 21 Adaptation of the FAIRplus data maturity levels (DSM). The original FAIRplus DSM consists of five levels (level 0 to 5). The levels are arranged in increasing complexity and requirements from bottom to top. Two additional levels (6 and 7) were introduced to align with the new FAIRification objective of automatic comparability. The original figure from FAIRplus (*Overview | FAIRplus Data Maturity*, n.d.) *was adapted and modified.*

4.2.2 Challenges and concept proposition

The automated comparability as the FAIR objective raises challenges for the previously introduced semantic model and the overall approach. Three challenges were identified after the adapted FAIRification objective. The interpretation of machine actionability from the FAIR definition is ambiguous since it depends on the machine's task. Increasing data and metadata through semantics is sufficient if machine actionability in the sense of findability is the FAIRification objective. Enabling the machine to decide on data set reuse requires comparable data. As a consequence, a reiteration and adaption of parts of the FAIRification process were required. The new FAIR objective triggered the requirement for the following three steps (creation of a self-developed semantic model, introduction of qualified

aggregated metadata, and automation of the SME comparability decision). The following challenges hinder the machine's comparability capability:

1. The machine cannot know how metadata differences of the chromatography components, e.g., increased pump pressures or longer injection times, influence the analytical result comparability. Two cases exemplify this finding: 1. The method-based metadata for two data sets are equal, but the chromatography peaks (result) look different. This can be caused by changes in the workflow/process or through human influence during experimental execution. 2. The chromatography peaks (result) look similar, although the metadata and method parameters are different. The machine cannot evaluate how the metadata differences in the method influence the overall result comparability. A pure method- and metadata-based semantic description of a complex analytical method, e.g., chromatography data, is insufficient to reach machine actionable data, following the comparability goal. This counts for a human- and machine-based evaluation.

2. The capture of all method-based metadata that is required to describe the chromatography device and the peaks (result) will be an unfeasible and time-consuming task. The fewest number of laboratories run fully autonomously with the capability to automatically capture all required metadata. For the ones running in a semi-automatic mode, the probability is high to rely on manual metadata capture through scientists.

3. Factors influencing biologics chromatography result in comparability, as described in 4.1.2.1, which are missing in the Allotrope HPLC model. Some process information in the model refers to how the chromatography method was executed, but laboratory workflow, e.g., how the biologics sample was treated and how the temperature stresses were executed, are not contained in the model. This information is essential since it substantially influences the comparability of the analytical results.

The three challenges could be overcome if the machine is provided with three solutions: 1. The human expertise (chromatography SME) on how and what metadata changes benefit or hinder the analytical comparability decision 2. Access to all relevant metadata on a higher abstracted (qualified aggregated metadata) level 3. Access to additional comparability influencing factors regarding error, workflow, method, and SST. The challenges and the proposed results are displayed in Figure 22. The proposed solutions were implemented in a result-based standardization concept that used the augmented, formalized, and automated SME comparability decision logic together with a semantic model that included all comparability-required qualified aggregated metadata. The previous goal and solutions of machine actionability focused solely on a semantic description of method-based metadata. The redefined comparability objective aims to make results from biologics comparable and is thus referred to as result based.

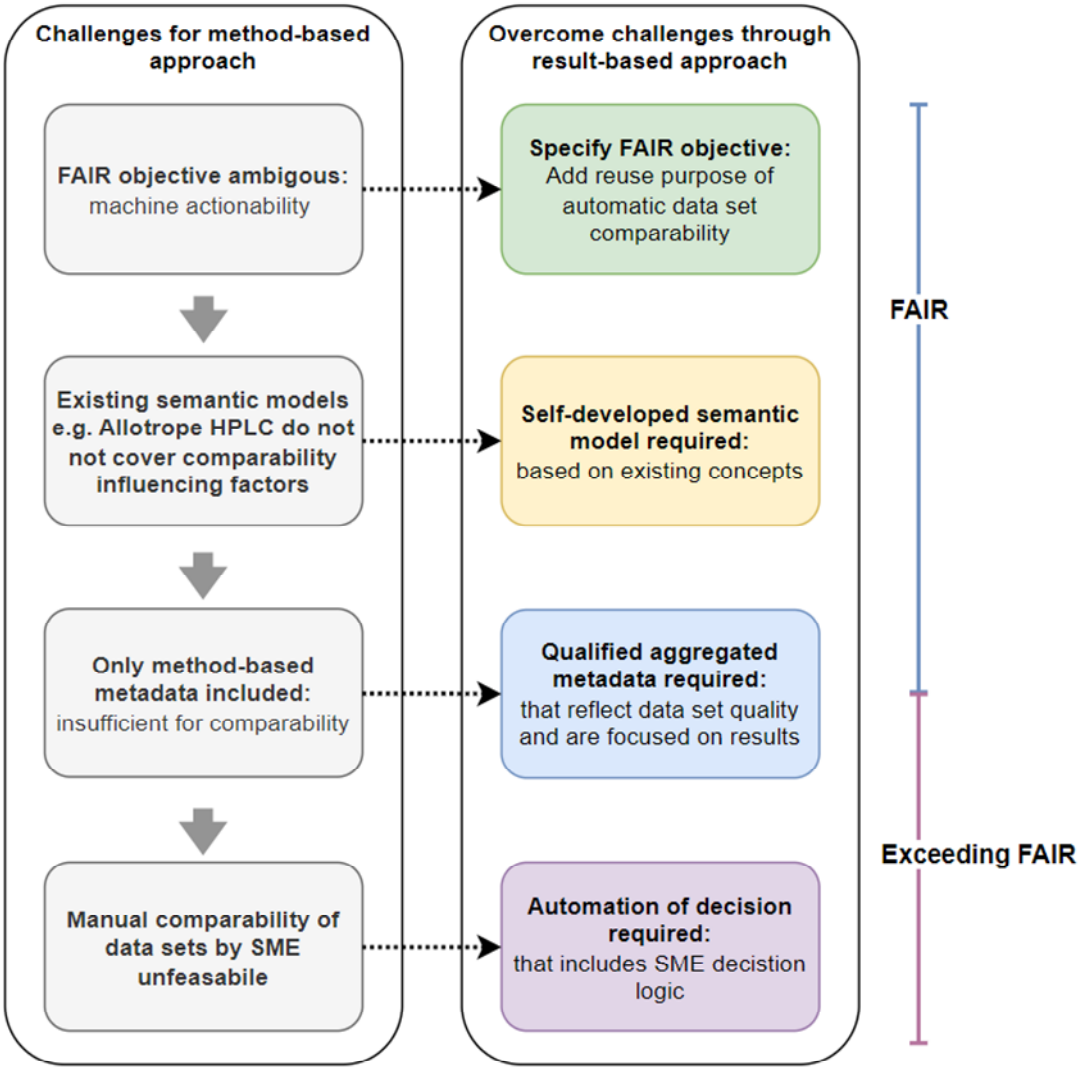


Figure 22 Challenges for the method-based approach and potential solutions. The redefinition of the FAIRification objective for an automated machine comparable approach raised three challenges (left side). Therefore, three suggestions to overcome the challenges in a result-based concept were proposed (bottom). The three suggested solutions are the creation of a self-developed semantic model, the introduction of qualified aggregated metadata, and the automation of the comparability decision.

4.3 Result-based standardization concept

This section describes the conceptualized result-based standardization required to achieve comparability. Current semantic models, such as the Allotrope HPLC model, predominantly focus on the reproducibility of experiments related to the chromatography device rather than on result comparability and data reuse. Consequently, this model is only partially suitable for the intended purpose of enabling machine actionable data. To overcome this limitation, relevant concepts and terms are extracted from these existing models and integrated into a self-developed semantic model. Together with qualified aggregated metadata, these allow for comparable data sets and automate decision-making. These three parts form a novel standardization concept for biologics data. The concept consists of three parts, as displayed in Figure 23.

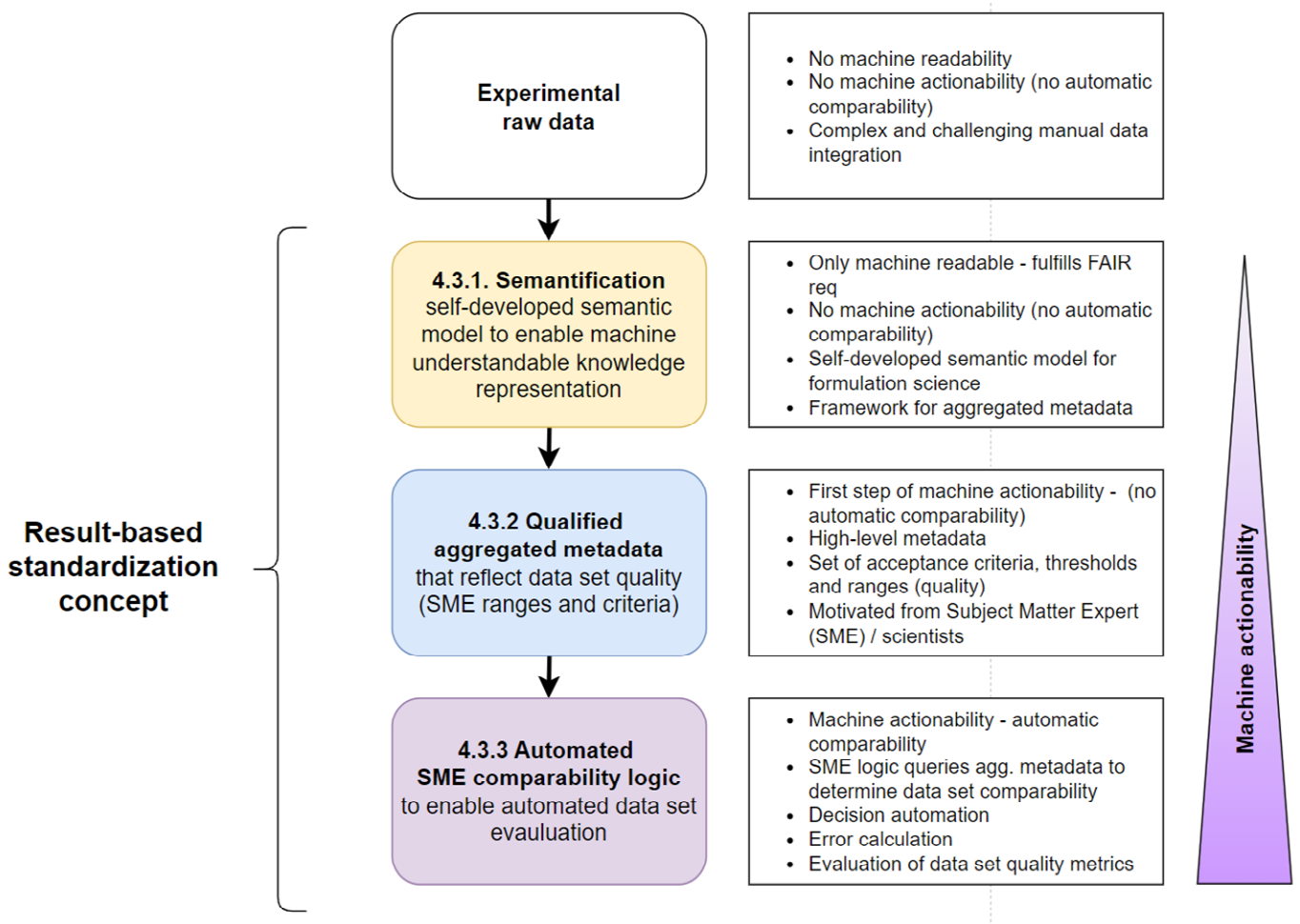


Figure 23 The Result-based standardization concept consists of three parts: 1. A self-developed semantic model 2. The qualified aggregated metadata, and 3. The SME comparability logic. The qualified aggregated metadata represent quality aspects through ranges and criteria of data sets, which are expressed in a semantic machine-understandable knowledge representation (semantic model). The scientist's logic to determine data set comparability is encoded into a flow chart. The logic automatically queries the semantic data and derives comparability decisions.

Experimental raw data is the input for the result-based standardization approach. The data consists of stability screening formulation data, including analytical results from HPLC (SEC) measurements. At this stage, the experimental raw data can be manually integrated and interpreted only with high effort by domain experts (scientists) familiar with the specific experiment setup. Moreover, they require knowledge about the requirements for integrating and comparing the raw data across different stability screenings, which include analytical chromatography data.

Semantification: It is necessary to transform tabular data into knowledge triples using web technology concepts such as ontologies to achieve machine readability. As there is currently no semantic model fully representing the formulation sciences, a new model for formulation sciences represented by stability screenings data was developed. This model incorporates analytical chromatography and is based on existing ontologies. Each sample and control, along with all relevant data, including formulation composition, stress conditions, workflow, process, and replicate information, is translated into knowledge triples during this process. The semantification allows for machine readable data, making the data meaningful and interpretable for machines. This satisfies FAIR requirements and classifies the data as reusable. However, it cannot yet be classified as machine actionable data, as defined in this thesis, since machines cannot automatically compare differently processed data sets, which is the main goal of the thesis. Raw data is transformed into a machine readable semantic format, giving the data meaning to humans and machines.

Qualified aggregated metadata are a set of rule-based acceptance criteria, thresholds, and ranges that determine data set quality. They are based on scientists' expertise and knowledge. One example is the expected performance range (stability) for the controls at a specific stress condition, e.g., 40°C 21 days (see 4.3.2.2). The control performance directly influences the samples and is used to validate or invalidate sample quality; therefore, it directly plays a key role during comparability assessment.

The qualified aggregated metadata represents the initial step beyond machine readability toward achieving machine actionability. These metadata are combined at a higher level and depend on lower-level metadata from the raw data set, such as formulation composition, analytical method, stress conditions, and experiment processing. Unlike metadata from the Allotrope HPLC model, these qualified aggregated metadata focus on the result comparability and what is required to achieve it. In comparison, the Allotrope HPLC model primarily focuses on achieving experiment reproducibility by modeling the analytical device (chromatography system) rather than the factors necessary for result comparability.

Qualified aggregated metadata represents the first step towards machine actionable data, although automation is not yet included. It allows for determining data quality criteria as a foundational framework for decision automation and data set evaluation.

SME comparability logic: The scientist's decision logic is formalized and automated as a comparability logic in a flow chart. This automation replicates how a scientist would manually assess the quality of a data set (qualified aggregated metadata), considering factors

such as formulation, analytical method, and process (including controls). Through automatic evaluation of the qualified aggregated metadata, data set quality can be assessed without human intervention. This is made possible by querying the qualified aggregated metadata in the semantic model. This automated evaluation outputs a comparability classification for each stability formulation screening (data set). This enables the identification of well-suited data sets for a later ML approach, which is fundamental and was not possible previously. The comparability is categorized into four groups: high comparability (green), medium comparability (yellow), low comparability (purple), and no comparability (red).

This approach enables the comparison of samples, stress conditions, and complete evaluations of HPLC (SEC) results across differently processed screenings. It goes beyond FAIR principles, as the data is not only machine readable but also machine actionable. The automation allows for fully automated machine actionability, as the machine can automatically evaluate data sets. The semantic representation facilitates interoperability beyond company boundaries.

4.3.1 Data semantification using a self-developed semantic model

This chapter describes the second part of the result-based FAIR concept: Encoding the qualified aggregated metadata (subtrees) into a machine understandable and accessible solution (semantic representation). First, the development process and RDF storage options are explained. Then, the relation to the OBI ontology is outlined, as well as to other ontologies that were used to build the model. Last, the data extraction through SPARQL and quality verification through reasoning and the storage in the triplestore is demonstrated. Figure 24 depicts how the semantic model fits into the overall standardization concept.

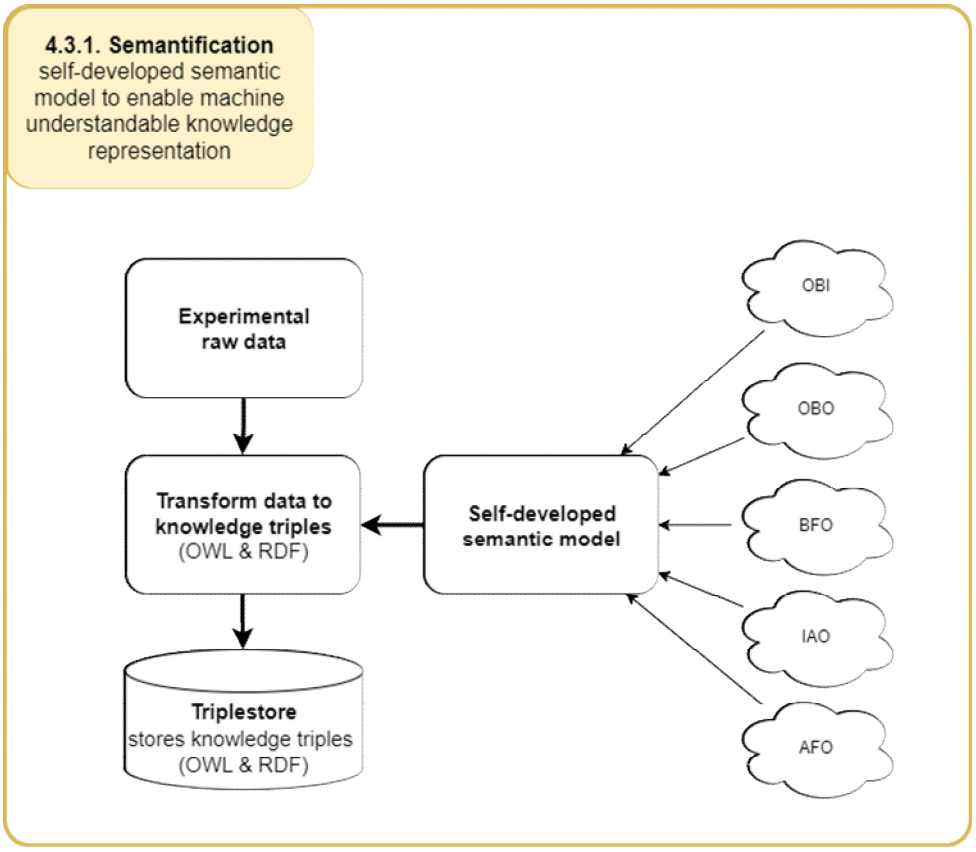


Figure 24 Semantification process of experimental raw data. Experimental data is transformed into knowledge triples (OWL and RDF) based on a self-developed semantic model. The semantic model was developed based on existing ontologies, using relevant terms and concepts. The resulting knowledge triples are stored in a triplestore.

4.3.1.1 Model development and deployment

A semantic model was developed based on semantic web concepts. It was created to encode all qualified aggregated metadata. The model was developed in Python through the RDFlib module (Carl, 2018). It is an abstract representation of the qualified aggregated metadata (subtrees). Without a link to the data, the model is comparable to a blank template that must first be filled with data (instantiation). This is similar to a class-object relation in software development, where the class represents the template, and the object is one concrete instance of that class. During the instantiation process, the semantic model is populated with data. Multiple stability screening data were used for this, including the SEC analytical results for every sample. These are converted as RDF triples. All sample information, including formulation, stress, analytical result, and analytical method information, were converted.

Three options were consecutively tested to store the RDF triples: During model development, the RDF triples were kept in memory through RDFlib and serialized on the hard drive into a Turtle file (1). The second option was the storage in the Blazegraph database (2). Blazegraph is an open-source triplestore that can store a large number of triples. Technically, it was a well-suited solution to store the RDF triples, but to align with company internal solutions, the affinity-internal available software MarkLogic (3) was chosen. Compared to Blazegraph,

MarkLogic offers advanced RDF storage and management capabilities. An additional benefit is the internal technical support that ensured a long-term stable and sustainable deployment of the semantic data.

4.3.1.2 Selection of ontologies

The semantic model was closely aligned to the OBI schema. OBI is an ontology for scientific investigations. This includes concepts for experimental assays, attributes, processes, analytical results, and experimental input/output. OBI uses concepts from four top-level ontologies:

The Open Biological and Biomedical Ontologies (OBO):
"The OBO project was initiated in the early 2000s, as it became clear that there was a community desire to expand ontologies beyond the scope of the Gene Ontology (GO) to tackle biological and biomedical problems more broadly (3). OBO was designed to organize and guide the development of ontologies according to common standards and principles (4), enabling modular composition of ontologies and providing guarantees of technical and scientific quality." (Jackson et al., 2021, p.2)

The Ontology for Biomedical Investigations (OBI):
"The OBI is an ontology that provides terms with precisely defined meanings to describe all aspects of how investigations in the biological and medical domains are conducted. OBI re-uses ontologies that provide a representation of biomedical knowledge from the Open Biological and Biomedical Ontologies (OBO) project and adds the ability to describe how this knowledge was derived." (Bandrowski et al., 2016, p.1)

The Basic Formal Ontology (BFO):
The BFO "… is an upper-level ontology developed to support integration of data obtained through scientific research. It is deliberately designed to be very small, in order that it should be able to represent in consistent fashion the upper level categories common to domain ontologies developed by scientists in different domains and at different levels of granularity." (Arp & Smith, 2008, p.1)

The Information artifact Ontology (IAO):
The IAO was developed out of OBI. The two ontologies are similar and have a close relation. OBI uses an OWL import mechanism to import all IAO terms. It contains information content entities (ICEs) like databases, documents, and digital images (Bandrowski et al., 2016)

Allotrope Foundation Ontologies (AFO):
The AFO "… provides a standard vocabulary and semantic model for the representation of laboratory analytical processes. The AFO suite is aligned at the upper layer to the Basic Formal Ontology (BFO). The core domains modeled include: Equipment, Material, Process, and Results." (Oberkampf, 2018, p.1)

In the BFO and OBI, some attributes or measurable features refer to quality. This is the concrete value of an independent continuant's characteristic, i.e., the weight from the BMI example or the value of a color. Both can be specified as quality following the BFO definition. The BFO quality must not be confused with the definition of data quality throughout this thesis. Here, data quality is understood as the excellence (classification) of something, i.e.,

how "good" or "reliable" a data set is. Furthermore, the qualified aggregated metadata represents the data quality required for a comparability decision.

4.3.1.3 Ontology class examples

A list and definition of some key classes used in the semantic model are presented in Table 6. Note that quality class follows the BFO definition.

Table 6 Ontology reference and description for each class.

Name	**OBO:Code**	**Explanation**
Material entity	BFO_0000040	An independent continuant that has some portion of matter. Input for material processing.
Process	BFO_0000015	A process has temporal proper parts for some time e.g., sealing of the well plate.
Role	BFO_0000023	The role of formulation e.g., buffer, sample or control.
Information content entity	IAO_0000030	A generically dependent continuant that is about something.
Size-exclusion chromatography	CHMO_0001013	Column chromatography where the separation is caused by differences in molecular size.
Assay	OBI_0000070	A planned process that produces information about the material entity, e.g., the SEC method.

RDF triples were used to serialize the entities and concepts. Listing 1 shows a code snippet from the semantic model's turtle file, including some key classes and the relation of the "size-exclusion chromatography" class to the parent classes. All classes are linked through the "subClassOf" relation. The "@prefix" defines the obo variable namespace. The connection of the obo prefix and the ontology class code enables access to the web URI. Through the web URI, a class definition can be retrieved in a way that is similar to the definition column in Table 5. The footnotes include the complete web URI to the "size-exclusion chromatography" class.[13]

[13] http://purl.obolibrary.org/obo/CHMO_0001013 (visited on 16/12/2023)

```
@prefix obo: <http://purl.obolibrary.org/obo/> .

        obo:BFO_0000015 a owl:Class ;
        rdfs:label "process"@en ;
        rdfs:subClassOf obo:BFO_0000003 ;

        obo:OBI_0000011 a owl:Class ;
        rdfs:label "planned process"@en ;
        rdfs:subClassOf obo:BFO_0000015 ;

        obo:OBI_0000094  a owl:Class ;
        rdfs:label "material processing" ;
        rdfs:subClassOf obo:OBI_0000011 ;

        obo:CHMO_0001013 a owl:Class ;
        rdfs:label "size-exclusion chromatography" ;
        rdfs:subClassOf obo:OBI_0000094 ;
```

Listing 1 Owl class definitions and the relation between the classes: process, planned process, material processing, and size-exclusion chromatography. Each class contains a human-readable label (description) and is connected to the parent class through the subClassOf relation.

4.3.1.4 Adaption of the OBI schema

The OBI schema was adapted to fit the needs of the semantic model. Figure 25 shows an overview of the adapted schema, including the classes, their relations, and their originating ontologies (*OBI Core Classes*, n.d.). The figure represents an experimental investigation that consists of multiple classes. The classes are connected through the "is-a relation" connection (thick blue arrow shape) or the "other relation" arrow (thin blue arrow shape). The classes are composed of three groups: material entity (green), process (light blue), and information content entity (grey). Some classes were unnecessary and removed from the schema (red). Removing the investigation class required establishing a connection between the "planned process" and the "study design execution" classes. The qualified aggregated metadata components are not directly listed in the schema because they are logic decisions, thresholds, or value acceptance ranges that are expressed properties or attributes, not ontology classes. An example is the different stress conditions, e.g., the (40C 7 days) stress from the workflow/process sample's history subtree. This stress and all other stress conditions are properties linked to the "study design execution" class through the "has_part" relation and, thus, are not depicted in Figure 25.

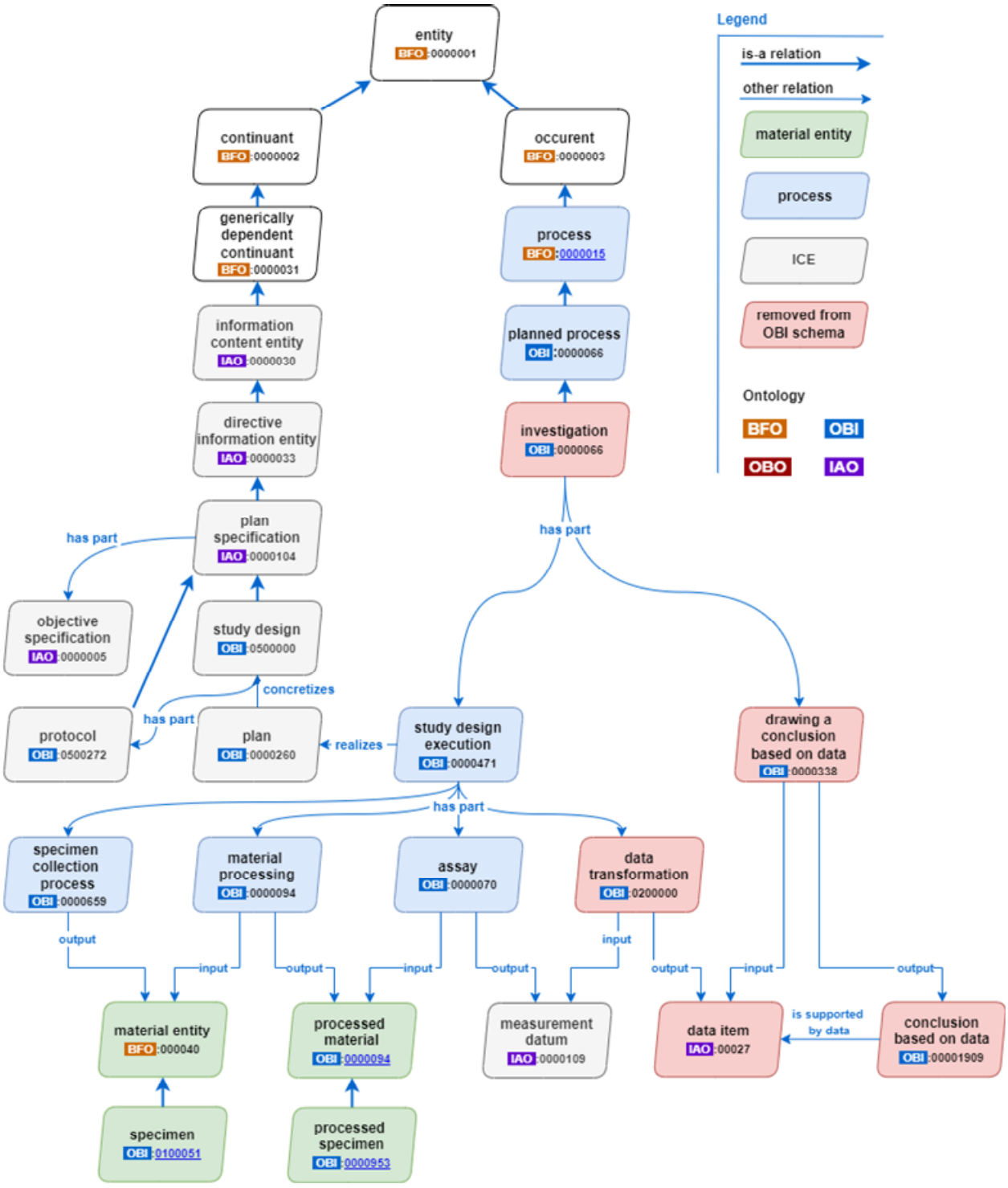


Figure 25 The adapted OBI schema. It is structured into three distinct groups: material entity (green), process (blue), and information content entity (ICE) (grey). The schema represents an experimental investigation, starting with the broad "entity" class at the top and increasing in detail down to the "measurement datum" class at the bottom. Figure modified from (*OBI Core Classes*, n.d.).

4.3.1.5 Expression of SEC sample measurement data

The basic principles of the OBI schema from Figure 25 were further modified to better reflect a SEC method for a liquid chromatography experiment. As shown in Figure 26, the ontology classes were implemented to represent a measurement and its quality. The “measurement datum” class (in purple) is a sub-class of “data item”, which is a sub-class of “information content entity”. The “is quality measurement of” relation is employed to link the "measurement datum" class to the "SEC monomer fraction quality" class, which is a sub-class of the “quality” class.

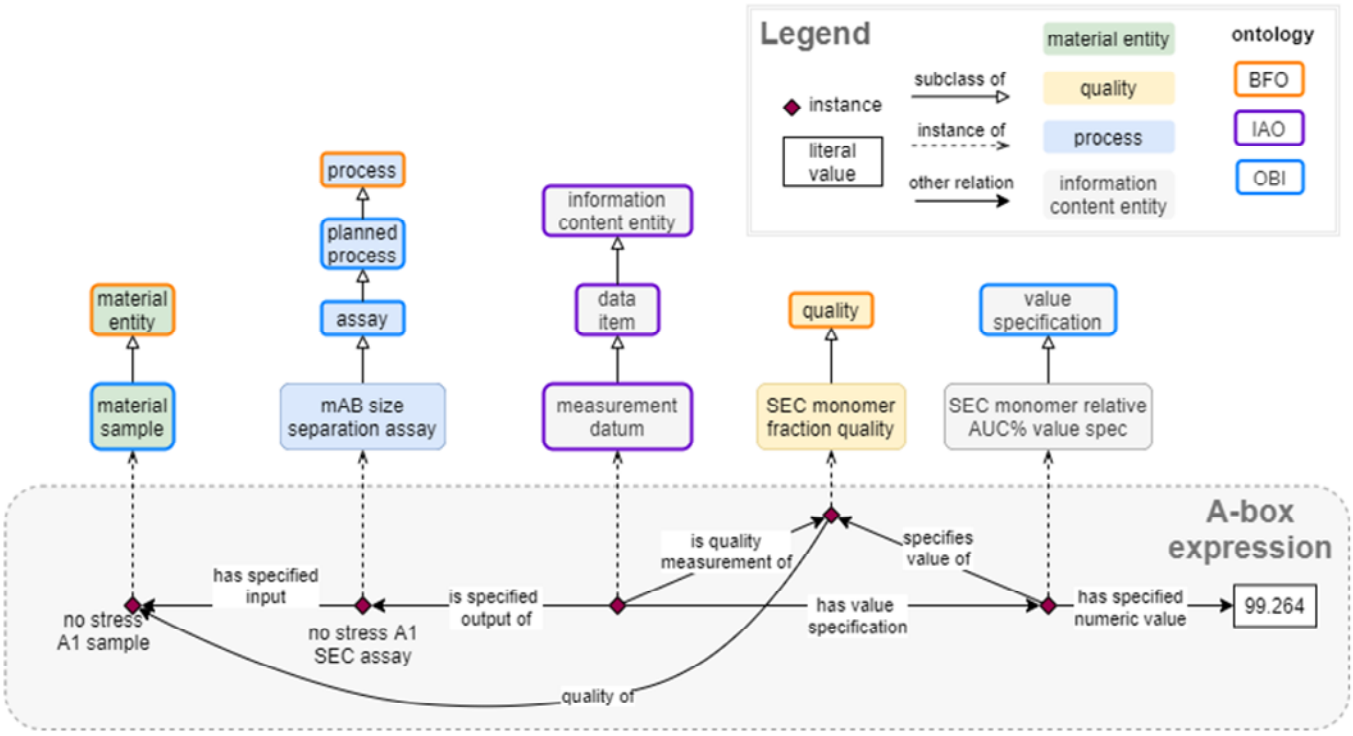


Figure 26 Diagram of a sample measurement semantic representation. The diagram shows the connected classes and their relations that lead to the numeric value of a SEC measurement. The external classes from the different ontologies have colored borders, which indicates the originating ontology.

4.3.2 Qualified aggregated metadata – decisions, threshold, and ranges

A close examination of the different types of metadata is required to categorize which types the Allotrope HPLC model uses and what would be required to achieve data set comparability. The following three metadata types exist:

- Structural metadata: Defines the relation of digital resource components so that they can be understood, e.g., pages in a book that are ordered and read in sequence.
- Descriptive metadata: It is used to find, identify, and discover a resource or object, e.g., author, language, and title.
- Administrative metadata: It supports the administration of a resource and its use, e.g., the provenance or permissions.

The method-based Allotrope HPLC model combines descriptive and structural metadata, which can be described individually as general metadata. Descriptive metadata is used to depict the different liquid chromatography components, such as pumps and injectors, and includes a detailed description of the settings of each component, such as pressures, injection times, or light scattering settings. Structural metadata describes the analytical results and the shape and form of the analytical peaks. Together, both can be summarized as method-based since all metadata is very detailed and on a low component- and device-dependent level, which is ideal for semantically describing the chromatography system and used for method reproducibility but insufficient to derive result comparability. A different type than method-based metadata is required to achieve comparable, and thus machine comparable liquid chromatography data. The focus should be shifted from the method and the device to the analytical result. All factors a chromatography SME considers when selecting different data sets based on their comparability must be available to the machine.

Furthermore, the metadata must be presented on a higher level. The comparability factors that influence the result must be included, such as the System Suitability Test (SST), analytical method, workflow, and error. These can be described as qualified aggregated metadata. This qualified aggregated metadata provides information about the quality of the comparability influencing factors and can, therefore, also be referred to as qualifying metadata. They represent a combination of structural and descriptive metadata but extend the list through qualifying metadata. Later, the qualified aggregated metadata is expressed in a semantic model, which is required for the final comparability decision. Table 7 shows the different types of metadata: general metadata, method-based metadata, and qualified aggregated metadata.

Table 7 Different metadata types and their characteristics.

Metadata type	**Description**	**Example**
General metadata	• Data that describe regular data	Any information describing data, providing additional context for the data.
Method-based metadata	• Focus on the applied analytical method and its reproducibility, e.g., the Allotrope HPLC model. • Enable machine and human readability	Enables reproducibility of analytical method, e.g., description of HPLC column in the Allotrope HPLC model
Qualified aggregated metadata	• Set analytical data into context • Include additional SME knowledge, e.g., threshold, criteria and ranges for data comparison • Serve a specific reuse purpose • Qualify a sample for manual or automatic comparison • Mandatory for machine actionability	Analytical result-focused metadata, e.g., an acceptance range for controls in a workflow (see 4.3.2.2)

Figure 27 illustrates how metadata types align with the FAIRplus data maturity levels (DSM). Method-based metadata can be categorized as standardized data (DSM level 3), semantically typed data (level 4), and managed data assets (level 5) since they can be used to describe analytical methods, e.g., Allotrope HPLC model, to enable reproducibility of the method. Until level 5, the different data types are considered machine readable, allowing machines to access the data but not interpret it like a human scientist. Qualified aggregated metadata is required to go beyond machine readability and enable machine actionability. This type of metadata combines method-based metadata with the definition of SME knowledge and expertise, including criteria, thresholds, and ranges that enable comparable data sets. At this stage, the data is considered comparable in that the comparable criteria are included. However, actual comparability is not yet determined, similar to defining a variable as an integer without specifying its numerical value. To achieve true machine actionability for humans and machines, the qualified aggregated metadata (including the SME criteria, thresholds, and ranges) are used to qualify the data. The abstract example of variable definition would involve assigning an actual integer value so that different integer variables can be compared and ordered. The SME comparability logic facilitates the order and rank of the data based on their quality, which enables the final level - automatic comparability.

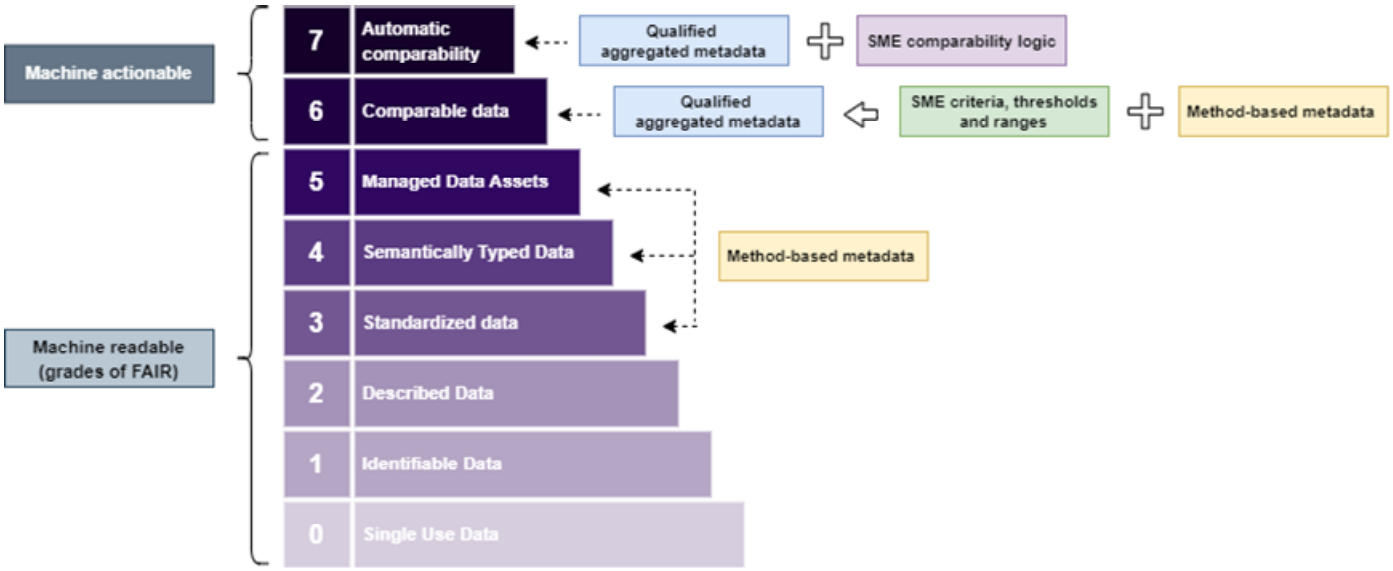


Figure 27 FAIRplus data maturity levels (DSM) and metadata hierarchy. Method-based metadata applies to levels 3 to 5. Combining SME criteria, threshold, and ranges with method-based metadata creates qualified aggregated metadata required to reach level 6 of comparable data. Extending qualified aggregated metadata and the SME comparability logic enables the final 7. level of automatic comparability. Figure modified from (*Overview | FAIRplus Data Maturity*, n.d.).

The qualified aggregated metadata are combined in a multi-level schematic decision tree that reflects criteria, threshold, and ranges for result comparability. The tree consists of four subtrees. Each subtree represents one of the comparability influencing factors groups (SST, analytical method, workflow, and error) that are further referred to as qualified aggregated metadata components. These are necessary to estimate chromatography data set comparability and are later transformed into a semantic model to allow access for the machine (agent). Each subtree is read from top to bottom and comprises nodes and leafs. The nodes represent a qualified aggregated metadata part required for a decision. The leafs depict the type of decision in the form of pass / no pass decisions, thresholds, or value acceptance ranges. Figure 28 shows the four top nodes on the highest level of each subtree.

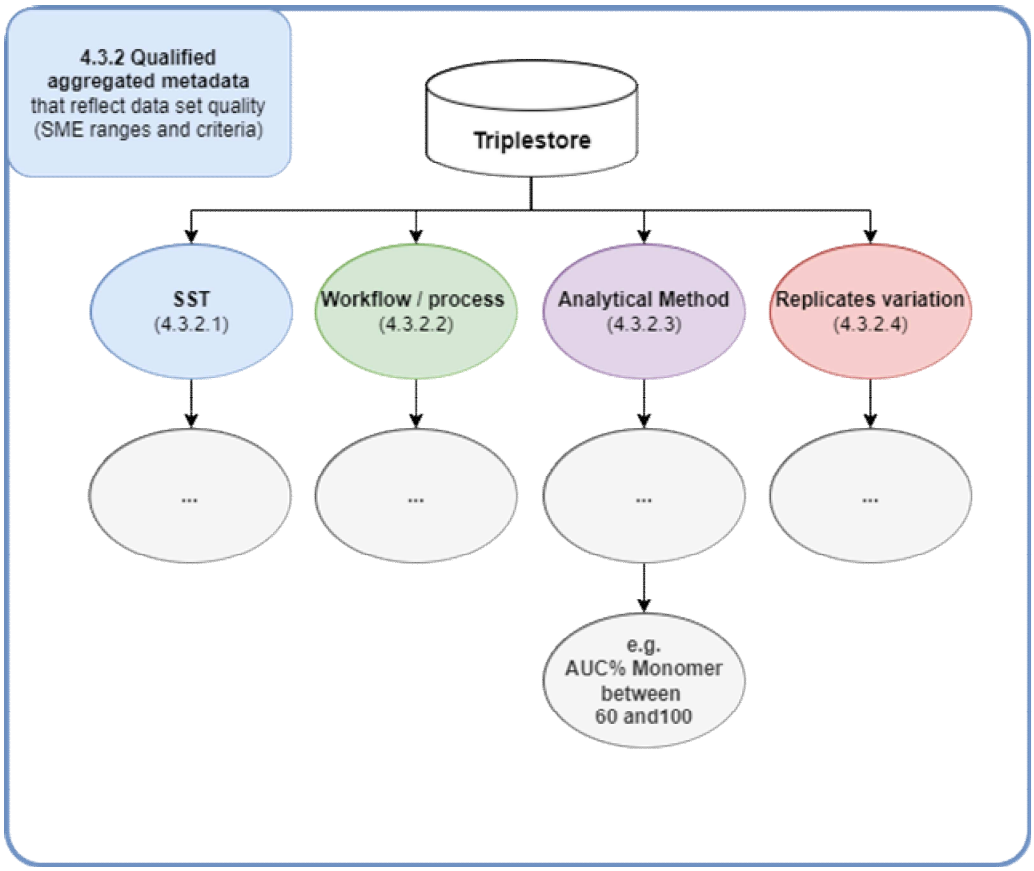


Figure 28 Top-level tree of the qualified aggregated metadata. The qualified aggregated metadata consists of multiple lower-level decisions grouped into four subtrees: SST, analytical method, workflow, and the replicates variation. Each subtree contains criteria, thresholds, and ranges related to the data set's quality (grey). One example is the valid monomer range between 60 and 100 percent monomer, which is part of the Analytical Method subtree. Samples within this range are classified as valid. Samples outside of this range are classified as invalid. If all subordinate decisions of one subtree, e.g., the analytical method.

4.3.2.1 System Suitability Tests

The SST are acceptance criteria to evaluate if a chromatographic system performs within an expected range before a real experiment is conducted. Analytical performance tests like the SST can be compared to a pH-meter calibration, which ensures that pH readings are accurately measured. During a several-week-long stability screening, the performance of the chromatography can vary due to external influences, e.g., room temperature, air pressure, and humidity. A standardized reference mAb measures six test runs before every actual experiment. The mean of these tests is calculated. The difference between the current SST mean and the expected SST mean results for the reference mAb are assessed. This enables the detection of performance variances over time. If the current SST mean varies from the expected SST mean results, the mechanic applies to the actual experiment sample results.

In some cases, a detected variance in an SST offset can correct the accurate experiment results by shifting the sample results by the difference between the SST test and the expected SST performance. The three essential parts are the tailing factor, precision test, and drift

control. A multitude of metadata describes each part. Instead of semantically describing all metadata, only the pass or no-pass decision is stored as qualified aggregated metadata for each of the three parts. Figure 29 shows the SST subtree including the three acceptance criteria: drift control in range, tailing factor monomer, and instrument precision (peak area/rt monomer).

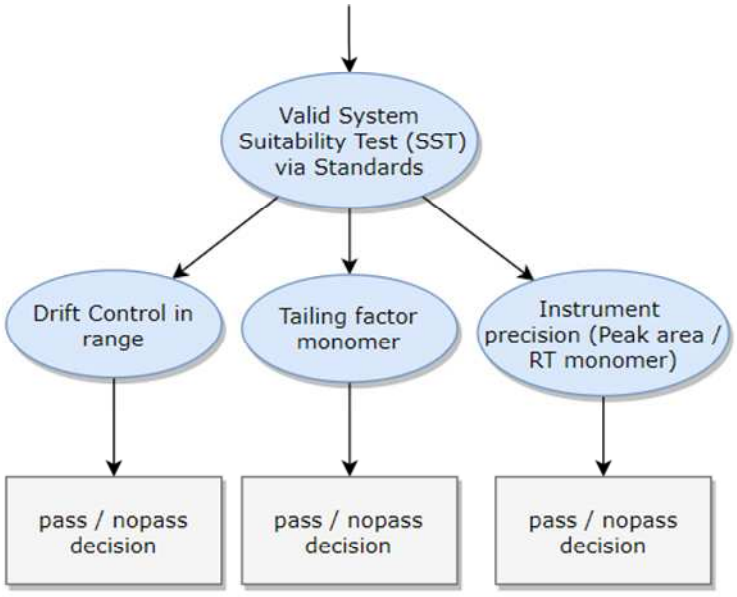


Figure 29 System Suitability Test (SST) qualified aggregated metadata. The SST exists of multiple components required for comparability evaluation: drift control, trailing factor, and instrument precision.

4.3.2.2 Workflow/process

The introduction of lab automation leads to a standardization of workflows and processes. In some cases, the standardized workflows and automation processes require adaptation to prepare for novel mAbs and anticipate liquid formulation issues. A highly viscous formulation, for example, which adheres to the sealing foil may require an additional centrifuging step. An unexpected workflow change may occur due to an automation malfunction. The samples may be subjected to unexpectedly higher temperatures than initially planned. This would lead to an increased protein degradation of all samples. Such a malfunction may go unnoticed in some cases.

The use of well-characterized molecules as positive and negative control is not just a good scientific practice, but a reliable one, which ensures the proper execution of the stress conditions. At higher stress conditions, the positive control shows minor protein degradation, serving as a reference point and resembling a stable mAb. On the other hand, the negative control starts degrading at low stresses, resembling an unstable mAb. During stability screenings, the controls and samples follow an equal workflow, as they are placed on the same well plate. Ideally, historical data in the form of experiments over several years exists for both controls at each stress condition, so that expected performance ranges can be spanned for the controls. Based on the historic performance, the expectance ranges can be calculated through the mean and the standard deviation of the controls. Both are calculated

for the positive and negative controls. To allow for a degree of freedom for the analytical performance, the standard deviation is multiplied by 2.5. The result is subtracted and added to span a 2.5 sigma range, which resembles the expected performance range for each control at a specific stress condition. Table 7 shows example data for the calculation of the upper and lower expected range for two controls at 40°C 7 days stress condition.

Table 8 Example calculation for the expected range of the positive and negative controls from historical data.

Control name	**mean SEC Monomer (40°C 7 days)**	**Standard deviation**	**Multiplicator**	**Upper expected range**	**Lower expected range**
Positive control	95	1	2.5	97.5	92.5
Negative control	88	2	2.5	93	83

This data is used to span an expected range for the controls behavior (degradation). In plate-based HTS concepts, the samples reside together with the controls on the same multi-well plate. If a control exceeds the expected range and thus does not perform as expected, it's a red flag. The explanation is that the stress was improperly applied due to some errors in the workflow. Since the samples and controls are located on the same well plate, this workflow malfunction equally affects the samples. In this case, the whole stress condition and all samples cannot be used for further evaluation and should be excluded from the data management. If the controls perform as expected, the samples are checked for additional criteria. For all samples under all stress conditions, a minimum sample performance of 90% percent area Monomer and a maximum of 5% HMW is required. These ranges are aligned with FDA guidelines for antibody drug development. Figure 30 shows the workflow subtree and all required acceptance criteria.

In summary, the use and performance of controls enables further data standardization and resembles good scientific practice. The consistent use of controls for each screening enables data benchmarking of the target molecule, and data set comparison between different stability screenings. Most importantly, the use of controls demonstrates how data set comparability can be achieved without the direct need for metadata. It is not the comparison of methods, workflows, processes, or stress metadata such as stress temperature (in Celsius), incubation time (in days), or centrifugation duration, but rather the determination of the correctness and comparability of a data set through the analytical results of well-characterized molecules.

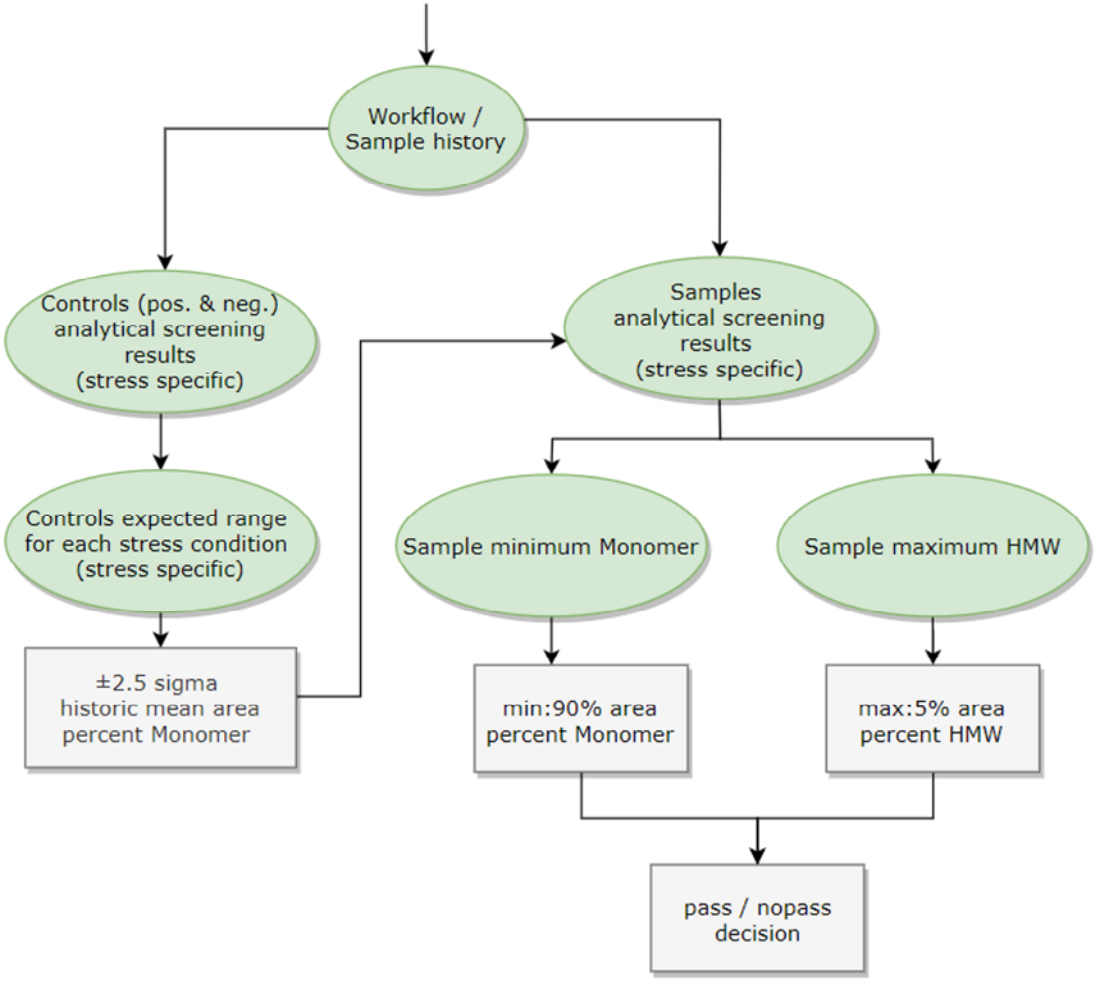


Figure 30 Workflow/process qualified aggregated metadata subtree describing the sample and control history.

4.3.2.3 Analytical method

The analytical method's suitability is tested before the stability of an unknown mAb can be assessed through a stability screening. This preparation phase is referred to as method verification. In some cases, platform methods must be adapted to new molecules to be able to characterize them correctly. In this process, an error in the analytical method can be estimated. Chromatography precision depends on multiple factors, e.g., device parameters, columns, and the analytical method. The analytical method error can be estimated at around two to three percent of the percent monomer in an automated laboratory setting that includes automatic workflows and a high-throughput screening concept (HTS). The error estimation and the error itself are essential to distinguish between different samples and their formulation effects during stability screenings.

During stability screenings, different stress conditions induce protein degradation and distinguish between the stabilizing formulation effects. The stress conditions. Some stress conditions have a greater impact on protein degradation than others. Within one screening,

it may occur that lower stress conditions, e.g., freeze-thawing, shaking stress (also referred to as mechanical stress), or low-temperature stress (5 degrees Celsius), induce less measurable stability differences. Consequently, the formulation effect is likely smaller than the method verification error. In other words, the noise (the analytical error) is greater than the signal (formulation effects). Consequently, no conclusions about formulation differences can be drawn at these conditions. Higher stress conditions, e.g., 25 and 40 degrees Celsius, induce a protein degradation greater than the analytical method error, enabling distinguishable formulation effects.

The analytical method error is essential to data comparison within one screening and the comparability assessment across different screenings. Furthermore, the analytical method defines a scientific threshold range for samples during workflow execution, as displayed in Figure 31. This is the range at which the analytical method performs well and can analyze the samples correctly. It is tested during the preparation phase of stability screenings (method verification). The range is between 60 and 100 % of the area percent Monomer. The analytical method cannot accurately determine the sample's Monomer content outside this range. A similar rational count for the protein concentrations valid method range. The analytical determination is accurate within 50 and 200 mg/ml protein concentration ranges.

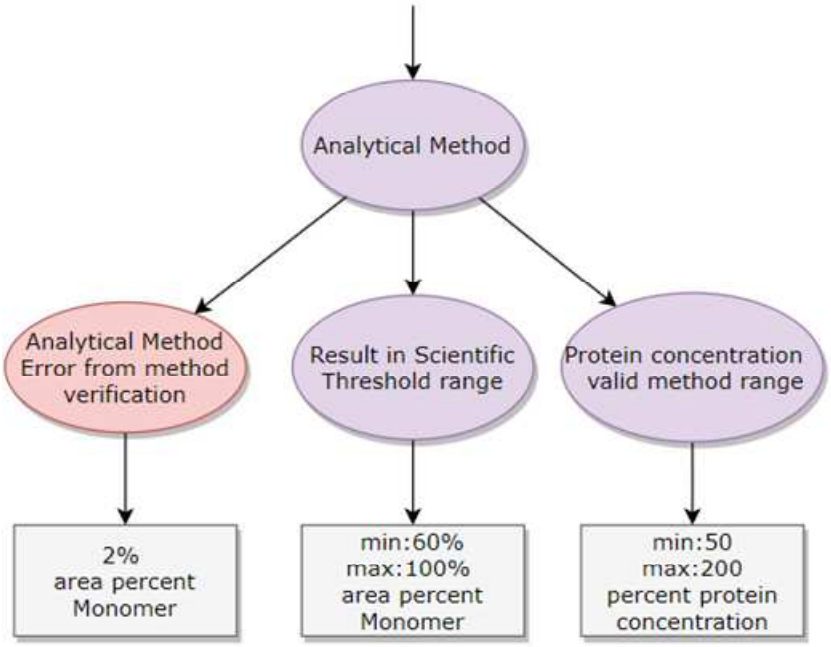


Figure 31 Analytical method qualified aggregated metadata subtree and components required for comparability evaluation.

4.3.2.4 Sample replication and errors

Sample replication is the multiplication of one or many samples. In formulation development, this resembles the multiple use of equally formulated samples (compositions) on a well plate. The advantage of sample replication is better estimating the overall error within one stability screening, which increases experimental quality, precision, and reliability. The disadvantage is that fewer formulations can be screened since fewer well positions are available on the

well plate. Therefore, the use of sample replication is in a constant area of tension between an increase in the number of samples and a quality increase through a more precise error estimation. External factors and minor workflow variations during the screening influence the samples and, thus, the analytical readouts. Additionally, the formulation composition influences the stability of each sample. Without sample replication, these factors are blurred by the analytical method error because the method error is greater than the measurable impact of these (the noise of the analytic is greater than the signal). Sample replication enables the quantification of these influences and allows for a precise determination of the method error. The classification and error calculations are performed on multiple levels:

- Sample-specific: Error calculation and comparability classification for each sample at each stress condition for each time point (approximately 720 decisions per data set).
- Stress-specific: Error calculation and comparability classification summarized from the sample results, representing all samples for a given stress condition (720 sample decisions grouped by the eight stress conditions per data set).
- Stability-Screening-specific: Overall error calculation and comparability score, assigning a single comparability score for the stability formulation screening based on the stress-specific errors and classifications.

This is achieved by calculating the replicate error for each replicate formulation group at each stress condition. For HTS concepts, the workflow/process + the analytical method summarizes a total error of around two to three percent Monomer. The analytical error can be precisely estimated through replication and ranges between 0.2 and 0.4 percent Monomer. This is significantly lower than a method error of two to three percent Monomer. Figure 32 shows a flow chart of the replicates subtree, and Figure 33 shows how the different levels of the tree are calculated for the different stress errors.

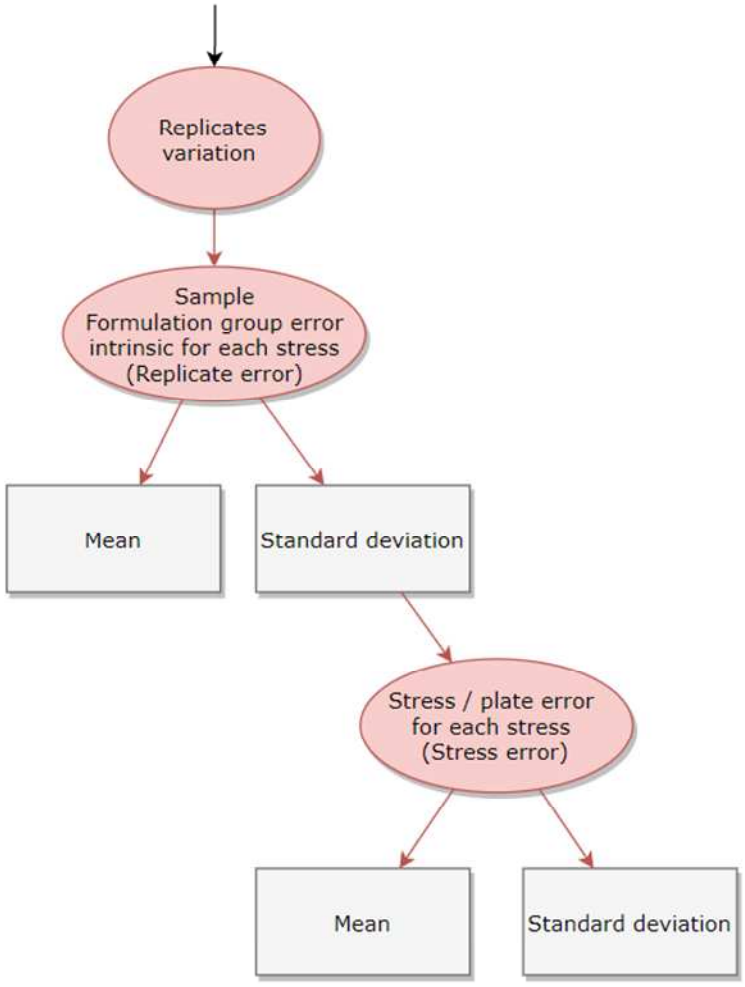


Figure 32 Replicates qualified aggregated metadata subtree and groups required for comparability evaluation of the error.

First, the replicate error is calculated for all formulation compositions within one stress. Usually, a sample replication factor of three is used as a compromise between statistically robust error calculation and remaining sample positions on the well plate. The mean ("Replicate performance (mean)") and the standard deviation ("Replicate error (std)") of the Monomer area percent of samples with equal formulation composition, e.g., "Formulation 1," are calculated for all replication groups. This first measure of the analytical performance of each sample is relative to the replicate group mean. Second, the mean and the standard deviation of all "Replicates error (std)" are calculated within one stress condition, e.g., "no stress." The result is the mean performance of all samples ("Replicate stress performance (mean)") and the error of all replicates for one stress condition ("Replicate stress error (std)"), also referred to as stress-specific error. The calculated stress error is especially important because it compares single stress conditions across screenings. In addition, it inherits all influences from the workflow and the analytical method within one measure.

Furthermore, it ensures the reproducibility of workflows. An additional benefit is increased screening robustness. If, for example, a single formulation composition is contaminated during a stability screening, it will be removed from the process and further evaluation. No

analytical conclusion can be drawn from this sample, and the stability data for that formulation composition is lost. If replicate samples of this formulation exist, the loss is less severe since a backup exists. The stress-specific error can later be used to calculate an overall stability-screening-specific error by calculating the mean of all stress-specific errors of one screening.

mAb	Formulation	Stress condition	Monomer area percent	Replicate performance (mean)	Replicate error (std)	Replicate stress performance (mean)	Replicate stress error (std)
mab 1	Formulation 1	no stress	98.1	98.0	0.2	98.3	0.2
mab 1	Formulation 1	no stress	97.8	98.0	0.2	98.3	0.2
mab 1	Formulation 1	no stress	98.2	98.0	0.2	98.3	0.2
mab 1	Formulation 2	no stress	98.2	98.5	0.2	98.3	0.2
mab 1	Formulation 2	no stress	98.5	98.5	0.2	98.3	0.2
mab 1	Formulation 2	no stress	98.7	98.5	0.2	98.3	0.2
mab 1	Formulation 1	Temperature 40°C 21 days	95.4	94.9	0.9	94.3	0.6
mab 1	Formulation 1	Temperature 40°C 21 days	93.7	94.9	0.9	94.3	0.6
mab 1	Formulation 1	Temperature 40°C 21 days	95.6	94.9	0.9	94.3	0.6
mab 1	Formulation 2	Temperature 40°C 21 days	94.6	93.7	0.8	94.3	0.6
mab 1	Formulation 2	Temperature 40°C 21 days	92.7	93.7	0.8	94.3	0.6
mab 1	Formulation 2	Temperature 40°C 21 days	93.8	93.7	0.8	94.3	0.6

Figure 33 Example table for the sample replicate calculation of two formulation groups over two different stress conditions.

Figure SI 1 depicts the complete assembly of the subtrees. The machine must be enabled to access the qualified aggregated metadata and traverse all criteria of each subtree programmatically. Therefore, the qualified aggregated metadata schemas and subtrees must be encoded into a machine understandable and accessible solution. This resembles the second part of the results-based concept, the semantic model, which will be described in the following chapter.

4.3.3 Inclusion of subject matter expert (SME) comparability logic

As the third and final part of the result-based concept, the machine must augment the comparability SME decision process based on the qualified aggregated metadata. High-quality data is a prerequisite for further reuse. The precision of the data quality determination is enhanced by the aggregated data quality defining metadata, which is based on the components: SST, analytical method, workflow / sample history and the sample replicate error represented in the previously described semantic model. The assessment categorizes datasets into four levels of comparability: none, low, medium, and high. These levels can also be quantified using numerical values: 0.0, 0.3, 0.6, and 1.0. The SST ensures that the chromatography system works as expected. This is one of the first two requirements to ensure ground-level data comparability. The second requirement is the correctness of the workflow. The workflow was correctly applied when the positive and negative controls were performed within a historically expected range for each stress. If these two requirements are fulfilled a data set reaches a "low" comparability. In the contrary case the comparability is classified as "none". Furthermore, if the suitability of the analytical method for a mAb is proven during method verification and the analytical method error is determined, "medium" comparability can be attained. Moreover, the use of replicates enables a precise error

estimation for each stress of a screening, which leads to a "high" comparability. Figure 34 shows all four data set comparability stages and their interdependence.

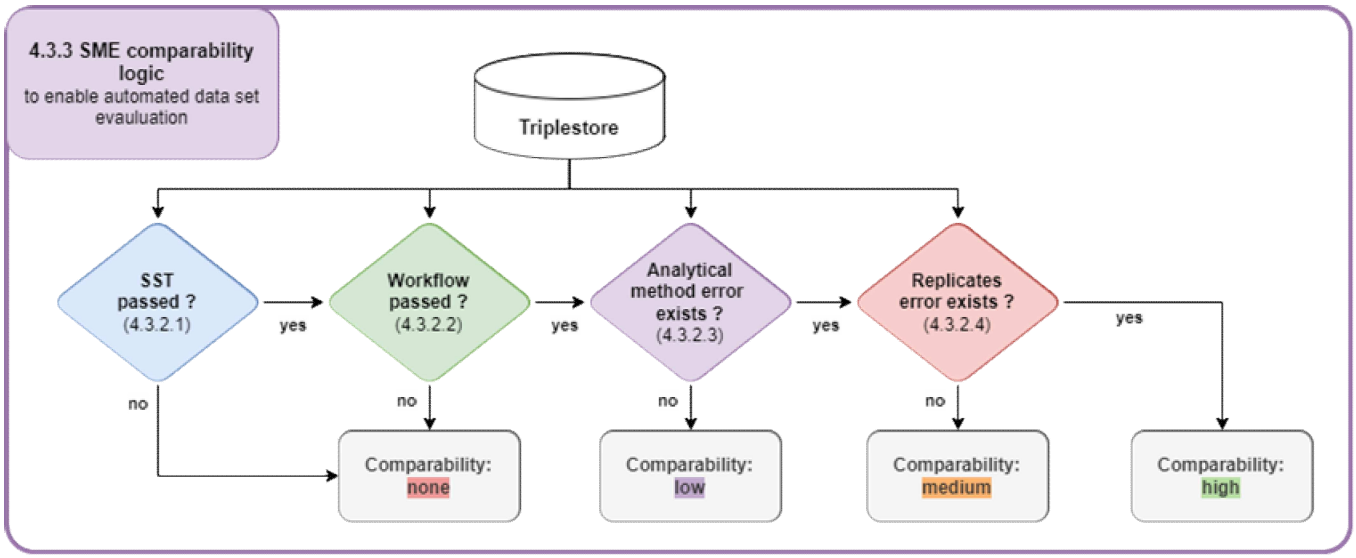


Figure 34 Flow chart of the SME comparability logic . The differently colored boxes resemble one group of qualified aggregated metadata. The chart is traversed from the left top to the bottom. Each leaf (grey) represents a comparability category.

The comparability assessment is a crucial step performed for all available stress conditions of a single stability screening. It not only provides a comprehensive understanding of the data comparability but also guides the selection of data sets for further reuse. In case data sets with a certain level of comparability are required or should be suggested for further reuse, the comparability assessment can provide a measure for those means. Moreover, the following chapter explains how different screenings including all stresses are ranked and how differently ranked screenings can be compared.

5 Evaluation

In this chapter, example formulation data is used to demonstrate, evaluate, and verify the application of the result-based standardization concept. Therefore, three stability screenings of different mAbs with different laboratory workflows are selected as experimental data. The differing workflows influence the result comparability. Each stability screening includes HPLC (SEC) analytical results. The comparability is automatically assessed for each screening by the machine, and the results are examined. The automatic comparability and how the different comparability parts play into the screening evaluation are displayed in a high-level flow chart in Figure 35.

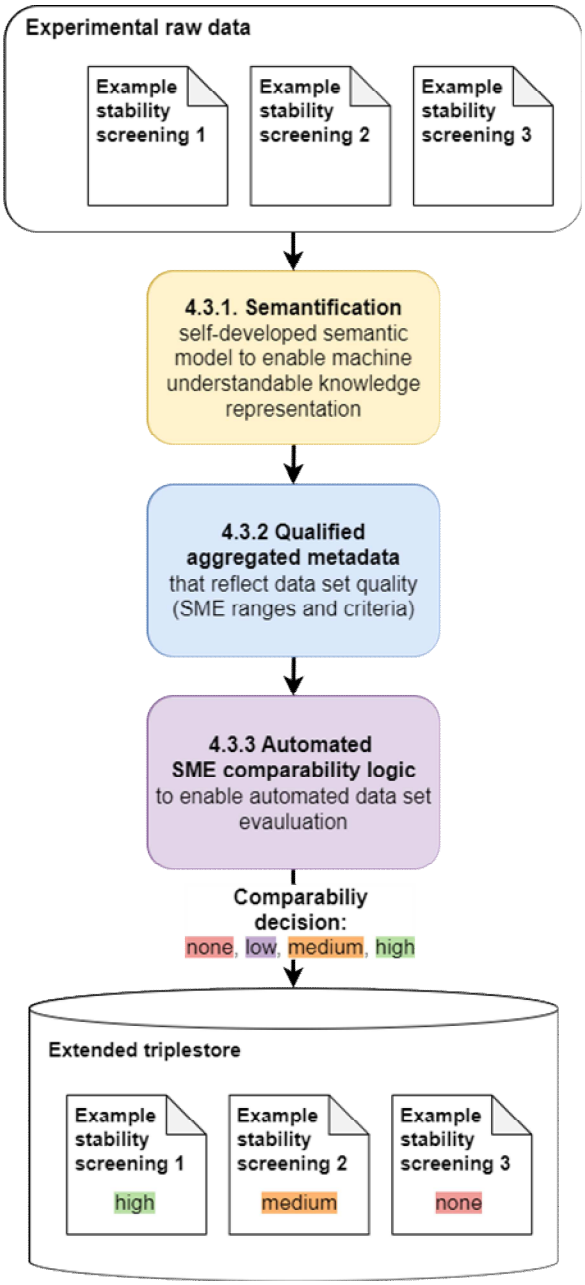
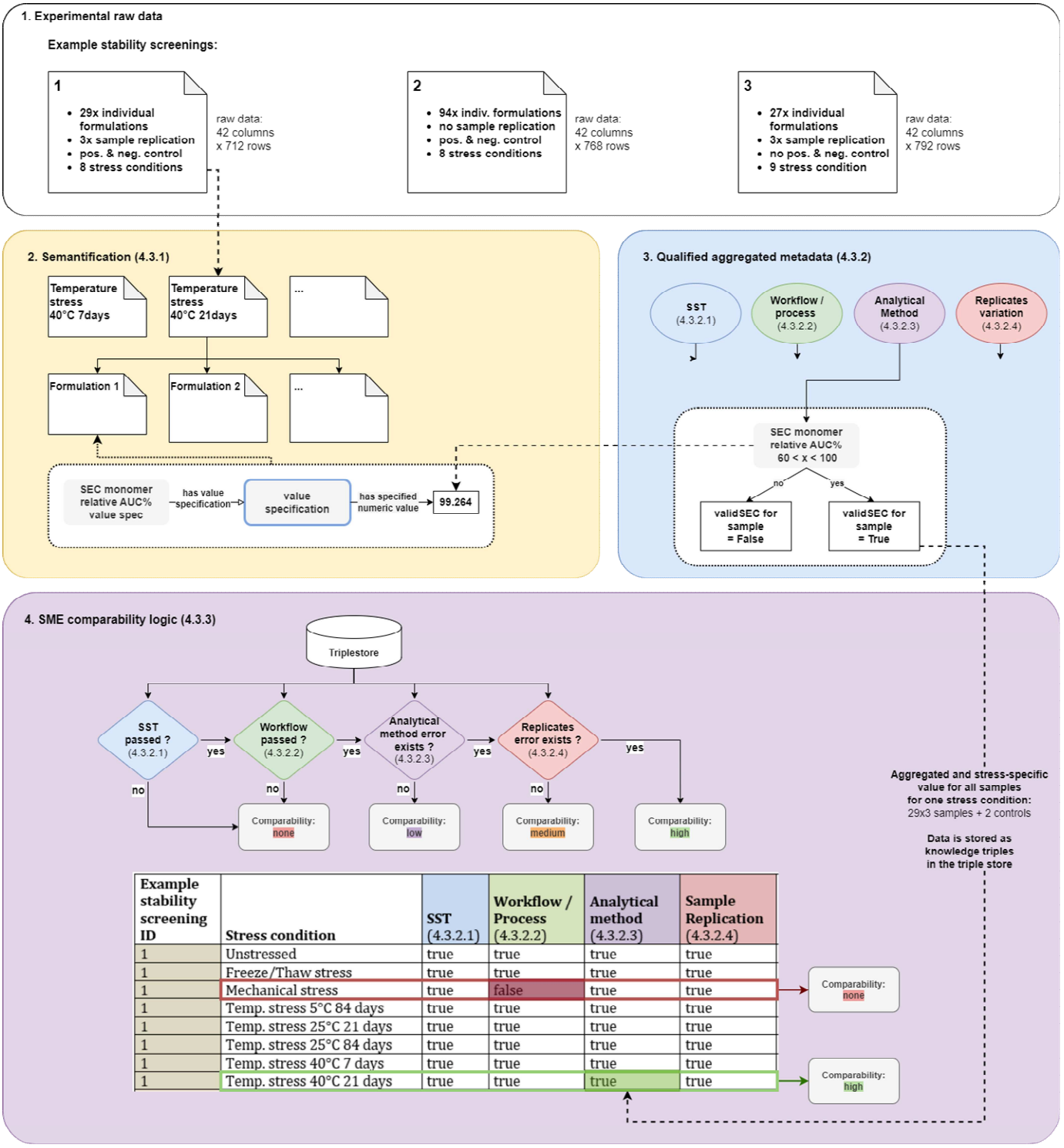


Figure 35 High-level workflow of data set comparability assessment.

First, the experimental raw data in the form of three stability screenings is described in detail. Then, the raw data is transformed into a semantic representation to enable a machine understandable knowledge representation. For this, a self-developed semantic model was developed as described in Chapter 4.3.1. The semantic data is stored in a triplestore. Based on the semantic data, the scientists' expertise, represented as decisions, thresholds, and ranges (qualified aggregated metadata), is used to express screening-specific quality metrics for each data set. These decisions extend the experimental data in the triplestore. In the final step, the quality metrics are not just prioritized, but carefully and intelligently ordered by the SME comparability logic. This crucial step enables the machine to automatically determine, prioritize, and order the individual stability screenings' comparability based on the semantically available quality metrics.

Each step is explained in detail in the following sections. Figure 36 provides a more detailed overview of each step that is mentioned in Figure 35.



Example stability screening ID	Stress condition	SST (4.3.2.1)	Workflow / Process (4.3.2.2)	Analytical method (4.3.2.3)	Sample Replication (4.3.2.4)
1	Unstressed	true	true	true	true
1	Freeze/Thaw stress	true	true	true	true
1	Mechanical stress	true	false	true	true
1	Temp. stress 5°C 84 days	true	true	true	true
1	Temp. stress 25°C 21 days	true	true	true	true
1	Temp. stress 25°C 84 days	true	true	true	true
1	Temp. stress 40°C 7 days	true	true	true	true
1	Temp. stress 40°C 21 days	true	true	true	true

Figure 36 Detailed workflow of the result-based standardization concept . Experimental data, including HPLC (SEC) results, starts a comparability assessment (1). This data is semantified (2) and stored in the triplestore to give the data semantic meaning. SPARQL Queries are used to inquire about data set quality aspects (qualified aggregated metadata) to return and store the results in the triplestore (3). The SME comparability logic prioritizes the quality results for each data set, errors are calculated, and the comparability is determined (4).

5.1 Experimental raw data

Experimental raw data from three stability screenings, including formulation data and corresponding HPLC (SEC) results, are used as input to evaluate the result-based standardization concept. These screenings contain information about ~96 liquid formulations (samples) for a specific protein (mAb). This includes information about the protein concentration, pH, and protein-stabilizing additives (excipients) such as solvents, stabilizers, preservatives, surfactants, and buffers. Additionally, the data contain the applied analytical method and the measurements for method verification. The formulations undergo various stress conditions, including Freeze/Thaw, Temperature, and Mechanical stress.

Furthermore, the data include applying these stress conditions over time (process), with specific time points, such as subjecting the samples to 40°C temperature stress for 21 days. Additionally, for each stress at each time point, negative and positive control in the form of well-characterized reference proteins are equally subjected to all stress conditions. Existing historical data on the controls and their negative and positive behavior allow us to draw conclusions if the stress conditions are correctly applied.

These stress conditions negatively impact the protein stability of the samples and the controls, leading to denaturation and indicating a less stable product depending on the formulation used. The goal is to identify the optimal liquid formulation stabilizes the protein during accelerated stress.

A liquid chromatography size-exclusion method (SEC) is used to measure the stability of each formulation at each time point. The SEC method determines three analytical results during each measurement, represented in an Area Under the Curve (AUC) with a total quantity of 100% for all components. The components of the AUC are as follows:

1. Monomer: The functional protein, which should ideally have a high percentage, typically above 95%, as the main isoform (Monomer).
2. Fragments: Low molecular weight components should be minimized and quantified around 0%.
3. Aggregates: As per FDA regulations, the percentage of aggregates must be below 5% to avoid potential life-threatening immune reactions after drug administration.

All screening-specific information can be seen in Figure 36 in the upper white box. Stability screening one, for example, contains 29 individual formulations. During the screening, each formulation is replicated three times. The multiplication results in 87 samples. However, a negative and positive control is also used during each stress condition, resulting in 89 samples being tested during each stress condition. Multiplied by the number of 8 stress conditions, this results in an overall 712 data points (rows). The formulation parameters such as pH, concentration, additives, stress conditions, as well as the corresponding analytical HPLC (SEC) readouts (Monomer, Fragments, and Aggregates) for each formulation (sample) are expressed as 42 columns. This sums up a data table with 712 rows and 42 columns for stability screening.

At this point, only domain experts (scientists) who are well-versed in experimental planning and execution can manually integrate and interpret the experimental raw data. The data is transformed into a semantic knowledge representation to make the data machine understandable and meaningful for the machine.

5.2 Semantification of raw data

To achieve machine readability the screening data in table format is transformed into knowledge triples through web technology concepts such as ontologies. The screenings are uploaded to a self-developed web app that transforms the raw screening data into knowledge triples according to the semantic model. A commercial triplestore transforms each formulation (data set rows) into the RDF triples. This is applied to all 712 rows and 42 columns as they are transformed into a semantic knowledge representation.

Figure 36 displays the data semantification on one formulation of the Temperature stress condition of 40°C at 21 days (yellow box). The self-developed semantic model (4.3.1) is used to create an SEC monomer relative AUC% value specification with a specified numeric value of 99.264. In parallel, the same semantification process is applied for the Aggregates and Fragments, which in total sum up with the Monomer to 100% AUC. Listing 2 displays the OWL code used to semantically represent the Monomer value.

```
ex:sample1_sec_rel_mono_md_valid
        a obo:IAO_0000109 ; # measurement datum

        has_value_specification: [
                a model:SEC_monomer_relareapercent_vs ;
                has_specified_numerical_value: 99.264;
        ];

        is_quality_measurement_of: [
                a model:SEC_monomer_fraction_quality ;
                quality_of: ex:sample1 ;
        ];

        is_specified_output_of: ex:my_sec .
```

Listing 2 Owl numerical value specification for a sample . The numerical value was specified with 99.264 monomer AUC percent.

The semantification enables machine readable and meaningful data, satisfying FAIR requirements and making the data reusable in the sense of findable. However, it does not yet qualify as machine actionable data, as machines cannot automatically compare differently processed data sets, which is the primary objective. Therefore, the qualified aggregated metadata are required to detect data set-specific quality metrics by the SME's threshold, ranges, and criteria (qualified aggregated metadata).

5.3 Querying qualified aggregated metadata

Qualified aggregated metadata, derived from the expertise of scientists, are applied in the next step to assess the quality of data sets (blue box). They represent a significant step towards machine actionability beyond machine readability. Based on the quality determination, the machine can later prioritize the importance of the qualified aggregated metadata to derive a comparability score for the stability screenings.

SPARQL is used to query the triplestore to check the RDF triples (screening data) against the qualified aggregated metadata decisions, criteria, and ranges. This can also be referred to as quality determination (infere quality). A following step used reasoning to verify the correct quality classification. The previous Monomer value validates the "scientific threshold range" from the "analytical method" part (see section 4.3.2.3). An example to demonstrate the application of SPARQL to determine the qualifier is displayed in Listing 3. It is used to perform ETL operations, i.e., to retrieve analytical data from the triplestore and check if samples exist that exceed the scientific range between 60 and 100 percent SEC Monomer content.

```
SELECT ?secMD ?validSEC
        {
                ?secMD
                a obo:IAO_0000109 ;  # measurement datum
                has_value_specification: ?secVS .

                ?secVS
                a model:SEC_monomer_relarea_percent_vs ;  # implies is_specified_output_of: some :SEC
                has_specified_numerical_value: ?val .

                BIND ((?val >= 60 && ?val <= 100) AS ?validSEC)
        }
```

Listing 3 SPARQL quality classification examples . SEC monomer percent values between 60 and 100 are classified as valid.

The SPARQL code demonstrates how the sample's analytical results (SEC Monomer content) are checked. In SPARQL a “?” marks a variable. Two variables are defined: ?secMD represents the "measurement datum" and ?validSEC is the quality classification result as an attribute. In this case, the ?validSEC was not serialized and only existed in memory, but it was later stored as a separate knowledge triple, which is added to the semantified raw data, extending the existing semantic data stored in the triplestore.

The returned value of the SPARQL from Listing 3 is displayed in Table 9. The first row in the table is the sample from Listing 2. A hypothetical second sample was added as a second row. A real data set would contain around 96 samples, resulting in 96 rows. The first column (?numericalValue) represents the numeric value of the SEC measurement of the samples. The second column (?secMD) represents the measurement datum class. The third column (?validSEC) is the variable that contains the sample's quality classification as a Boolean variable. The valid range was defined between 60 and 100. Therefore, the first sample with

a value 80 is classified as valid. The second sample with a value of 102 is outside of the valid range and, therefore, classified as invalid.

Table 9 SPARQL query results for the analytical method part . The value from the first row is the example from Listing 2, and the resulting classification is true. The second row is a hypothetical example to demonstrate the returned value outside the valid range of 60 to 100 percent Monomer. The quality classification is represented by the last column (?validSEC).

?numericalValue (SEC Monomer AUC%)	**?secMD**	**?validSEC**
99.264	obo:IAO_0000109	true
102.0	obo:IAO_0000109	false

The dashed arrow in Figure 36 between the qualified aggregated metadata (blue) and the SME comparability logic (purple) depicts the storage of the classification results in table form. This is done for a simplistic reason. The classifications are stored in the triplestore as knowledge representation and extend the existing semantic data. The above SPARQL demonstrates that the SEC Monomer value of 99.264 is valid, which is indicated by setting the Analytical method column for the 42°C 21 days temperature stress to true (green overlay). This is applied to all formulations over all stress conditions.

One alternative to utilizing SPARQL for assigning a temporary variable is to employ reasoning to validate or invalidate samples based on class assignment. This can be achieved by assigning the "InvalidMonomerAUCPercentageMD" class for formulations. Listing 4 shows an example of achieving this based on the analytical method range from the previous example. The reasoning code is written in OWL and resembles a query for the inverse ranges from the SPARQL from Listing 3. The conditions use the "EquivalentTo" relation and query for data points of type "measurement datum" that contain the value specification of "SEC Monomer_area_relareapercent" with a specific numerical value that is smaller than 60 or greater than 100. The difference between the SPARQL from Listing 3 and the reasoning from Listing 4 is that SPARQL returns a query. Reasoning assigns a class or condition through implicit information without the return of values. Listing 3 and 4 demonstrate different ways to validate or invalidate samples.

```
Class: :InvalidMonomerAUCPercentageMD
                Annotations: rdfs:label "Invalid monomer AUC percentage SEC measurement datum"

                EquivalentTo:
                        obo:IAO_0000109     # measurement datum
                        and obo:OBI_0001938  # has_value_specification
                                some (model:SEC_monomer_relareapercent_vs
                        and (obo:OBI_0001937 # has_specified_numerical_value
                                some (xsd:decimal[< 60.0] or xsd:decimal[> 100.0])))
```

Listing 4 Reasoning example . The code classifies samples with a monomer percent lower than 60 or greater than 100 as invalid. It is the exact opposite of Listing 3.

In the next step, to achieve true machine comparability, the machine must augment the human-like (chromatography SME) data set comparability decision based on the quality metrics.

5.4 Application of the SME comparability logic

The SME's comparability logic flow chart is applied to prioritize the qualified aggregated metadata. This is achieved by automatically evaluating the qualified aggregated metadata in the triplestore. Depending on the qualified aggregated metadata, each stress condition of each screening is classified into four categories: high, medium, low, and no comparability (represented by the colors green, yellow, purple, and red). This is achieved by evaluating all subordinate qualified aggregated metadata decisions of one qualified aggregated metadata component (SST, Workflow, Analytical method, and sample replicates). In the last step, a total comparability score is determined for the whole screening based on the compatibility score of the individual stress conditions.

Additionally, error calculations are performed for the classification. The error is the discrepancy between a sample's true and observed values. It includes variety and uncertainty from the analytical method, which is influenced by the structural complexity and, thus, variability of biologics, as well as external influences such as temperature, time, humidity, and sample handling during the workflow. The errors are calculated on different levels, as described in chapters 4.3.2.3 and 4.3.2.4.

The SME comparability logic part at the bottom of Figure 36 (purple) shows how the different qualified aggregated metadata are prioritized by applying the SME comparability logic flow for the related stability screening. All qualified aggregated metadata for the temperature stress condition at 40°C 21 days are fulfilled (specified as true in the table), and therefore, this stress condition is classified with a high comparability. The qualified aggregated metadata for the mechanical stress indicates that the workflow/process could not be validated due to a subordinate decision failing the query. The passing of the workflow leads to the classification not being comparable to the related stress condition.

5.5 Overall comparability assessment results

An overall comparability assessment was conducted for three different stability screenings, as depicted in Figure 36. The results for each screening are presented in Table 10. The table shows the resulting comparability assessment for each stress condition of the three screenings. All values for all qualified aggregated metadata decisions are stored in the triplestore in semantic form. The table representation was chosen for simplistic reasons. The table, especially the "stress condition" column, shows that the screenings were conducted unequally due to differing stress conditions. Example screening 3, for example, includes temperature stress at 5°C, but only for 7 and 21 days of incubation time. Compared to example screenings 1 and 2, the 84-day incubation pull point is missing. The analytical method and the SST are passed for all stress conditions. The workflow criteria for screening 3 are not fulfilled since neither positive nor negative controls are used. Without these, no references for the samples exist. These are important to ensure a successful workflow application, including the stresses, which is an overall drawback for comparability. Example screening 1, specifically the mechanical stress ("mech"), indicates that the workflow stress

condition does not meet the required criteria. This indicates that during the mechanical stress execution, some processing issues occur that affect the controls or scientific reasons exist to drop the controls. Either way, this directly impacts the stress's comparability. Furthermore, replicate criteria are only used, for example, in screenings 1 and 3. In screening 2, no replicates are used, therefore, this criterion is not passed, as indicated through the "false" values.

Furthermore, the stress error is calculated from the mean of all standard deviations from the formulation replicates described in chapter 4.2.1.5. No controls are used during screening 2; therefore, the 2 percent Monomer error from the method verification is assigned as stress error. The comparability assessment of screening 1 indicates the highest classification for all stress conditions except for mechanical stress due to invalid controls. Screening 2 is ranked with medium comparability, although no replication is used. Consequently, the 2 percent Monomer error from the method verification is assigned as stress error for all samples. Screening 3 is categorized as incomparable. Although sample replication allows a precise stress error calculation, resulting in an error range between 0.05 and 0.52 Monomer percent for the samples, the workflow criteria are not fulfilled because no controls are used. If the workflow is incomparable, the screening can solely be used to distinguish formulation (sample) differences within the screening but not about other screenings. Table 11 shows a summary of the stress-specific comparability decisions for each screening.

Table 10 Detailed comparability results for the different stress conditions of the three stability screenings. The value "true" indicates that all subordinate-related qualified aggregated metadata decisions for one component (SST, Workflow/process, Analytical method, and Sample Replication) were successfully passed. The value false indicates that at least one subordinate decision was not met.

Example stability screening ID	**Stress condition**	**SST** (4.3.2.1)	**Workflow / Process** (4.3.2.2)	**Analytical method** (4.3.2.3)	**Sample Replication** (4.3.2.4)	**Error source**	**Comparability result**	**Stress error** (Monomer area percent)
1	Unstressed	true	true	true	true	replicates	high	0.02
1	Freeze/Thaw stress	true	true	true	true	replicates	high	0.07
1	Mechanical stress	true	false	true	true	replicates	none	
1	Temp. stress 5°C 84 days	true	true	true	true	replicates	high	0.06
1	Temp. stress 25°C 21 days	true	true	true	true	replicates	high	0.09
1	Temp. stress 25°C 84 days	true	true	true	true	replicates	high	0.21
1	Temp. stress 40°C 7 days	true	true	true	true	replicates	high	0.14
1	Temp. stress 40°C 21 days	true	true	true	true	replicates	high	0.30
2	Unstressed	true	true	true	false	method verification	medium	2.00
2	Freeze/Thaw stress	true	true	true	false	method verification	medium	2.00
2	Mechanical stress	true	true	true	false	method verification	medium	2.00
2	Temp. stress 5°C 84 days	true	true	true	false	method verification	medium	2.00
2	Temp. stress 25°C 21 days	true	true	true	false	method verification	medium	2.00
2	Temp. stress 25°C 84 days	true	true	true	false	method verification	medium	2.00
2	Temp. stress 40°C 7 days	true	true	true	false	method verification	medium	2.00
2	Temp. stress 40°C 21 days	true	true	true	false	method verification	medium	2.00
3	Unstressed	true	false	true	true	replicates	none	0.11
3	Freeze/Thaw stress	true	false	true	true	replicates	none	0.52
3	Mechanical stress	true	false	true	true	replicates	none	0.20
3	Temp. stress 5°C 7 days	true	false	true	true	replicates	none	0.05
3	Temp. stress 5°C 21 days	true	false	true	true	replicates	none	0.10
3	Temp. stress 25°C 7 days	true	false	true	true	replicates	none	0.08
3	Temp. stress 25°C 21 days	true	false	true	true	replicates	none	0.11
3	Temp. stress 40°C 7 days	true	false	true	true	replicates	none	0.06
3	Temp. stress 40°C 21 days	true	false	true	true	replicates	none	0.31

Table 11 Overall example screenings comparability results . The high-level results combine stress and sample-specific comparability results into one screening-specific for each example screening (rows). The subtrees of the aggregated qualified metadata are displayed as columns.

Example stability screening ID	**SST** (4.3.2.1)	**Workflow / Process** (4.3.2.2)	**Analytical method** (4.3.2.3)	**Sample Replication** (4.3.2.4)	**Overall screening comparability result**
1	true	true	true	true	high
2	true	true	true	false	medium
3	true	false	true	true	none

5.6 Error influence within and across screenings

Figure 37 illustrates how errors affect comparability assessments across two screenings. The left screening (example screening 2) involves single samples (red), while the right screening (example screening 3) includes two formulation groups (blue), each with three samples. The Monomer was measured using an SEC method for all samples.

In example screening 2 (red), both samples were analyzed using the regular two percent analytical method error for the ambient temperature (nostress) and the increased temperature stress conditions (25°C at 14 days). In example screening 3 (blue), the replicate error is calculated for each individual formulation due to sample replication during the screening. The use of sample replication allows to calculate a precise error for the sample group, significantly reducing the analytical error, as shown by the smaller vertical error bars ranging from 0.4 to 0.6 percent.

The results indicate that in screening 2, the impact of temperature stress on sample stability is smaller than the two percent stress error (red). This means that a differentiation between formulation stabilities cannot be observed because the differences are overshadowed by the measurement noise (2 percent analytical method error) at ambient and higher temperature stress conditions.

In screening 3, the replicate error (blue) can be calculated due to sample replication. This error is significantly smaller than the stability differences induced by increased temperature stress, indicating that a formulation's stabilizing influence can be detected.

Overall, screening 3 demonstrates an improved quality compared to screening 2, due to the use of sample replication, as indicated by a high comparability ranking in Table 11. The data from Figure 37 and Table 11 emphasizes the importance of errors for comparison between and within stability screenings and highlights its importance during comparability assessments.

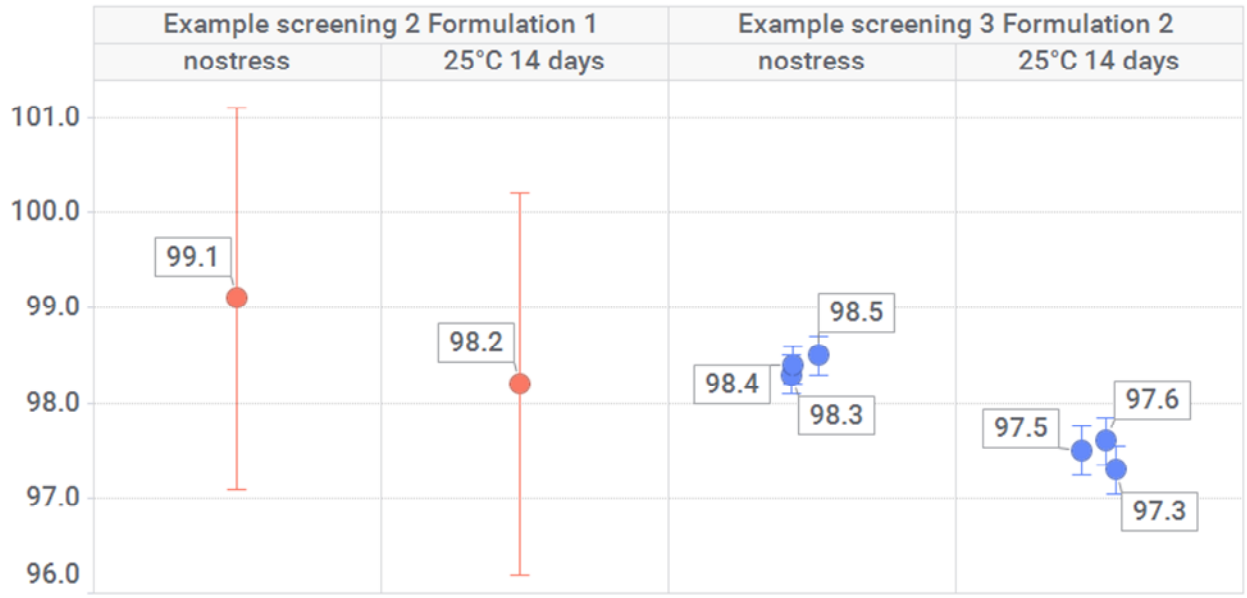


Figure 37 Screening and stress-specific error comparison. The left side (red) shows two sample formulations from example stability screening 2 at ambient (nostress) temperature and after 14 days at 25°C temperature stress condition. The vertical error bars (red) indicate the regular 2 percent analytical method error. The right side (blue) shows two groups of replication formulations (example screening 3) for the same stress conditions as from the right side. The error bars (blue) indicate a reduced error due to sample replication.

6 Discussion

This chapter is focused on the challenges and concepts of data reuse and standardization in R&D and life sciences. Additionally, how the requirements outlined in Chapter 2.4 are fulfilled is elaborated. The importance of data standardization concepts such as the FAIR principles, and the challenges in implementing them are highlighted. A transition from readable data to machine actionable data and the inclusion of quality aspects in FAIR for biologics data is discussed. Furthermore, community-agreed data standardization and implications, such as the need for FAIR experiments, comparability-driven data sets, and collaboration, are discussed. The advantages of semantic data representation are compared to those of table-based formats. Moreover, a generalization of the standardization concept as potential applications in clinical use cases is eluted. Finally, limitations and future research directions are elaborated.

6.1 Data reuse challenges

Data reuse and standardization is a big challenge for R&D and life sciences. Both thrive towards increased data reusability and increased data value. Desperate for a solution, they jump on the old and well-known train in the form of the old solutions they have learned and established over the years. The effort to collect data in classical databases and dashboards. These solutions may work for simple data sets but are only partially sufficient for more complex data types like biologics data. As demonstrated in Chapter 2, data standardization is key to reusing complex data efficiently for advanced analytics.

Recent promising industry-driven solutions to solve the analytical result standardization were only partially successful since they tried to solve the problem with just a greater collection of data – more metadata to describe the most complex laboratory processes and methods. To overcome the increased number of metadata, higher abstracted metadata was used. It was demonstrated that controls can aid in overcoming the required number of metadata. Moreover, biologics data requires a rethinking of the data collection strategy. It was highlighted that any complex data collection, e.g., biologics data collection, must be guided by a purpose before the actual data collection process is started. Furthermore, the goal of the data and metadata collection efforts was not the exact expression of the method and processes but rather the important factor that a scientist requires to make analytical results comparable, which guided our development efforts.

However, translating these learnings into a standardized concept was difficult. The FAIR guiding principles appeared to be a promising solution. Raised from the scientific field in the form of 14 guiding principles, these were broadly accepted by the community and enabled scientific data reuse. However, a concrete implementation has not been suggested and was not intended by the authors since FAIR is not a standard (Mons et al., 2017).

6.2 Verification of the requirements

This section explains how the standardization concept fulfills the requirements from Chapter 2.4.

Before developing the concept, the term and meaning of machine actionability were examined. Machine actionability was redefined, and the FAIRification objective was set to aim for automatic data reuse similar to that of a human operator. This definition aligns with Requirement 1: “Machine actionability so that the machine has the capability to reuse biologics data similar to a human.”

To enable the machine to have an adequate context to set chromatography biologics data sets into context, a data collection strategy concentrated on the purpose of data collection before the actual collection process was initiated. Furthermore, the semantic modeling concentrated on what was necessary to provide a sufficient context to make chromatographic results comparable.

This aligned with Requirement 2: “The standardization concept should provide a sufficient context for the machine to set biologics data results from different data sets in relation to be able to compare different data sets.”.

To fulfill Requirement 3: “A sufficient collection of relevant metadata that enables manageable data integration.” the qualified aggregated metadata were introduced. These enable an adequate set of relevant metadata that reduces the required amount of chromatography data set metadata to a manageable size.

Furthermore, these fulfill Requirement 4: “The metadata should reflect data quality to identify if a data set is useful for the application of predictive concepts so that comparability with other data sets can be determined.” Since they reflect the quality of a data set, it is necessary to evaluate whether a data set is comparable and has sufficient quality for further reuse, e.g., the application of predictive concepts.

The RDF knowledge triples, and their storage allow machine access and comply with Requirement 5:” Data representation in an ideal format to store complex data so that the machine can automatically access it.”

Moreover, existing ontologies were used and adapted during the development of the semantic domain model. These covered Requirement 6: “Instead of the development of new concept, the reuse of existing concepts is a desired goal.”.

To meet Requirement 7 – “The concept should enable to automate the scientist’s logic for data set comparability, so that the machine can automatically act on the data, without human intervention” – a logic for data set comparability assessment was developed into a flow chart. This enabled an automated assessment similar to an autonomously acting scientist so that the machine could automatically decide on the data set comparability without human intervention.

6.3 Interpretation of FAIR

FAIR implementations in the community are increasingly identified to end in machine readable data. In contrast, a machine actionable state is required to enable the machine to automatically act on the data. The perception is shifting from a solely readable state towards automated reuse intended through the machine. The provided FAIRification examples from the initial publication (Wilkinson et al., 2016) mainly focus on data discoverability. A revised comment further describes that machine actionable resources "to maximally fulfill the FAIR guidelines must utilize a widely-accepted machine readable framework for data and knowledge representation" (Commission High Level Expert Group on the European Open Science Cloud, 2016). Chen et al. have identified the need to create FAIR AI-ready training datasets (Chen et al., 2022). Barend Mons commented that FAIR enables more effective AI and could therefore be seen as "fully AI ready" (Mons, 2020) (Wilkinson et al., 2016). McKinstry et al. understand "AI-ready" as a self-explanatory state (McKinstry et al., 2021) so that the machine can understand the data. However, they end before autonomous reuse of the data through the machine, which would be required to generate an AI-ready data.

This thesis goes a step further and adds the "Fully" to "AI-ready" by demonstrating how standardization concepts like FAIR can be used to not only make the machine understand the data but automatically make use of and act on it to enable true machine actionability. Similarly, in Chapter 4, the purpose for reuse was identified to be greater than solely findable data. Machine actionability was specified as the goal for humans and machines so that the machine can automatically reuse the data.

Therefore, the machine must be able to identify the quality of biologics data sets to determine their suitability and comparability. Harrow et al. have identified that FAIR only indirectly covers data quality aspects, which are required to increase the data value of research data (Harrow et al., 2022). This is particularly true for biologics data because the machine needs additional information, e.g., quality aspects, incentive schemes, and privacy regulation metrics, to enhance the reuse of biological data through the machine, which extends the initial FAIR principles (Holub et al., 2018). The qualified aggregated metadata were introduced in Chapter 4 to represent the aspects of biologics quality. The qualified aggregated metadata obscures the distinction between metadata and results.

On the one hand, they can be considered metadata since they serve as metadata necessary for making comparability decisions. On the other hand, they are experimentally measured readouts obtained from analytical experiments, such as the SST, and thus can be considered results. However, determining how to incorporate quality aspects in a FAIR manner was far from straightforward.

6.4 FAIR implementations

There is no clear path that demonstrates how to correctly implement FAIR. Organizations For our purpose and domain, no semantic model existed. Similar to others (Garabedian et al., 2022)(van der Velde et al., 2022)(Schoening et al., 2022) and to converge with existing solutions, the decision was made to self-develop a biologics semantic model under the use of parts from existing ontologies. The approach adds to the plethora of implementations. Semantically aligning multiple ontologies is a challenging task that may be resolved through

semantic interoperability (Nagowah et al., 2018). Achieving interoperability between ontologies and systems can potentially increase data value worldwide (Obrst, 2003). A community-agreed path to implement scientific data standards can potentially establish FAIR as the gold standard (as a data standardization concept).

Comparable examples could be the Fraunhofer Society's mp3 standard in the 90s that revolutionized the audio industry and turned into the audio file standard or Google's recent AlphaFold breakthrough to reliably predict protein folding, which could evolve as the new norm for protein structure modeling. Such a standard could guide future implementations and help to align the community. A mentionable example of a community alignment is the FAIR Cookbook[14] from the FAIRplus initiative[15]. The Cookbook collects practical examples, implementation choices, best practices, and tutorials throughout the FAIR community (Rocca-Serra et al., 2023). It is an ideal source to start any FAIRification journey. This thesis tried to enable convergence with the FAIR ecosystem by using and adapting existing solutions such as the RDA indicators FAIR assessment from the FAIR Cookbook. Moreover, this manuscript can be seen as a detailed FAIR recipe describing a FAIR journey from start to end. Some parts, e.g., the OBI adaptation, will likely be converted into a FAIR cookbook recipe shortly.

The question is: Is this yet another FAIR implementation in the plethora of solutions, or is the presented standardization concept something beyond the FAIR environment? Clearly, the concept is influenced by the FAIR principles, but it exceeded the scope of FAIR implementations for a solely semantical representation of the domain through metadata. However, this work also did not present a concrete industry standard or format for biologics data. The here presented concept resembles a strategy to solve complex data integration within a specific niche. But how can the concept be transferred to other domains that struggle with complex data reuse? The development efforts were directed towards achieving result comparability, which prompted the inclusion of the scientist's decisions (SME) in determining the comparability of the data sets. Although, the scientist's decision is based on his experience, may depend on the environment (company, department, function) and could potentially change in another setting, the approach to augment the scientists decision, can be transferred to other domains. This can be applied in any field to guide a standardization strategy independent of the domain.

6.5 Implications

The presented concept has implications apart from an increased value for scientific data. Additionally, it sharpens the internal perception of experiments sustainability and enables not only FAIR data but FAIR experiments and workflows. As Borycz et al. pointed out, automated FAIR-compliant workflows are required for future data reuse (Borycz & Carroll, 2020). Ideally, generating these leads to highly comparable data sets, a desired goal. Through

[14] https://faircookbook.elixir-europe.org/

[15] https://fairplus-project.eu/

the presented concept, these comparability influencing factors were made available in a tangible form for the first time.

Furthermore, the assumption that lab automation automatically results in highly comparable data sets is misleading. Not all experimental data sets are categorized with high comparability due to pipeline molecule or project constraints. In many cases, experiments are executed to yield a single scientific conclusion without further reuse in mind. The concept highlights the requirements to generate comparable data sets. Therefore, it can guide upcoming experimental planning and enable us to reflect on the decision to which extent comparability-driven experiments for further reuse are desired.

These comparability-driven data sets are required to demonstrate the return on investment of such a concept. Advanced analytics accuracy will increase with increased number of these high-quality data sets. In addition, value will be generated in the long run. Therefore, cultural and infrastructure change within organizations is required (Borycz & Carroll, 2020)(Wise et al., 2019)(Alharbi et al., 2021). The value of data must be highlighted, especially in pharmaceutical R&D organizations. This comes hand in hand with the will to invest and provide the necessary resources to accomplish this. In addition, data strategies must reflect this will to highlight the importance of more standardized data. This requires a change from both sides. 1.) From top-down management through a reliable commitment to hire or develop more standardization-aware data stewards and to provide them with the required resources. Furthermore, to lay out and follow a clear vision throughout the organization. 2.) From the bottom up, to lay aside the siloed and restrictive data sharing mindset and enable data convergence of solutions across organizations through collaboration.

Moreover, collaborating closely with other pharmaceutical companies to elaborate the best standardization strategy and best practices in pre-competitive consortia will bring additional benefits. Alignment with academic partners will help follow the latest trends and be profitable for all sides.

6.6 Benefits of semantics

Semantic knowledge representations offer multiple benefits compared to table-based and relational data storage. These include:

1. Increased data reuse: Semantic representations express data and metadata in a standardized format that is understandable for both humans and machines. Each concept and entity is unambiguous, identifiably, and findable using identifiers, resulting in increased data reuse.

2. Reduced manual data integration effort: The standardized representation of data using ontologies and semantic models reduces the need for manual data integration and preparation, saving time and effort when integrating data from different sources.

3. Enhanced data interoperability: Semantic models and ontologies use RDF and OWL as a common language, which benefits data interoperability. This enables easier data exchange between different systems. Moreover, it facilitates semantic mapping, allowing different semantic models to be mapped to a common one by aligning concepts, terms, and relationships.

4. Easy data extension: Semantic data models and ontologies have a hierarchical structure that promotes data extension with minimal effort. New data can be added as nodes without changing the underlying semantic model, making it easier to adapt to changing requirements and modify existing data.

5. Support of inheritance and subtyping: Ontologies support inheritance and subtyping, allowing new concepts to inherit properties and characteristics from existing concepts. This promotes data extension and reusability of existing concepts, allowing for quick inclusion of new concepts.

6. Reasoning: Semantic data models and ontologies enable reasoning, which involves automatic inference of knowledge based on defined rules and relationships. This helps derive new insights and make logical inferences from the available data. Additionally, reasoning can help fill in missing gaps in data.

7. Improved data quality: Semantic representations enhance data consistency and accuracy by providing standardized sets of concepts and relationships, improving data quality and reliability.

Overall, the use of semantic representations and ontologies brings numerous benefits. However, it is also essential to mention disadvantages in comparison to table-based or relational representations, such as:

1. Complexity: Semantic concepts and ontologies can be more complex to design and implement. Furthermore, they require a specific domain knowledge and expertise during the creation process.

2. Performance: In certain scenarios, relational systems can outperform semantic representations due to the optimization and maturity of relational-based systems.

Relational systems are more mature because they are historically older and more established.

3. Data volume: The data volume for semantic representations may increase more than that for relational concepts due to the triple concept of subject, predicate, and object as knowledge representation. Large-scale semantic data sets with increased data volume can negatively impact performance.

4. Ontology ecosystem: Numerous ontologies exist for specific domains and problems. It can be challenging to find and choose the best ontology to solve the problem at hand.

As semantic technologies are increasingly adopted, technological disadvantages are evolving and improving. Wider adaptation may promote semantic concepts and present a viable option to traditional table-based and relational solutions. However, recent advances enable the transformation from relational solutions to knowledge-based semantic ones by database-to-ontology mappings (Spanos et al., 2010).

6.7 Generalization

Apart from the scientific domain, standardization concepts draw increasing attention in the health care community, since it also faces the challenge of complex data integration. The value of reusable data is already recognized (Kalendralis et al., 2021)(Sinaci et al., 2020). Understanding a patient's likeliness to be affected by certain illnesses is simple due to genome sequencing technology. These results do not require a greater context for integration since they are exact and reproducible. Understanding and integrating patient health care data across a diverse clinical ecosystem, including questionnaires and analytical results, adds to the complexity of the problem. The introduction of CDE was an attempt to increase clinical and health care data value and reuse through standardization. These standardized questions can be answered at sites, studies, or clinical trials. Kush et al. analyzed reasons for the poor use of CDEs and suggested an improved use of these in alignment with FAIR to increase data use and safe costs (Kush et al., 2020). The proposed concept could be generalized and aligned with their work.

A generalization of the concept could guide standardization efforts for clinical use cases, i.e., enabling a machine to automatically suggest patients for a medical study cohort inclusion based on their medical records. CDEs can provide relevant context information about a patient and the criteria and conditions for study inclusion, like the qualified aggregated metadata (Sheehan et al., 2016). Existing ontologies such as CMDO can help build a machine accessible semantic representation to store the CDEs (Kim et al., 2019). An alternative option may be the representation through FAIR Digital Objects (FDOs) (Queralt-Rosinach et al., 2022). Furthermore, the health care professional's comparability logic (SME) can be augmented into a flow chart that automatically categorizes and suggests patients, which is suitable for inclusion in a medical study.

In the last step, the health care professional can decide on the inclusion of the patient. Demonstrating the additional value of increased data reuse through standardization may help to transform siloed and protective data mentalities into an increased data-sharing mindset. The biggest challenge in recent years, the COVID-19 pandemic, and the delayed data

reuse, which led to questionable decisions, further highlight the importance of reusable healthcare system data (Queralt-Rosinach et al., 2022). Further investigation on the application of the presented standardization concept in health care data and clinical settings would be beneficial.

6.8 Limitations and future work

The presented concept relies on popular implementation choices to fulfill FAIR requirements. It was developed in a generalizable way to use the approach in other domains and organizations for their data sets. Nevertheless, some limitations must be outlined, and adaptations must be performed before the method is suitable for implementation in other fields. One example is that the augmentation of the SME is subjective. Although the choice of comparability influencing factors is aligned with scientific best practices within the chromatography field, other chromatography SMEs could choose different influencing factors based on their experience. This is not a problem within one department because the scientists should consent to a standardized evaluation, but it could instead require SMEs across departments to agree on certain comparability factors.

Furthermore, the comparability factors, including the quality criteria, are valid for biologics focusing on chromatography results but can vary for other domains or analytical methods apart from the chromatography and require adaptation. Consequently, this implies the need to change the semantic model and would require selecting a different, better-suited ontology. This could, again, lead to challenging semantic interoperability between different stakeholders (Harrow et al., 2019). However, the process of SME augmentation of the comparability factors and the decision is the mandatory step to enable the machine to take human-like actions to enable true machine actionability. Nevertheless, these considerations could be explored in research efforts following this thesis.

Moreover, the comparability factors and SME decisions were solely based on classical mAbs. Other mAb variants, e.g., highly engineered mAb variants such as DVDs, ADCs, and fusion proteins, were not tested for comparability. Certain aspects of the concept need to be adapted to effectively determine the comparability of data sets derived from these compounds. Likewise, company-internal well-known mAbs were used as positive and negative protein controls. These controls allow data comparability decisions within the company. However, sharing internal controls with competing companies will lead to legal concerns. Externally, organizations must use their well-known company internal molecules as controls. So far, no legally agreed upon universal control (standard molecule) exists to be used as a reference without these concerns. Such an industry-standard molecule would be a valuable key element in applying the presented concept to allow comparability between organizations.

In addition to establishing an industry-standard molecule, a standardized container format that includes the data, its semantic expression, and the SME logic could be beneficial. The analytical data is physically separated from their semantic representation through RDF triples in the triplestore. The semantic model based on the adapted OBI ontology is intrinsically represented through the data but can be serialized as a Turtle file. These resources should be encapsulated in one place in a container. This would also be beneficial for their long-time archive. The ADF could be a potential solution since it can include raw and semantic data, ontology models, and descriptions in an arbitrary format.

7 Summary

The pharmaceutical industry aims to reuse scientific data for machine learning and artificial intelligence. The most critical analytical method in biologics drug development and characterization is high-performance liquid chromatography (HPLC). This thesis underscores the challenge of data integration due to the complexity of biologics and HPLC technology, seeking to bridge this gap by adapting standardization concepts like the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles to make data interpretable for machines as well as humans.

The work focuses on the development of a generic standardization concept that enables machine actionable biologics formulation data. Machine actionability refers to the complete machine-based interpretation of dataset quality, facilitating automatic comparability of datasets. Achieving this requires an adaptation of existing standardization concepts such as FAIR to enable the automation of decision-making and data set comparability evaluation. Furthermore, current semantic models, such as the semantic HPLC model from the Allotrope Foundation for chromatography devices, are only partially suitable for the intended purpose of enabling machine-actionable data. This is because they predominantly focus on the reproducibility of experiments rather than focusing on data reuse in terms of result comparability and autonomous data evaluation. To overcome this limitation, a new standardization concept was developed that consists of three consecutive steps:

1. The creation of a semantic model to convert experimental raw data into a semantic format using web technology concepts such as ontologies, making it machine-readable and giving it meaning for machines and humans. The model focuses on formulation sciences data, particularly stability screening data, and incorporates analytical chromatography results.

2. The definition of aggregated metadata as a set of rule-based acceptance criteria, thresholds, and ranges that allow for the determination of dataset quality. They are based on the subject matter expert (SME), which resembles a scientific expert for HPLC. The aggregated metadata represent the initial step beyond machine readability towards achieving machine actionability. These metadata are combined at a higher level and depend on lower-level metadata from the raw dataset, including formulation composition, analytical method, stress conditions, and experiment processing.

3. The formalization and automation of the SME comparability logic. In assessing the quality of datasets, the logic replicates the manual data assessment process of an HPLC SME. By automatically evaluating aggregated metadata expressed in the semantic model, dataset quality can be assessed without human intervention. The output of this evaluation is a comparability classification, which categorizes datasets into four groups based on their quality and suitability for machine learning. The classification and error calculations are performed on multiple levels.

In conclusion, this thesis presents a novel concept to standardizing biologics formulation data, making it machine-readable and actionable, thus enhancing data comparability and reusability in the pharmaceutical industry. The concept has been developed in a generic way, to allow for the implementation in other domain

8 Supplementary information

List of supplementary information

Figure SI 1 Overarching qualified aggregated metadata decision tree *98*
Figure SI 2 RDA indicators FAIR assessment of the HTS data set *99*
Figure SI 3 RDA indicators FAIR assessment of the BiologicsDashboard data set *100*
Figure SI 4 RDA indicators FAIR assessment of the PharmaCircle data set *101*

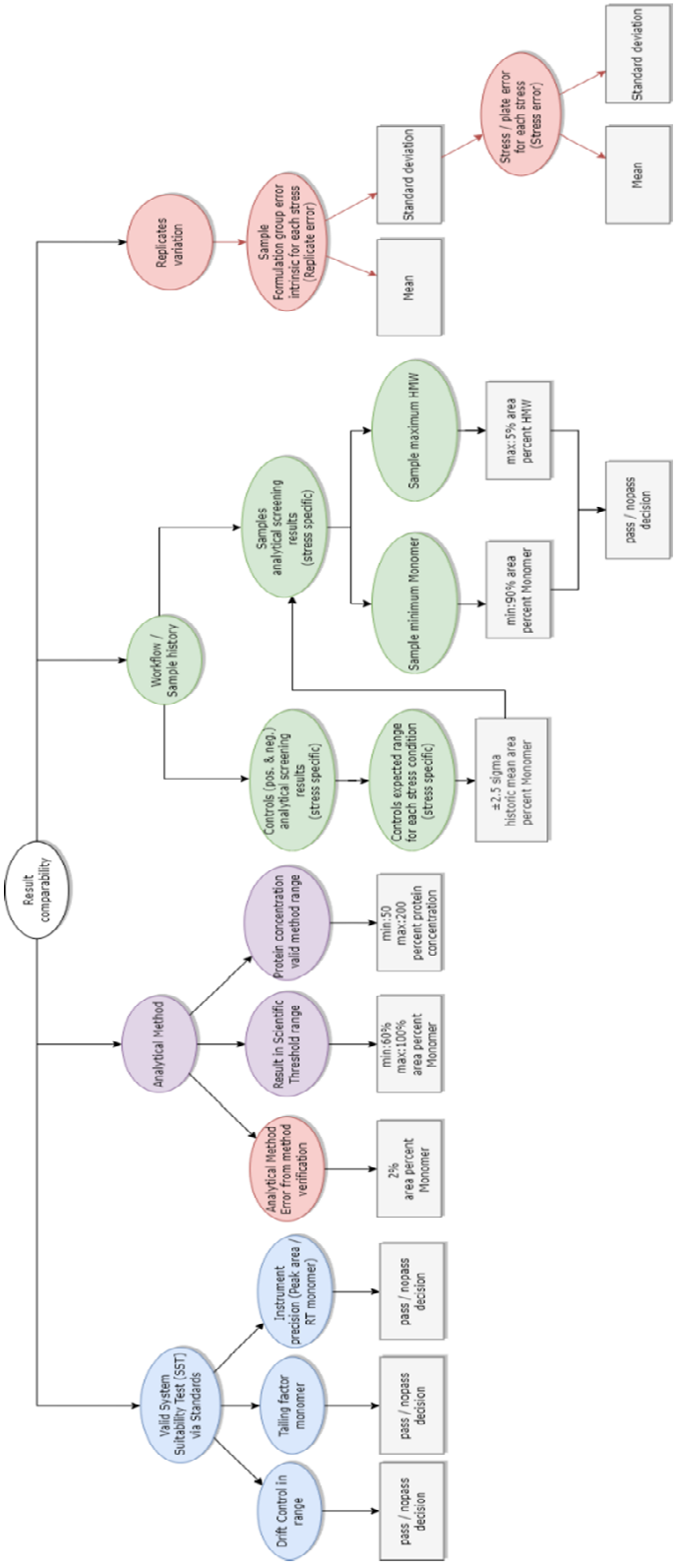


Figure SI 1 Overarching qualified aggregated metadata decision tree. The tree displays the four sub trees combined into one. It includes all threshold, criteria and ranges which influence a SEC analytical result comparability.

	Sub-principle	ID	Indicator	Priority	HTS Assessment overall	HTS Assessment Essential	HTS Assessment non-essential
Findable	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	0	0	-
Findable	F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	0	0	-
Findable	F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1	1	-
Findable	F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	0	0	-
Findable	F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	1	1	-
Accessible	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important	0	-	0
Accessible	A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential	1	1	-
Accessible	A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	1	1	-
Accessible	A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	0	0	-
Accessible	A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	0	0	-
Accessible	A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	1	1	-
Accessible	A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	1	1	-
Accessible	A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	0	-	0
Accessible	A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	0	0	-
Accessible	A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	0	-	0
Accessible	A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful	1	-	1
Accessible	A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential	0	0	-
Interoperable	I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	0	-	0
Interoperable	I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	0	-	0
Interoperable	I3	RDA-I3-01M	Metadata includes references to other metadata	Important	0	-	0
Interoperable	I3	RDA-I3-01D	Data includes references to other data	Useful	0	-	0
Interoperable	I3	RDA-I3-02M	Metadata includes references to other data	Useful	0	-	0
Interoperable	I3	RDA-I3-02D	Data includes qualified references to other data	Useful	0	-	0
Interoperable	I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	0	-	0
Interoperable	I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	0	-	0
Reusable	R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	1	1	-
Reusable	R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	0	0	-
Reusable	R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	0	-	0
Reusable	R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	0	-	0
Reusable	R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important	1	-	1
Reusable	R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful	0	-	0
Reusable	R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	1	1	-
Reusable	R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	1	1	-
Reusable	R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	0	0	-
Reusable	R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important	0	-	0

	Sum	11	9	2
Score total = # fullfilled / # total (in relevant category)	Score total	26.83%	21.95%	4.88%
Score applicable = # fullfilled / (# total - # NA) (in relevant category)	Score applicable	26.83%	21.95%	4.88%
%NA = # NA / # total (in relevant category)	% NA	0.00%	0.00%	0.00%
	Findable sub sum	2	2	0
Findable subscore		28.57%	28.57%	0.00%
Findable subscore applicable		28.57%	28.57%	0.00%
Findable subscore %NA		0.00%	0.00%	0.00%
	Accessible sub sum	5	4	1
Accessible subscore		41.67%	33.33%	8.33%
Accessible subscore applicable		41.67%	33.33%	8.33%
Accessible subscore %NA		0.00%	0.00%	0.00%
	Interoperable sub sum	0	0	0
Interoperable subscore		0.00%	0.00%	0.00%
Interoperable subscore applicable		0.00%	0.00%	0.00%
Interoperable subscore %NA		0.00%	0.00%	0.00%
	Reusable sub sum	4	3	1
Reusable subscore		40.00%	30.00%	10.00%
Reusable subscore applicable		40.00%	30.00%	10.00%
Reusable subscore %NA		0.00%	0.00%	0.00%

Figure SI 2 RDA indicators FAIR assessment of the HTS data set including the FAIR assessment results

	Sub-principle	ID	Indicator	Priority	BiologicsDashboard Assessment overall	BiologicsDashboard Assessment Essential	BiologicsDashboard Assessment non-essential
Findable	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	0	0	-
Findable	F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	0	0	-
Findable	F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1	1	-
Findable	F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	0	0	-
Findable	F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	0	0	-
Accessible	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important	1	-	1
Accessible	A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential	0	0	-
Accessible	A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	0	0	-
Accessible	A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	0	0	-
Accessible	A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	0	0	-
Accessible	A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	0	0	-
Accessible	A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	0	0	-
Accessible	A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	0	-	0
Accessible	A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	0	0	-
Accessible	A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	0	-	0
Accessible	A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful	1	-	1
Accessible	A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential	0	0	-
Interoperable	I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	0	-	0
Interoperable	I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	0	-	0
Interoperable	I3	RDA-I3-01M	Metadata includes references to other metadata	Important	1	-	1
Interoperable	I3	RDA-I3-01D	Data includes references to other data	Useful	1	-	1
Interoperable	I3	RDA-I3-02M	Metadata includes references to other data	Useful	1	-	1
Interoperable	I3	RDA-I3-02D	Data includes qualified references to other data	Useful	0	-	0
Interoperable	I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	0	-	0
Interoperable	I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	0	-	0
Reusable	R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	0	0	-
Reusable	R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	0	0	-
Reusable	R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	0	-	0
Reusable	R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	0	-	0
Reusable	R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important	1	-	1
Reusable	R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful	0	-	0
Reusable	R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	1	1	-
Reusable	R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	1	1	-
Reusable	R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	0	0	-
Reusable	R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important	0	-	0

	Sum	9	3	6
Score total = # fullfilled / # total (in relevant category)	Score total	21.95%	7.32%	14.63%
Score applicable = # fullfilled / (# total - # NA) (in relevant category)	Score applicable	21.95%	7.32%	14.63%
%NA = # NA / # total (in relevant category)	% NA	0.00%	0.00%	0.00%
	Findable sub sum	1	1	0
Findable subscore		14.29%	14.29%	0.00%
Findable subscore applicable		14.29%	14.29%	0.00%
Findable subscore %NA		0.00%	0.00%	0.00%
	Accessible sub sum	2	0	2
Accessible subscore		16.67%	0.00%	16.67%
Accessible subscore applicable		16.67%	0.00%	16.67%
Accessible subscore %NA		0.00%	0.00%	0.00%
	Interoperable sub sum	3	0	3
Interoperable subscore		25.00%	0.00%	25.00%
Interoperable subscore applicable		25.00%	0.00%	25.00%
Interoperable subscore %NA		0.00%	0.00%	0.00%
	Reusable sub sum	3	2	1
Reusable subscore		30.00%	20.00%	10.00%
Reusable subscore applicable		30.00%	20.00%	10.00%
Reusable subscore %NA		0.00%	0.00%	0.00%

Figure SI 3 RDA indicators FAIR assessment of the BiologicsDashboard data set including the FAIR assessment results

	Sub-principle	ID	Indicator	Priority	PharmaCircle Assessment overall	PharmaCircle Assessment Essential	PharmaCircle Assessment non-essential
Findable	F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	0	0	-
Findable	F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	0	0	-
Findable	F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	0	0	-
Findable	F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	1	1	-
Findable	F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	0	0	-
Findable	F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	0	0	-
Accessible	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important	0	-	0
Accessible	A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential	0	0	-
Accessible	A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	0	0	-
Accessible	A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	0	0	-
Accessible	A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	0	0	-
Accessible	A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	1	1	-
Accessible	A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	1	1	-
Accessible	A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	0	-	0
Accessible	A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	0	0	-
Accessible	A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	0	-	0
Accessible	A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful	1	-	1
Accessible	A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential	0	0	-
Interoperable	I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important	0	-	0
Interoperable	I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	0	-	0
Interoperable	I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	0	-	0
Interoperable	I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	0	-	0
Interoperable	I3	RDA-I3-01M	Metadata includes references to other metadata	Important	1	-	1
Interoperable	I3	RDA-I3-01D	Data includes references to other data	Useful	1	-	1
Interoperable	I3	RDA-I3-02M	Metadata includes references to other data	Useful	1	-	1
Interoperable	I3	RDA-I3-02D	Data includes qualified references to other data	Useful	0	-	0
Interoperable	I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	0	-	0
Interoperable	I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	0	-	0
Reusable	R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	0	0	-
Reusable	R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential	0	0	-
Reusable	R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	0	-	0
Reusable	R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	0	-	0
Reusable	R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important	1	-	1
Reusable	R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful	0	-	0
Reusable	R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	1	1	-
Reusable	R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	0	0	-
Reusable	R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential	0	0	-
Reusable	R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important	0	-	0

	Sum	9	4	5
Score total = # fullfilled / # total (in relevant category)	Score total	21.95%	9.76%	12.20%
Score applicable = # fullfilled / (# total - # NA) (in relevant category)	Score applicable	21.95%	9.76%	12.20%
%NA = # NA / # total (in relevant category)	% NA	0.00%	0.00%	0.00%
	Findable sub sum	1	1	0
Findable subscore		14.29%	14.29%	0.00%
Findable subscore applicable		14.29%	14.29%	0.00%
Findable subscore %NA		0.00%	0.00%	0.00%
	Accessible sub sum	3	2	1
Accessible subscore		25.00%	16.67%	8.33%
Accessible subscore applicable		25.00%	16.67%	8.33%
Accessible subscore %NA		0.00%	0.00%	0.00%
	Interoperable sub sum	3	0	3
Interoperable subscore		25.00%	0.00%	25.00%
Interoperable subscore applicable		25.00%	0.00%	25.00%
Interoperable subscore %NA		0.00%	0.00%	0.00%
	Reusable sub sum	2	1	1
Reusable subscore		20.00%	10.00%	10.00%
Reusable subscore applicable		20.00%	10.00%	10.00%
Reusable subscore %NA		0.00%	0.00%	0.00%

Figure SI 4 RDA indicators FAIR assessment of the PharmaCircle data set including the FAIR assessment results

9 References

Ahuja, S., Roth, M., Gangadharaiah, R., Schwarz, P., & Bastidas, R. (2016). Using Machine Learning to Accelerate Data Wrangling. *IEEE International Conference on Data Mining Workshops, ICDMW*. https://doi.org/10.1109/ICDMW.2016.0055

Alharbi, E., Skeva, R., Juty, N., Jay, C., & Goble, C. (2021). Exploring the current practices, costs and benefits of fair implementation in pharmaceutical research and development: A qualitative interview study. *Data Intelligence*, *3*(4), 507–527. https://doi.org/10.1162/dint_a_00109

Allotrope-HPLC-JSON-Schema. (n.d.). *Allotrope GitLab Repository - Liquid Chromatography JSON Schema File*. Retrieved April 25, 2024, from https://gitlab.com/allotrope-public/asm/-/blob/main/json-schemas/adm/liquid-chromatography/REC/2024/03/liquid-chromatography.tabular.embed.schema.json

Allotrope-HPLC-SampleJSON. (n.d.). *Allotrope GitLab Repository - Liquid Chromatography Sample JSON File*. Retrieved April 25, 2024, from https://gitlab.com/allotrope-public/asm/-/blob/main/test/adm/liquid-chromatography/REC/2024/03/liquid-chromatography.json

Aloulen, Z., Belhajjame, K., Grigori, D., & Acker, R. (2019). A Domain-Independent Ontology for Capturing Scientific Experiments. *Communications in Computer and Information Science*, *1040*, 53–68. https://doi.org/10.1007/978-3-030-30284-9_4

Alt, N., Zhang, T. Y., Motchnik, P., Taticek, R., Quarmby, V., Schlothauer, T., Beck, H., Emrich, T., & Harris, R. J. (2016). Determination of critical quality attributes for monoclonal antibodies using quality by design principles. *Biologicals*. https://doi.org/10.1016/j.biologicals.2016.06.005

Arp, R., & Smith, B. (2008). Function, Role, and Disposition in Basic Formal Ontology. *Nature Precedings*. https://doi.org/10.1038/npre.2008.1941.1

Aucamp, J. P., Cosme, A. M., Lye, G. J., & Dalby, P. A. (2005). High-throughput measurement of protein stability in microtiter plates. *Biotechnology and Bioengineering*, *89*(5), 599–607. https://doi.org/10.1002/BIT.20397

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., … Zheng, J. (2016). The Ontology for Biomedical Investigations. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0154556

Beyan, O., Emam, I., Rocca-Serra, P., Sansone, S.-A., Juty, N., Alharbi, E., Wood, C., Henderson, D., Burdett, T., & Konopko, M. (2021). D2.5 FAIRplus FAIR Data Maturity Framework. *Zenodo*. https://doi.org/10.5281/ZENODO.5040592

Borman, P. (2021). Reducing Uncertainty of an Analytical Method through Efficient Use of Replication. *Pharmaceutical Technology Europe*, *33*(4), 37–42. https://www.researchgate.net/publication/350995290

Borycz, J., & Carroll, B. (2020). Implementing FAIR data for people and machines: Impacts and implications-Results of a research data community workshop. *Information Services and Use*, *40*(1–2), 71–85. https://doi.org/10.3233/ISU-200083

Carl, B. (2018). *rdflib: A high level wrapper around the redland package for common rdf applications*. Zenodo. https://doi.org/10.5281/zenodo.1098478

Chan, A. C., & Carter, P. J. (2010). Therapeutic antibodies for autoimmunity and inflammation. *Nature Reviews Immunology*, *10*(5), 301–316. https://doi.org/10.1038/nri2761

Chapman, T. (2003). Lab automation and robotics: Automation on the move. *Nature*. https://doi.org/10.1038/421661a

Chen, Y., Huerta, E. A., Duarte, J., Harris, P., Katz, D. S., Neubauer, M. S., Diaz, D., Mokhtar, F., Kansal, R., Park, S. E., Kindratenko, V. V., Zhao, Z., & Rusack, R. (2022). A FAIR and AI-ready Higgs boson decay dataset. *Scientific Data 2022 9:1*, *9*(1), 1–10. https://doi.org/10.1038/s41597-021-01109-0

Chirino, A. J., & Mire-Sluis, A. (2004). Characterizing biological products and assessing comparability following manufacturing changes. *Nature Biotechnology 2004 22:11*, *22*(11), 1383–1391. https://doi.org/10.1038/nbt1030

Commission High Level Expert Group on the European Open Science Cloud. (2016). *Realising the European Open Science Cloud: first report and recommendations*. *June*, 1–19. https://doi.org/10.2777/940154

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, *22*(2), 151–185. https://doi.org/10.1016/S0167-6296(02)00126-1

Driving Innovation with the Allotrope Framework - Astrix. (n.d.). Retrieved February 28, 2023, from https://astrixinc.com/driving-innovation-allotrope-framework/

Drug-Dev.com. (2021). *Drug-Dev.com - PharmaCircle definition*. 2021. https://drug-dev.com/company-profiles/6070/

DuCharme, B., & Beijing. (2013). *Learning SPARQL Querying and Updating with SPARQL 1.1*. https://books.google.de/books?hl=de&lr=&id=j2kXeNeZ00YC&oi=fnd&pg=PR7&dq=sparql&ots=-JXG9bkkTu&sig=GnzJWy4KQsiZZsXR_xl3IwZNxbc#v=onepage&q=sparql&f=false

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., & Deane, C. M. (2014). SAbDab: The structural antibody database. *Nucleic Acids Research*, *42*(D1), D1140–D1146. https://doi.org/10.1093/nar/gkt1043

European Commission. (2019). *Cost-benefit analysis for FAIR research data*. https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

FAIRplus. (n.d.). *FAIRification Process - GO FAIR*. Retrieved March 18, 2023, from

https://www.go-fair.org/fair-principles/fairification-process/

Fekete, S., Beck, A., Veuthey, J. L., & Guillarme, D. (2014). Theory and practice of size exclusion chromatography for the analysis of protein aggregates. *Journal of Pharmaceutical and Biomedical Analysis*, *101*, 161–173. https://doi.org/10.1016/j.jpba.2014.04.011

Fiser, A. (2010). Template-based protein structure modeling. In *Methods in molecular biology (Clifton, N.J.)*. https://doi.org/10.1007/978-1-60761-842-3_6

Fürber, C. (2016). Semantic Technologies. In *Data Quality Management with Semantic Technologies* (pp. 56–68). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12225-6_4

Garabedian, N. T., Schreiber, P. J., Brandt, N., Zschumme, P., Blatter, I. L., Dollmann, A., Haug, C., Kümmel, D., Li, Y., Meyer, F., Morstein, C. E., Rau, J. S., Weber, M., Schneider, J., Gumbsch, P., Selzer, M., & Greiner, C. (2022). Generating FAIR research data in experimental tribology. *Scientific Data 2022 9:1*, *9*(1), 1–11. https://doi.org/10.1038/s41597-022-01429-9

Garcia, L., Bolleman, J., Gehant, S., Redaschi, N., Martin, M., Bateman, A., Magrane, M., Orchard, S., Raj, S., Ahmad, S., Alpi, E., Bowler, E., Britto, R., Bursteinas, B., Bye-A-Jee, H., Dogan, T., Garmiri, P., Georghiou, G., Gonzales, L., … Zhang, J. (2019). FAIR adoption, assessment and challenges at UniProt. *Scientific Data 2019 6:1*, *6*(1), 1–4. https://doi.org/10.1038/s41597-019-0180-9

Goswami, S., Wang, W., Arakawa, T., & Ohtake, S. (2013). Developments and Challenges for mAb-Based Therapeutics. *Antibodies 2013, Vol. 2, Pages 452-500*, *2*(3), 452–500. https://doi.org/10.3390/ANTIB2030452

Guizzardi, G. (2020). Ontology, ontologies and the "i" of fair. *Data Intelligence*, *2*(1–2), 181–191. https://doi.org/10.1162/dint_a_00040

Hanna, J., Joseph, E., Brochhausen, M., & Hogan, W. R. (2013). Building a drug ontology based on RxNorm and other sources. *Journal of Biomedical Semantics*, *4*(1), 1–9. https://doi.org/10.1186/2041-1480-4-44

Harrow, I., Balakrishnan, R., Jimenez-Ruiz, E., Jupp, S., Lomax, J., Reed, J., Romacker, M., Senger, C., Splendiani, A., Wilson, J., & Woollard, P. (2019). Ontology mapping for semantically enabled applications. *Drug Discovery Today*, *24*(10), 2068–2075. https://doi.org/10.1016/j.drudis.2019.05.020

Harrow, I., Balakrishnan, R., Küçük McGinty, H., Plasterer, T., & Romacker, M. (2022). Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. *Drug Discovery Today*, *27*(5), 1441–1447. https://doi.org/10.1016/J.DRUDIS.2022.01.006

Haynie, C., Gardiner, S., & Della Corte, D. (2024). Rise of the Allotrope Simple Model: Update from 2023 Fall Allotrope Connect. *Drug Discovery Today*, *29*(4), 103944. https://doi.org/10.1016/j.drudis.2024.103944

Heald, A., Bramham-Jones, S., & Davies, M. (2021). Comparing cost of intravenous infusion and subcutaneous biologics in COVID-19 pandemic care pathways for rheumatoid arthritis and inflammatory bowel disease: A brief UK stakeholder survey. *International Journal of Clinical Practice*, *75*(9), e14341. https://doi.org/10.1111/IJCP.14341

Hendrikx, J. J. M. A., Haanen, J. B. A. G., Voest, E. E., Schellens, J. H. M., Huitema, A. D. R., & Beijnen, J. H. (2017). Fixed Dosing of Monoclonal Antibodies in Oncology. *The Oncologist*. https://doi.org/10.1634/theoncologist.2017-0167

Holub, P., Kohlmayer, F., Prasser, F., Mayrhofer, M. T., Schlünder, I., Martin, G. M., Casati, S., Koumakis, L., Wutte, A., Kozera, Z., Strapagiel, D., Anton, G., Zanetti, G., Sezerman, O. U., Mendy, M., Valík, D., Lavitrano, M., Dagher, G., Zatloukal, K., … Litton, J. E. (2018). Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. *Biopreservation and Biobanking*. https://doi.org/10.1089/bio.2017.0110

Hong, P., Koza, S., & Bouvier, E. S. P. (2012). A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *Journal of Liquid Chromatography and Related Technologies*, *35*(20), 2923–2950. https://doi.org/10.1080/10826076.2012.743724

Hutchinson, L., & Kirk, R. (2011). High drug attrition rates—where are we going wrong? *Nature Reviews Clinical Oncology 2011 8:4*, *8*(4), 189–190. https://doi.org/10.1038/nrclinonc.2011.34

Jackson, R. C., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D., Duncan, W., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., … Peters, B. (2021). OBO Foundry in 2021: Operationalizing Open Data Principles to Evaluate Ontologies. *Database : The Journal of Biological Databases and Curation*, *2021*. https://doi.org/10.1093/database/baab069

Jacobsen, A., Azevedo, R. de M., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). Fair principles: Interpretations and implementation considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Jacobsen, A., Kaliyaperumal, R., Santos, L. O. B. da S., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data fairification process. *Data Intelligence*, *2*(1–2), 56–65. https://doi.org/10.1162/dint_a_00028

Jenke, D. R. (1996). Chromatographic method validation: A review of current practices and procedures. III. Ruggedness, revalidation and system suitability. *Journal of Liquid Chromatography and Related Technologies*, *19*(12), 1873–1891. https://doi.org/10.1080/10826079608014012

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021).

Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, *45*(4), 184–192. https://doi.org/10.1145/505248.506007

Kalendralis, P., Sloep, M., van Soest, J., Dekker, A., & Fijten, R. (2021). Making radiotherapy more efficient with FAIR data. *Physica Medica*, *82*, 158–162. https://doi.org/10.1016/J.EJMP.2021.01.083

Kanehisa, M., & Subramaniam. (2002). The KEGG database. *Novartis Foundation Symposium*, *247*, 91–103. https://doi.org/10.1002/0470857897.ch8

Kim, H. H., Park, Y. R., Lee, K. H., Song, Y. S., & Kim, J. H. (2019). Clinical MetaData ontology: A simple classification scheme for data elements of clinical data based on semantics. *BMC Medical Informatics and Decision Making*, *19*(1), 1–11. https://doi.org/10.1186/S12911-019-0877-X/TABLES/3

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., & Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, *19*(2), 120–127. https://doi.org/10.1016/J.SBI.2009.03.004

Kolluri, S., Lin, J., Liu, R., Zhang, Y., & Zhang, W. (2022). Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. *AAPS Journal*, *24*(1), 1–10. https://doi.org/10.1208/s12248-021-00644-3

Kopácsi, S., Hudak, R., & Ganguly, R. (2017). Implementation of a classification server to support metadata organization for long term preservation systems. *VOEB-Mitteilungen*, *70*(2), 225–243. https://doi.org/10.31263/voebm.v70i2.1897

Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., Pétavy, F., Galvez, J., Becnel, L. B., Zhou, F. L., Harmon, N., Jauregui, B., Jackson, T., & Hudson, L. (2020). FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics*, *107*, 103421. https://doi.org/10.1016/j.jbi.2020.103421

Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., Jalaludin, B., Yeo, A. E. T., & Talaei-Khoei, A. (2013). Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, *82*(1), 10–24. https://doi.org/10.1016/j.ijmedinf.2012.10.001

Lin, P. H., Liao, C., Chen, W., Vanderbruggen, T., Emani, M., & Xu, H. (2022). Making Machine Learning Datasets and Models FAIR for HPC: A Methodology and Case Study. *Proceedings - 2022 4th International Conference on Transdisciplinary AI, TransAI 2022*, 128–134. https://doi.org/10.1109/TransAI54797.2022.00029

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., & Wu, H. C. (2020). Development of therapeutic antibodies for the treatment of diseases. In *Journal of Biomedical Science* (Vol. 27, Issue 1). https://doi.org/10.1186/s12929-019-0592-z

Machado, G., Mendoza, M. R., & Corbellini, L. G. (2015). What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Veterinary Research*, *46*(1). https://doi.org/10.1186/S13567-015-0219-7

Machina, H. K., & Wild, D. J. (2013). Electronic Laboratory Notebooks Progress and Challenges in Implementation. *Journal of Laboratory Automation*. https://doi.org/10.1177/2211068213484471

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today*, *24*(3), 773–780. https://doi.org/10.1016/J.DRUDIS.2018.11.014

Matentzoglu, N., Goutte-Gattat, D., Tan, S. Z. K., Balhoff, J. P., Carbon, S., Caron, A. R., Duncan, W. D., Flack, J. E., Haendel, M., Harris, N. L., Hogan, W. R., Hoyt, C. T., Jackson, R. C., Kim, H., Kir, H., Larralde, M., McMurry, J. A., Overton, J. A., Peters, B., … Osumi-Sutherland, D. (2022). Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database*, *2022*. https://doi.org/10.1093/DATABASE/BAAC087

McKinstry, A., Boydell, O., Le, Q., Preet, I., Hanafin, J., Fernandez, M., Warde, A., Kannan, V., Griffiths, P., McKinstry, A., Boydell, O., Le, Q., Preet, I., Hanafin, J., Fernandez, M., Warde, A., Kannan, V., & Griffiths, P. (2021). AI-Ready Training Datasets for Earth Observation: Enabling FAIR data principles for EO training data. *Eguga*, EGU21-12384. https://doi.org/10.5194/EGUSPHERE-EGU21-12384

Millecam, T., Jarrett, A. J., Young, N., Vanderwall, D. E., & Della Corte, D. (2021). Coming of age of Allotrope: Proceedings from the Fall 2020 Allotrope Connect. *Drug Discovery Today*, *26*(8), 1922–1928. https://doi.org/10.1016/j.drudis.2021.03.028

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*(7796), 491. https://doi.org/10.1038/d41586-020-00505-7

Mons, B., Neylon, C., Velterop, J., Dumontier, M., Da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Moradi-Sardareh, H., Moradi, M., Bordbar, E., Malekpour, M., Bagheri, S., Nakhodazadeh, N., Rahbar, S., & Farhadian Asgarabadi, J. (2016). The Use of Monoclonal Antibodies in the Treatment of Alzheimer Disease. *Asian Pacific Journal of Cancer Biology*, *1*(3). https://doi.org/10.31557/apjcb.2016.1.3.59-66

Muhammed, M. T., & Aki-Yalcin, E. (2019). Homology modeling in drug discovery: Overview, current applications, and future perspectives. In *Chemical Biology and Drug Design* (Vol. 93, Issue 1, pp. 12–20). Blackwell Publishing Ltd. https://doi.org/10.1111/cbdd.13388

Nagowah, S. D., Ben Sta, H., & Gobin-Rahimbux, B. (2018). An overview of semantic interoperability ontologies and frameworks for IoT. *Proceedings - 2018 6th International Conference on Enterprise Systems, ES 2018*, 82–89.

https://doi.org/10.1109/ES.2018.00020

Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M., & Arosio, P. (2021). Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. In *Trends in Pharmacological Sciences*. https://doi.org/10.1016/j.tips.2020.12.004

Neal, R., Robbins, J., Dankers, R., Mitra, A., Jayakumar, A., Rajagopal, E. N., & Adamson, G. (2020). Deriving optimal weather pattern definitions for the representation of precipitation variability over India. *International Journal of Climatology*, *40*(1), 342–360. https://doi.org/10.1002/joc.6215

Nisbet, R., Miner, G., & Yale, K. (2017). *Handbook of statistical analysis and data mining applications*. Handbook of Statistical Analysis and Data Mining Applications. https://doi.org/10.1016/c2012-0-06451-4

Oberkampf, H. (2018). Allotrope Framework: Semantischer Datenstandard für das Lab 4.0. *Iuta 3. AnalytikTag*. https://www.iuta.de/wp-content/uploads/2018/09/Oberkampf-Allotrope-Framework.pdf

OBI Core Classes. (n.d.). Retrieved March 2, 2023, from http://obi-ontology.org/docs/core-classes/

Obrst, L. (2003). Ontologies for Semantically Interoperable Systems. *Proceedings of the Twelfth International Conference on Information and Knowledge Management - CIKM '03*. https://doi.org/10.1145/956863

Olson, J. E. (2003). Data Quality: The Accuracy Dimension. In *Data Quality: The Accuracy Dimension*. Elsevier. https://doi.org/10.1016/B978-1-55860-891-7.X5000-8

Onufriev, A. V., & Alexov, E. (2013). Protonation and pK changes in protein-ligand binding. *Quarterly Reviews of Biophysics*, *46*(2), 181. https://doi.org/10.1017/S0033583513000024

Overview | FAIRplus Data Maturity. (n.d.). Retrieved March 21, 2023, from https://fairplus.github.io/Data-Maturity/

Paton, N. W. (2007). Automation everywhere: Autonomics and data management. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4587 LNCS*, 3–12. https://doi.org/10.1007/978-3-540-73390-4_2/COVER

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). *Artificial intelligence in drug discovery and development*. Drug Discovery Today. https://doi.org/10.1016/j.drudis.2020.10.010

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Pedro Manuel Díaz Ortuño. (2005). *W3C SPARQL 1.1 Query Language*. https://skos.um.es/TR/rdf-sparql-query/

Queralt-Rosinach, N., Kaliyaperumal, R., Bernabé, C. H., Long, Q., Joosten, S. A., van der Wijk, H. J., Flikkenschild, E. L. A., Burger, K., Jacobsen, A., Mons, B., & Roos, M. (2022). Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. *Journal of Biomedical Semantics*, *13*(1), 12. https://doi.org/10.1186/s13326-022-00263-7

Rathore, A. S., & Winkle, H. (2009). Quality by design for biopharmaceuticals. *Nature Biotechnology 2009 27:1*, *27*(1), 26–34. https://doi.org/10.1038/nbt0109-26

RDA FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model: specification and guidelines. *Research Data Alliance*, *June*, 2019–2020. https://doi.org/10.15497/rda00050

Redman, T. C. (1998). Impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82. https://doi.org/10.1145/269012.269025

Redman, T. C. (2001). *Data Quality: The Field Guide - Thomas C. Redman - Google Books*. https://books.google.de/books?hl=de&lr=&id=iE2qsxDyEEYC&oi=fnd&pg=PR11&dq=redman+2001&ots=6lmX06H0KB&sig=YfX403h57s-1LjLHEuY4HHoeXRg&redir_esc=y#v=onepage&q=redman 2001&f=false

Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., Kong, S., Long, X., Hei, B., Liu, M., Liu, Y., Naumov, V., Shneyderman, A., Ozerov, I. V, Wang, J., … Zhavoronkov, A. (2023). *AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor †*. https://doi.org/10.1039/d2sc05709c

Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., Splendiani, A., Burdett, T., Giessmann, R. T., Henderson, D., Batista, D., Emam, I., Gadiya, Y., Giovanni, L., Willighagen, E., Evelo, C., Gray, A. J. G., Gribbon, P., Juty, N., … Sansone, S. A. (2023). The FAIR Cookbook - the essential resource for and by FAIR doers. *Scientific Data 2023 10:1*, *10*(1), 1–12. https://doi.org/10.1038/s41597-023-02166-3

Schoening, T., Durden, J. M., Faber, C., Felden, J., Heger, K., Hoving, H. J. T., Kiko, R., Köser, K., Krämmer, C., Kwasnitschka, T., Möller, K. O., Nakath, D., Naß, A., Nattkemper, T. W., Purser, A., & Zurowietz, M. (2022). Making marine image data FAIR. *Scientific Data 2022 9:1*, *9*(1), 1–10. https://doi.org/10.1038/s41597-022-01491-3

Sheehan, J., Hirschfeld, S., Foster, E., Ghitza, U., Goetz, K., Karpinski, J., Lang, L., Moser, R. P., Odenkirchen, J., Reeves, D., Rubinstein, Y., Werner, E., & Huerta, M. (2016). Improving the value of clinical research through the use of Common Data Elements. *Clinical Trials*, *13*(6), 671–676. https://doi.org/10.1177/1740774516653238

Siedler, M., Eichling, S., Huelsmeyer, M., & Angstenberger, J. (2020). Chapter 13: Formulation Development for Biologics Utilizing Lab Automation and In Vivo Performance Models. *AAPS Advances in the Pharmaceutical Sciences Series*, *35*, 299–

341. https://doi.org/10.1007/978-3-030-31415-6_13

Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M. L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J. M., Pérez-Pérez, M. M., Laleci Erturkmen, G. B., Hernández-Pérez, T., Méndez-Rodríguez, E., & Parra-Calderón, C. L. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. *Methods of Information in Medicine*, *59*(6), E21–E32. https://doi.org/10.1055/S-0040-1713684/ID/JR19020007-24

Sjuts, H., Schreuder, H., Engel, C. K., Bussemer, T., & Gokarn, Y. (2020). Matching pH values for antibody stabilization and crystallization suggest rationale for accelerated development of biotherapeutic drugs. *Drug Development Research*. https://doi.org/10.1002/ddr.21624

Snyder, L. R., Kirkland, J. J., & Dolan, J. W. (2010). Introduction to Modern Liquid Chromatography. In *Introduction to Modern Liquid Chromatography*. John Wiley \& Sons. https://doi.org/10.1002/9780470508183

Spanos, D.-E., Stavrou, P., & Mitrou, N. (2010). *Bringing Relational Databases into the Semantic Web: A Survey*. 1–17. http://www.w3.org/2001/sw/rdb2rdf/

Taylor, S. (2021). Understanding large and complex biological data sets using visualization. *The Biochemist*, *43*(5), 54–58. https://doi.org/10.1042/BIO_2021_165

Touré, V., Krauss, P., Gnodtke, K., Buchhorn, J., Unni, D., Horki, P., Raisaro, J. L., Kalt, K., Teixeira, D., Crameri, K., & Österle, S. (2023). FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. *Scientific Data 2023 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41597-023-02028-y

van der Velde, K. J., Singh, G., Kaliyaperumal, R., Liao, X. F., de Ridder, S., Rebers, S., Kerstens, H. H. D., de Andrade, F., van Reeuwijk, J., De Gruyter, F. E., Hiltemann, S., Ligtvoet, M., Weiss, M. M., van Deutekom, H. W. M., Jansen, A. M. L., Stubbs, A. P., Vissers, L. E. L. M., Laros, J. F. J., van Enckevort, E., … Swertz, M. A. (2022). FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. *Scientific Data 2022 9:1*, *9*(1), 1–13. https://doi.org/10.1038/s41597-022-01265-x

Vesteghem, C., Brøndum, R. F., Sønderkær, M., Sommer, M., Schmitz, A., Bødker, J. S., Dybkær, K., El-Galaly, T. C., & Bøgsted, M. (2020). Implementing the FAIR Data Principles in precision oncology: Review of supporting initiatives. *Briefings in Bioinformatics*, *21*(3), 936–945. https://doi.org/10.1093/bib/bbz044

Wang, R. Y. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–34. https://doi.org/10.1080/07421222.1996.11518099

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and

stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Sansone, S. A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M. E., & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, *6*(1). https://doi.org/10.1038/s41597-019-0184-5

Wilkinson, M. D., Verborgh, R., da Silva Santos, L. O. B., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., van Mulligen, E. M., Ciccarese, P., Kuzniar, A., Gavai, A., Thompson, M., Kaliyaperumal, R., Bolleman, J. T., & Dumontier, M. (2017). Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, *2017*(4). https://doi.org/10.7717/peerj-cs.110

Wise, J., de Barron, A. G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R. C., Winnenburg, R., Plasterer, T., Gupta, V., & Hedley, V. (2019). Implementation and relevance of FAIR data principles in biopharmaceutical R&D. In *Drug Discovery Today* (Vol. 24, Issue 4, pp. 933–938). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2019.01.008

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

10 Curriculum vitae

Personal Details

First name:	Axel
Family name:	Wilbertz
Title:	M.Sc.
Date of birth:	25 November 1988
Place of birth:	Eberbach
Nationality:	German

Education

2018 - present	**Doctor scientiarum humanarum (Dr. sc. hum.)**, *Medizinische Fakultät Mannheim der Ruprecht-Karls-Universität zu Heidelberg*
2013 - 2016	**Master of Science (M.Sc.),** Grade: 1.4 in Computer Science, *Mannheim University of Applied Sciences,*
2009 - 2013	**Bachelor of Science (B.Sc.),** Grade: 2.5 in Medical Computer Science, *Mannheim University of Applied Sciences,*
2008 - 2009	**Bachelor of Science (B.Sc.),** (unfinished) Business Informatics, *University of Mannheim*
1999 - 2008	**Abitur,** *Hohenstaufen-Gymnasium Eberbach*

Work experience

2022 - present	**Senior Business Engineer**, *AbbVie*
2016 - 2022	**Data Scientist**, *AbbVie*
2013 - 2015	**Student research assistant (Bioinformatics Scientist)**, *Department of Hematology and Oncology, University Medical Centre Mannheim*
2010 - 2013	**Student research assistant (Sleep Apnea)**, *Sleep Disorders Center, University Medical Centre Mannheim*

Conferences & publications

2023 **Conference talk** *M2Aind Symposium Künstliche Intelligenz,* October 10, Mannheim, Germany

2022 **Conference talk (virtual)** *FAIRplus Innovation and SME Forum,* May 17

2022 **Conference talk** *Future Labs live 2022*, June 7-8, Basel, Switzerland

2021 **Conference talk (virtual)** *PRISME Forum Spring 2021 Technical Meeting*, May 19-20

2019 **Conference talk** *11th Annual Bioprocessing Summit*, August 12-16, Boston, US

2017 **Conference talk** *16th Annual Laboratory Informatics Summit 2017,* September 19-21, Amsterdam, Netherlands

2016 **Conference talk** *15th Annual Laboratory Informatics Summit 2016,* September 21-22, Brussels, Belgium

2013 **Paper** Mossner, M., Wilbertz, A, … & Nowak, D. (2013), Array-Based Integrative Analysis Of Epigenomic and Transcriptomic Alterations In CD71+ Bone Marrow Erythroprogenitor Cells From Patients With Myelodysplastic Syndromes. *Blood*, *122*(21), 1561.

11 Acknowledgement

Firstly, I would like to thank Prof. Dr. Jan Stallkamp for the kind supervision of my thesis and for supporting me throughout my entire PhD study with his expertise. I thank you for accepting me as your student. Despite the non-linear development of the thesis, you helped shape it into an understandable outline. Furthermore, I would like to thank you for providing words of motivation and giving me enough freedom to independently work on this topic.

A huge thanks goes to Prof. Dr. Ivo Wolf for the comprehensive support and overwhelming amount of feedback, especially during the end phase of the PhD. Your support enabled me to create an increasingly understandable outline. Moreover, I thank you for our inspiring discussions, which sometimes lasted until late in the evening.

Dr. Jonas Angstenberger, thank you for not only being a colleague but also a good and constant friend along the way. I am especially thankful for your scientific guidance, personal development, and support. I am very grateful for reviewing my results and providing constant feedback beyond work hours.

I thank Dr. Michael Siedler, for his vision of the PhD topic and for providing the scientific basis for the thesis. Moreover, I thank you for the scientific guidance, my personal development, and for constantly pushing me to my limits, motivating and supporting me throughout the time.

Dr. Thomas Merdan, thank you for your dedication and motivation to drive forward reusable scientific data and the vision of a data-driven organization.

I would like to express my appreciation and thanks to Tanja Meyer for sharing her knowledge and deep understanding in the field of liquid chromatography.

I thank Dr. Rainer Winnenburg & Dr. Matthias Negri for their expertise and support during the development and optimization of the semantic model.

I would like to acknowledge Dr. Mark Griffiths for his help and expertise regarding the random forest approaches and evaluating the training and testing data.

I thank Dr. Martin Huelsmeyer for his expertise and guidance during the antibody modeling and protein property calculation at the beginning of the thesis.

I would like to thank all the colleagues at AbbVie who supported me and contributed to the thesis by generating data or having an open ear or sharing a word of motivation.

Sebastian Schöning and Dr. Stefan Scheuermann, I would like to thank for giving guidance, support, and advice during the PhD and the course at Fraunhofer IPA Mannheim.

I would also like to thank the Medical Faculty Mannheim of the Ruprecht-Karls-University Heidelberg for accepting this PhD thesis. In particular, I thank Susanne Volz for her organizational support.

I am very thankful to my family, friends, and loved ones for their continuing loving support and patience throughout my PhD thesis.

Finally, I would like to express my deepest gratitude to my mother, Christine Wilbertz, and my father, Andreas Wilbertz, for their unconditional love, encouragement, and support. Without you, I would not be where and who I am today.