STRENDA DB: enabling the validation and sharing of enzyme kinetics data

Neil Swainston[1], Antonio Baici[2], Barbara M. Bakker[3], Athel Cornish-Bowden[4], Paul F. Fitzpatrick[5], Peter Halling[6], Thomas S. Leyh[7], Claire O'Donovan[8], Frank M. Raushel[9], Udo Reschel[10], Johann M. Rohwer[11], Santiago Schnell[12], Dietmar Schomburg[13], Keith F. Tipton[14], Ming-Daw Tsai[15], Hans V. Westerhoff[16], Ulrike Wittig[17], Roland Wohlgemuth[18], Carsten Kettner[10,*]

[1]Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals, Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, United Kingdom.

[2]Department of Biochemistry, University of Zürich, Zürich, Switzerland.

[3]University of Groningen, University Medical Center Groningen, Hanzeplein 1, NL-8713 GZ Groningen, The Netherlands.

[4]Aix Marseille Univ, CNRS-IMM-BIP, 31 chemin Joseph-Aiguier, F-13009 Marseille, France.

[5]The University of Texas Health Science Center, 7703 Floyd Curl Dr., San Antonio, TX 78229-3900, USA.

[6]WestCHEM, Department of Pure & Applied Chemistry, University of Strathclyde, 16 Richmond Street, Glasgow G1 1XQ, United Kingdom.

[7]The Albert-Einstein-College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA.

[8]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus Hinxton, Cambridge CB10 1SD, United Kingdom.

[9]Texas A&M University, Department of Chemistry, College Station, TX 77843-3255, USA.

[10]Beilstein-Institut, Trakehner Straße 7–9, D-60487 Frankfurt am Main, Germany.

[11]Department of Biochemistry, University of Stellenbosch, Stellenbosch, South Africa.

[12]Department of Molecular & Integrative Physiology, and Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, 1000 Wall Street, Ann Arbor, MI 48109, USA.

[13]Technical University of Braunschweig, Bioinformatics and Systems Biology, Langer Kamp 19b, D-38106 Braunschweig, Germany.

[14]Trinity College Dublin, School of Biochemistry and Immunology, College Green, Dublin 2, Ireland

[15]Academia Sinica, Institute of Biochemical Sciences, 128 Sec. 2., Academia Rd., Nankang, Taipei, 115, Taiwan.

[16]Manchester Centre for Integrative Systems Biology, and School for Chemical Engineering and Analytical Science, University of Manchester, Manchester M1 7DN, United Kingdom; Synthetic Systems Biology and Nuclear Organization, Swammerdam Institute for Life Science, University of Amsterdam, The Netherlands; Molecular Cell Biology, Faculty of Sciences, Vrije Universiteit Amsterdam, The Netherlands.

[17]Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany.

[18]Sigma-Aldrich, Member of Merck Group, Industriestraße 25, CH-9470 Buchs, Switzerland.

*Corresponding author: Beilstein-Institut, Trakehner Straße 7–9, D-60487 Frankfurt am Main, Germany; ckettner@beilstein-institut.de; ORCID: 0000-0002-8697-6842, Tel.: +49 69 7167 3221.

Running title: STRENDA DB: enabling enzyme kinetics data sharing

Article type : Commentary

Keywords: enzyme, kinetics, enzymology, database, enzymology.

Conflict of interest: the authors declare no conflict of interest

Summary

STRENDA DB, freely available at http://www.strenda-db.org, is an online validation and storage system for functional enzyme data that aims at being integrated into the publication practices of the scientific community and into the publication processes of journals. It provides a simple-to-use web submission tool and searchable database allowing the sharing, comparison and accurate reporting of enzyme kinetics data.

The submission tool incorporates the STandards for Reporting ENzymology DAta (STRENDA), Guidelines which specify minimum information requested in the reporting of enzyme function data, including kinetic parameter values and full experimental conditions under which they were acquired. STRENDA DB checks the manuscript data entered by the author for compliance with the STRENDA Guidelines. If data is submitted prior to or during the publication process, the submission tool aids the author of a manuscript in the submission of kinetic parameters, ensuring that all required data and metadata are supplied. Data sets compliant with the Guidelines are assigned a STRENDA Registry Number and registered a Direct Object Identifier (DOI), which provides a perennial and resolvable identifier for each dataset. The data will normally be publicly available in STRENDA DB only after the corresponding article has been peer-reviewed and published in a journal. Data can also be submitted after publication.

By promoting the practice of simultaneously submitting articles to journals and kinetics data to STRENDA DB, reviewers of journal articles as well as authors and consumers of data will benefit from the availability of standardised data in multiple ways.

Introduction

Enzyme kinetics is important to many fields within the biological sciences and is a discipline practiced by a large number of researchers. The study of enzyme functions has led to important developments for the sustainable production of a wide variety of compounds in the food, pharmaceutical, flavour and fragrance, agro- and chemical industries [1], and the discovery of novel enzyme functions. These activities cross frontiers for both fundamental and applied research. If biology is to be understood as a dynamical process, then researchers need quantitative data on the regulation and energetics of enzymes.

To date, enzymology data is available in repositories such as BRENDA [2] and SABIO-RK [3]. While these resources are extensively curated by experts, the quality and completeness of the data depends on the quality of data available in the scientific literature. All too often, however, essential metadata about the conditions under which kinetic parameters were obtained (e.g. temperature, pH, ionic strength, enzyme and substrate concentrations, presence of activators and inhibitors) are not comprehensively reported in papers. Such omissions make compiling of necessary metadata, and therefore reuse and comparison of datasets, difficult [4, 5]. These difficulties become even more acute for those wishing to use published data to model the behaviour of metabolic systems, cellular behavior or the interaction of cells within tissues and organs. This is the case in particular for systems biologists, who require reliable data for enzymes from many Enzyme Classification (EC) classes to be able to produce accurate predictive models. Specialized repositories for specific enzyme classes, such as the CAZy database [6] that focuses on structural and functional information about enzymes which assemble, modify and break down oligo- and polysaccharides, are limiting their datasets on their topics, while systems biocatalysis and systems biology approaches need to collect data in different formats from specialized repositories and the scientific literature.

To mitigate this problem, the STandards for Reporting ENzymology DAta (STRENDA) guidelines were developed [7, 8], following a community-based discussion of the currently accepted best approaches for data reporting in enzyme research. The goal of these guidelines is to improve the quality of data reporting in the scientific literature, enabling readers and reviewers to interpret, evaluate and corroborate the experimental findings. Since their approval, more than 50 biochemistry journals have recommended that their authors follow the STRENDA Guidelines when reporting functional enzymology data (see http://www.beilstein-institut.de/en/projects/strenda/journals).

However, despite the existence of the STRENDA Guidelines, many publications still do not describe the experimental conditions and results in sufficient detail to allow the experiment to be reproduced, a topic that has recently attracted considerable attention [9, 10]. Furthermore, it is clear that not only researchers could benefit from having a resource that indicates best practices for the reporting of enzyme kinetics data, but that the value and impact of published work in biocatalysis could be increased, thereby promoting increased citations and further growth of applications [11].

It is now common practice for scientists to submit experimental data to public repositories as a result of policies established by journals and funding agencies. There are multitudinous databases and repositories for ‘omics data, such as ArrayExpress [12], PRIDE [13], MetaboLights [14], and PDB [15]. These resources provide user-interfaces enabling researchers to share transcriptomics, proteomics, metabolomics and protein structure data. However, to date, there is no similar resource to encourage the user submission of enzyme kinetics data for biological molecules.

This paper describes a functional enzyme database, STRENDA DB. In contrast to the available enzyme resources such as BRENDA and SABIO-RK, STRENDA DB has been designed specifically to accept data submissions directly from the research community, ensuring that newly-acquired enzyme kinetics data are collected with appropriate metadata as it enters the literature. STRENDA-DB implements the STRENDA Guidelines in an intuitive and easy-to-use web-based form, facilitating the submission and sharing of data, aiding the literature review process, and increasing the visibility, accessibility and impact of enzyme kinetics publications. The system provides a community-driven and continually-updated enzyme kinetics resource supporting enzymology research. Currently, more than ten journals already recommend their authors both to apply STRENDA DB to validate their manuscript data on completeness and to deposit this data in the database. A related initiative, BioCatNet [16] also accepts kinetic data from authors, particularly raw data on reaction progress and initial rates. It uses an Excel sheet for data entry, and handles some complications found in applied biocatalysis.

Description of the components of STRENDA DB

The STRENDA DB web-based interface, hosted by the Beilstein-Institut, is freely available at http://www.strenda-db.org, and offers two tools: i) data submission; and ii) data query. The design of the user interface fulfils the requirements of a responsive design that allows the user to submit and query data from any device connected with the internet. The web application has been implemented using Primefaces 4.0 and JSF (Java Servlets) 2.1, and the data is stored in an Oracle 12C database.

Data submission

The data submission tool collects data and metadata from users. The goal is to collect data during the preparation of conventional journal submissions to improve the quality of enzyme data reported in the literature. On the basis of the STRENDA Guidelines the data is collected in a common and standardised format. This will simplify the review process, reproducibility of enzyme assays as well as the accessibility of information to the community.

Authors enter the relevant functional enzyme data from their manuscript into the data submission system. The data entry requires the description of the minimum information on materials, methods and assay conditions, as well as the experimental results based on the corresponding experimental conditions. The minimum information is defined by the STRENDA Guidelines, and determines the compulsory fields in the entry section of STRENDA DB. The system validates automatically the data entered in the compulsory fields against completeness and formal correctness (e.g. pH range, defined temperature range). When required information is missing the user receives detailed warning information. After the successful finalization of the data input, the author receives a STRENDA Registry Number (SRN) for each data set, providing an unambiguous identifier comparable to the UniProt AC for protein data sets [17]. In addition, each dataset is assigned a DOI that allows data referencing and access. The data becomes publicly available in the database only after the corresponding article has been peer-reviewed and accepted for publication in a journal.

The STRENDA Guidelines require a full description of the identity of the catalytic or binding entity (enzyme, protein, nucleic acid or other molecule). This information should include the origin or source of the molecule, its purity, composition and other characteristics, such as post-translational modifications, mutations and any modifications made to facilitate expression or purification. The assay methods and exact experimental conditions of the assay must be fully described if it is a new assay or provided as a reference to previously published work, with or without modifications. The temperature, pH and pressure (if other than atmospheric) of the assay must always be included, even if previously published.

The data submission to STRENDA DB is possible only after registration and login into the submission system. This allows the user to interrupt the entry process without losing data already entered. In addition, it identifies researchers responsible for the data input to the database development and curation team.

The data submission tool was designed to streamline the data collection process. The web-based tool allows simple navigation through the submission system, providing extensive help tooltips and hints, as well as autofill functionality when specifying enzymes and small molecules by making use of UniProt [17] and PubChem [18], respectively.

The overall concept of the submission system of STRENDA DB reflects the structure of a manuscript, i.e. introduction, materials and methods, results, discussion and references. For the data input, the materials and methods as well as the results section are most relevant. The submission tool therefore acts as a structural support for the author guided by this general manuscript structure when entering data into STRENDA DB. In addition, the design around the STRENDA Guidelines allows authors to identify required data for entering in the database.

In STRENDA DB the top level of structure is a “Manuscript”, typically containing all the data that might ultimately appear in a published paper, specified by its title and authors. A Manuscript can contain data for one or more “Experiments”, each of which involves the study of one specific protein as the active enzyme (Figure 1). This structure allows the user to enter data from the comparison of, for example, the activity of two isozymes, such as two mutant proteins, each of which would be a different Experiment. The core of the definition of an Experiment is the basic data on the protein, such as protein identification, sequence modifications (PSMs), post-translational modifications (PTMs), source and the typical reaction which it catalyses. For each Experiment, there will be one or more “Datasets”. Each Dataset consists of one defined assay condition linked to the experimental result(s), for example, the determination of kinetic parameters at a defined pH. The effects of changes in conditions that can be summarised by kinetic parameters, such as different substrate or inhibitor concentrations, are captured within a single Dataset. But changes in substrate identity or temperature, for example, would require a different Dataset. In the case of a pH profile, the Experiment will contain several Datasets each with different pH values but with the same assay components connected to the pH dependent kinetic parameters. In consequence, when entering tabular data (pH profiles can be represented in tables), the author needs to enter the description of the enzyme assay only once and only varies the specific parameters for the subsequent assay conditions.

The following examples may illustrate the concept:

1. The kinetics of human hexokinase is explored using varied initial concentrations of substrate ‘A’ (Figure 2). The methods used and techniques applied are described in a specific text box and the protein assayed is defined (with UniProt definitions and EC numbers), e.g. hexokinase. The single Dataset includes the data pair of the components used in this assay (with the varied initial substrate 'A' concentrations ranging from a minimum to a maximum value) and the experimental results, i.e. the corresponding kinetics parameters.

2. The kinetics of the yeast pyruvate kinase 1 (PYK1) is investigated at various pH values (3 to 9). Again, methods used and techniques applied are described, followed by the description of the pyruvate kinase assayed (as the Experiment). The first assay starts at pH 3 and the corresponding kinetic results are added. This makes the first $Dataset_1$. For each subsequent assay, most components of $Dataset_1$ remain constant, with only the pH parameter being changed as the corresponding kinetic results are entered (Figure 3). Similarly, such an approach is applicable to represent the kinetics at various assay temperatures. In principle, any modification in the assay conditions can affect the kinetic parameters and thus this data is kept in the ‘container’ defined as a Dataset.

   For the experiment that studies the pH profile of pyruvate kinase 1, the scheme reads as follows:

   $Dataset_1$: components used in this assay (Assay Conditions) at pH 3 and corresponding kinetics parameters (Results).

   $Dataset_2$: components used in the assay conditions from $Dataset_1$ (just copied and pasted from here) but at pH 4 and corresponding kinetics parameters.

   An additional five Datasets can be similarly input (one for each pH step from 5 to 9).

3. The kinetics of the yeast pyruvate kinases, PYK1 and 2 are investigated and compared at various pH (3 to 9). For the input of these data, two Experiments are defined, one for PYK1 and one for PYK2, since two different proteins need to be described. The Datasets are entered for each protein as described in example 2 above. As is displayed in Figure 4, this data representation builds a tree of Experiments and Datasets. The complexity of such a data tree can grow to support increasingly specific experimental designs.

In addition, if inhibitors or activators are used in the experiment, the first Dataset would include the kinetic parameters without the inhibitor or activator. The subsequent Dataset provides the kinetics parameters that are dependent on the added inhibitor or activator. If several inhibitors are tested the number of Datasets corresponds to the number of inhibitors.

It should be noted that the data input does not result in the simple completion of a checklist. Rather the details that must be included depend on the nature of the enzyme, the type of experiment performed and what results are to be reported. The STRENDA DB system already recognises these complexities, in particular by providing expandable sections for details only required under particular circumstances. Thus, kinetic parameters for activators or inhibitors can be only entered if activators or inhibitors have been defined in the description of the assay conditions. As the system further develops, it is envisaged that more sophisticated automated validation steps will be introduced. Similarly, over time further expandable sections will be added to support more complex experiments.

Successful data input results in assignment of both the STRENDA Registry Number (SRN) and a DOI, which are identifiers for the data within an Experiment on the functional properties of a single enzyme. Thus, multiple SRNs and DOIs can be linked analogously to one manuscript containing one or more Experiments. The user can therefore subsequently query the database for a given publication using a PubMed identifier (PMID) and obtain the number of SRNs and Experiments along with the assay conditions and experimental results respectively. The DOIs are automatically registered with DataCite (https://www.datacite.org) to enable users not only to search the metadata of datasets but also to support the community by providing a perennial, resolvable identifier for each dataset in STRENDA DB (Figure 5).

Data query

The query interface is accessed via the ‘Query’ button in the menu. The interface has been kept straightforward and simple by following the search mask of major search engines such as Google. For querying STRENDA DB neither registration nor login is required. The user can search in the database using key terms such as protein name, EC number, UniProt accession number, organism, author name, PMID, SRN or DOI. For an initial overview the search mask can be left empty and all datasets published are displayed.

The hit list consists of a table that displays entries for all the key terms mentioned above plus a column with hyperlinks that provide access to: i) the experimental overview; ii) the fact sheet downloadable as a PDF file; and iii) an experimental XML file (Figure 6). The experimental overview is accessed via the ‘Show’ button in the right hand column of the hit list table. This page displays the header data such as the manuscript title and the names of the authors as well as the identifiers of this data set (SRN and DOI) along with the most important data on the protein studied. The header data are followed by the list of Datasets, which include the assay conditions with the calculated kinetic parameters (Figure 7).

The fact sheet contains all input data in a human-readable format (Supporting Information File S1) and contains far more information than the experimental overview page, including the sequence of the protein, identifiers of chemical compounds used in the assay, and concentration of enzyme in the assay and data on how this was measured. The fact sheet can be extended by additional data such as International Union of Pure and Applied Chemistry (IUPAC) names and the IUPAC International Chemical Identifier (InChI) of the compounds used in the assay. Authors are encouraged to submit the fact sheet to the journal as supplementary information along with the main manuscript to the journal. The supplementary information is not only considered for publication, but also indicates that the reporting of the enzyme assays is in compliance with the STRENDA Guidelines; the SRN assigned indicates that all relevant information is provided in the manuscript or its supplementary information.

Since all data sets are assigned a DOI and can be cited elsewhere, there is an alternative way to search and directly access datasets deposited in STRENDA DB; clicking on a hyperlinked DOI leads the user to the corresponding hit page, which is linked to both the Experiment overview page and the data fact sheet PDF (Figure 8).

Workflow

The STRENDA Commission strongly encourages the scientific community to incorporate the STRENDA DB in the general publication workflow. It is proposed to authors to submit their enzyme function data to STRENDA DB, where this data is automatically validated on compliance with the STRENDA Guidelines. A successful formal compliance is confirmed by the awarding of a SRN and documented in a fact sheet (in PDF format) containing all input data that can be submitted with the manuscript to the journal. Once the corresponding article has been peer-reviewed and published in the journal, the bibliographic data, in the form of a PMID, is added and the experimental data is made publicly accessible in STRENDA DB (Figure 9).

The direct electronic submission of data by the authors prior to or during the publication has proven to be the gold standard for comprehensive data acquisition for protein structures in PDB [19]. We expect that the STRENDA DB would become the analogous tool to PDB for enzyme functional data.

Discussion

STRENDA DB is the first database adhering to community-based guidelines for ensuring reproducibility of enzyme kinetics data. It is designed to aid the data provider in publishing and sharing data, the manuscript reviewer in interpreting data during the review process, the data consumer in finding, comparing and utilising publicly available kinetics data, and the funding agency increasing research impact and availability of data. The checking for completeness and validation offered by the STRENDA DB system benefits all involved in the process of reporting and publishing. Authors will be assured that they have comprehensively recorded all essential details of the experiment – and hence reduce problems that currently can occur with data reproducibility. Journal reviewers and editors can be assured that the data and metadata underlying a publication has been reported fully and will eventually be available to the whole scientific community. Readers of a published paper will know that a comprehensive description of the experiments and results is available in a standardised format.

Supporting the review process is a key consideration of STRENDA DB, although it will of course be for individual journals to decide if and how to incorporate STRENDA DB into their review and publication policies. It is hoped the catalysis community and its journals will move towards a model in which authors would be required to submit the underlying data to STRENDA DB at the point of manuscript submission. This would be a logical extension of the current state in which journals request that authors follow the written STRENDA Guidelines in preparing a manuscript. Journals could also require that the dataset be made publicly available at the point of publication. This mirrors the approach taken with a range of 'omics data types, including that of protein structure data and the PDB. However, validation of a dataset as STRENDA compliant is not intended to replace the general review process. STRENDA DB merely checks that an enzyme function experiment has been comprehensively described, and makes no judgment on the scientific quality. Reviewers and editors will still need to evaluate the importance of the topic studied, the experimental design and the reliability of the results. The review process may be aided by access to the PDF summary fact sheet generated by STRENDA DB, which shows in a standardised format all data and metadata. As STRENDA DB develops, it may include additional automated checks on the submitted data based on appropriate validation criteria, but the final judgment on the integrity of the data will always be left to expert reviewers and editors.

All datasets in STRENDA DB are assigned a persistent DOI, which allows for their direct access via web browsers. Authors will be able to quote these to allow readers immediate access to the data once a paper has been published. Such an approach will increase the accessibility of experimental data, in accordance with the general trend of increasing data reuse and ensuring reproducibility. Through the DOI it will also be possible for authors and others to track the use of their datasets, and hence support the trend of rewarding data providers for sharing data in addition to the traditional performance metrics based upon citations of publications.

To facilitate the finding and reuse of datasets submitted to STRENDA DB, the system includes numerous cross references to well-used, publicly available databases, such as UniProt for definition of enzymes, ExplorENZ for the definition of EC numbers and reactions catalysed by the enzyme [20], and PubChem for specification of any small molecule compounds present in an assay mixture, such as substrates, products, buffers, salts, and inhibitors. At the time of data entry, such links are provided by searchable fields in the submission tool to aid the user. Including such facilities in the interface provide the advantage of reducing the amount of data that the submitter must supply manually and increasing the accuracy of supplied metadata. Furthermore, linking this metadata to external database identifiers facilitates data retrieval and integration with external applications and related data resources such as KEGG [21] and ChEBI [22, 23]. One such consumer of enzyme kinetics data is the systems biology community, who will be greatly aided by the availability of reliable enzyme activity data in a standardised and annotated format, from which realistic and predictive models of signalling and metabolic pathways may be built [24, 25, 26].

The STRENDA Commission is aware that the data entry process must be as simple as possible to minimise burden to authors, in particular those who are first-time users of the database. Apart from the data input process which reflects the schema of the common structure of a manuscript, the user is guided through the input process by tool tips associated with most of the input fields. During the data entry process, users receive specification of data required, and steps to take to continue data input. In addition, a comprehensive and downloadable user guide is available online, which provides the reader with a description of both the STRENDA DB and the step-by-step data input process. Finally, video tutorials (freely accessible at http://www.beilstein.tv/categories/strenda/) demonstrate step-by-step the data entry process in STRENDA DB.

Future developments

The Beilstein-Institut and the STRENDA Commission will support the upkeep and development of the database over the coming years, including the provision of data curation of entries submitted by the community. In time, it is hoped that the STRENDA DB will provide access to kinetics data covering a multitude of enzymes from prokaryotic and eukaryotic proteomes. It is recognised, however, that the current release of STRENDA DB is an initial version, and as such only handles the most common experimental procedures. Over time, and with the benefit of user feedback, STRENDA DB will improve its functionality and to cover a broader range of experimental methods and data types. Additional features will be introduced in such a way that limits additional demands on the user. For example, fields specific to a particular experiment type will be hidden in expandable sections when not required. The current system already makes extensive use of such facilities; for example, it provides details of protein sequence and post-translational modifications. Many future developments are envisaged, and their implementation will be prioritised in consultation with the user community, who are encouraged to provide their feedback.

The STRENDA Commission envisions a series of improvements for the database. The system could accept a more complete description of the kinetic equation to which data has been fitted to estimate parameters. This may be offered as a selection from a standard list, perhaps utilising existing resources and ontologies. Such enhanced definitions may incorporate methods to report rates of the formation of multiple products formed from the same substrate, such as two enantiomers of a given product. Specifying a kinetic equation would simplify the interface, limiting the parameter values required from the user, and would also allow validation algorithms to flag possible mistakes in the set of user-specified parameters. For example, a warning can be issued if a K_m value entered falls outside the reported range of substrate concentrations studied, violating conditions for validity of the kinetic equation [27], or if a reported rate would convert all substrate present in a few seconds (1000-fold mistakes in units are not uncommon). Similarly, values entered for effector concentrations studied could be automatically compared with kinetic parameters reported for those effectors. A description of the software used for data analysis could be included along with calculated errors for all parameters.

An extended system for the specification of macromolecular ingredients (other than the enzyme) in an assay mixture may be implemented, generating links to appropriate databases and utilising existing ontologies where appropriate. This will be especially relevant in considering special cases of data, including multi-component and multi-EC number enzymes. This may be extended to accommodate protein descriptions that differ from the wild-type description in UniProt, considering issues such as the presence of pro- and signal sequences and of zymogen peptides that have been cleaved in the actual protein studied, hetero-oligomer proteins made up of multiple UniProt entries, enzymes that are studied with tightly bound metal ions and prosthetic groups, especially where more than one variant is possible. (This may all be best solved by help text explaining how to describe these possibilities.) Automated cross-checking between the specification of PTMs against the protein sequence may also be introduced, ensuring that PTMs are only ever assigned to “allowed” residues.

The collection of additional metadata may be offered, including introduction of more structured fields to capture some items that currently go into the “Experimental Methods” free text box. Such fields may include, “What compound was monitored to follow the reaction?” and “What analytical / spectroscopic method was used to monitor it?” Again, such details will be determined in consultation with the user community. In instances where catalytic activity or binding cannot be detected, an estimate of the limit of detection based on the sensitivity and error analysis of the assay could be asked for.

Previous work has illustrated the feasibility of integrating the analysis of initial rate data or even progress curve data with the submission of enzyme kinetics data [28]. As such, introduction of more sophisticated data, including those on bisubstrate reactions, grid data sets, or time-course data will

also be investigated. Over time this may develop into a downloadable tool that can be used locally in labs at the time of experiment, incorporating analysis of raw experimental data, collection of appropriate metadata, performance of STRENDA validation, and seamless data transfer to the database.

Another consideration is the development of improved methods for data retrieval, including combined and complex queries, and the incorporation of a programmatically accessible API, allowing for both the submission and the extraction of data to be integrated with existing LIMS and electronic lab notebook systems.

Acknowledgements

NS acknowledges the funding from the Biotechnology and Biological Sciences Research Council (BBSRC) under grants BB/M017702/1, “Centre for synthetic biology of fine and speciality chemicals (SYNBIOCHEM)”, BB/K019783/1, “Continued development of ChEBI towards better usability for the systems biology and metabolic modelling community”, and BB/M006891/1, “Enriching Metabolic PATHwaY models with evidence from the literature (EMPATHY)”. This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM). SS acknowledges the funding from NIH / NIDDK under grant R25 DK088752.

STRENDA and STRENDA DB are completely funded by the Beilstein-Institut.

Author Contributions

CK initiated and drove the project. UR developed the STRENDA DB under guidance from CK. All authors contributed towards the design and testing of STRENDA DB, contributed initial data sets, and wrote and approved the manuscript.

References

1. Carbonell P, Currin A, Jervis AJ, Rattray NJ, Swainston N, Yan C, Takano E & Breitling R (2016) Bioinformatics for the synthetic biology of natural products: integrating across the Design-Build-Test cycle. *Nat Prod Rep*. 33, 925-32.

2. Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW & Schomburg D (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res*., 43, D439-46.

3. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A., Sauer-Danzwith H, Mir S, Krebs O, Bittkowski M, Wetsch E, Rojas I & Müller W (2012) SABIO-RK--database for biochemical reaction kinetics. *Nucleic Acids Res*., 40, D790-6.

4. Wittig U, Rey M, Kania R, Bittkowski M, Shi L, Golebiewski M, Weidemann A, Müller W & Rojas I (2014) Challenges for an enzymatic reaction kinetics database. *FEBS J*. 281:572-82.

5. Wittig U, Kania R, Bittkowski M, Wetsch E, Shi L, Jong L, Golebiewski M, Rey M, Weidemann A, Rojas I & Müller W (2014) Data extraction for the reaction kinetics database SABIO-RK. *Perspectives in Science*. 1(1-6), 33-40.

6. Lombard V, Golaconda RH, Drula E, Coutinho PM & Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490-5.

7. Apweiler R, Cornish-Bowden A, Hofmeyr JH, Kettner C, Leyh TS, Schomburg D & Tipton K (2005) The importance of uniformity in reporting protein-function data. *Trends Biochem Sci*. 30, 11-2.

8. Tipton KF, Armstrong RN, Bakker BM, Bairoch A, Cornish-Bowden A, Halling PJ, Hofmeyr J-H, Leyh TS, Kettner C, Raushel FM, Rohwer J, Schomburg D & Steinbeck C (2014) Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Perspectives in Science* 1(1-6), 131-137.

9. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM & Haendel MA. (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*. 1, e148.

10. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U & Silberberg SD (2012) A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012, 490, 187-91.

11. Gardossi L, Poulsen PB, Ballesteros A, Hult K, Svedas VK, Vasic-Racki D, Carrea G, Magnusson A, Schmid A, Wohlgemuth R & Halling PJ (2010) Guidelines for reporting of biocatayltic reactions. *Trends Biotechnol*. 28, 171-180.

12. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U & Brazma A (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res*., 43, D1113-6.

13. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R & Hermjakob H (2016) 2016 update of the PRIDE database and related tools. *Nucleic Acids Res*., 2016, 44, D447-D456.

14. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, González-Beltrán A, Sansone SA, Griffin JL & Steinbeck C (2013) MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res*., 41, D781-6.

15. Rose WP, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE & Burley SK (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res*. 43, D345-56.

16. Buchholz PC, Vogel C, Reusch W, Pohl M, Rother D, Spieß AC & Pleiss J (2016) BioCatNet: A Database System for the Integration of Enzyme Sequences and Biocatalytic Experiments. *Chembiochem*. 17, 2093-8.

17. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*. 43, D204-12.

18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J & Bryant SH (2016) PubChem Substance and Compound databases. *Nucleic Acids Res*. 44, D1202-13.

19. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* 20, 391-96

20. McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res*. 37, D593–7.

21. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H & Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27, 29-34.

22. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P & Steinbeck C (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. 44, D1214-9.

23. Swainston N, Hastings J, Dekker A, Muthukrishnan V, May J, Steinbeck C & Mendes P (2016) libChEBI: an API for accessing the ChEBI database. *J Cheminform*. 8, 11.

24. Krause F, Schulz M, Swainston N & Liebermeister W (2011) Sustainable model building the role of standards and biological semantics. *Methods Enzymol*. 500, 371-95.

25. Li P, Dada JO, Jameson D, Spasic I, Swainston N, Carroll K, Dunn W, Khan F, Malys N, Messiha HL, Simeonidis E, Weichart D, Winder C, Wishart J, Broomhead DS, Goble CA, Gaskell SJ, Kell DB, Westerhoff HV, Mendes P & Paton NW (2010) Systematic integration of experimental data and models in systems biology. *BMC Bioinformatics*. 11, 582.

26. Van Eunen K, Kiewiet JAL, Westerhoff HV & Bakker BM (2012) Testing Biochemistry Revisited: How In Vivo Metabolism Can Be Understood from In Vitro Enzyme Kinetics. *PLOS Comp. Biol*. 8, e1002483.

27. Schnell S (2014) Validity of the Michaelis-Menten equation—steady state or reactant stationary assumption: that is the question. *FEBS J*. 281, 464-72.

28. Swainston N, Golebiewski M, Messiha HL, Malys N, Kania R, Kengne S, Krebs O, Mir S, Sauer-Danzwith H, Smallbone K, Weidemann A, Wittig U, Kell DB, Mendes P, Müller W, Paton NW & Rojas I (2010) Enzyme kinetics informatics: from instrument to browser. *FEBS J*. 277, 3769-79.

Supporting material

Table S1: Example of an Experiment fact sheet from a real data set (doi:10.22011/strenda_db.KJWIQY), a PDF document containing data submitted to STRENDA DB in a human-readable format, which may itself be submitted as supporting information alongside a journal publication.

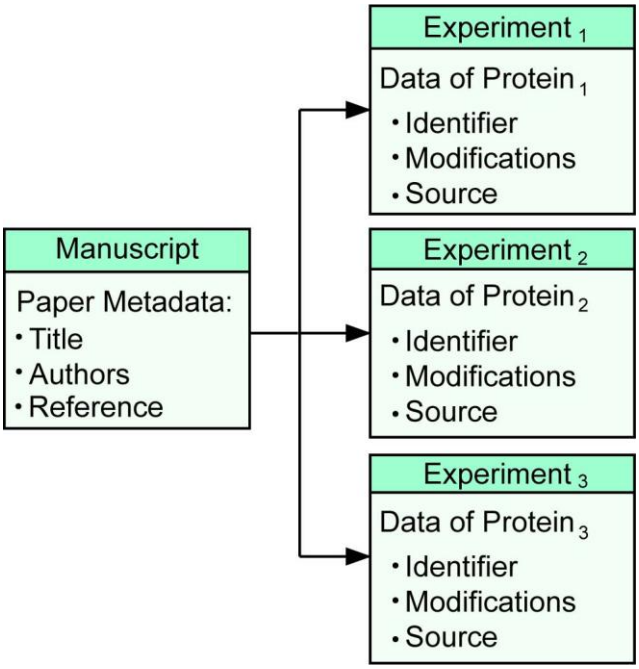

Figure 1: Creation of Experiment data container. Each Experiment is defined by the precise protein used in the assay. Before entering any assay data, the protein has to be identified unambiguously by identifier, protein sequence, etc.

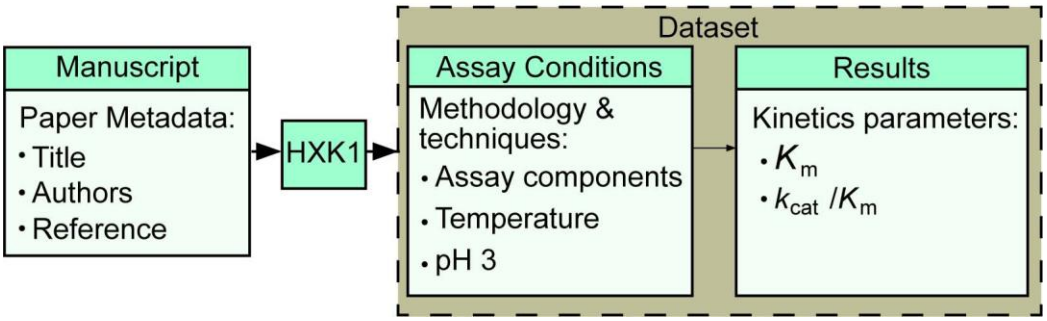

Figure 2: Creation of Dataset data containers. Since the experimental results are dependent on both the methodology applied and the assay components used, assay conditions and results form a pair, the Dataset. Each modification in either of the parts of this pair requires review or modification of the other one.

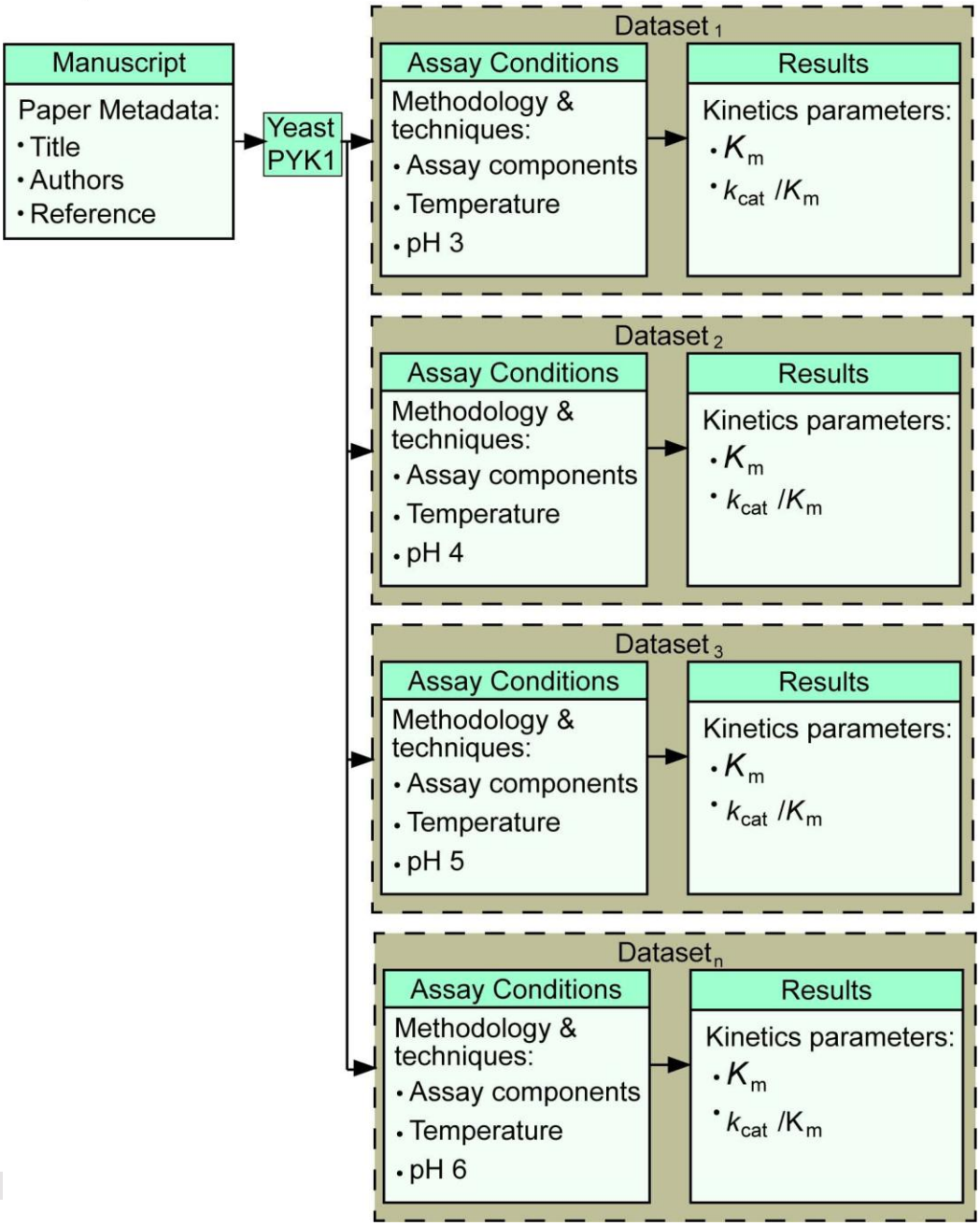

Figure 3: Input of tabular data. If, for example, series of data are published such as pH or temperature profiles, substrate specificities etc., the assay condition is entered just once for the first Dataset and copied and modified correspondingly for the subsequent Datasets. A series of Datasets is generated resulting in a quick input of tabular data.

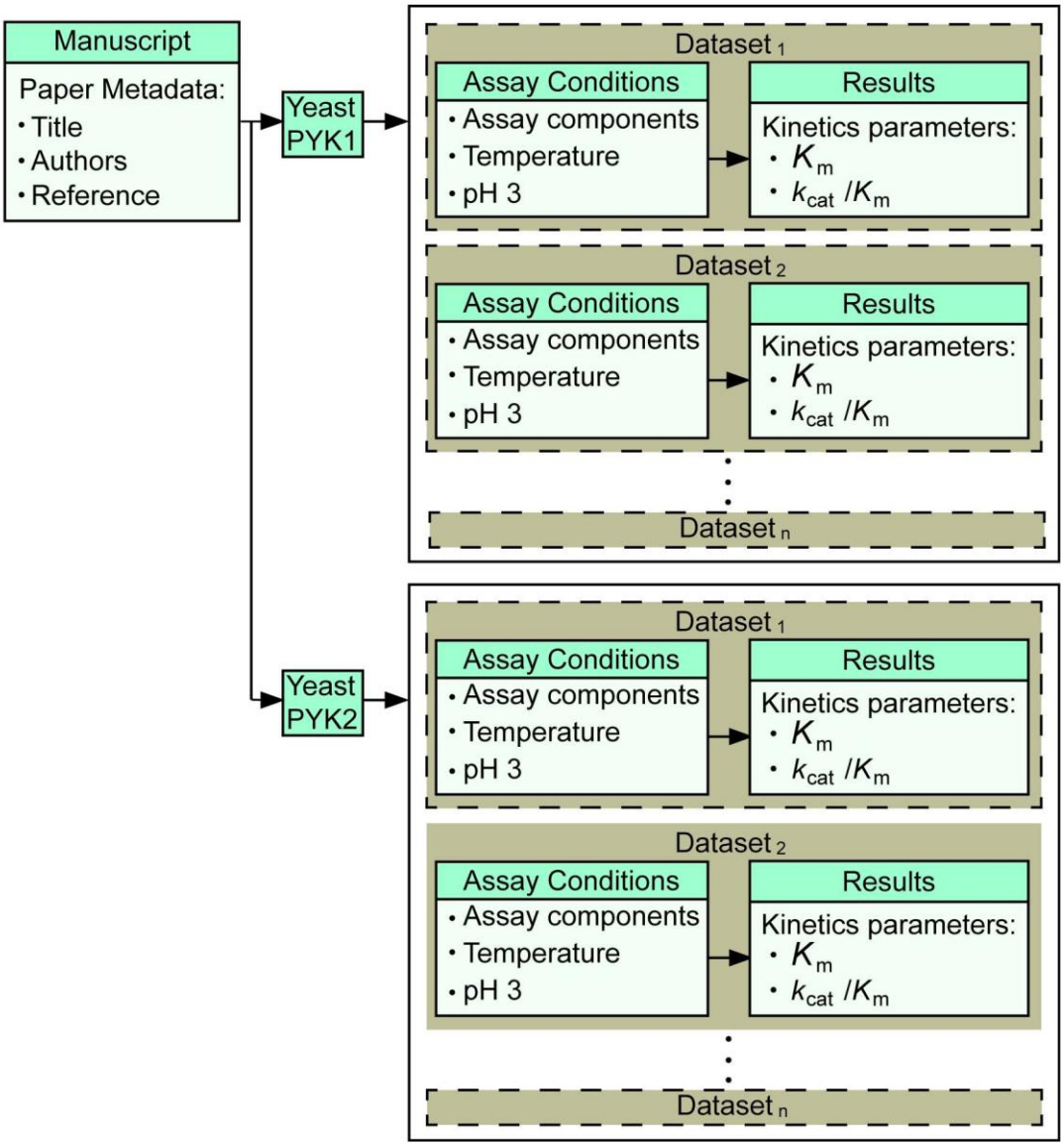

Figure 4: Input of the activity of more than one protein. Each protein activity characterized requires the definition of a new Experiment (see also Fig. 1) followed by the creation of corresponding Datasets. This input structure enables the user to enter experimental data from, e.g. the comparison of a native and a modified protein.

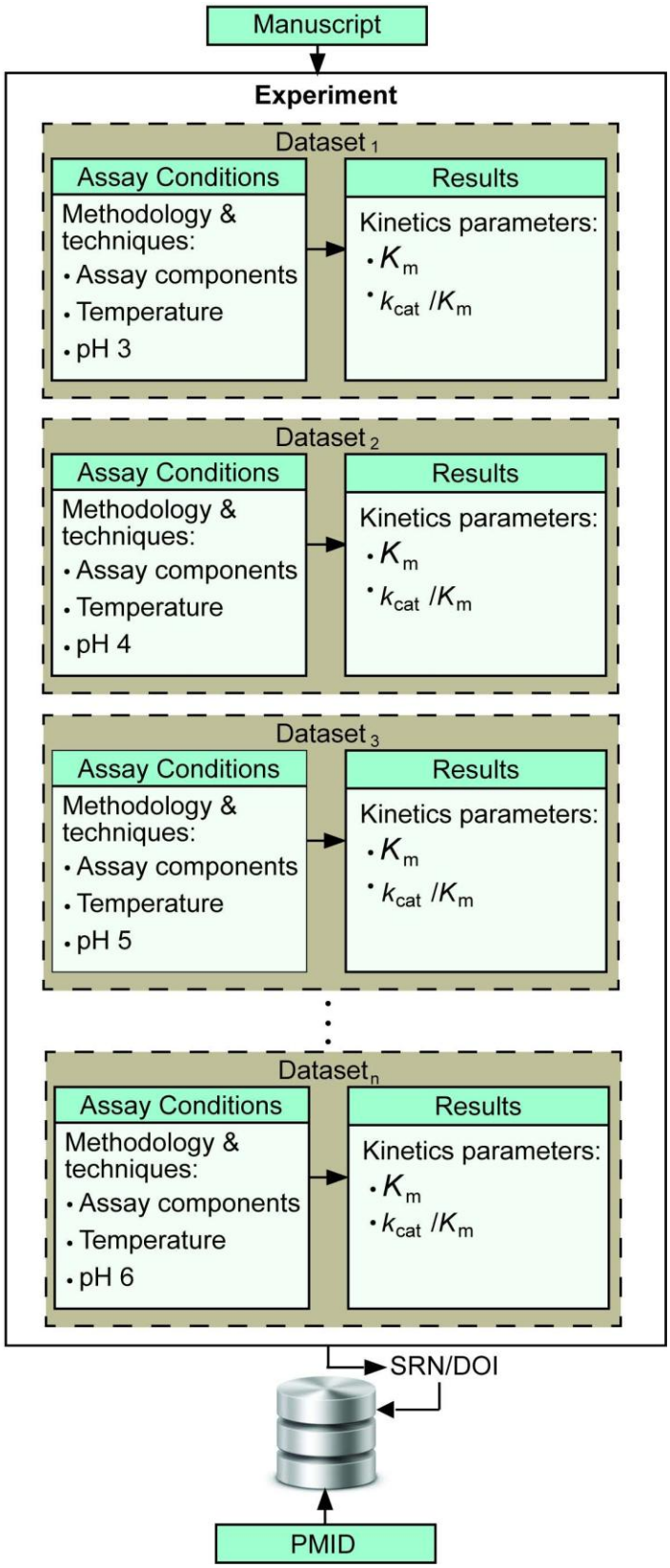

Figure 5: After the manuscript has been peer-reviewed and published in a journal, the bibliographic data (PubMed Identifier, PMID) is added and the experimental data is made publicly accessible in STRENDA DB.

Search STRENDA DB

Search Clear

Protein	UniProtKB AC	Organism	SRN	PMID	DOI	Actions
Putative transferase	Q0P8J6	Campylobacter jejuni	AJQQFK	26650156	10.22011/strenda_db.AJQQFK	Show Export PDF Export XML
Phenylalanine-4-hydroxylase (PAH) (Phe-4-monooxygenase)	P04176	E.coli	4ICZMX	25453233	10.22011/strenda_db.4ICZMX	Show Export PDF Export XML
Myeloblastin (AGP7) (C-ANCA antigen) (Leukocyte proteinase 3) (PR-3) (PR3) (Neutrophil proteinase 4) (NP-4) (P29) (Wegener autoantigen)	P24158	Homo sapiens (Human)	D9YPHC	9067256	10.22011/strenda_db.D9YPHC	Show Export PDF Export XML
Neutrophil elastase (Bone marrow serine protease) (Elastase-2) (Human leukocyte elastase) (HLE) (Medullasin) (PMN elastase)	P08246	Homo sapiens (Human)	5V5MWU	9067256	10.22011/strenda_db.5V5MWU	Show Export PDF Export XML
TrpB2o from Arabidopsis thaliana	N.A	Arabidopsis thaliana	3IZNOK	25184516	10.22011/strenda_db.3IZNOK	Show Export PDF Export XML
TrpB2o from Thermotoga maritima	Q97TX6	Thermotoga maritima	LQZ1C3	25184516	10.22011/strenda_db.LQZ1C3	Show Export PDF Export XML
Intestinal-type alkaline phosphatase (IAP) (Intestinal alkaline phosphatase)	P19111	Bos taurus (Bovine)	Y9CBQW	20025997	10.22011/strenda_db.Y9CBQW	Show Export PDF Export XML
Schistosoma hematobium SULT	N.A	E coli	FFZAQ9	28536265	10.22011/strenda_db.FFZAQ9	Show Export PDF Export XML
Schistosoma japonicum SULT	N.A	E. coli	IH7FNY	28536265	10.22011/strenda_db.IH7FNY	Show Export PDF Export XML
Myeloblastin (AGP7) (C-ANCA antigen) (Leukocyte proteinase 3) (PR-3) (PR3) (Neutrophil proteinase 4) (NP-4) (P29) (Wegener autoantigen)	P24158	Homo sapiens (Human)	3JIILJ	9067256	10.22011/strenda_db.3JIILJ	Show Export PDF Export XML
RNA lariat debranching enzyme, putative	C4M1P9	E. coli	DVE8AT	27930312	10.22011/strenda_db.DVE8AT	Show Export PDF Export XML
Sulfotransferase (Sulfotransferase oxamniquine resistance protein)	V9PWX7	E. coli	STPKKC	28536265	10.22011/strenda_db.STPKKC	Show Export PDF Export XML
Tryptophan 2-monooxygenase	P06617	E. coli	KJWIQY	7893667	10.22011/strenda_db.KJWIQY	Show Export PDF Export XML
Tryptophan 2-monooxygenase	P06617	E coli	WZOV5O	7893667	10.22011/strenda_db.WZOV5O	Show Export PDF Export XML
2-hydroxymuconate tautomerase (4-oxalocrotonate tautomerase) (4-OT)	Q01468	Escherichia coli	XG8UGZ	28684981	10.22011/strenda_db.XG8UGZ	Show Export PDF Export XML
2-hydroxymuconate tautomerase (4-oxalocrotonate tautomerase) (4-OT)	B1Y2E4	Escherichia coli	OPC8OQ	28684981	10.22011/strenda_db.OPC8OQ	Show Export PDF Export XML
D-amino-acid oxidase (DAAO) (DAMOX) (DAO)	P00371	Escherichia coli BL21(DE3)	TOI0T2	28355481	10.22011/strenda_db.TOI0T2	Show Export PDF Export XML

Figure 6: Screenshot of the hit list after querying STRENDA DB.

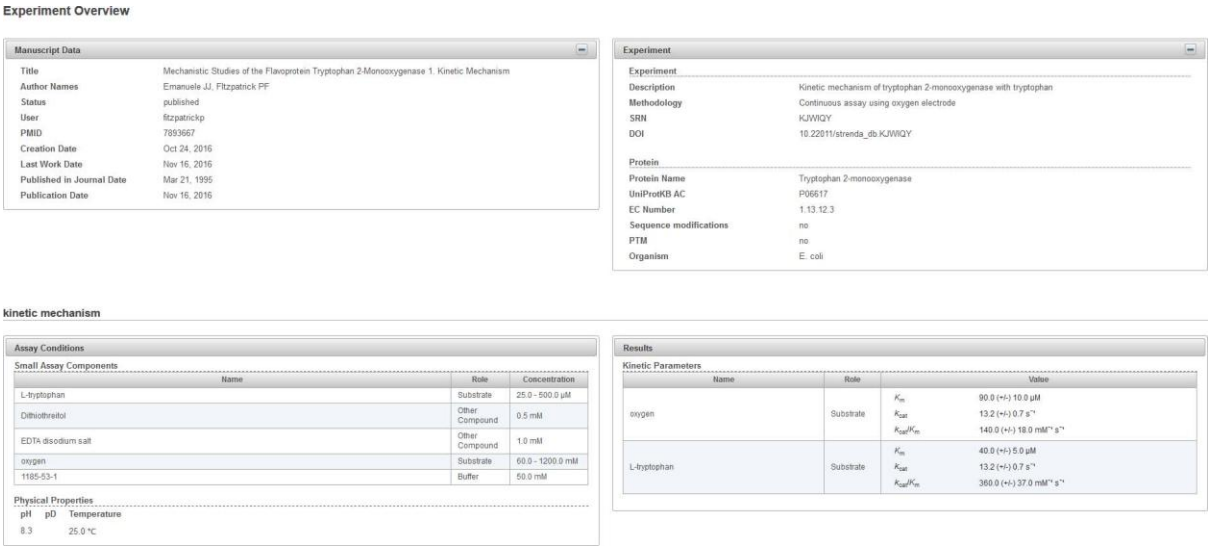

Figure 7: Screenshot of the Experiment overview as accessed by clicking on an entry in the hit list.

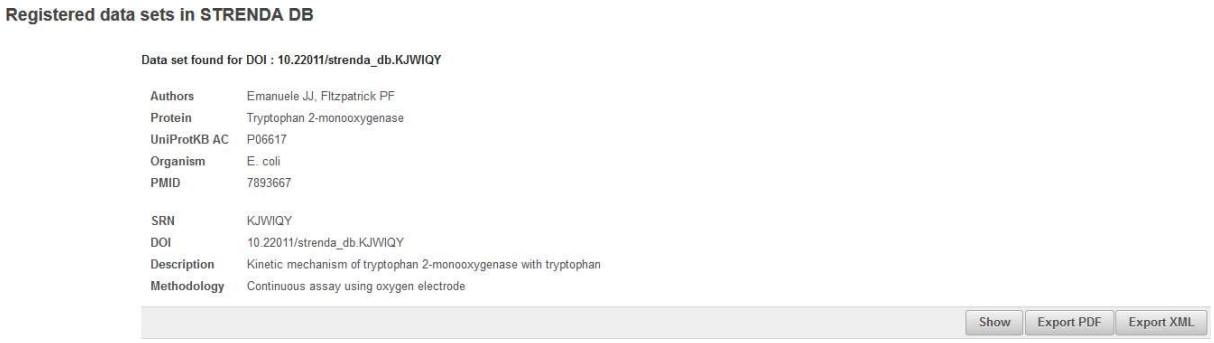

Figure 8: Screenshot of the hit page after clicking on a hyperlinked DOI published elsewhere.

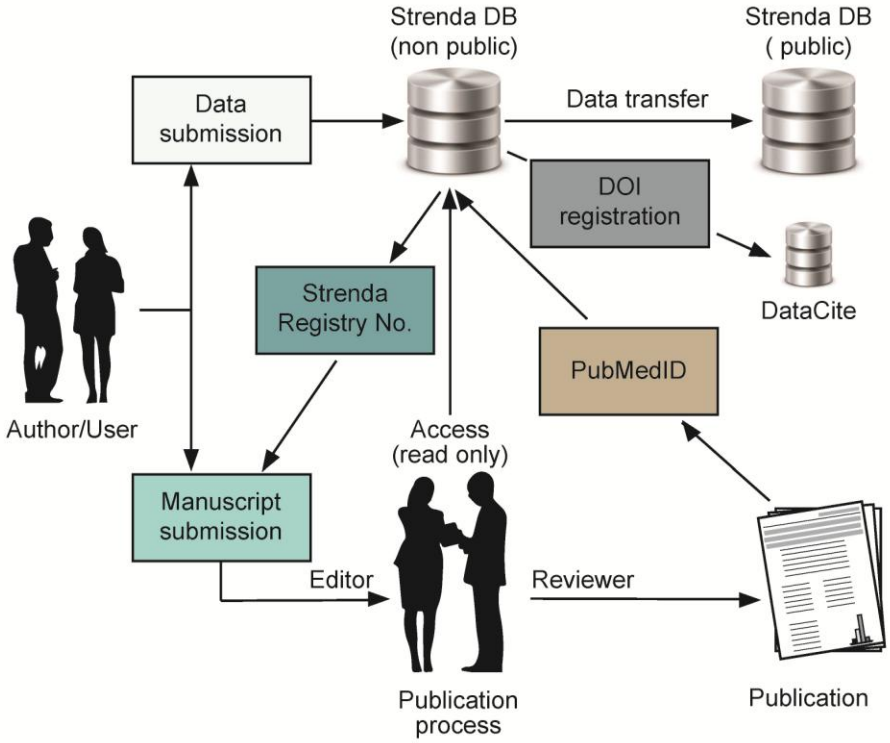

Figure 9: Incorporation of STRENDA DB in the publishing workflow. After data submission, the data remain in a private part of the database, which is only accessible by the editor and reviewers of the manuscript. Upon publication, the data is made publicly available.