EnzymeML – a data exchange format for biocatalysis and enzymology

Jan Range[1], Colin Halupczok[1], Jens Lohmann[1], Neil Swainston[2], Carsten Kettner[3], Frank T. Bergmann[4], Andreas Weidemann[5], Ulrike Wittig[5], Santiago Schnell[6], Jürgen Pleiss[1*]

[1] Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

[2] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom

[3] Beilstein-Institut, Trakehner Str. 7 – 9, 60487 Frankfurt am Main, Germany

[4] BioQUANT/COS, Heidelberg University, INF 267, Heidelberg, Germany

[5] Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

[6] Department of Molecular & Integrative Physiology, and Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA

* Corresponding author:
Jürgen Pleiss
Institute of Biochemistry and Technical Biochemistry
University of Stuttgart
Allmandring 31
70569 Stuttgart, Germany
E-mail: Juergen.Pleiss@itb.uni-stuttgart.de
ORCID: 0000-0003-1045-8202

Abstract

EnzymeML is an XML–based data exchange format that supports the comprehensive documentation of enzymatic data by describing reaction conditions, time courses of substrate and product concentrations, the kinetic model, and the estimated kinetic constants. EnzymeML is based on the Systems Biology Markup Language, which was extended by implementing the STRENDA Guidelines. An EnzymeML document serves as a container to transfer data between experimental platforms, modelling tools, and databases. EnzymeML supports the scientific community by introducing a standardised data exchange format to make enzymatic data findable, accessible, interoperable, and reusable according to the FAIR data principles. An Application Programming Interface in Python and Java supports the integration of applications. The feasibility of a seamless data flow using EnzymeML is demonstrated by creating an EnzymeML document from a structured spreadsheet or from a STRENDA DB database entry, by kinetic modelling using the modelling platform COPASI, and by uploading to the enzymatic reaction kinetics database SABIO-RK.

1. Introduction

Enzyme catalysis and enzymology provide a powerful toolbox for sustainable synthesis routes and innovative solutions for bio-based chemistry. A better understanding of cellular biochemistry and the comprehensive biochemical characterization of the desired enzyme-catalyzed reaction enable novel approaches in enzyme engineering and process development.[1] Standardization of reporting of enzymatic data and metadata is considered to be pivotal to accelerating bioprocess development and reducing costs[2], facilitating sharing, analysis, and reuse of data and thus enabling quality control and reproducibility of experiments[3]. Therefore, a major challenge for enzymology and biocatalysis lies in the current practices of dealing with experimental data in academic laboratories[4]. In most academic research groups, data acquisition, curation, and documentation are performed manually without a universally accepted standard across laboratories. Data and metadata are typically stored in *ad hoc* repositories, such as paper lab notebooks, spreadsheets in different formats, and semi-structured text files containing custom annotations. Experimental or computational data is often poorly annotated, lacking a complete description of the acquisition and analysis procedures, or associated metadata. Despite previous efforts to address these issues[5], raw data are rarely available in machine-readable, even less in machine-actable format, preventing their further analysis and third-party validation. As it stands, the process of data acquisition, data analysis, and documentation is time consuming and error-prone, as is the recovery and interpretation of legacy data in most academic laboratories. Consequently, both the quality and the completeness of data and metadata solely relies on the experimenter's expertise and care.

Meta-research studies suggest the lack of standardization to report and share experimental protocols, results and data as one of the causes of the reproducibility crisis in the biomedical sciences[6,7]. This is also true for enzymology and biocatalysis. An empirical analysis of published papers investigating enzyme function illustrates how critical information for the reproducibility of experimental finding is missing in the literature[8]; the missing information includes the concentration of enzyme and/or substrates, the composition of the entire buffer systems including the identity of counter-ions, pH values and assay temperatures.

The incompleteness of metadata prevents the interpretation of inconsistent data arising from different studies. An example of such variability is demonstrated in a large global benchmark study[9], in which the variability of a dissociation constant for a protein-protein interaction determined by 150 participants using a general protocol exceeded its average value. When

investigators were given detailed fixed protocols, the dissociation constants still varied up to 20%[10,11]. This kind of irreproducibility is commonplace in enzymology and has an essential impact on subsequent research.

In response to the reproducibility crisis, the scientific community is developing and adopting new guidelines for reporting experimental protocols and statistical analysis. Scientific journals are responding accordingly[12], and there has been a recommendation to modify the academic reward system by recognising scientists who aligned with best practices for reproducible research[13]. Initiatives such as the German National Research Data Infrastructure develop an infrastructure for standardised research data exchange[14], the Standards in Laboratory Automation consortium (SiLA) provide a framework for the exchange, integration, sharing, and retrieval of electronic laboratory information (https://sila2.gitlab.io/sila_base/), and data repositories such as Zenodo and Dataverse enable data sharing[15] . Efforts in standardization and data reproducibility have been long established in other 'omics fields, with standard exchange formats for transcriptomics[16], proteomics[17], and metabolomics[18] data becoming increasingly developed and adopted over the last twenty years. However, in biocatalysis and enzymology exchange standards or software support to aid data analysis, management, and sharing is still absent, and raw experimental data such as the time dependency of substrate or product concentration, derived data such as kinetic parameters, and metadata such as reaction conditions or the kinetic model are typically reported in plain text, figures, or tables[19]. Currently, kinetic parameters and corresponding information about the reactions, enzymes, and experimental conditions are extracted and annotated manually from scientific publications and inserted into databases such as SABIO-RK[20] or BRENDA[21] to structure and standardise the data. Missing information such as unambiguous external identifiers is added manually by database curators. As a first step for the standardised reporting of enzyme function data, the enzymology and biocatalysis community has established the Standards for Reporting Enzymology Data (STRENDA) Guidelines, which provide the minimum information necessary to describe assay conditions and enzyme activity data[22,23]. Currently, more than 55 international biochemistry journals have included adherence to the STRENDA Guidelines in their instructions for authors reporting enzymology data. STRENDA DB has been established as a public database to support authors checking the completeness of their data upon submission of their manuscript and to provide public access to data on reaction conditions and kinetic parameters of an experiment[24]. However, the upload of data is performed manually via a graphical user interface, and the process from data acquisition to kinetic modelling and

publication is still time consuming and error prone. Most importantly, original data such as the measured time course of substrate and product concentrations is not reported or has to be extracted from figures, thus preventing the reuse of original data for kinetic modelling. Not only is published data incomplete and inaccessible, but also unpublished research data and metadata are stored by research group members with insufficient documentation and annotation. In addition, the current data management prevents researchers from upscaling their experimental designs to high-throughput biocatalytic approaches by using pipetting robots[25] or flow reactors[26], and hinders the comprehensive study of the multidimensional parameter space of biocatalytic reactions.

Here, we introduce EnzymeML, a data exchange format for biocatalysis and enzymology, which makes enzyme data findable, accessible, interoperable, and reusable in accordance to the FAIR data principles[27]. An application programming interface (API) provides Python and Java libraries to integrate applications and databases and to enable a seamless data flow from the bench to kinetic modelling tools and publication platforms. The machine-actable EnzymeML document on data and metadata of an enzymatic reaction could serve as a micropublication, supplementing the respective scientific paper.

2. Principles of EnzymeML

EnzymeML has been designed to support data acquisition, data analysis, and sharing of data by providing a standardised exchange format for enzymatic data (**Fig. 1**). EnzymeML is written in eXtensible Markup Language (XML) and comprises the most relevant data and metadata from measurement and modelling. Given the ubiquity of XML, vast amounts of software are available that read, write, manipulate, and process XML documents. More importantly, XML allows for the specification of a machine-actable schema which ensures interoperability. The central core of EnzymeML is the Systems Biology Markup Language (SBML), an established data format in systems biology for sharing, evaluating, and developing models of biochemical reaction networks[28]. Interoperability with existing software tools and databases is achieved by applying a common terminology and vocabulary that allow the integration of data from various sources for subsequent processing, because many of the concepts supported by SBML – educts, products, reactions, modifiers, reaction rates – are common to enzymology and biocatalysis. However, EnzymeML goes beyond SBML, because it serves to describe the effect of enzyme sequence and reaction medium to an enzymatic reaction.

EnzymeML implements the STRENDA Guidelines: For the complete machine-actable description of an enzymatic experiment, the STRENDA Guidelines were incorporated. In addition, metadata on the experiments and the kinetic model were included, resulting in a comprehensive data exchange format that comprises 71 attributes (**Tab. S1**). The current version of EnzymeML includes all STRENDA fields with a controlled vocabulary or values and excludes fields with plain text such as experiment methodology, in order to make EnzymeML structured and machine actable.

EnzymeML was built within the framework of several internationally recognised standards: SBML is a widely used XML-based markup language and describes almost 50% of the attributes (**Tab. S1**). MathML was applied to describe the equation of the kinetic model,[28] and the guidelines on Minimal Information Required in the Annotation of Models (MIRIAM)[29] were applied for the consistent annotation of components such as reactants, products, and enzymes, using terms from external data repositories such as ChEBI[30] and Uniprot[31]. A controlled, relational vocabulary of terms, the Systems Biology Ontology (SBO)[32], was used to define reactants, inhibitors, activators, parameters, and the kinetic model. All files are combined into a single document using the OMEX format[33]. Furthermore, EnzymeML uses the Distributions package for SBML Level 3 (*http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/distrib*) to support the specification of ranges of initial concentrations.

EnzymeML is extensible: EnzymeML-specific attributes are added to SBML using the "annotation" element, which supports metadata specific to enzymology to be added to the XML document whilst maintaining compatibility with SBML. EnzymeML documents are valid SBML files and can therefore be used and manipulated by many software tools that support the SBML format.

EnzymeML is platform independent: XML has been designed to store and transfer data, and is fully agnostic to the operating system and supported by different programming languages. Comma-Separated Values (CSV) is a platform-independent text file format, which was designed for storing and transporting data structured in tables. CSV-formatted files can be read by the modelling platform COPASI[34] and by spreadsheet editors such as Excel. All components

of EnzymeML are self-descriptive (SBML, MathML, OMEX), which makes EnzymeML human readable and machine actable.

EnzymeML is modular: EnzymeML was developed as a container for experimental and modelling data, supporting a seamless data flow between different applications (**Fig. 2**). Data obtained from an experiment and metadata on experimental conditions can be stored by the experimentalist in a spreadsheet, which is convertible into EnzymeML using the API. Longer term, it is hoped that electronic lab notebooks, laboratory information management systems, and enzymology software will support the format. The EnzymeML document contains sufficient experimental data to allow for the estimation of the kinetic parameters by modelling platforms such as COPASI[34], BioCatNet[35], or Matlab™. Kinetic parameters can then be included in the EnzymeML document. As a consequence, enzyme assay data may be easily reanalyzed and checked with a range of data fitting algorithms, increasing reusability and confidence in both the experimental data and reported kinetic parameters.

EnzymeML enables data publication in compliance with FAIR principles: An EnzymeML document stores comprehensive information about data and metadata of an enzymatic experiment: the experimental conditions, the time course of substrate and product concentration, the kinetic model, and the estimated kinetic parameters, thus making the experiment and its analysis reproducible. Upon publication, it is recommended to use EnzymeML documents as supplementary material. By depositing EnzymeML documents on platforms such as FAIRDOMHub[36] or Dataverse[37] using a digital object identifier, EnzymeML documents are findable and accessible. EnzymeML documents also include references to the scientific publications from which they arose, providing contextual information.

3. Structure of EnzymeML documents

An EnzymeML document is a ZIP container in the widely used OMEX format.[33] It consists of three file types: a file using SBML to describe the experimental reaction conditions, the kinetic model, and the kinetic parameters, CSV (comma-separated values)-formatted files to store the time courses of substrate and product concentrations, and a manifest file lists the content of the ZIP container (**Figure 1**).

The experimental conditions are reported according to the STRENDA recommendations, the kinetic model is described by using MathML and SBML in the experiment file. This file also describes the format of the CSV-formatted file which contains the raw time course data. Instead of using headers to describe columns, the complete CSV-formatted file description is done within the SBML file. This approach has the advantage of enabling a comprehensive description of each column, such as measured species, units and data types, instead of a single header. The SBML file uses two elements, notes and annotation. A notes tag contains human-readable information as plain text, whereas an annotation tag contains structured, machine-actable information. Notes and annotation tags are used to add information which is required by the STRENDA Guidelines, but not included in SBML, such as protein sequence, pH, or temperature. Thus, this file is a valid SBML document, which contains additional information on enzyme-catalyzed reactions. An extensive description of the EnzymeML document structure is available in the Supporting Information.

4. EnzymeML application programming interface (API)

Although EnzymeML is semi-human-readable, the user is not expected to read or write EnzymeML documents directly, but to use software to generate EnzymeML documents, which can then be used as a standardised exchange format to transfer data between applications (**Figure 2**). APIs to read, write, edit, and visualise EnzymeML have therefore been developed, using the popular programming languages Python and Java, to support the development of such software tools. The library PyEnzyme was built based on its respective SBML counterpart libSBML. To simplify the implementation of the libraries for enzyme-catalyzed reactions, the terminology of enzymology and biocatalysis is used, hiding the more systems biology focused SBML terms, while maintaining full compatibility with the SBML format.

The adaption of the API to an application is enabled by an additional thin layer, which maps the objects of the API to the equivalent objects defined within the respective application. Thus, by editing a template, the functionality of reading and writing of EnzymeML can be easily incorporated into an application without the need to modify the API. For five applications (COPASI import/export, STRENDA-DB export, BioCatNet export, SABIO-RK import, simulation of time course data), application-specific thin API layers are provided (TL_COPASI, TL_STRENDAML and TL_BioCatNet, respectively). Because the API enables batch processing, management of enzymatic data is scalable, and high throughput strategies of

experimentation and data analysis become feasible. By data export in formats such as Pandas DataFrame, large datasets could be analyzed by novel analysis methods based on machine learning.

Upon reading, writing, and visualization of EnzymeML documents, the API controls data completeness and consistency, such as checking the definition of reactants and proteins upon reading or writing of a reaction, or by checking that scalar properties such as pH are within the necessary range. A specific validation tool guarantees compatibility with SBML. Further application-specific validation tools have been added, such as a STRENDA DB validator to check for compatibility with the STRENDA Guidelines. For more details, readers can find a description of API below and the Supporting Information.

5. Application of EnzymeML

To illustrate the power of EnzymeML, we illustrate selected applications for experimental enzymologists, system biology modelers, and software developers.

5.1 Creating EnzymeML documents from structured spreadsheets

In the absence of a standard format, experimentalists typically store their experimental time course data in a spreadsheet following an *ad hoc* structure. Recently, a CSV-formatted spreadsheet, the BioCatNet template[35], was proposed to store and report experimental data on enzyme-catalyzed reactions according to the STRENDA Guidelines. The API was used to convert the BioCatNet spreadsheet, containing time course data on substrate and product concentration and comprehensive information as the reaction conditions, to EnzymeML. Initially, each field of the respective spreadsheet template was extracted via a thin API layer (TL_BioCatNet) and further processed by the API to an object layer. Finally, the objects were written to an EnzymeML document (see SI 3.1).

5.2 Creating EnzymeML documents from STRENDA DB entries

STRENDA DB is a database on enzyme-catalyzed reactions, which covers the most important information on reaction conditions and kinetic parameters.[24] The API was used to create an EnzymeML document from a STRENDA DB entry via a STRENDA DB-specific thin API layer (TL_STRENDA) to the object layer using the PyEnzyme library. The resulting EnzymeML document was then created by the API (see SI 3.2).

5.3 Upload of EnzymeML documents to SABIO-RK

SABIO-RK is a curated database that contains information about biochemical reactions, their kinetic rate equations with parameters, and experimental conditions. [20] An already existing SBML parser for the upload of SBML models in SABIO-RK was extended to read the additional annotations in EnzymeML to allow the import of EnzymeML documents and to create a new SABIO-RK entry in the internal curation interface (see SI 3.3). SABIO-RK curators check the new SABIO-RK entries for consistency and completeness according to the SABIO-RK requirements before they are finally submitted to the public SABIO-RK database.

5.4 Editing of EnzymeML: simulation of time course data from kinetic parameters

STRENDA-DB entries provide for an enzyme-catalyzed reaction the kinetic parameters K_M and k_{cat} assuming a Michaelis-Menten model and the concentration range of the substrate. However, they are lacking information on the product and on the time course of substrate or product concentrations. PyEnzyme was used to add the product and time course data to the EnzymeML document (see SI 3.4). By a single function in the API, the time course of substrate concentrations was simulated from the kinetic parameters for initial concentrations from 0 to 0.5 mM for a time interval of 200 seconds to visualise kinetic behavior and study the effect of kinetic parameters

5.5 Kinetic modelling of EnzymeML data by COPASI

COPASI is a modelling and simulation environment, which supports the OMEX format.[34] Using the PyEnzyme library and a COPASI-specific thin API layer (TL_COPASI), the time course data (measured concentrations of substrate or product) are loaded into COPASI. Within COPASI, different kinetic laws are applied, kinetic parameters are estimated, and plots are generated to assess the result. The selected kinetic model and the estimated kinetic parameters are then added to the EnzymeML document (see SI 3.5).

6. Outlook

For many years, researchers worldwide from various disciplines have recognised that data published in the literature is not reliable unless the full set of information required is provided[23]. Therefore, the FAIR principles were introduced to encourage the comprehensive documentation of structured metadata in all stages of their life cycle in order to guarantee reproducibility of experiments and to enable reuse of results. A discipline-specific standard data

exchange format such as EnzymeML therefore provides three functionalities to optimise research in biocatalysis and enzymology: it allows the experimentalist to collect data and metadata in a structured format for data analysis; it allows project partners to transfer data and metadata between different sites and different applications; and it enables findable and reusable publication and archiving of data and metadata[38].

Currently, data flow from laboratory to publication is a challenging and complex process involving diverse processing stages, and numerous steps of data reformatting and manual input. Such manual approaches are becoming increasingly unsustainable, especially in light of recent advances in miniaturization and robotics which have enabled the intensive, high-throughput screening of enzymes and process conditions.[39] Such technological advances foster the discovery of novel enzymatic systems and the (retro-)synthetic design of enzyme-catalyzed reaction cascades through integration of systematic data acquisition, data analysis, and simulation.[40]

In a fully digitalised biocatalytic laboratory, an electronic lab notebook supports researchers at the bench to plan experiments and to collect experimental data and metadata,[41,42] all laboratory devices are connected by a common standard,[43] various modelling and data analysis tools are combined to analyze the data[34,35,44], and the results are uploaded to searchable repositories without manual intervention[24,20].

With the integration of EnzymeML the interoperability and compatibility of the tools and databases will be improved, and possible current limitations and inconsistencies in the data models of the repositories will be resolved. In the future, EnzymeML will be combined with other standards to enrich the data model and to connect disciplines that are relevant to enzymology. Incorporating AniML[43] or SiLA enables access to laboratory devices, and ThermoML[42] offers a comprehensive description of the reaction medium.

The introduction of EnzymeML as a uniform transport container for experimental data and metadata, will encourage the development of software infrastructure built on this standardised format to greatly simplify the process of analyzing and publishing enzymology data, supporting the increasing experimental throughput, and ultimately promoting the digitalization of the fields of enzymology and biocatalysis[14].

7. Code availability

The XML Schema, the API, templates of the thin API layer, and all files mentioned in the Application section are available at *https://github.com/EnzymeML* and *https://zenodo.org/record/5021263#.YNQPtS223BI*.

Acknowledgements

The authors acknowledge Michael Hucka (California Institute of Technology) for inspiring discussions and constructive comments during the meetings of the EnzymeML Development Team and Patrick Buchholz (University of Stuttgart) for his support with BioCatNet. JP acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, grants EXC310 and EXC2075). NS acknowledges funding from the Biotechnology and Biological Sciences Research Council (BBSRC) under grant "GeneORator: a novel and high-throughput method for the synthetic biology-based improvement of any enzyme" (BB/S004955/1) and from the University of Liverpool. AW and UW acknowledge funding from the Klaus Tschira Foundation and the German Federal Ministry of Education and Research within de.NBI (031A540). FTB acknowledges funding from the German Federal Ministry of Education and Research within de.NBI (031L0104A). We are grateful for the support of Beilstein-Institut zur Förderung der Chemischen Wissenschaften by supporting discussions through its Beilstein Enzymology Symposia and STRENDA Commission Meetings.

References

1 A. Pellis, S. Cantone, C. Ebert and L. Gardossi, *N. Biotechnol.*, 2018, **40**, 154–169.

2 T. Decoene, B. De Paepe, J. Maertens, P. Coussement, G. Peters, S. L. De Maeseneire and M. De Mey, *Crit. Rev. Biotechnol.*, 2018, **38**, 647–656.

3 V. Lapatas, M. Stefanidakis, R. C. Jimenez, A. Via and M. V. Schneider, *J. Biol. Res.*, 2015, **22**, 1–16.

4 C. Kettner and A. Cornish-Bowden, *Perspect. Sci.*, 2014, **1**, 1–6.

5 N. Swainston, M. Golebiewski, H. L. Messiha, N. Malys, R. Kania, S. Kengne, O. Krebs, S. Mir, H. Sauer-Danzwith, K. Smallbone, A. Weidemann, U. Wittig, D. B. Kell, P. Mendes, W. Müller, N. W. Paton and I. Rojas, *FEBS J.*, 2010, **277**, 3769–3779.

6 P. B. Stark, *Nature*, 2018, **557**, 613.

7 M. Baker and D. Penny, *Nature*, 2016, **533**, 452–454.

8 P. Halling, P. F. Fitzpatrick, F. M. Raushel, J. Rohwer, S. Schnell, U. Wittig, R.

Wohlgemuth and C. Kettner, *Biophys. Chem.*, 2018, **242**, 22–27.

9 R. L. Rich, G. A. Papalia, P. J. Flynn, J. Furneisen, J. Quinn, J. S. Klein, P. S. Katsamba, M. B. Waddell, M. Scott, J. Thompson, J. Berlier, S. Corry, M. Baltzinger, G. Zeder-Lutz, A. Schoenemann, A. Clabbers, S. Wieckowski, M. M. Murphy, P. Page, T. E. Ryan, J. Duffner, T. Ganguly, J. Corbin, S. Gautam, G. Anderluh, A. Bavdek, D. Reichmann, S. P. Yadav, E. Hommema, E. Pol, A. Drake, S. Klakamp, T. Chapman, D. Kernaghan, K. Miller, J. Schuman, K. Lindquist, K. Herlihy, M. B. Murphy, R. Bohnsack, B. Andrien, P. Brandani, D. Terwey, R. Millican, R. J. Darling, L. Wang, Q. Carter, J. Dotzlaf, J. Lopez-Sagaseta, I. Campbell, P. Torreri, S. Hoos, P. England, Y. Liu, Y. Abdiche, D. Malashock, A. Pinkerton, M. Wong, E. Lafer, C. Hinck, K. Thompson, C. Di Primo, A. Joyce, J. Brooks, F. Torta, A. B. Bagge Hagel, J. Krarup, J. Pass, M. Ferreira, S. Shikov, M. Mikolajczyk, Y. Abe, G. Barbato, A. M. Giannetti, G. Krishnamoorthy, B. Beusink, D. Satpaev, T. Tsang, E. Fang, J. Partridge, S. Brohawn, J. Horn, O. Pritsch, G. Obal, S. Nilapwar, B. Busby, G. Gutierrez-Sanchez, R. Das Gupta, S. Canepa, K. Witte, Z. Nikolovska-Coleska, Y. H. Cho, R. D'Agata, K. Schlick, R. Calvert, E. M. Munoz, M. J. Hernaiz, T. Bravman, M. Dines, M.-H. Yang, A. Puskas, E. Boni, J. Li, M. Wear, A. Grinberg, J. Baardsnes, O. Dolezal, M. Gainey, H. Anderson, J. Peng, M. Lewis, P. Spies, Q. Trinh, S. Bibikov, J. Raymond, M. Yousef, V. Chandrasekaran, Y. Feng, A. Emerick, S. Mundodo, R. Guimaraes, K. McGirr, Y.-J. Li, H. Hughes, H. Mantz, R. Skrabana, M. Witmer, J. Ballard, L. Martin, P. Skladal, G. Korza, I. Laird-Offringa, C. S. Lee, A. Khadir, F. Podlaski, P. Neuner, J. Rothacker, A. Rafique, N. Dankbar, P. Kainz, E. Gedig, M. Vuyisich, C. Boozer, N. Ly, M. Toews, A. Uren, O. Kalyuzhniy, K. Lewis, E. Chomey, B. J. Pak and D. G. Myszka, *Anal. Biochem.*, 2009, **386**, 194–216.

10 M. J. Cannon, G. A. Papalia, I. Navratilova, R. J. Fisher, L. R. Roberts, K. M. Worthy, A. G. Stephen, G. R. Marchesini, E. J. Collins, D. Casper, H. Qiu, D. Satpaev, S. F. Liparoto, D. A. Rice, I. I. Gorshkova, R. J. Darling, D. B. Bennett, M. Sekar, E. Hommema, A. M. Liang, E. S. Day, J. Inman, S. M. Karlicek, S. J. Ullrich, D. Hodges, T. Chu, E. Sullivan, J. Simpson, A. Rafique, B. Luginbühl, S. N. Westin, M. Bynum, P. Cachia, Y.-J. Li, D. Kao, A. Neurauter, M. Wong, M. Swanson and D. G. Myszka, *Anal. Biochem.*, 2004, **330**, 98–113.

11 D. G. Myszka, Y. N. Abdiche, F. Arisaka, O. Byron, E. Eisenstein, P. Hensley, J. A. Thomson, C. R. Lombardo, F. Schwarz, W. Stafford and M. L. Doyle, *J. Biomol. Tech.*, 2003, **14**, 247–69.

12 M. McNutt, *Science*, 2014, **346**, 679.
13 J. P. A. Ioannidis, *PLoS Med.*, 2014, **11**, e1001747.
14 C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S. A. Schunk, K. Wagemann and D. Linke, *ChemCatChem*, , DOI:10.1002/cctc.202001974.
15 M. D. Wilkinson, R. Verborgh, L. O. B. da Silva Santos, T. Clark, M. A. Swertz, F. D. L. Kelpin, A. J. G. Gray, E. A. Schultes, E. M. van Mulligen, P. Ciccarese, A. Kuzniar, A. Gavai, M. Thompson, R. Kaliyaperumal, J. T. Bolleman and M. Dumontier, *PeerJ Comput. Sci.*, 2017, **2017**, e110.
16 P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, A. Brazma and A. Brazma, *Genome Biol.*, 2002, **3**, RESEARCH0046.
17 P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu and R. Aebersold, *Nat. Biotechnol.*, 2004, **22**, 1459–1466.
18 M. Larralde, T. N. Lawson, R. J. M. Weber, P. Moreno, K. Haug, P. Rocca-Serra, M. R. Viant, C. Steinbeck and R. M. Salek, *Bioinformatics*, 2017, **33**, 2598–2600.
19 U. Wittig, R. Kania, M. Bittkowski, E. Wetsch, L. Shi, L. Jong, M. Golebiewski, M. Rey, A. Weidemann, I. Rojas and W. Müller, *Perspect. Sci.*, 2014, **1**, 33–40.
20 U. Wittig, R. Kania, M. Golebiewski, M. Rey, L. Shi, L. Jong, E. Algaa, A. Weidemann, H. Sauer-Danzwith, S. Mir, O. Krebs, M. Bittkowski, E. Wetsch, I. Rojas and W. M??ller, *Nucleic Acids Res.*, 2011, **40**, D790–D796.
21 I. Schomburg, A. Chang and D. Schomburg, *Nucleic Acids Res.*, 2002, **30**, 47–49.
22 R. Apweiler, R. Armstrong, A. Bairoch, A. Cornish-Bowden, P. J. Halling, J.-H. S. Hofmeyr, C. Kettner, T. S. Leyh, J. Rohwer, D. Schomburg, C. Steinbeck and K. Tipton, *Nat. Chem. Biol.*, 2010, **6**, 785.
23 K. F. Tipton, R. N. Armstrong, B. M. Bakker, A. Bairoch, A. Cornish-Bowden, P. J. Halling, J.-H. Hofmeyr, T. S. Leyh, C. Kettner, F. M. Raushel, J. Rohwer and D. Schomburg, *Perspect. Sci.*, 2014, **1**, 131–137.
24 N. Swainston, A. Baici, B. M. Bakker, A. Cornish-Bowden, P. F. Fitzpatrick, P. Halling,

T. S. Leyh, C. O'Donovan, F. M. Raushel, U. Reschel, J. M. Rohwer, S. Schnell, D. Schomburg, K. F. Tipton, M.-D. Tsai, H. V. Westerhoff, U. Wittig, R. Wohlgemuth and C. Kettner, *FEBS J.*, 2018, **285**, 2193–2204.

25 M. Dörr, M. P. C. Fibinger, D. Last, S. Schmidt, J. Santos-Aberturas, D. Böttcher, A. Hummel, C. Vickers, M. Voss and U. T. Bornscheuer, *Biotechnol. Bioeng.*, 2016, **113**, 1421–1432.

26 R. H. Ringborg, A. Toftgaard Pedersen and J. M. Woodley, *ChemCatChem*, 2017, **9**, 3285–3288.

27 M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. . 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. a. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.

28 M. Hucka, F. T. Bergmann, A. Dräger, S. Hoops, S. M. Keating, N. Le Novère, C. J. Myers, B. G. Olivier, S. Sahle, J. C. Schaff, L. P. Smith, D. Waltemath and D. J. Wilkinson, *J. Integr. Bioinform.*, , DOI:10.1515/jib-2017-0081.

29 N. Le Novère, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence and B. L. Wanner, *Nat. Biotechnol.*, 2005, **23**, 1509–1515.

30 J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, *Nucleic Acids Res.*, 2016, **44**, D1214-9.

31 T. UniProt Consortium, *Nucleic Acids Res.*, 2018, **46**, 2699.

32 M. Courtot, N. Juty, C. Knüpfer, D. Waltemath, A. Zhukova, A. Dräger, M. Dumontier, A. Finney, M. Golebiewski, J. Hastings, S. Hoops, S. Keating, D. B. Kell, S. Kerrien, J. Lawson, A. Lister, J. Lu, R. Machne, P. Mendes, M. Pocock, N. Rodriguez, A. Villeger, D. J. Wilkinson, S. Wimalaratne, C. Laibe, M. Hucka and N. Le Novère, *Mol. Syst. Biol.*, 2011, **7**, 543.

33 F. T. Bergmann, R. Adams, S. Moodie, J. Cooper, M. Glont, M. Golebiewski, M. Hucka, C. Laibe, A. K. Miller, D. P. Nickerson, B. G. Olivier, N. Rodriguez, H. M. Sauro, M. Scharm, S. Soiland-Reyes, D. Waltemath, F. Yvon and N. Le Novère, *BMC*

Bioinformatics, 2014, **15**, 369.

34 S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes and U. Kummer, *Bioinformatics*, 2006, **22**, 3067–3074.

35 P. C. F. Buchholz, R. Ohs, A. C. Spiess and J. Pleiss, *Biotechnol. J.*, 2019, **14**, 1–8.

36 K. Wolstencroft, O. Krebs, J. L. Snoep, N. J. Stanford, F. Bacall, M. Golebiewski, R. Kuzyakiv, Q. Nguyen, S. Owen, S. Soiland-Reyes, J. Straszewski, D. D. Van Niekerk, A. R. Williams, L. Malmström, B. Rinn, W. Müller and C. Goble, *Nucleic Acids Res.*, 2017, **45**, D404–D407.

37 M. Crosas, *D-Lib Mag.*, , DOI:10.1045/january2011-crosas.

38 J. Pleiss, *ChemCatChem*, , DOI:10.1002/CCTC.202100822.

39 P. Fernandes, *Int. J. Mol. Sci.*, 2010, 11, 858–879.

40 K. S. Rabe, J. Müller, M. Skoupi and C. M. Niemeyer, *Angew. Chemie - Int. Ed.*, 2017, 56, 13574–13589.

41 C. Barillari, D. S. M. Ottoz, J. M. Fuentes-Serna, C. Ramakrishnan, B. Rinn and F. Rudolf, *Bioinformatics*, 2016, **32**, 638–40.

42 P. Tremouilhac, A. Nguyen, Y.-C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung and S. Bräse, *J. Cheminform.*, 2017, **9**, 54.

43 H. Bär, R. Hochstrasser and B. Papenfuß, *J. Lab. Autom.*, 2012, **17**, 86–95.

44 C. D. Christensen, J. H. S. Hofmeyr and J. M. Rohwer, *Bioinformatics*, 2018, **34**, 124–125.

Figures

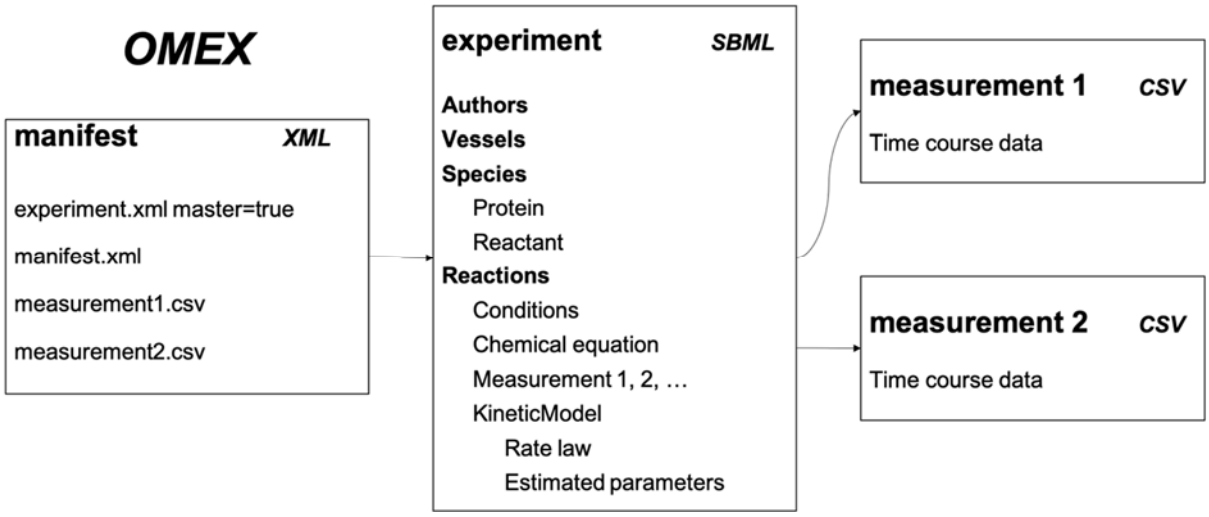

Fig. 1: Structure of an EnzymeML document. An EnzymeML document is a ZIP container in OMEX format and contains the experiment file (SBML) with the metadata of the experiment, the kinetic model, and the estimated kinetic parameters, and the measurement files (CSV) with the time courses of substrate and product concentrations. The manifest file (XML) lists the content of the ZIP container.

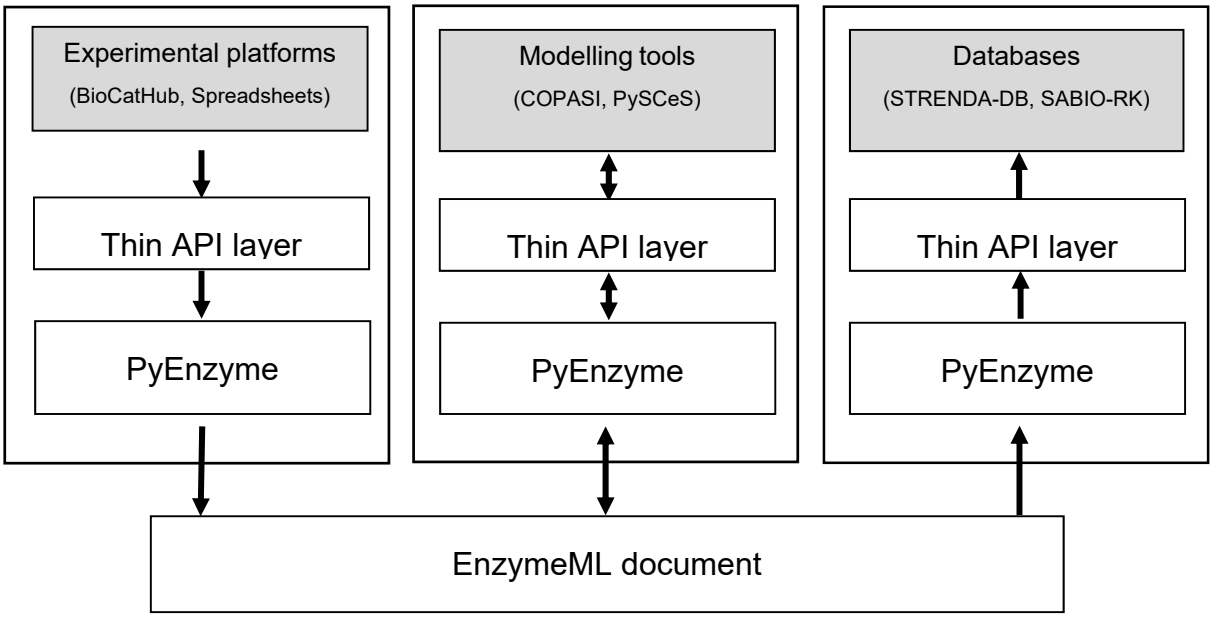

Fig. 2: Integration of applications. The EnzymeML document serves as a container to transfer data between applications such as experimental platforms, modelling tools, and databases for publication of enzymatic experiments. The EnzymeML API consists of a Python library PyEnzyme and provides read and write functionalities to the applications. The API is adapted to each application by an application-specific thin API layer.