**Hugo Daniel Cepeda Mochão**

Degree in Computer Science

# Improvement of KiMoSys framework for kinetic modelling

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
**Computer Science and Engineering**

Adviser: Rafael Costa, Researcher/Professor,
NOVA University of Lisbon

Co-adviser: Pedro Barahona, Full Professor,
NOVA University of Lisbon

Examination Committee

| | |
|---|---|
| Chair: | Margarida Mamede |
| Rapporteur: | Jorge Cruz |
| Member: | Rafael Costa |

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

**February, 2021**

**Improvement of KiMoSys framework for kinetic modelling**

This document was created using the (pdf)LaTeX processor, based on the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

# Acknowledgements

First, I would like to thank my supervisors Rafael Costa and Pedro Barahona, for all the help, constructive feedback, suggestions, and all the availability shown whenever I needed anything, even in this pandemic time. They helped me pushing my boundaries and raising my standards. I also thank Frank Bergmann for all the availability to help me with the COPASI integration.

I thank FCT-UNL and all the teachers for providing me all the necessary help in these five years. Their high standards and the knowledge they passed to me was instrumental in developing this project.

I would also like to thank my family, especially my mother and brother, for always pushing me to be a better person and professional. Throughout my academic and school path, they have always provided me all the necessary tools. They allowed me to be in a privileged position, where I could pursue studies in university education. Without their hard work, this could not be possible.

Finally, I would like to thank all my friends from school and FCT-UNL. They helped to shape the person I am today and incentivize me to do better always. Thank you for these incredible years.

# Abstract

Over the past years, an increasing amount of biological data produced shows the importance of data repositories. The databases ensure an easier way to reuse and share research data between the scientific community. Among the most important features are the quick access to data, described by metadata and available in standard formats, and the compliance with the FAIR (Findable, Accessible, Interoperable and Reusable) Data Principles for data management.

KiMoSys (https://kimosys.org) is a public domain-specific repository of experimental data, containing concentration data of enzymes, metabolites and flux data. It offers a web-based interface and upload facility to publish data, making it accessible in standard formats, while also integrating kinetic models related to the data.

This thesis is a contribution to the improvement and extension of KiMoSys. It includes the addition of more downloadable data formats, the introduction of data visualization, the incorporation of more tools to filter data, the integration of a simulation environment for kinetic models and the inclusion of a unique persistent identifier system.

As a result, it is provided a new version of KiMoSys, with a renewed interface, multiple new features, and an enhancement of the previously existing ones. These are in accordance with all FAIR data principles. Therefore, it is believed that KiMoSys v2.0 will be an important tool for the systems biology modeling community.

**Keywords:** Repository, kinetic models, experimental data, standard formats, data visualization, FAIR principles, data management, modeling, systems biology

# Resumo

Nos últimos anos, uma quantidade crescente de dados biológicos produzidos atesta a importância dos repositórios de dados. As bases de dados garantem uma maneira mais fácil de reutilizar e partilhar dados de pesquisa entre a comunidade científica. Entre as características mais importantes estão o rápido acesso aos dados, descritos por metadados e disponíveis em formatos padrão, e o cumprimento dos Princípios FAIR (*Findable*, *Accessible*, *Interoperable* e *Reusable*) para a gestão de dados.

KiMoSys (`https://kimosys.org`) consiste num repositório público de domínio específico de dados experimentais, contendo dados de concentração de enzimas, metabolitos e dados de fluxo. Oferece uma interface para a web e uma ferramenta de carregamento de dados, tornando-os acessíveis em formatos padrão, além de integrar modelos cinéticos relacionados aos dados.

Esta tese contribui para o melhoramento e extensão do KiMoSys. Inclui a adição de mais formatos de dados para descarga, a introdução de visualização de dados, a incorporação de mais opções para filtrar os dados, a integração de um ambiente de simulação para modelos cinéticos e a inclusão de um sistema de identificador único persistente.

Como resultado, é apresentada uma nova versão do KiMoSys, com uma interface renovada, várias novas características e um aprimoramento das anteriormente existentes. Estas estão de acordo com todos os princípios de dados FAIR. Portanto, acredita-se que o KiMoSys v2.0 será uma ferramenta importante para a comunidade de modelagem de sistemas biológicos.

**Palavras-chave:** Repositório, modelos cinéticos, dados experimentais, formatos padrão, visualização de dados, princípios FAIR, modelagem, gestão de dados, sistemas biológicos

# Contents

# List of Figures

# List of Tables

# Acronyms

| | |
|---|---|
| BioPAX | biological pathway exchange |
| ChEBI | chemical entities of biological interest |
| COPASI | complex pathway simulator |
| CSS | cascading style sheet |
| CSV | comma-separated values |
| DOI | digital object identifier |
| FAIR | findable accessible interoperable reusable |
| FAQ | frequently asked questions |
| FBA | flux balance analysis |
| HTML | hypertext markup language |
| ISA | investigation-study-assay |
| KEGG | kyoto encyclopedia of genes and genomes |
| KiMoSys | KInetic MOdels of biological SYStems |
| MATLAB | matrix laboratory |
| MIAME | minimum information about a microarray experiment |
| MIBBI | minimum information for biological and biomedical investigations |
| MS | mass spectrometry |
| MVC | model view container |
| NMR | nuclear magnetic resonance |
| ODE | ordinary differential equation |

| | |
|---|---|
| RDF | resource description framework |
| SBGN | systems biology graphical notation |
| SBML | systems biology markup language |
| SBW | systems biology workbench |
| UniProt | universal protein resource |
| XML | extensible markup language |

# Chapter 1

# Introduction

## 1.1 Motivation

With the amount of biological data produced increasing each year, databases are a crucial tool to store, share and maintain data, improving the quality and reproducibility of work done by the scientific community. Having diverse organisms available, categorized by their different data types, and the kinetic models associated with them in a repository is very important to the systems biology community. Being able to have this information agglomerated in a single web platform could spare a lot of time. It is much simpler to access this kind of data than to have to search for it in supplementary material, articles, journals or books.

In the last few years, there has been an increasing need to improve the infrastructure that supports the storing of biological data [1]. An easier way to share and reuse data could facilitate the work of the biological community, possibly leading to faster discoveries in fields of study that are of major importance.

Several public data repositories emerged to answer this need. However, to uniformize and standardize the way data is published and accessed, there were also developed rules that these repositories are encouraged to follow. These rules ensure, for example, that the different repositories work with the same data formats which make the job easier for the community. Scientific Data journal (https://www.nature.com/sdata) states that data in repositories should be public, datasets maintained in their published form and have permanent identifiers, among others recommendations [2]. PLOS publisher (https://www.plos.org) has similar suggestions [3], and encourages repositories to have an entry in FAIRsharing[4].

## 1.2 Overview of KiMoSys' Original Version

KiMoSyS (www.kimosys.org) is a domain-specific data repository and was developed in Ruby on Rails (https://rubyonrails.org) framework. Rails uses the MVC (Model-View-Container) pattern and is a friendly and simple tool to develop web applications. By using easy commands in the terminal, the user can create, for example, a Controller, and Rails automatically generates other files, like Views or Helpers, facilitating the job of the developer. It allows constructing applications in a short amount of time due to being easy to use and provides an easy way to migrate databases. On the other hand, the runtime and boot speed are slow. Some popular sites are built in Ruby on Rails, such as GitHub (https://github.com), Airbnb (www.airbnb.com), Shopify (https://www.shopify.com) and Twitch (https://www.twitch.tv).

KiMoSys is defined as a "*user-friendly platform that includes a public data repository of published experimental data, containing concentration data of metabolites and enzymes and flux data. It was designed to ensure data management, storage and sharing for a wider systems biology community*" [5]. It is a recommended repository by Scientific Data (https://www.nature.com/sdata/) in the mathematical & modelling resources field [6], and is also recommended by PLOS (https://www.plos.org), eLife (https://elifesciences.org), BioMed Central (https://www.biomedcentral.com), and GigaScience (https://academic.oup.com/gigascience). KiMoSys is also part of FAIRSharing (https://fairsharing.org/biodbcore-000631/).

KiMoSys 1.0, in addition to keeping datasets (public or private), has also the ability to store the kinetic models associated with them. It is possible to see the metadata related to a dataset or model, such as the publication, which can be downloaded, and its authors. The dataset is also available to download. If a dataset has a model associated, there is the option to navigate to the model page, access to its information, and extract it. Data can be searched by typing a name of an organism, a data type or a strain.

This repository also supports the kinetic modelling of metabolic networks, providing tools to:

- Model reduction based on the conjunctive method described by Machado et al. [7]. The reduction is done by choosing an SBML file, a txt file with the flux values, and the metabolites to remove.

- Setting metabolite concentration values in the model, by choosing an SBML file and a txt file with the metabolite concentration values.

- Converting metabolic networks in kinetic models by simply choosing an SBML file.

- Setting flux values in the model, by choosing an SBML file and a txt file with the flux values.

The user can choose Systems Biology Markup Language (SBML) files of his own or select data already present in the repository. Then, an XML output file is created and presented to the user.

KiMoSys experimental data is mainly focused on four types: enzyme/protein concentrations, fluxes measurements, metabolites at steady-state, and time-series data of metabolites. To submit data, KiMoSys offers two options:

1. Quick Submit, which consists of submitting a dataset in any file format, a description of it, and an email. Afterward, the KiMoSys team will check the file submitted, and register the data in the repository.

2. Electronic Data-Submission, that can only be used by registered users. First, it is necessary to download and complete an Excel template available on the website. There are four templates available, being those for each type of data present in KiMoSys. After completing the template, it is necessary to introduce the metadata related to it: general information about the data (publication, PubMed ID, Keywords, etc), experiment description and details, and finally insert the template. The last step of this process is to choose the availability of this dataset: *public*, *hidden from everyone except me*, or *confidential peer review* (only available to reviewers). A curation process will only start if the data is made public.

It is also possible to fill the template and send it by email to kimosys@kdbio.inesc-id.pt, and the data will be introduced later by the KiMoSys team.

To register on the website, one has to fill some fields (name, email, etc). After that, a confirmation email will be sent, and, upon acceptance, the user can log in. Besides the Electronic Data-Submission functionality, a registered user has an additional tab in the navigation bar, called *My Repository*, which is a list of data submitted by him.

Despite these features, KiMoSys cannot be considered FAIR (Findable, Accessible, Interoperable, and Reusable).

## 1.3 Contributions and Thesis Outline

The purpose of this thesis is to improve the KiMoSys platform, adding some key functionalities, and offering a distinct set of features. The database was built using a similar architecture to that of the KiMoSys v1.0 [5]. Ultimately, the objective is to enhance this tool, implementing a distinct set of capabilities, both in the field of experimental data and in kinetic modeling, making it more useful, thus increasing the number of users and datasets. In general, the main characteristics to implement are: improve user interface, add more options to export files (for example XML and RDF), introduce a way to preview excel files on the website, integrate a simulation environment, introduce data visualization and add a unique persistent identifier to each EntryID of a dataset and associated kinetic model. The implementation of all FAIR principles is crucial to achieve this goal. A more detailed list of key features to implement, and its description, is listed below.

### 1.3.1 Improve Web Interface

Improve the User Interface: for example, introduce new search and filter options. In the original KiMoSys version it is only possible to search by typing the organism name, its strain, or data type. One objective is to search for an organism and be able to see its associated models. Other interesting feature would be to implement checkboxes to see, for example, all organisms of a single data type or strain.

Additionally, introduce a way to preview excel files on the website. It would be more user friendly to dispose the information online than to force the user to download it to check the contents of the file.

Another goal is to make KiMoSys' datasets and models Entry IDs citation easier, with multiple citation formats available.

### 1.3.2 More Downloadable Data Formats

Currently, KiMoSys supports some data standards like SBML and excel files. It is a goal to add more exportable formats to feed a variety of modeling frameworks. For example, in the computational models perspective, BioModels repository [8], in addition to SBML, has the option to export some models in CellML. RDF (https://www.w3.org/RDF/) and .txt formats could also be introduced to export metadata.

### 1.3.3 New Simulation Environment

The introduction of a simulation environment for kinetic models enhances KiMoSys' abilities concerning the modeling resources. It is intended to provide elaborated modeling support, such as a time-course simulation.

### 1.3.4 Add a Unique Persistent Identifier System

As seen before, a persistent identifier of data is commonly referenced as a crucial attribute to have in repositories. It is one of the FAIR principles, relevant scientific journals and publishers recommend it [2, 3], and other publications describe it as essential [9].

The Findable principle will be implemented, increasing data FAIRness in KiMoSys.

### 1.3.5 New Data Visualization

One objective is, from data in excel files, to introduce libraries that provide tools to visualize data, showing to the user, for example, plots describing diverse information of the data, such as the concentration variation of a certain substance over time, or a graph depicting the influence an organism has in another.

### 1.3.6 Thesis Structure

This thesis is organized as follows: Chapter 1 makes an introduction to the general problem, giving also a description of KiMoSys v1.0 [5] and the goals to implement. Then, Chapter 2 follows, offering quick access to some of the major concepts related to this repository data, such as omics data, data management and kinetic modeling. Chapter 3 depicts several web repositories from two fields: experimental data and mathematical models, providing the main differences between them. Other related tools are addressed, as well as the data standards, with special attention to the FAIR Data Principles. Chapter 4 has detailed information about the tools used to develop this project. The results are present in Chapter 5, with multiple illustrations about the new KiMoSys 2.0 version, and the conclusions are detailed in Chapter 6.

# Chapter 2

# Background

## 2.1 Omics Data

KiMoSys is focused on omics data, more specifically metabolomics, fluxomics, and proteomics. Metabolomics is the large-scale study of metabolites [10], and there are two main techniques to acquire data: Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS). Fluxomics studies the rates of metabolic reactions, and there are two main technologies to analyze fluxes: Flux Balance Analysis (FBA) and $^{13}$C-fluxomics [11]. Proteomics is the study of proteomes, which are sets of proteins produced in a biological context [12]. In the Bioinformatics field, proteomics data can be collected through Mass Spectrometry [13].

## 2.2 Standards

It is essential that the experimental data is stored according to standards well known by the community. This guarantees a common environment of work, in which the scientists are already familiarized. For experimental data the Minimum Information standards, created in 2008, is used and called the Minimum Information for Biological and Biomedical Investigations (MIBBI) project [14]. Minimum Information About a Microarray Experiment (MIAME), part of MIBBI, was created to describe microarrays with the minimum information possible [15]. There are a lot more standards in MIBBI that deal with different kinds of organisms, such as molecular interactions, genome sequences, proteomic experiments, etc. For storing mathematical models, Systems Biology Markup Language (SBML) is the most common standard, and it is based in XML. CellML, also based in XML, and Biological Pathway Exchange (BioPAX) are other choices to store the computational models related to the biological processes. Metadata is also a very important aspect to take into account when storing experimental data. Experience details and description, the journal where data was published, its authors, and other important information can help to understand more about the experimental data. ISA (Investigation-Study-Assay) framework is an open source tool that allows producing detailed metadata information, enhancing the reproducibility and reusability of the data [16].

To complement data information it is important to have data visualization, to provide a better understanding of it. Systems Biology Graphical Notation (SBGN), developed by some members of the team responsible for creating SBML [17], was built to try to supply a standard to graphical notations present in maps of biological processes. SBGN allows constructing 3 different diagrams. The process diagrams represent the molecular interactions between biochemical organisms; the entity relationship diagrams correspond to the effects a certain organism has in another organism, and the transformations associated with it; activity flow diagrams allow to observe the flow of information between biochemical organisms [18]. SBGNview is an R package developed to provide a way to visualize omics data on pathway maps, being that pathway designed in SBGN [19]. Given an SBGN pathway file (SBGN-ML, which is a special XML format), and an omics data table, SBGNview has the ability to produce image files that contain omics data on glyphs [19]. SBGN-ML has two main data types: a glyph, which can be a high-level SGBN node or a just sub-node; and an arc, that represents an SBGN arc between two SBGN nodes [20].

Scientific journals have been making an effort to encourage researchers to make data shareable and public, endorsing public data repositories [21]. They advise the repositories to follow certain specifications of how these should work, being the FAIR Principles [1] the main reference. McQuilton et al propose criteria that strengthen data FAIRness [9]. Two types of criteria exist: they can either be Essential, or Desirable. Essential criteria can be seen as a must-have. Persistent Data Identifiers and User Support can be found in this category. Desirable criteria are also important but are a low priority compared to Essential criteria. It is more related to features that can be added in the future, after introducing all the Essential ones. By agreeing to these specifications, the chances of repositories being recommended or certified by journals and publishers raise. However, in 2017, as Husen et al. show [21], the process of certification is still in maturation, as less than 6% of the recommended repositories (evaluated in that study) had a certification.

### 2.2.1 FAIR Principles

The FAIR principles are indicators of how repositories' data should be built, and its rules are the following [1]:

**Findable** - data defined by a persistent identifier and detailed metadata

- F1: (meta)data are assigned a globally unique and persistent identifier

- F2. Data are described with rich metadata (defined by R1 below)

- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

**Accessible** - well-defined license and access conditions

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

**Interoperable** - ready to be combined with other data by humans and machines; standardised formats and vocabulary

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

**Reusable** - ready to be reused in future research and processed using computational methods

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

FAIRshake framework [22] allows to check the FAIRness of data. This toolkit uses the principles described above, and evaluates if projects agree to FAIR Principles. Table 2.1 shows the steps to evaluate a project (repository) in FAIRshake [22].

Table 2.1: Steps to Perform and Visualize FAIR Assessments with FAIRshake (adapted from the original article) [22]

| **Step** | **Instructions** |
|---|---|
| Sign up | Fill in a registration form. |
| Log in | Enter user name and password. |
| Start a project | Fill out a form that describes the project. |
| Register digital objects | Register digital objects in FAIRshake and associate them with the project. |
| Add a FAIR metric | Fill out a form to set up the FAIR metric question and possible answers. |
| Add a FAIR rubric | Associate a collection of FAIR metrics with a new rubric. |
| Associate rubrics with digital objects | Associate each registered digital object from the project with a registered rubric. |
| Perform assessments | Answer each FAIR metric question to fill in the FAIR evaluation questionnaire. |
| Visualize the FAIR results with an insignia | Hosting websites can use a JavaScript library to visualize FAIR assessments of the digital objects they host.<br>Alternatively, the insignia can be visualized via a browser extension or a bookmarklet. |

## 2.3 Data Management

The increasing rate of scientific work produced, together with the continuous adoption of Open Access policies, gives great importance to data repositories [23]. These should preserve the data, provide open access, and data reusability through data citation [24]. Currently, it is not easy to provide a citation to a dataset, therefore scientists choose to publish papers along with the dataset [25], and the conditions of the experiment. These are described by metadata, which is also essential for data reusability.

### 2.3.1 Metadata

The purpose of metadata is to provide meaningful information and context about the data, allowing for a better perception of it [26]. In the context of experimental research, metadata should have information about procedures [27], as detailed and concise as possible. Duval et al. [28] propose four principles that metadata should follow: Modularity, where different metadata vocabularies should be able to be combined semantically and syntactically, in order to produce more metadata reusability; Extensibility, that says that metadata should easily be extended by additional elements if needed (and correctly processed by applications); Refinement, that is focused on having defined vocabularies to standardize metadata; and Multilingualism, based on the idea that metadata should respect multicultural diversity, such as date representation.

### 2.3.2 Generic Data Production Cycle

The data production life cycle can generally be composed of seven different steps [29], described as follows:

1. Idea - Conceptualize a topic in which would be relevant to do any kind of scientific research.
2. Project - Assemble a team, identify what are the tools needed, and design the project steps.
3. Collect, create, and search for data - This step demands an extensive research about related data in literature or web repositories.
4. Store/extract data - Store the relevant data from the previous search.
5. Clean/verify data - The process of data validation, where previous collected data might be released.
6. Data analysis - Make a thorough analysis of the data, ensuring that it can be reproduced and replicated.
7. Data publication - The final step, where data is published through a scientific paper, or a web repository (or even both).

### 2.3.3 Data Repositories

Repositories play an important part in:

- Data management, as it is possible to interact with different kinds of data and to explore datasets.
- Data access and storage, facilitated by a centralized database with different data types, where scientists can deposit their work.
- Data maintenance and validation, with curation systems ensuring the quality of data, that is described by detailed metadata.

When constructing a repository, it is essential to have in mind several factors. First, it is crucial to precise what type of data will be stored, and its range. A large amount of data types could mean a broader community of users interacting with the repository, but at the same time could difficult the introduction of features such as data visualization, as several different kinds of datasets would have to be addressed. Secondly, it is necessary to choose the architecture. This depends on what is expected to be the number of entries in the database and the number of requests that the server has to process. As some tools and frameworks provide better scalability, others do not scale as well but grant a better performance in smaller environments. The final choice is where to host the repository. From a university perspective, it could recur to external maintenance, possibly having bigger costs, or to host the repository in its own services. Continuous maintenance of repositories is crucial, and, unfortunately, is not rare to see projects being left behind. DIPSBC [30] is an example of a discontinued platform.

Validation and curation of data is key; it is of the utmost importance to verify the repositories' data, assuring its veracity to not mislead the scientists, ensuring that it can be reproduced and replicated. Replicability, in the computational experiments perspective, is accomplished when computations can be repeated, with the results being the same [31]. Reproducibility is achieved when an independent team can obtain the same result, but using their developed artifacts during the experience [31]. Some repositories, like SABIO-RK [32], have a curation team that is responsible to verify if the data inserted in the repository are truly reliable and correct. If the data is validated, then it is annotated as curated, which means it was verified.

## 2.4 Mathematical Modeling Approaches

Stoichiometric and kinetic models are the two main metabolic approaches concerning the modeling and optimization of biological systems [33]. Stoichiometric models are applied at genome scale and can access the feasibility of steady states, while kinetic models are used to simulate the metabolite and flux concentrations over time, hence requiring more experimental information to be built [34]. These models can be used to improve the cell design ability to produce and optimize industrial relevant products [35].

Kinetic models offer a better perspective about the metabolism, but the knowledge needed to build them is not big enough at the moment, making them harder to construct. Constraint-based stoichiometric modeling does not try to find a single solution; instead, an array of solutions is defined. If a solution fails any constraint, it automatically drops out of the solution set. Mass balance, thermodynamics, and energy balance are the usual stoichiometric constraints used in this type of modelling [36].

The construction of a kinetic model can be described in four steps, that proceed iteratively:

1. Model building and calibration, where ODEs usually represent the rate reactions;

2. Simulation;

3. Model validation and analysis;

4. Model applications.

The estimation of unknown kinetic parameters is very challenging, and can be successfully done with a relevant amount of experimental data [37].

Approximated kinetic representations are important when there is not enough data to make a correct estimation of certain parameters [38]. Lin-log is an example of an approximated kinetic format, and is based on the concept that a reaction rate is associated with the thermodynamic driving force [39]. Moreover, the rate is proportional to the enzyme activity, and the metabolite levels depicted by the linear sum of non-linear logarithmic concentration terms [40].

Experimental data takes a big part in constructing kinetic models. Ordinary Differential Equations (ODEs) are used to model metabolism kinetic models, giving detailed information about biochemical interactions, their dynamics and the system components [41]. Fig. 2.1 represents a simple metabolic network, where metabolites are the nodes (A to D), and reactions are the arcs [42]. These reactions can be modeled by ODEs. Assuming that, for A, the resultant ODE would be:

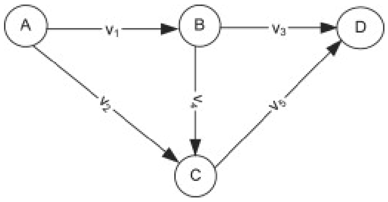


Figure 2.1: Example of a simple metabolic network.

$$\frac{\partial A}{\partial t} = -v_1 B - v_2 C.$$

However, this data must be extremely precise. To build an accurate and trustworthy kinetic model, one has to take in consideration diverse factors that could contaminate the data, as stated by Vasilakou et al. [43]: central carbon metabolism effluxes are difficult to precisely quantify; it is hard to distinguish reaction cycles and parallel reaction rates; the low information of intracellular concentration measures makes it hard to correctly choose the parameters of the model. Also, during the experiments, it is necessary to take into account certain rules to not corrupt the data. Usually, kinetic parameters are obtained by *in vitro* experiments, which means they are made outside of a living organism. However, they differ from the *in vivo*, where is difficult to identify with precision the parameters, as the organisms in these experiments can only be perturbed by genetic modifications or by extracellular stimuli [43]. So, after obtaining the parameters from *in vitro* experiments, it is necessary to use *in vivo* data to calibrate them. There are two approaches to build kinetic models: bottom-up and top-down. Bottom-up consists of forming separately subparts of the model and then putting all together, while in top-down all the parts are fitted simultaneously [44]. Top-down offers a better prediction but needs larger datasets and more complicated mathematical methods [44]. After constructing the kinetic model, it should be capable to represent the experimental observations and also predict the genetic or environmental perturbation [43].

# Chapter 3

# Related Work

In this chapter, the main repositories of experimental data and mathematical models of metabolism will be addressed, as well as their main characteristics and different focuses. They are commonly accepted and respected within the scientific community, largely follow the FAIR principles and the recommendations of publishers and journals about how data should be integrated into repositories. Table 3.1 gives a summary of the experimental data repositories, having a brief description of each one, their addresses and data types present.

## 3.1 Experimental Data Repositories

In this section the main repositories for experimental data are described. Their main characteristics are explained, as well as their compliance with the FAIR Principles.

### 3.1.1 SABIO-RK

SABIO-RK (`http://sabio.h-its.org`) [32] is a database that stores information about biochemical reactions and their kinetic properties. Data can be submitted from laboratory experiments, or manually inserted from literature, and has automated consistency checks. It can be accessed in two ways:

1. Web-based user interfaces

2. Web services that allow data access by other tools

The amount of data in this repository has been increasing at a steady rate, and as of July 2020, it had 6563 curated publications, 64732 curated entries, and more than 110000 kinetic parameters. The most common organisms in SABIO-RK are Eukaryota

(63.7%), followed by bacteria (33.5%), and with less expression are Archae (1.8%) and viruses (1%) [45]. SABIO-RK offers filter options while searching for organisms, such as environmental conditions (pH and temperature), submission date or enzyme type (wildtype, mutant or recombinant), among others. It is also possible to make an advanced search, combining the data that is being searched with other parameters, which can be an entry ID, kinetic data, enzymes, among many others.

### 3.1.2 BRENDA

BRENDA (https://www.brenda-enzymes.org) is a relational database that contains all enzymes classified according to the system of the Enzyme Commission numbers, which is a numerical classification for enzymes. It has approximately 46 000 references of 40 000 enzymes from more than 6900 organisms. It stores information on specific organisms about molecular properties, reactions, stability, ligands, and links to other databases [46]. BRENDA is recommended by ELIXIR (https://elixir-europe.org).

BRENDA offers a diverse set of features: text-based and structured based queries, an explorer that functions similarly to a file explorer that divides the enzymes according to their EC number, a map describing several metabolic pathways (that can be downloaded), among others. Although having distinct and important characteristics, sometimes it can be hard to navigate through the website, due to its interface not being user-friendly or clear enough. Nonetheless, BRENDA complies with FAIR principles, and it is a reference when it comes to enzymes.

### 3.1.3 CeCaFDB

CeCaFDB (http://www.cecafdb.org) is a fluxomics database, having more than 500 flux distributions between 36 organisms, and its related information, taken from literature, such as growth conditions or genotype, among others [47].

It offers ways to search and visualize metabolic pathways, being possible to interact with them and to download in diverse formats. It is also possible to browse for metabolites and reactions, which have an external link to KEGG (https://www.genome.jp/kegg/). Data is available to download in excel files.

CeCaFDB has a simple website, easy to use and to understand. The possibility to interact with the metabolic pathways, adjusting them to the user desire is a great advantage. However, as KiMoSys, the Findable principle is not present in this repository.

### 3.1.4 MassIVE

MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) is a repository for mass spectrometry, and it is part of the ProteomeXchange Consortium (http://www.proteomexchange.org). It has 298.19 terabytes of information, including 10741 public datasets with more than five million files, 20 370 proteins, 5 849 004 peptides and 15 855 998 peptide variants.

It is possible to access and download datasets, and some of them have also links to entries in ProteomeXchange repository. Additionally, in a dataset page, there are options to reanalyze spectra, add a reanalysis, or to comment. MassIVE also provides a way to compare datasets, in terms of proteins, peptides and spectrum identifications, using Venn diagrams. It is also possible to explore all the remaining data present on the website (proteins, peptides, etc) and its information.

Sometimes, the website can be a little slow to load datasets (probably because of its large amount of data) or to redirect to other pages. In terms of the FAIR principles, *Findable* is still missing.

### 3.1.5 MetaboLights

MetaboLights (https://www.ebi.ac.uk/metabolights) is a repository for metabolomics experiments and it is recommended by ELIXIR [48]. It has 722 metabolomic studies with its data available to download, 26 017 compounds and their descriptions, with some of them having links to ChEBI (https://www.ebi.ac.uk/chebi/), and 2820 species.

On the website, it is possible to search for studies, compounds, and species. When searching for studies or compounds, there are a lot of filters that help the process of scrutiny data. Each study has a description, metadata, a set of files that can be downloaded, the protocols used and other relevant data. A compound can be seen in 2D or 3D and has details about its reaction, the species it where can be found and a brief chemical description. When it comes to search for species, MetaboLights provides a tree of the biological kingdoms, that can be expanded by clicking in the nodes. When clicking in a leaf, it will show the compounds and studies of that species that are present in the repository.

Overall, it is very simple to navigate and to search for data in MetaboLights repository, being its user-friendly interface a great advantage. Although being recommended by ELIXIR, it does not have the *Findable* principle implemented.

### 3.1.6 PeptideAtlas

PeptideAtlas (http://www.peptideatlas.org) is a repository for peptides present in mass spectrometry proteomics experiments, and it is part of the ProteomeXchange Consortium [49]. There are exactly 1532 samples of raw data available to download in this repository, being possible to see the publication that the data belongs to, if there is one.

PeptideAtlas offers a simple way to search through its database, having a few filters. This repository provides data visualization of pathways, which are originally from KEGG, and annotated with information from SRM Atlas (http://www.srmatlas.org). It is also possible to browse for proteins present in neXtProt (https://www.nextprot.org) and to compare proteins from different builds. One can also see how the database is structured, as it provides access to its database schema.

This repository does not have a unique persistent identifier, therefore the *Findable* principle is missing. Its interface is a little confusing to use, some links lead to pages not found, and sometimes the loading times can take a long time.

### 3.1.7 PRIDE

PRIDE (http://wwwdev.ebi.ac.uk/pride) is a repository for proteomics data, it is part of the ProteomeXchange Consortium, and recommended by ELIXIR [50]. Its datasets can be divided into four categories: human, mouse, rat, and horse. The number of submissions per year was increasing at a steady state since it launched in 2004, until 2019 when it reached its peak with 2298 submissions. 2020 is expected to have a considerable decrease, as only 1283 submissions were registered up until November. The USA is where most of the submissions come from, followed by Germany and the United Kingdom.

This repository is divided in two branches: PRIDE Archive, which contains the datasets and all its related information, and PRIDE Peptidome, that contains information about peptides. In the Archive it is possible to search for datasets and see their characteristics. Each dataset is composed by a brief summary, its properties, publication, and raw data available to download, as well as links to similar studies. Peptidome has detailed information about the peptide, including a visualization of its consensus spectrum, and links to datasets where it is present.

PRIDE complies with the FAIR data principles, and uses the DOI system to identify its data. It has a clean interface, it is easy to navigate through the website and to search for data. It also presents some interesting statistics of the database about the provenience and the content of its data.

### 3.1.8 ProteomeXchange

ProteomeXchange Consortium (http://www.proteomexchange.org) was created to provide a standard in mass spectrometry proteomics data submissions and dissemination [51]. This consortium is composed by six repositories: PRIDE (http://wwwdev.ebi.ac.uk/pride), PeptideAtlas (http://www.peptideatlas.org), MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), jPOST (http://jpost.org), iProx (https://www.iprox.org) and Panorama Public (https://panoramaweb.org/project/home/begin.view).

Datasets present in this repository come from the members of the Consortium. It is possible to see to what repository belongs to the dataset, and it has links to its original publication and to where it can be downloaded. It is possible to group the datasets by repository, and ProteomeXchange offers ways to search and filter data by species or by keywords.

ProteomeXchange does not have a unique persistent identifier system, failing the *Findable* principle. While it can be a major advantage to have all datasets centralized in one repository, sometimes it can be easier to go to the original repository of the dataset, due to it having more features than to just download data. Its interface sometimes is not responsive enough and some of the other repositories offer clearer ways to see the datasets' information.

Table 3.1: Main curated single-data type repositories.

| Name | URL | Description | Data types |
|---|---|---|---|
| SABIO-RK | http://sabio.h-its.org | A database that stores information about biochemical reactions and their kinetic properties. | Kinetic rate equations with parameters |
| BRENDA | https://www.brenda-enzymes.org | A relational database that contains all enzymes classified according to the system of the Enzyme Commission numbers. Stores information on specific organisms about molecular properties, reactions, stability, and ligands. | Enzymes |
| CeCaFDB | http://www.cecafdb.org | A fluxomics database that offers ways to search and visualize metabolic pathways, being possible to interact with them, and to browse for metabolites and reactions. | Fluxomics |
| MassIVE | https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp | A repository for mass spectrometry, part of the ProteomeXchange Consortium, having ways to compare datasets using Venn diagrams and to reanalyze spectra. | Mass spectrometry |

*Continued on next page*

Table 3.1 – *Continued from previous page*

| **Name** | **URL** | **Description** | **Data types** |
|---|---|---|---|
| MetaboLights | https://www.ebi.ac.uk/metabolights | A repository for metabolomics experiments where is possible to search for studies, compounds, and species. Its user-friendly interface is a great advantage. | Metabolomics |
| PeptideAtlas | http://www.peptideatlas.org | A repository for peptides present in mass spectrometry proteomics experiments, part of the ProteomeXchange Consortium. Provides data visualization of pathways, and is possible to compare proteins. | Peptides |
| PRIDE | https://www.ebi.ac.uk/pride | A repository for proteomics data, part of the ProteomeXchange Consortium. Is divided in two branches: PRIDE Archive, which contains the datasets, and PRIDE Peptidome, that contains information about peptides. | Proteomics |
| ProteomeXchange | http://www.proteomexchange.org | Created to provide a standard in mass spectrometry proteomics data submissions and dissemination, and composed by six repositories. Datasets present in this repository come from the members of the Consortium. | Mass Spectrometry |

*Continued on next page*

Table 3.1 – *Continued from previous page*

| Name | URL | Description | Data types |
|---|---|---|---|
| KiMoSys | https://kimosys.org | Public data repository of published experimental data, containing concentration data of metabolites and enzymes and flux data. In addition to keeping datasets, has also the ability to store the kinetic models associated with them. | Metabolomics, Proteomics and Fluxomics |

## 3.2 Repositories for Standardized Mathematical Models

This section addresses the principal repositories for mathematical models, describing their main features and purposes.

### 3.2.1 Bio Models

BioModels is a central repository of mathematical models that represent biological processes, and it is recommended by ELIXIR. It is composed by two branches [8]:

1. Models derived from literature.

2. Models generated automatically from pathway resources, such as KEGG (https://www.kegg.jp), BioCarta (https://mitelmandatabase.isb-cgc.org), MetaCyc (https://metacyc.org), PID (https://wiki.nci.nih.gov/pages/viewpage.action?pageId=315491760) and SABIO-RK (http://sabio.h-its.org), or from individual patients [52].

Recently, BioModels introduced a supported format different from SBML, CellML. Models submitted to BioModels are curated by a curation team, which can receive help from the authors as well [8]. Data can be searched from each branch separately, or from all models present in the database. When searching, it is possible to filter data by the format of the model, organisms, disease, or the curation status. Each dataset has a summary, model files available to preview or to download, the version history, the components, and if there is one, the curation results. BioModels also offers a way to view pathways.

BioModels does not have a persistent identifier system yet implemented, failing the *Findable* principle, and its interface sometimes is not intuitive but is one of the most important repositories in the field of computational models.

### 3.2.2 JWS Online

JWS Online (https://jjj.bio.vu.nl) is a database for kinetic models and provides tools to its simulation [53]. It is a powerful tool, SMBL compliant, offering ways to create models and to modify existing ones. It is divided into three branches: Model Database, Simulation Database, and Manuscript Database.

In the Model Database, each model has clear and extensive metadata, including the publication, the model parameters, species, and reactions. A model can be simulated, originating a schema of the model, and it is possible to interact with it. This tool also offers ways to see the time evolution of the different species, display metabolite and fluxes information, scan a particular parameter and see a plot of a reaction. In the Simulation Database, is possible to run the simulations, generating some plots. The Manuscript Database contains the publications associated with the data.

JWS Online is an important tool in the computational model's field. It offers an extensive and distinct set of computations over models and simulations. However, does not have a persistent data identifier system, not complying with the *Findable* principle.

### 3.2.3 Physiome Model Repository

Physiome Model Repository (https://models.physiomeproject.org/welcome) is a repository for kinetic models, storing mainly CellML and FieldML models, but having the capability to store any file format [54]. In some formats, it is necessary to install plug-ins to see the representation of the models.

Models available in CellML are divided by categories, while FieldML files are all displayed since there are only 6 models. Each model has a brief description, its metadata, curation status, the mathematics associated with it and a schema depicting the model. In CellML models, is possible to see its file converted into code. This conversion is done using CellML API and can generate, for example, C, Python or MATLAB files. There is also a section that shows how to correctly cite data from this repository.

This repository does not have the *Findable* principle implemented. Although having interesting features, such as being capable to store any file formats or see the mathematics of the model components, its interface is not clear sometimes. In addition to that, some models have empty information in certain fields.

## 3.3 Generic Data Repositories

More generic data repositories are covered in this section. These are not focused on a specific area; instead, they aggregate data from different fields.

### 3.3.1 myExperiment

myExperiment (https://www.myexperiment.org/home) is inspired by social networks like Facebook [55], offering ways to share bioinformatics workflows. Currently, there are 11 055 members, 427 groups, 3932 workflows, 1448 files, and 482 packs. Groups allow sharing data between its members. Files and packs can be of diverse things: tools supporting the development in the bioinformatics field and tutorials of how to use them, or XML files representing data.

In the website it is possible to search for users and see who has the most friends, the most credited, and the highest rated. Similarly, one can see the groups with the most members, most shared data, and most credited, as well as the list of their members. Workflows, files and packs can be searched using diverse filters. Workflows have metadata describing the experiment and are available to download.

It is a very interesting and innovative idea to build a kind of social network that allows the sharing of data. Groups of scientists working in the same field could benefit a lot. Similarly to other repositories seen above, it does not have a unique persistent identifier system and, consequently, the *Findable* principle is not present.

### 3.3.2 figshare

In figshare (https://figshare.com) is possible to find all different kinds of data, from Agriculture to Arts [56]. It currently has 5 482 061 entries in its database, with the vast majority being from scientific fields, such as Biology, Medicine or Biochemistry.

Although having a large amount of data diversity, it is easy to search in figshare, as several categories are available to facilitate this process. This tool encourages the users to submit results from failed experiments, as those can be important to colleagues from the same area.

figshare is a particularly interesting tool due to its versatility. Its interface is modern and well designed, facilitating the search of data in a database with a large number of entries. In addition, it has the ability to preview multiple file formats, including programming languages code files.

### 3.3.3 Dryad

Dryad (https://datadryad.org/stash) is a repository for evolutionary biology data [57]. Currently this database has 37 154 entries, where Holocene is the area with the most representation, with a total of 2405. The French National Centre for Scientific Research (https://www.cnrs.fr/en/cnrs) is the institute that holds more entries, while Molecular Ecology (https://onlinelibrary.wiley.com/journal/1365294x) is the most represented scientific journal.

It is easy to explore data in this repository, as it offers a search box and some categories of filters. These are: Subject Area, Geographical Location, Scientific Journal, and Institution. Ahead of each filter, there is the total number present in the database. It is possible to select multiple filters, and when clicking on one, it happens a cascade effect, meaning that the filters and their numbers refresh according to the previous filter selection.

Each dataset page presents detailed metadata, some statistics about its views, downloads, and citations, and the dataset files available to download. Dryad has a curation team that verifies all submitted datasets. These are under the CC0 license, meaning that they can be used and modified by anyone, not having copyright.

Dryad is well designed, having a clean and intuitive interface. It is easy to navigate through the website and to search for data. Each dataset is richly described and has external links for its original publication. Additionally, it is possible to download the dataset files separately or together.

## 3.4 ELIXIR Core Data Resources

ELIXIR Core Data Resources (https://elixir-europe.org/platforms/data/core-data-resources) is a set of the most important European resources in the life-science field [58]. It was created to enhance the development of trustable resources that are a reference in their field.

Multiple repositories seen above are part of this set, which means they comply with ELIXIR indicators of how repositories should be built. There are five categories of indicators: scientific focus and quality of science, community served by the resource, quality of service, legal and funding infrastructure and governance, and impact and translational stories [58]. Each indicator corresponds to one or more FAIR principles.

To become part of ELIXIR Core Data Resources it is necessary to fill an application, which initially is checked for completeness. If it goes through, then it will be evaluated by a committee that meets annually, alongside a panel of independent experts and the ELIXIR Scientific Advisory Board [58]. In this annual meeting, it is discussed which resources will be included in ELIXIR Core Data Resources, as well as potential members that can leave due to not complying with the ELIXIR indicators.

## 3.5 Tools for Kinetic Models Simulation

This section will describe well known simulation tools, listing their abilities concerning kinetic models simulation, and providing the main advantages and differences between them.

### 3.5.1 COPASI

COPASI (COmplex PAthway SImulator) [59] is a simulation software with a vast set of functionalities. It has two versions: a graphic interface, called CopasiUI, and a command line version, CopasiSE. It is built using C++ and has its source code available. Additionally, there are Java, Python, and C# language bindings also available to download, generated by SWIG software (http://www.swig.org).

COPASI has multiple simulation algorithms, such as LSODAR, Stochastic Runge-Kutta or Gillespie's algorithm [60], a vast set of analysis methods, data visualization, and the ability to export models in diverse formats. Supports models in SBML format and can simulate their behavior with stochastic kinetics or ODEs. Additionally, it is capable to create plots, to represent matrices in 3D bar charts, among other ways to visualize data [61].

### 3.5.2 CellDesigner

CellDesigner is a tool for modeling biochemical networks, while also having the capacity to simulate them [62]. Similarly to COPASI, it has the ability to visualize data and to create, export, and import models. It has support for standard formats, having the capacity to work with SBML models, and SBGN.

While the main feature of CellDesigner is the biochemical network modeling, it also allows model simulation, which is assured by the integration with the SBML ODE Solver and COPASI. Additionally, it provides linkage with the Systems Biology Workbench (SBW). This wide range of functionalities gives CellDesigner an important place within the Systems Biology Community.

### 3.5.3 JDesigner

JDesigner (http://jdesigner.sourceforge.net/Site/JDesigner.html) is similar to CellDesigner, being a tool for biochemical network modeling. It also integrates SBW, is SBML compliant, and can export models in diverse formats, such as Java or Matlab. Regarding model simulation, it offers tools for time course and steady-state simulations. These can be done using libRoadRunner [63] or Jarnac [64].

As CellDesigner, JDesigner is more focused on biochemical networks modeling. Nonetheless, it also has the ability to simulate kinetic models, therefore offering a variety of interesting features.

# Chapter 4

# Methodology

This chapter will address the tools and methodologies used to achieve the goals specified in Chapter 1, leading to the results described in chapter 5. The chosen options will be depicted, as well as other libraries that were studied, but eventually left out.

## 4.1 Ruby on Rails

KiMoSys (1.0 and 2.0) is built using Ruby on Rails, which is one of the most popular web development frameworks in the world, being recognized for its ability to quickly develop and deploy web applications. It follows the Model-View-Container framework, and it is often considered user-friendly, being easy to work with even for beginners.

The original version of KiMoSys was deployed in 2014, using Rails 3.2 and Ruby 1.8.7. As technologies evolve and change at an impressively fast pace nowadays, updating the software was a very demanding and thorough task. The most recent Rails version is 6.0.3.2, while Ruby's is 2.7.1. To update Rails, it was necessary to read and follow the major upgrades notes, available at https://guides.rubyonrails.org/upgrading_ruby_on_rails.html. It was required to take small steps: first, the upgrade to the 4.0 version was done, then to 4.1, followed by 4.2, 5.0, 5.1, 5.2, and finally to Rails 6.0. Every detail was taken into account, as even the file structure has suffered changes.

Rails is also well known for its *gems*, which are mostly developed by the community. In the long term, this means that many of them stop receiving updates, and possibly not even supporting Rails latest version. Due to this, it was necessary to delete or replace an important amount of *gems*. On the other hand, some of the maintained *gems* received updates that had a direct impact on the way they were used, hence the code had to be adjusted accordingly.

In this upgrade process, some deprecated *gems* were not deleted, causing some problems in later development stages. For example, when trying to create new users, the database was doing an *UPDATE* instead of an *INSERT*. Before doing this operation, Rails calls the *new_record?* function, which returns *true* if that record does not exist in the database (therefore using an *INSERT*), and *false* otherwise (doing an *UPDATE*). Strangely, it was returning *nil*. This erratic behaviour was occurring due to *protected_attributes_continued* usage. This gem was not initializing some models' constants and variables, causing this unexpected result. After being removed, other errors disappeared, for instance in the mailer, which shows how impactful a deprecated *gem* can be. Other specific updates are described in the next sections.

## 4.2 Storage

KiMoSys 2.0 uses an SQLite3 database, and Active Storage to handle file storage. While the database is the same from the original version, Active Storage is relatively new in Rails projects, and a novelty in KiMoSys 2.0, which previously used Paperclip (`https://github.com/thoughtbot/paperclip`).

### 4.2.1 SQLite3

SQLite provides fast, secure, and reliable operations [65]. Moreover, it is the Rails default database. Concerning database changes, some column names had to be modified, since they were using nomenclature reserved to Rails. In Organisms table a new column, called Organisms Involved, was introduced. This column has the ChEBI [66], KEGG [67], and UNIPROT [68] IDs, that are present in the Organisms' Excel files, and now are available to be searched. Due to the introduction of other libraries, new tables were created. These particular changes will be addressed in the following sections.

### 4.2.2 Rails Active Storage

Active Storage, introduced in Rails 5.2 (April 2018), is a tool that facilitates file uploading, while also linking these files to Active Storage objects [69]. It provides an easy way to upload files to cloud storage services or a local disk.

To include Active Storage it was necessary to add two new tables to the database: *active_storage_blobs*, which has information about the attachment, like filename, content type, file, size, or its key (file location); and *active_storage_attachments*, that stores the class name of the model, and the *blob_id*, which corresponds to the attachment ID in the *active_storage_blobs* table.

### 4.2.3 From Paperclip to Active Storage

As Paperclip was deprecated [70], it was necessary to migrate to Active Storage. To achieve this, several guidelines had to be followed [71–75]. First, the two new database tables were filled with the Paperclip information, which was present in specific Paperclip columns from model tables. Other data specifically needed for these tables, such as the key and checksum columns, were randomly filled. After this, the file location was moved: Paperclip stores its files in *public/system/* directory, while Active Storage uses *storage/*. Also, as previously said, *active_storage_blobs* has a column with the file location (key), which means that new directories were created to correctly match the information of that table. For example, an attachment with a key *e2d1a4eb-4c32-4f09-9621-ac735e61dcc0*, was placed in *storage/e2/d1/*. The first two characters indicate the first folder, and the next two the second folder. Finally, after assuring that the new directories were matching the data in the new tables, an update of the code was made: Paperclip functions were replaced by the Active Storage ones.

## 4.3 Web Interface

Concerning the user interface, KiMoSys 1.0 used mostly its custom CSS code, with exceptions for the Active Admin pages and the repository table, that used external sources. The new version includes Bootstrap and continues to use Datatables software for the main repository table, although with some changes.

### 4.3.1 Bootstrap

Bootstrap is one of the most popular libraries in the world regarding HTML, JavaScript, and CSS, having 144000 stars in GitHub as of today (`https://github.com/twbs/bootstrap`). It provides appealing and user-friendly interfaces. The Bootstrap website (`https://getbootstrap.com`) has in-depth guides and documentation, which facilitate its integration even for beginners. Its large community helps immensely in the debugging process, since the problems one encounters, probably have already been faced by someone else, and the answer can be found in sites like Stack Overflow (`https://stackoverflow.com`). KiMoSys 2.0 introduced Bootstrap through Webpacker.

### 4.3.2 DataTables

DataTables is a flexible and developer-friendly tool for constructing tables. It provides multiple interesting features, such as instant search, pagination, multi-column ordering, or the possibility to use several data sources [76]. Additionally, it supports various useful extensions and plug-ins.

DataTables software was already used in KiMoSys' original version, but now is imported through Webpacker, whereas before this was done statically, using a CSS file downloaded from the DataTables website. Besides this version update, the SearchPanes extension was also introduced. It is a powerful and flexible tool, providing multiple ways to filter data interactively, and allowing filter customization.

Before choosing to work with DataTables, other libraries were considered. Bootstrap Table (https://bootstrap-table.com) was the initial one, but some problems emerged: although being a user-friendly tool and having the Bootstrap appealing interfaces, it was not very developer-friendly, in a way that it was too strict, not providing the flexibility seen in DataTables. After trying this library, other options were studied, such as Dynatable, but DataTables seemed the best choice, which was proven later.

## 4.4 Data Visualization

In order to allow file preview and plot visualization, it was necessary to include multiple file parsing and chart libraries. KiMoSys 2.0 has four different data types, and consequently four differently constructed excel files, which hardened this task. For each data type, a different approach was taken.

### 4.4.1 Chart libraries

Chartkick (https://chartkick.com) is the *gem* that deals with charts. It has the impressive ability to create plots with a single line of Ruby, and can work with three different types of libraries: Highcharts (https://www.highcharts.com), Google Charts (https://developers.google.com/chart), and Chart.js (https://www.chartjs.org). Its main advantage is the capacity to build plots from Active Record models. Additionally, it provides ways to interact directly with the libraries' specific options, being possible to use almost their full potential.

These attributes were employed to create a new statistics page about the repository data, and to make plots from the organisms' Excel files, now available to preview. The libraries used in KiMoSys are Highcharts and Chart.js.

### 4.4.2 File Parsing

Concerning CSV and XLSX files, roo (https://github.com/roo-rb/roo) is the *gem* used for parsing, while rubyXL (https://github.com/weshatheleopard/rubyXL) is involved in the process of creating new XLSX files. As for XML, REXML (https://github.com/ruby/rexml) parses the file, and Google's prettify (https://github.com/googlearchive/code-prettify) makes an understandable display of it, providing line numbers and color highlighting of the code. Finally, Nokogiri (https://nokogiri.org) is used to obtain relevant information from kinetic models XML files, such as its reactions or species.

XLSX files have multiple empty cells, and the rows and columns where data start vary a lot. Hence, when constructing the plots, it is necessary to keep in mind these constraints.

As chartkick is excellent to build plots from Active Record models, it is trickier when dealing with raw data.

Fig. 4.1 shows how Chartkick accepts data to build line charts. *Name* is the plot element name, and *data* represents the dataset array.

```
line_chart [
   {name: "MetaboliteA", data: series_metabolite_a},
   {name: "MetaboliteB", data: series_metabolite_b}
]
```

Figure 4.1: How to Initialize Chartkick graphs.

As *time-series data of metabolites* XLSX files correlate the time with metabolites concentration, each *data* had to have tuples [TIME, CONCENTRATION] associated with a metabolite *name*. Fig. 4.2 depicts how graphs of this data type were constructed, using a complex array. Then, using the *map* function, it is possible to match the concentrations with the respective time.

```

graph =
[
   [ ["METABOLITE_NAME_0"], [TIME_00, CONCENTRATION_00], [TIME_01, CONCENTRATION_01] ],
   [ ["METABOLITE_NAME_1"], [TIME_10, CONCENTRATION_10], [TIME_11, CONCENTRATION_11] ],
   [ ["METABOLITE_NAME_2"], [TIME_20, CONCENTRATION_20], [TIME_21, CONCENTRATION_21] ]
]

line_chart graph.map { |line|
   {name: line[0], data: line[1..]}
}
```

Figure 4.2: Building time-series graphs.

*Flux measurements* and *metabolites at steady-state* bar graphs are similarly built when compared with *time-series data of metabolites*, but are differently parsed because time is not needed.

## 4.5 Kinetic Models Simulation

Model simulation is done by COPASI Python language bindings, which are available through a library called python-copasi (`https://pypi.org/project/python-copasi/`), that can be downloaded via pip. This tool allows accessing COPASI functionalities from Python code. Mathematica [77] (the software used by JWS Online) and CellDesigner were also considered but discarded. The first because it is not a free resource, and the second for the reason that is more focused on biochemical networks modeling, not offering the amplitude of features that COPASI [59] does regarding model simulation.

Three simulations were introduced: time course for metabolites, time course for fluxes, and steady-state. For each simulation a script was created. Fortunately, the COPASI language bindings also have various code examples of different simulations, which helped immensely to understand how the library works. The time course simulations were mainly adapted from one of these examples, however, there was none focused in a simple steady-state. Since this tool lacks in documentation, it was necessary to look for answers in the COPASI User Forum google group, or to contact directly via e-mail Dr. Frank Bergmann, who is part of the developing team. Dr. Bergmann, in addition to all the insightful suggestions and answers, also provided an example of a steady-state simulation.

After having the scripts skeletons, it was necessary to adapt them to KiMoSys 2.0. As the library was working, the time course results could only be seen through a created file. Hence, after creating and parsing the file, it has to be deleted, otherwise it would occupy space unnecessarily. To make sure that files are created with different names, they are assigned a random number with nine digits. The conditions of time course simulations, such as initial time, duration, or the number of steps, were also adjusted in order to be set by the users. The steady-state simulation does not take any additional arguments nor creates a file, therefore it was easier to integrate.

## 4.6 DOI Assignment

The DOI assignment was done through FCT-NOVA University of Lisbon services (Salima Rehemtula, Head of Impact Department), that have a partnership with DataCite (https://datacite.org). For this purpose, a new database column was inserted in Organisms and Associated Model tables.

## 4.7 Other Relevant Tools

This section describes other relevant tools used in this thesis.

### 4.7.1 Webpacker

Webpacker (https://github.com/rails/webpacker), introduced in Rails 5.1 and the default JavaScript compiler since Rails 6.0, facilitates the integration of JavaScript libraries. The libraries can be easily installed via yarn (https://yarnpkg.com), and Webpacker does the rest.

In KiMoSys, Bootstrap, DataTables (and its extensions), Chart.js, Highcharts, and jQuery (https://jquery.com) are imported through Webpacker. The ability to easily integrate JavaScript libraries in Rails projects brings great advantages, but configuring Webpacker was not easy. As it is relatively new in Rails, there are not many guides or information available. More time than expected was lost trying to import some of these libraries, mostly due to minimal errors, such as just a misplaced capital letter.

### 4.7.2 Active Admin

Active Admin (`https://activeadmin.info`) generates administrator web pages where is possible to manage the system, having the options to create, delete, or edit database data. It was already part of KiMoSys' first version, and no major updates were done.

### 4.7.3 Devise

Devise (`https://github.com/heartcombo/devise`) is a *gem* that manages users' authentications, sessions, and registrations. Several options can be configured, such as a timeout for a user session, or the necessity to confirm the account after registration. Devise is not new in KiMoSys 2.0, but some upgrades were done. Besides minor code updates, the integration with Google reCAPTCHA was introduced.

### 4.7.4 Google reCAPTCHA

KiMoSys 1.0 was experiencing an increasingly large amount of account creation by bots. Therefore, Google reCAPTCHA v2 (`https://www.google.com/recaptcha/about/`) was included to prevent malware attacks on registrations. Furthermore, Google offers an admin web page where is possible to see statistics of the reCAPTCHA usage, such as approved and disapproved tries, or the average time response.

### 4.7.5 CanCanCan

CanCanCan (`https://github.com/CanCanCommunity/cancancan`) is an authorization library, and a continuation from the deprecated CanCan (`https://github.com/ryanb/cancan`), which was used in KiMoSys' first version. With it is possible to define what users can access.

Fortunately, CanCanCan is not significantly different from its previous version, so just minor code updates were done. Its main advantage is the ability to easily set all the different types of users' permissions in only one file.

### 4.7.6 Impressionist

Impressionist (`https://github.com/charlotte-ruby/impressionist`) *gem* provides pieces of information about the access to the application resources, called *impressions*. In KiMoSys 2.0 it was used to know the total visits number of an Associated Model or Organism page, as well as the exact number of times that a particular dataset or kinetic model was downloaded.

To integrate Impressionist was necessary to run a database migration, that introduced a new table. This table has the information of the *impressions*, such as the *impressionable* *type* (Active Record object), *controller name*, or the *action name* (in this case it can be *view* or *download*). Furthermore, to be able to have impressions, the intended controllers and models were changed accordingly to this library specifications.

### 4.7.7 Simple Form

Simple Form (https://github.com/heartcombo/simple_form) facilitates forms creation with few lines of Ruby. Additionally, it has the remarkable option to use Bootstrap layouts. This library was introduced due to the lack of maintenance of Formtastic (https://github.com/formtastic/formtastic), which was used in the previous version.

### 4.7.8 MathJax

MathJax (https://www.mathjax.org) is a JavaScript engine that allows the display of mathematical formulas and equations in pretty and understandable formats. In KiMoSys 2.0 it improves the readability of kinetic models' reactions equations.

Chapter 5

# Results and Discussion

## 5.1 Preliminary Work

Before starting the KiMoSys 2.0 development, first it was necessary to update the project. KiMoSys, created in 2014, was developed in version 3.2 of Rails and 1.8.7 of Ruby language. Currently Ruby is in version 2.6.5 and Rails in 6.0, which means a considerable amount of evolution of this framework. It was necessary to check all the documentation of version changes and adapt the code to it. It was also necessary to run some migrations in the database due to this. The source code of KiMoSys 2.0 is available in GitHub (https://github.com/HugoMochao/KiMoSys-2.0), as well as the previous version (https://github.com/rs-costa/KiMoSys).

## 5.2 KiMoSys 2.0 Update

This section describes how every goal proposed in section 1.3 was implemented in KiMoSys 2.0. In addition to the new features, the user manual (https://kimosys.org/documentation/Documentation_Kimosys_v2.pdf) was also upgraded, according to the new interface and capabilities of the repository.

### 5.2.1 Unique Persistent Identifier System

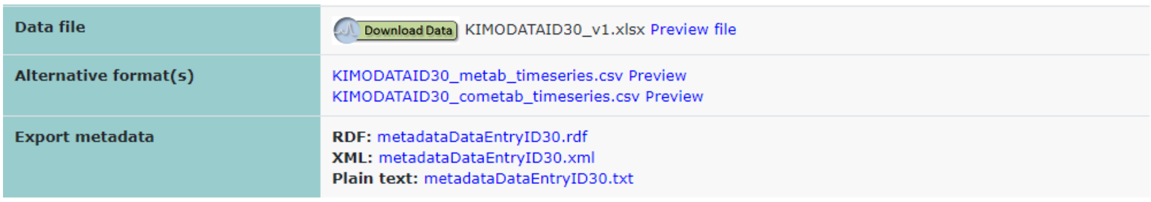
The process of creating DOIs is done manually, which means that for each DataEntry and ModelEntry ID is necessary to fill a form with their information, such as metadata or URL, and send it to FCT-UNL services (Head of Impact Department). After creation, the DOIs are introduced in the database.

For datasets submitted in the future, the procedure will be similar: the faculty services are contacted and given the necessary information to create a DOI. Afterward, a

KiMoSys administrator introduces the generated DOI for the EntryID. The University's DOI generating services are still in an initial phase, but in the future it will be possible to create batches of DOIs, instead of the current approach that can only produce one at a time.

### 5.2.2 More Downloadable Data Formats

Metadata is fundamental to describe data. KiMoSys 1.0 already provided detailed metadata regarding datasets and associated kinetic models. Hence, there was an opportunity to have this information available to download in multiple web-standard formats. As fig 5.1 shows, RDF, XML, and Plain Text were introduced.



| | |
|---|---|
| **Data file** | Download Data KIMODATAID30_v1.xlsx Preview file |
| **Alternative format(s)** | KIMODATAID30_metab_timeseries.csv Preview<br>KIMODATAID30_cometab_timeseries.csv Preview |
| **Export metadata** | **RDF:** metadataDataEntryID30.rdf<br>**XML:** metadataDataEntryID30.xml<br>**Plain text:** metadataDataEntryID30.txt |

Figure 5.1: Screenshot of the partial view from the Data Entry ID 30 page, (`https://kimosys.org/repository/30`) with the three different formats of metadata available to download.

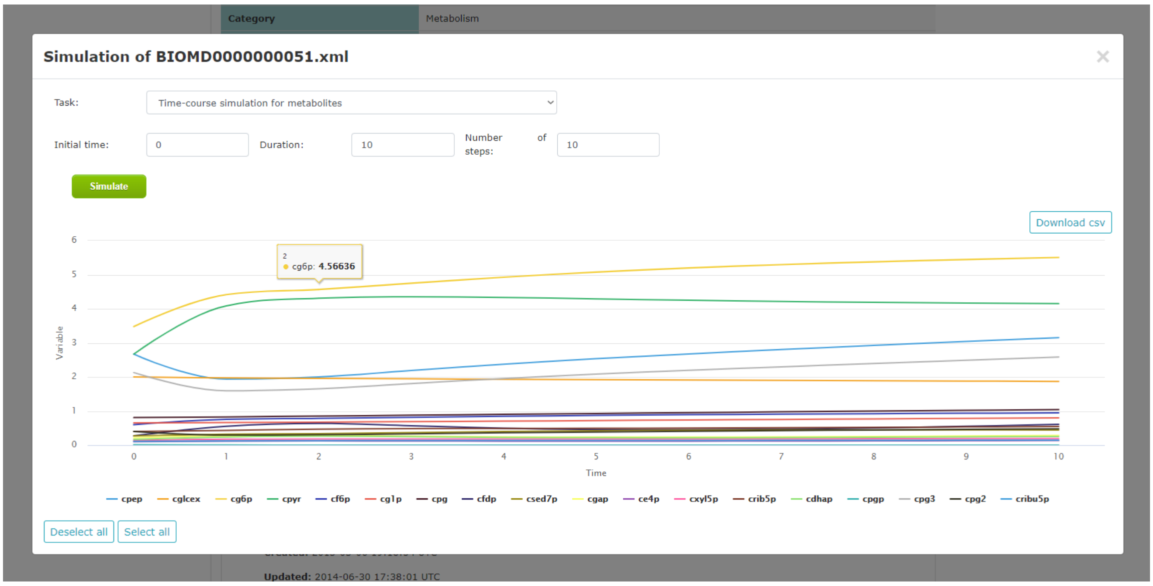
To achieve this goal, it was necessary to comprehend the structure of the different formats in order to build it correctly. For example RDF uses the *<data:organism>Escherichia coli</data:organism>* format, while the right structure of an XML file is *<organism>Escherichia coli</organism>*. Knowing this, then it was just a matter of completing the file with the data/model information available in the database.

### 5.2.3 New Simulation Environment

The introduction of a simulation environment facilitates the use of the associated kinetic models. As previously explained, the simulations are done through the COPASI Python language bindings, available to download via pip, in a package called *python-copasi*.

Fig 5.2 shows the simulation interface, where it is possible to choose the simulation type (time-course or steady-state), and in the case of a time-course, the user can also choose the simulation initial time, duration, and number of steps. Fig 5.3 shows the simulation interface available in the tools tab, which is slightly different from the previous. It allows to choose any SBML model file available in the database, or even an uploaded file from the user.

The steady-state simulation results are presented in two tables: one for the species, with their concentrations and rates, and another for the reactions with their fluxes values. The time-course simulation shows the plot results of the simulation. It is possible to select/deselect elements by clicking on them, or just press the select/deselect all buttons.

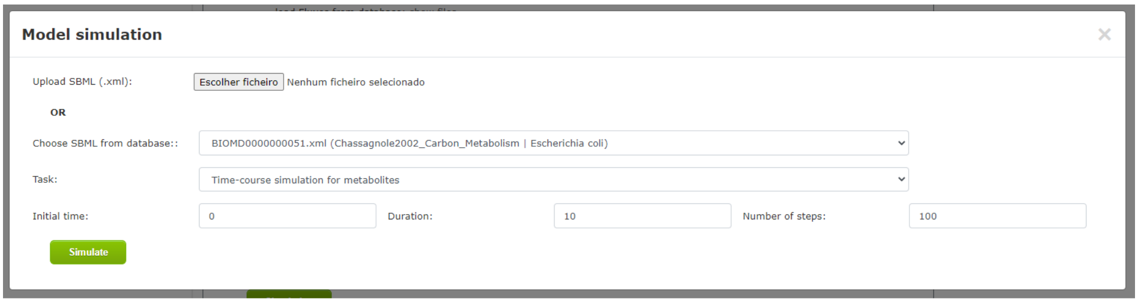


Task: Time-course simulation for metabolites

Initial time: 0 Duration: 10 Number of steps: 10

Updated: 2014-06-30 17:38:01 UTC

Figure 5.2: Example of a metabolite time-course simulation for Model Entry ID 13 (`https://kimosys.org/repository/30/associated_models/13`). Users can select/deselect elements by clicking on them or by pressing the select/deselect all buttons. Additionally, it is possible to check the values by hovering the plot lines.



Upload SBML (.xml): Nenhum ficheiro selecionado

Choose SBML from database:: BIOMD0000000051.xml (Chassagnole2002_Carbon_Metabolism | Escherichia coli)

Task: Time-course simulation for metabolites

Initial time: 0 Duration: 10 Number of steps: 100

Figure 5.3: The tools tab simulation interface. Here, the user can choose his own uploaded kinetic model to simulate, or any of the models available in the database.

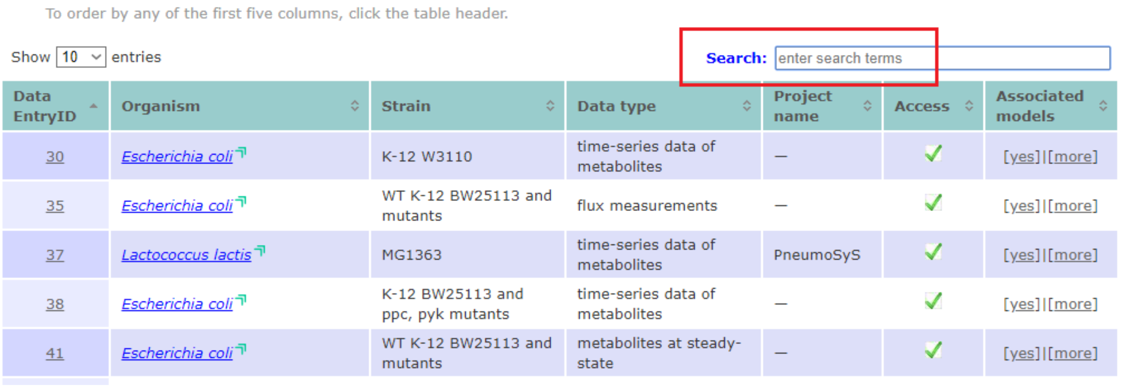
### 5.2.4 Improve Web Interface

Concerning user interface, the main objectives were improving the process of searching datasets, introducing a preview of files, making it easier to cite datasets and associated models, and facilitating the data submission process.

#### 5.2.4.1 Improved Queries and a new Filter Panel

When searching for data, it is important to have a wide range of queries, covering the most relevant data attributes, therefore facilitating the process to the user. To facilitate the searching process, improved queries and a new filter panel were introduced.

In the previous KiMoSys version, when exploring and searching for data, the only option available was to use the search box (fig 5.4a). Metadata search was already available, and it was enhanced with a few more attributes available to explore. Now the user can search for data and model elements (metabolites, reactions, proteins) by the standard cross-references ChEBI, KEGG and UniProt IDs, as shown in fig 5.4b. This manual search option can be combined with the filter panel described below.



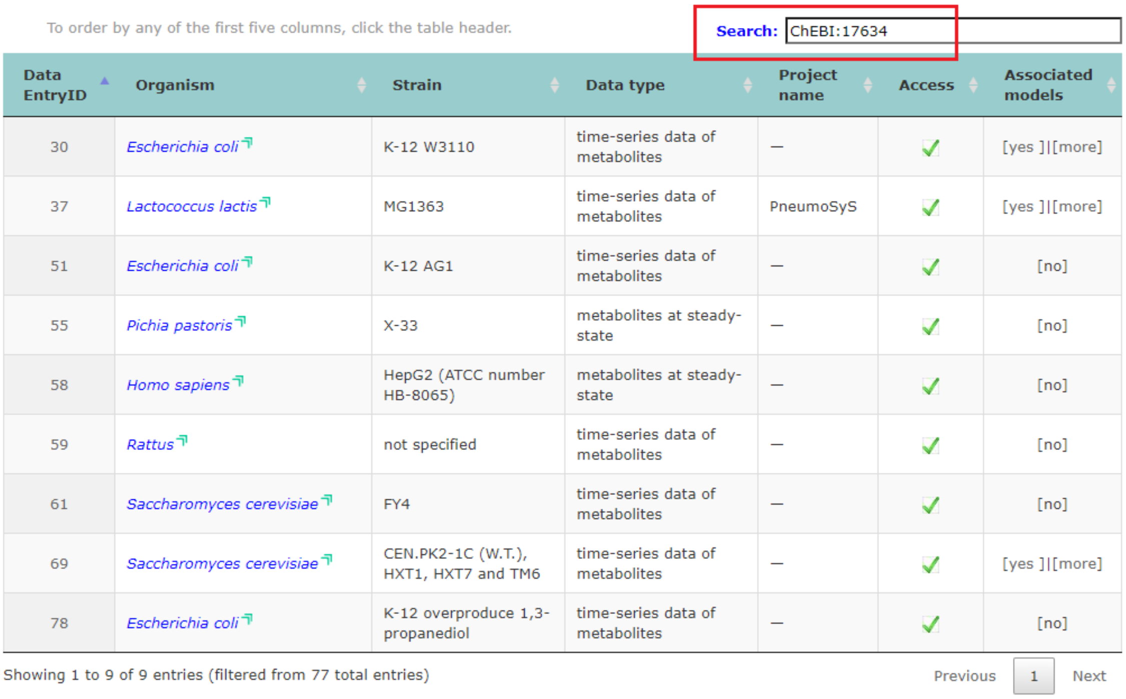
To order by any of the first five columns, click the table header.

Show 10 entries

Search: enter search terms

| Data EntryID | Organism | Strain | Data type | Project name | Access | Associated models |
|---|---|---|---|---|---|---|
| 30 | *Escherichia coli* | K-12 W3110 | time-series data of metabolites | – | ✓ | [yes]\|[more] |
| 35 | *Escherichia coli* | WT K-12 BW25113 and mutants | flux measurements | – | ✓ | [yes]\|[more] |
| 37 | *Lactococcus lactis* | MG1363 | time-series data of metabolites | PneumoSyS | ✓ | [yes]\|[more] |
| 38 | *Escherichia coli* | K-12 BW25113 and ppc, pyk mutants | time-series data of metabolites | – | ✓ | [yes]\|[more] |
| 41 | *Escherichia coli* | WT K-12 BW25113 and mutants | metabolites at steady-state | – | ✓ | [yes]\|[more] |

(a) The repository view (https://kimosys.org/repository) from the previous version. It was only possible to search in the input box.



To order by any of the first five columns, click the table header.

Search: ChEBI:17634

| Data EntryID | Organism | Strain | Data type | Project name | Access | Associated models |
|---|---|---|---|---|---|---|
| 30 | *Escherichia coli* | K-12 W3110 | time-series data of metabolites | – | ✓ | [yes ]\|[more] |
| 37 | *Lactococcus lactis* | MG1363 | time-series data of metabolites | PneumoSyS | ✓ | [yes ]\|[more] |
| 51 | *Escherichia coli* | K-12 AG1 | time-series data of metabolites | – | ✓ | [no] |
| 55 | *Pichia pastoris* | X-33 | metabolites at steady-state | – | ✓ | [no] |
| 58 | *Homo sapiens* | HepG2 (ATCC number HB-8065) | metabolites at steady-state | – | ✓ | [no] |
| 59 | *Rattus* | not specified | time-series data of metabolites | – | ✓ | [no] |
| 61 | *Saccharomyces cerevisiae* | FY4 | time-series data of metabolites | – | ✓ | [no] |
| 69 | *Saccharomyces cerevisiae* | CEN.PK2-1C (W.T.), HXT1, HXT7 and TM6 | time-series data of metabolites | – | ✓ | [yes ]\|[more] |
| 78 | *Escherichia coli* | K-12 overproduce 1,3-propanediol | time-series data of metabolites | – | ✓ | [no] |

Showing 1 to 9 of 9 entries (filtered from 77 total entries)

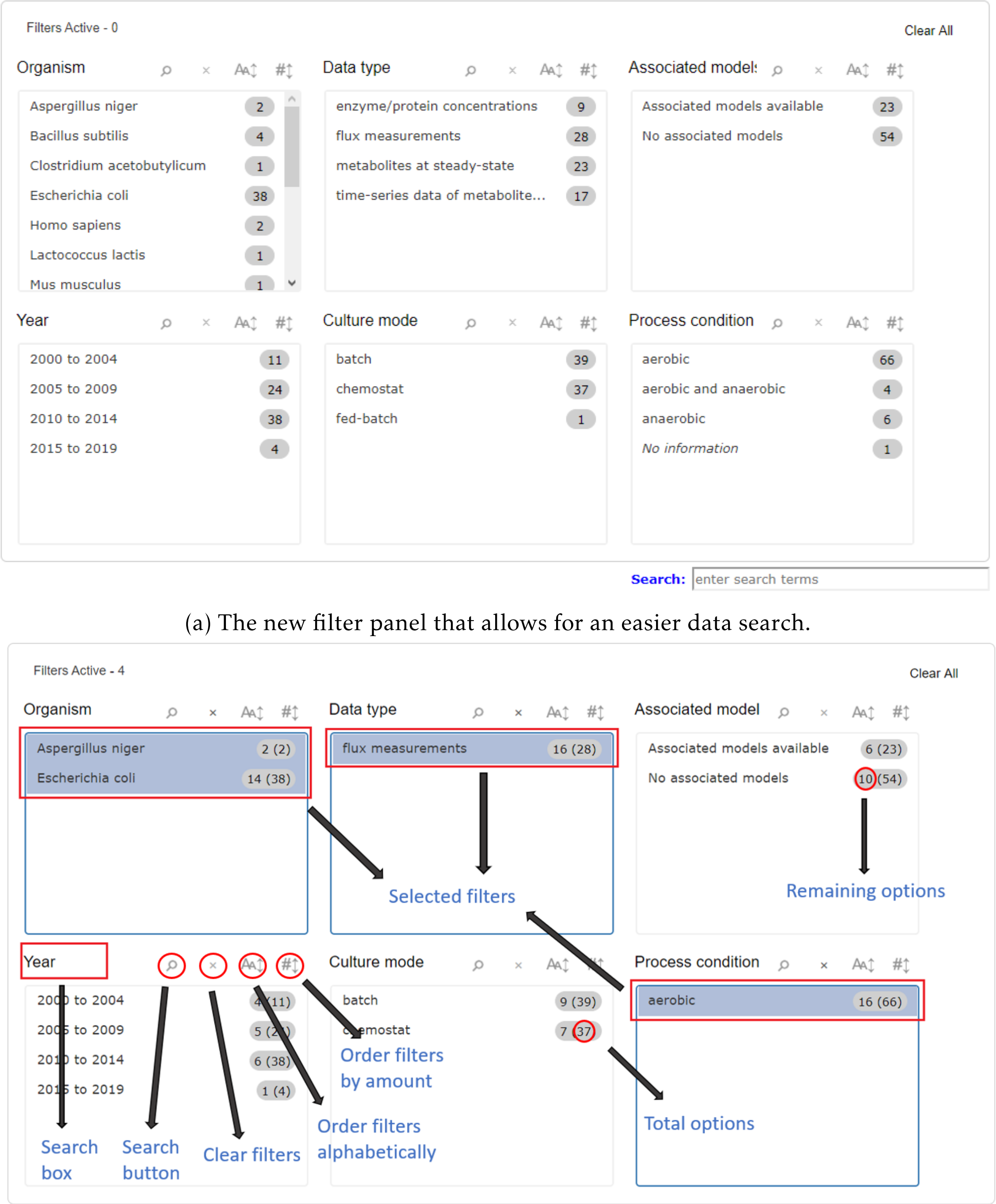
(b) The new view of the repository, and the new option to search by the standard cross-references ChEBI, KEGG and UniProt IDs that are present in the Data Entry IDs Excel files.

Figure 5.4: Screenshots from the old and new repository view, and its new functionalities.

As fig 5.5a shows, there are six categories available, being possible to select multiple options in the same or from different boxes, that can be viewed as Boolean expressions. For example, the selection of *Aspergillus niger*, *Escherichia coli*, *flux measurements* and *aerobic* (fig 5.5b) produces the following Boolean expression:

$$(\textit{Aspergillus niger} \vee \textit{Escherichia coli}) \wedge \textit{flux measurements} \wedge \textit{aerobic}$$

The result of the query will be all Data Entry IDs whose data type is *flux measurement*,



Filters Active - 0

Clear All

| Organism | |
|---|---|
| Aspergillus niger | 2 |
| Bacillus subtilis | 4 |
| Clostridium acetobutylicum | 1 |
| Escherichia coli | 38 |
| Homo sapiens | 2 |
| Lactococcus lactis | 1 |
| Mus musculus | 1 |

| Data type | |
|---|---|
| enzyme/protein concentrations | 9 |
| flux measurements | 28 |
| metabolites at steady-state | 23 |
| time-series data of metabolite... | 17 |

| Associated model: | |
|---|---|
| Associated models available | 23 |
| No associated models | 54 |

| Year | |
|---|---|
| 2000 to 2004 | 11 |
| 2005 to 2009 | 24 |
| 2010 to 2014 | 38 |
| 2015 to 2019 | 4 |

| Culture mode | |
|---|---|
| batch | 39 |
| chemostat | 37 |
| fed-batch | 1 |

| Process condition | |
|---|---|
| aerobic | 66 |
| aerobic and anaerobic | 4 |
| anaerobic | 6 |
| *No information* | 1 |

Search: enter search terms

(a) The new filter panel that allows for an easier data search.

Filters Active - 4

Clear All

| Organism | |
|---|---|
| Aspergillus niger | 2 (2) |
| Escherichia coli | 14 (38) |

| Data type | |
|---|---|
| flux measurements | 16 (28) |

| Associated model | |
|---|---|
| Associated models available | 6 (23) |
| No associated models | 10 (54) |

| Year | |
|---|---|
| 2000 to 2004 | 4 (11) |
| 2005 to 2009 | 5 (24) |
| 2010 to 2014 | 6 (38) |
| 2015 to 2019 | 1 (4) |

| Culture mode | |
|---|---|
| batch | 9 (39) |
| chemostat | 7 (37) |

| Process condition | |
|---|---|
| aerobic | 16 (66) |

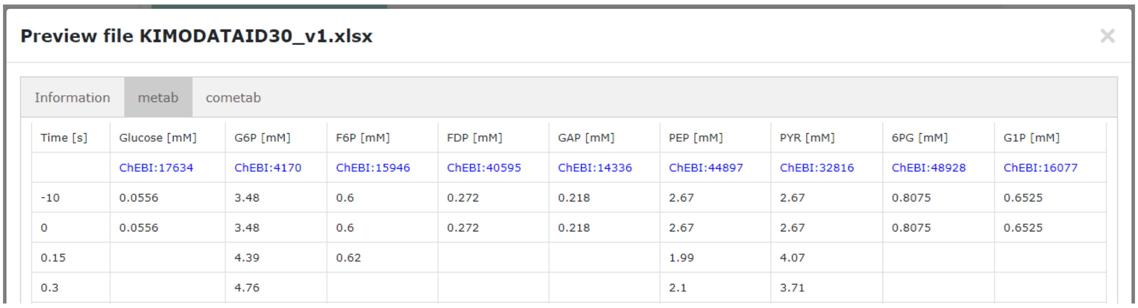
(b) Detailed information about how the panel works after a filter selection.

Figure 5.5: Screenshots of the new filter panel and its interactions.

the process condition *aerobic*, and the Organism either *Aspergillus niger* or *Escherichia coli*. Ahead of each element it is represented its total number present in the database. As one filter is selected, the panel refreshes, showing the remaining options and its numbers, according to the previous option(s) selected (depicted in fig 5.5b). Additionally, inside each box it is possible to sort its elements alphabetically or numerically, and even to type and search for an element. The selections can be restored using the "*Clear All*" button in the top-right corner. The user also has available the option to just clear the filters of one box, by pressing its "X" button.

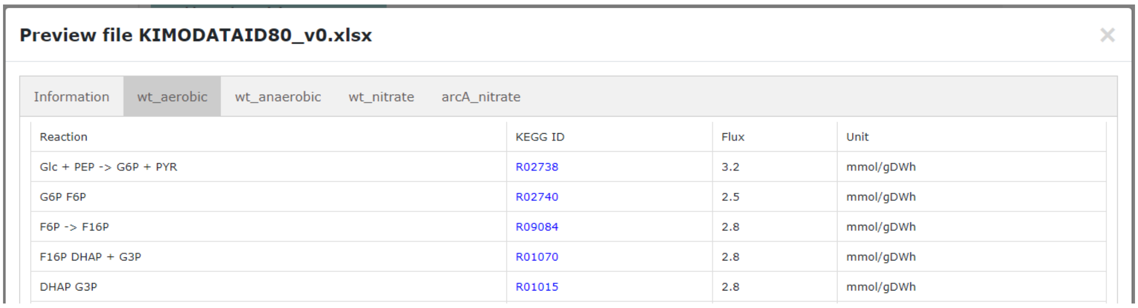
#### 5.2.4.2 File Preview

Accessing and viewing data without having to download it provides a more user-friendly and engaging interface to the users. The new KiMoSys' version allows to preview XLSX, CSV, and SBML files. Fig 5.6 shows the preview of the different XLSX files, where is possible to navigate through the tabs, as if the user were in Excel. Adding to that, each ChEBI, KEGG and UniProt ID, has its respective external link (figs 5.6a, 5.6b and 5.6c), providing more insightful information to the user.



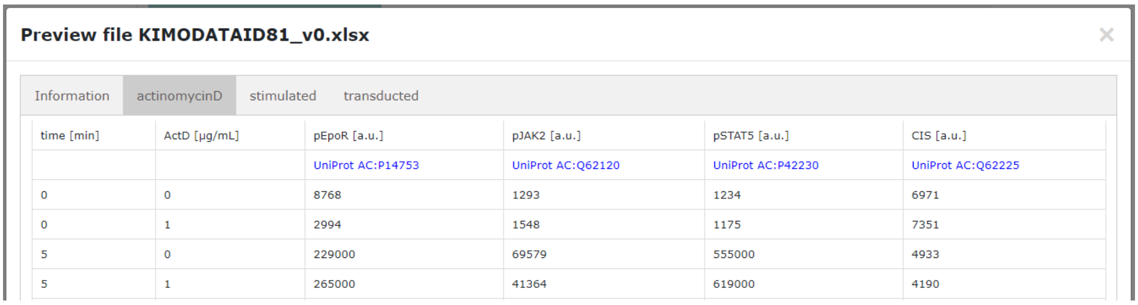
| Time [s] | Glucose [mM] | G6P [mM] | F6P [mM] | FDP [mM] | GAP [mM] | PEP [mM] | PYR [mM] | 6PG [mM] | G1P [mM] |
|---|---|---|---|---|---|---|---|---|---|
| | ChEBI:17634 | ChEBI:4170 | ChEBI:15946 | ChEBI:40595 | ChEBI:14336 | ChEBI:44897 | ChEBI:32816 | ChEBI:48928 | ChEBI:16077 |
| -10 | 0.0556 | 3.48 | 0.6 | 0.272 | 0.218 | 2.67 | 2.67 | 0.8075 | 0.6525 |
| 0 | 0.0556 | 3.48 | 0.6 | 0.272 | 0.218 | 2.67 | 2.67 | 0.8075 | 0.6525 |
| 0.15 | | 4.39 | 0.62 | | | 1.99 | 4.07 | | |
| 0.3 | | 4.76 | | | | 2.1 | 3.71 | | |

(a) Preview of a time-series of metabolites XLSX file (Data Entry ID 30).



| Reaction | KEGG ID | Flux | Unit |
|---|---|---|---|
| Glc + PEP -> G6P + PYR | R02738 | 3.2 | mmol/gDWh |
| G6P F6P | R02740 | 2.5 | mmol/gDWh |
| F6P -> F16P | R09084 | 2.8 | mmol/gDWh |
| F16P DHAP + G3P | R01070 | 2.8 | mmol/gDWh |
| DHAP G3P | R01015 | 2.8 | mmol/gDWh |

(b) Preview of a flux measurements XLSX file (Data Entry ID 80).



| time [min] | ActD [µg/mL] | pEpoR [a.u.] | pJAK2 [a.u.] | pSTAT5 [a.u.] | CIS [a.u.] |
|---|---|---|---|---|---|
| | | UniProt AC:P14753 | UniProt AC:Q62120 | UniProt AC:P42230 | UniProt AC:Q62225 |
| 0 | 0 | 8768 | 1293 | 1234 | 6971 |
| 0 | 1 | 2994 | 1548 | 1175 | 7351 |
| 5 | 0 | 229000 | 69579 | 555000 | 4933 |
| 5 | 1 | 265000 | 41364 | 619000 | 4190 |

(c) Preview of an enzyme/protein XLSX file (Data Entry ID 81).

Figure 5.6: Screenshots of the previews available for the different data types.

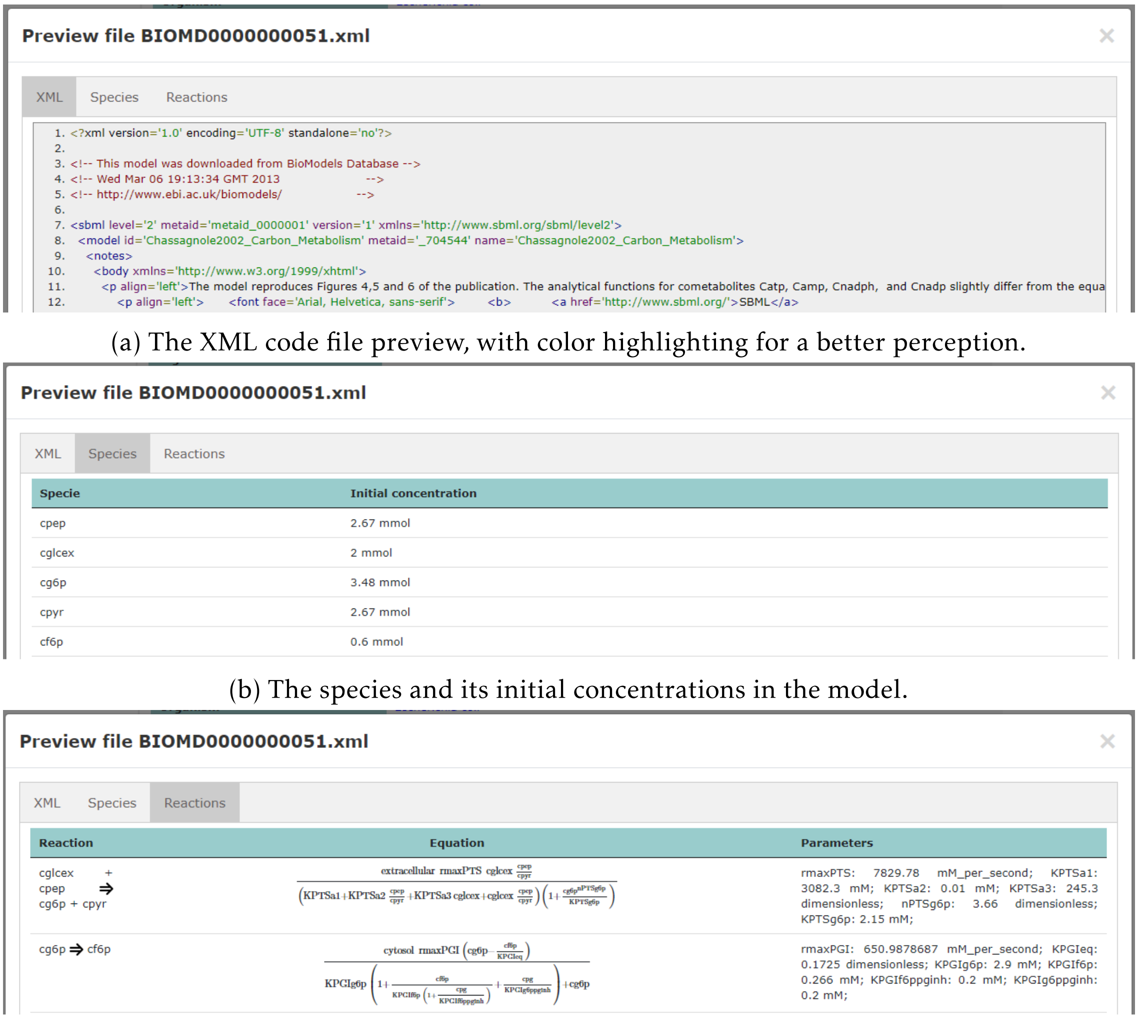
In fig 5.7 it is depicted an SBML file preview for an associated kinetic model. Its content is separated in three tabs that the user can navigate through:

1. The XML code, with different colors that help to better understand the raw code

(fig 5.7a).

2. The species present in the kinetic model and their respective initial concentrations (fig 5.7b).

3. The reactions of the model (fig 5.7c). Each reaction has its mathematical equation disposed in a readable way. By default, when double clicking the equation image it zooms in. MathJax allows to change this configuration by right-clicking in the equation. Finally, the parameters of the reaction are presented with their respective values and units.



Preview file BIOMD0000000051.xml
×
XML
Species
Reactions

```
<?xml version='1.0' encoding='UTF-8' standalone='no'?>

<!-- This model was downloaded from BioModels Database -->
<!-- Wed Mar 06 19:13:34 GMT 2013                      -->
<!-- http://www.ebi.ac.uk/biomodels/                  -->

<sbml level='2' metaid='metaid_0000001' version='1' xmlns='http://www.sbml.org/sbml/level2'>
 <model id='Chassagnole2002_Carbon_Metabolism' metaid='_704544' name='Chassagnole2002_Carbon_Metabolism'>
  <notes>
   <body xmlns='http://www.w3.org/1999/xhtml'>
    <p align='left'>The model reproduces Figures 4,5 and 6 of the publication. The analytical functions for cometabolites Catp, Camp, Cnadph,  and Cnadp slightly differ from the equa
      <p align='left'>     <font face='Arial, Helvetica, sans-serif'>       <b>        <a href='http://www.sbml.org/'>SBML</a>
```

(a) The XML code file preview, with color highlighting for a better perception.

Preview file BIOMD0000000051.xml
×
XML
Species
Reactions

| Specie | Initial concentration |
|---|---|
| cpep | 2.67 mmol |
| cglcex | 2 mmol |
| cg6p | 3.48 mmol |
| cpyr | 2.67 mmol |
| cf6p | 0.6 mmol |

(b) The species and its initial concentrations in the model.

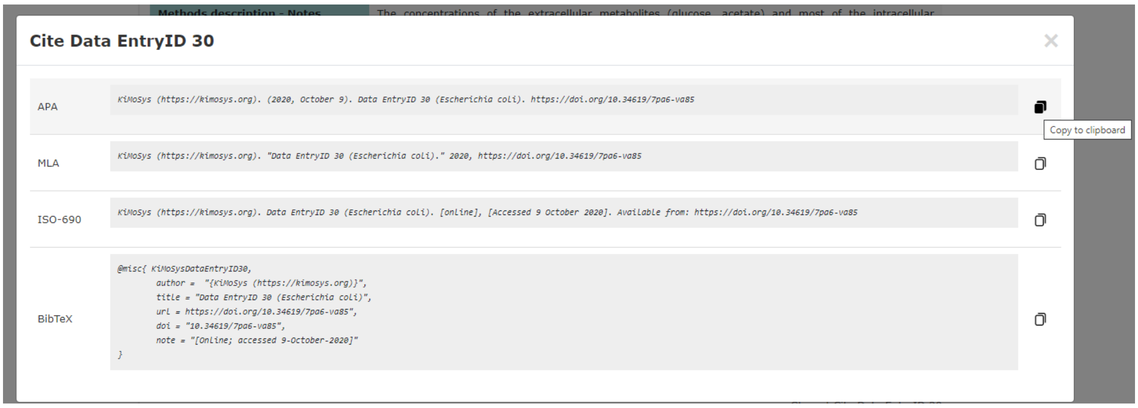
| Reaction | Equation | Parameters |
|---|---|---|
| cglcex + cpep ⇒ cg6p + cpyr | $\frac{extracellular\ rmaxPTS\ cglcex\frac{cpep}{cpyr}}{\left(KPTSa1+KPTSa2\frac{cpep}{cpyr}+KPTSa3\ cglcex+cglcex\frac{cpep}{cpyr}\right)\left(1+\frac{cg6p^{nPTSg6p}}{KPTSg6p}\right)}$ | rmaxPTS: 7829.78 mM_per_second; KPTSa1: 3082.3 mM; KPTSa2: 0.01 mM; KPTSa3: 245.3 dimensionless; nPTSg6p: 3.66 dimensionless; KPTSg6p: 2.15 mM; |
| cg6p ⇒ cf6p | $\frac{cytosol\ rmaxPGI\left(cg6p-\frac{cf6p}{KPGIeq}\right)}{KPGIg6p\left(1+\frac{cf6p}{KPGIf6p\left(1+\frac{cpg}{KPGIf6ppginh}\right)}+\frac{cpg}{KPGIg6ppginh}\right)+cg6p}$ | rmaxPGI: 650.9878687 mM_per_second; KPGIeq: 0.1725 dimensionless; KPGIg6p: 2.9 mM; KPGIf6p: 0.266 mM; KPGIf6ppginh: 0.2 mM; KPGIg6ppginh: 0.2 mM; |

(c) The reactions of the model and its equations and parameters.

Figure 5.7: Preview of Model Entry ID 13.

#### 5.2.4.3 Data Citation

Data citation is important because it makes data and metadata more Accessible and Reusable. This new version introduced the ability to cite data easily, by simply clicking in a button that copies the citation information into the clipboard. Four generic citation formats were implemented: APA, MLA, ISO-690, and BibTeX, as shown in fig 5.8.



Methods description - Notes The concentrations of the extracellular metabolites (glucose, acetate) and most of the intracellular

**Cite Data EntryID 30** ×

| | |
|---|---|
| APA | KiMoSys (https://kimosys.org). (2020, October 9). Data EntryID 30 (Escherichia coli). https://doi.org/10.34619/7pa6-va85 |
| MLA | KiMoSys (https://kimosys.org). "Data EntryID 30 (Escherichia coli)." 2020, https://doi.org/10.34619/7pa6-va85 |
| ISO-690 | KiMoSys (https://kimosys.org). Data EntryID 30 (Escherichia coli). [online], [Accessed 9 October 2020]. Available from: https://doi.org/10.34619/7pa6-va85 |

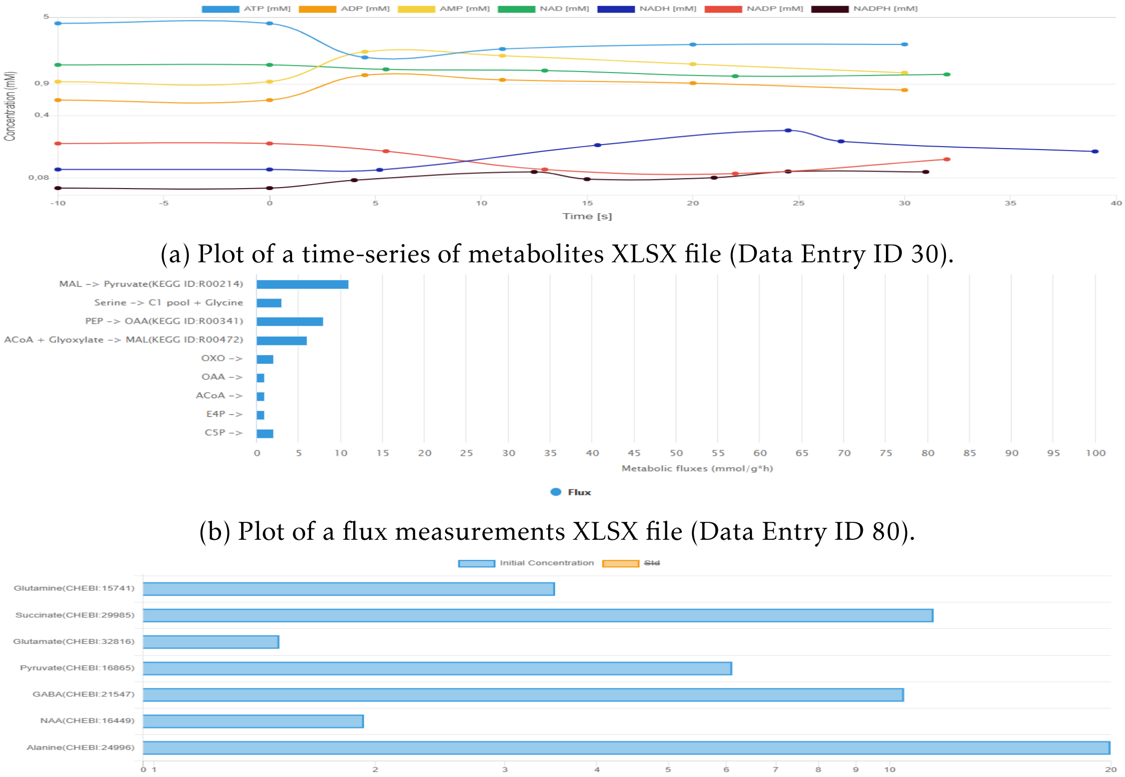
BibTeX

```
@misc{ KiMoSysDataEntryID30,
        author =  "{KiMoSys (https://kimosys.org)}",
        title = "Data EntryID 30 (Escherichia coli)",
        url = https://doi.org/10.34619/7pa6-va85",
        doi = "10.34619/7pa6-va85",
        note = "[Online; accessed 9-October-2020]"
}
```

Figure 5.8: The four citation options available, with the copy to clipboard button.

### 5.2.5 New Data Visualization

Data visualization is crucial for a more intuitive interpretation of experimental results. In addition to the simulation graphs seen previously, plots for the different data types were also introduced. These are built with the data present in the Excel files. Figs 5.9a, 5.9b



(a) Plot of a time-series of metabolites XLSX file (Data Entry ID 30).

(b) Plot of a flux measurements XLSX file (Data Entry ID 80).

(c) PLot of an enzyme/protein XLSX file (Data Entry ID 81).

Figure 5.9: Snapshots of the plots available for the different data types.

and 5.9c show the different plots available. These are interactive, the user can choose the elements that appear, by clicking on them.

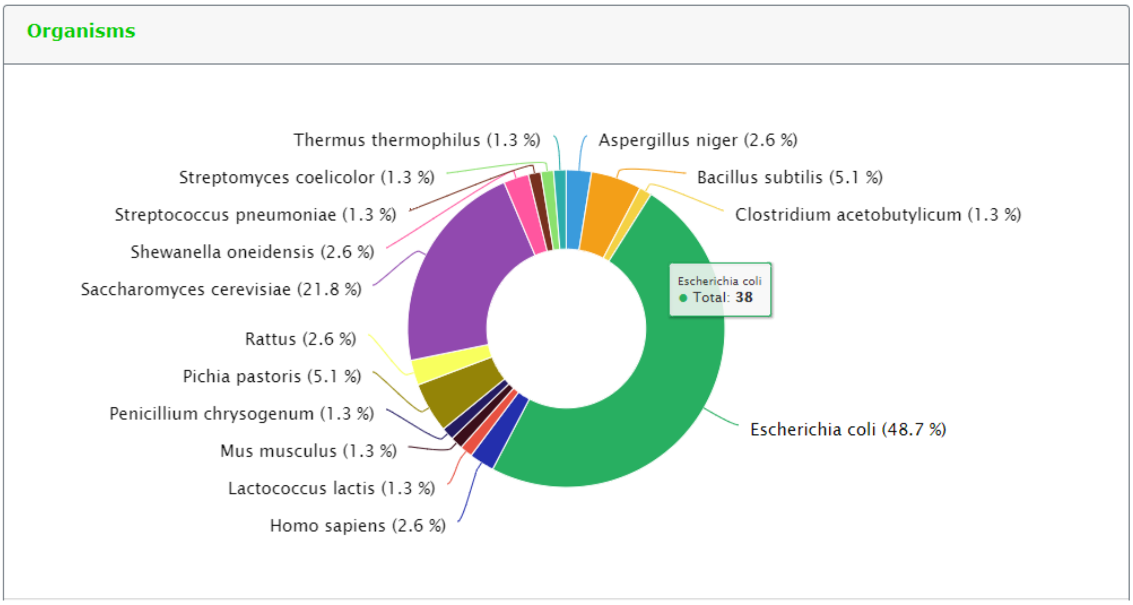
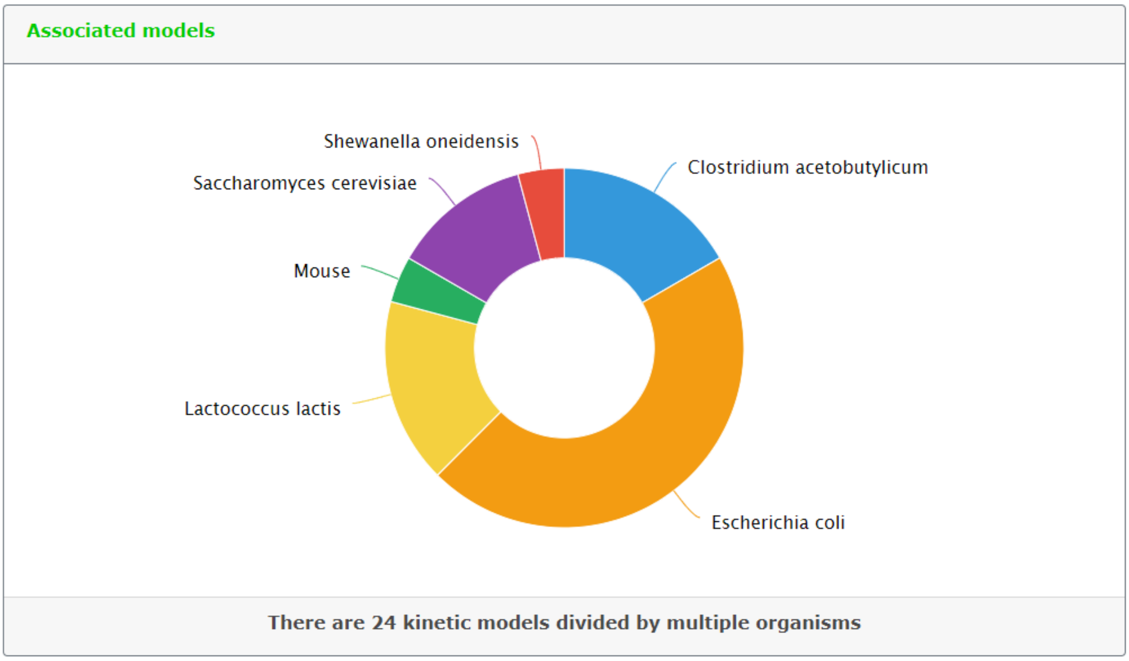


Figure 5.10: The graph that depicts the Organisms present in KiMoSys repository.



There are 24 kinetic models divided by multiple organisms

Figure 5.11: The information graph about the Associated Models.

In addition to these plots, a new statistics page with the repository data was introduced in KiMoSys' new version. This page provides a better overview of the repository data, with detailed graphs about Organisms (fig 5.10), Associated Models (fig 5.11), or Data Availability (fig 5.12), among others. By hovering the cursor in a certain element it

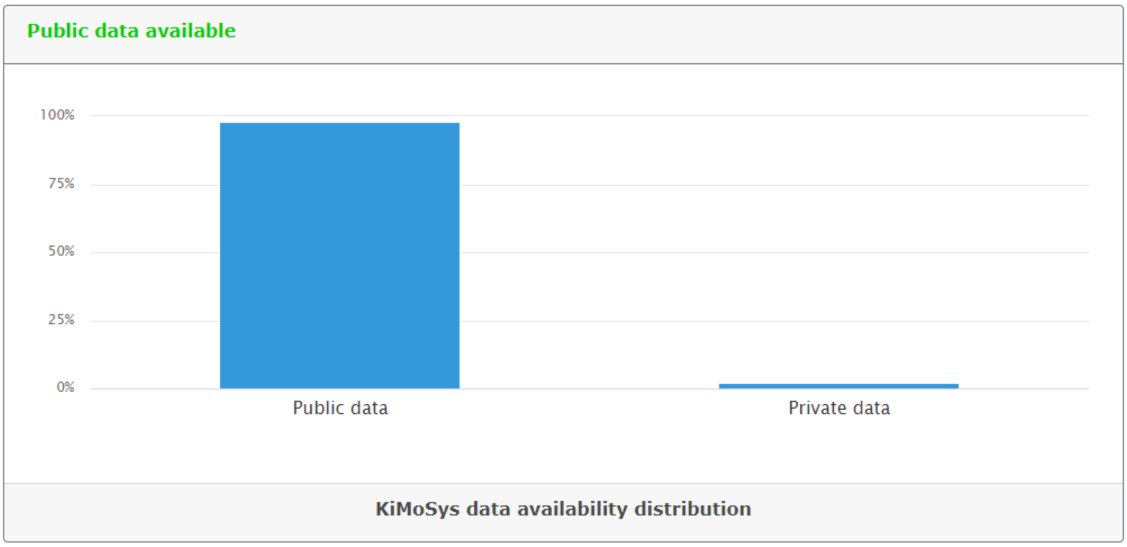


Figure 5.12: Data availability: almost 100% of the data in KiMoSys is public.

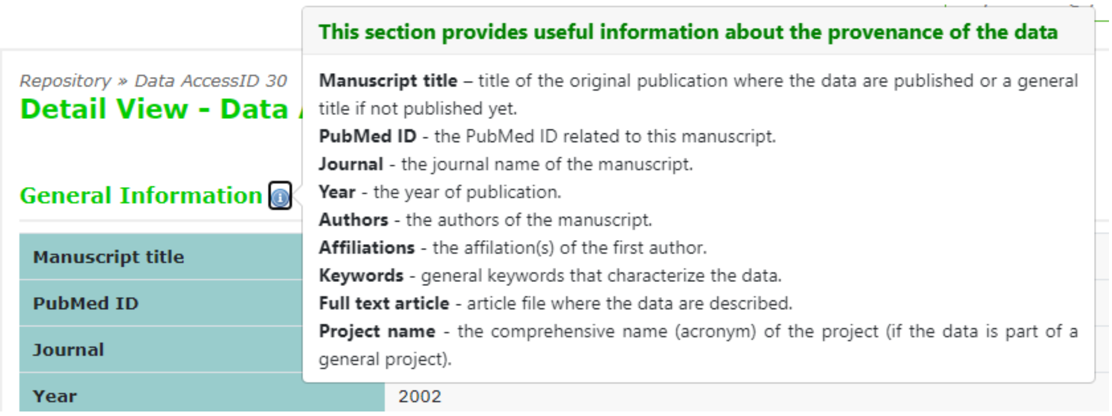
is possible to see its exact amount.

### 5.2.6 Minor Interface Changes

Bootstrap allowed to modernize KiMoSys' 2.0 interface. Several minor changes were introduced with the help of this tool. Regarding this, tooltips (figs 5.13 and 5.14) have now an improved visual. In figs 5.16 and 5.15 is possible to see the differences between the previous version and the new concerning the Entry IDs pages and forms, respectively. These modifications allow for a better read of data.

### 5.2.7 An Improved Submission Tool

In KiMoSys' original version, when submitting a new Data Entry ID, it was necessary to fill an Excel file with the metadata, and then repeat this information in the online submission form. To eliminate this inconvenience, the process was automated. When

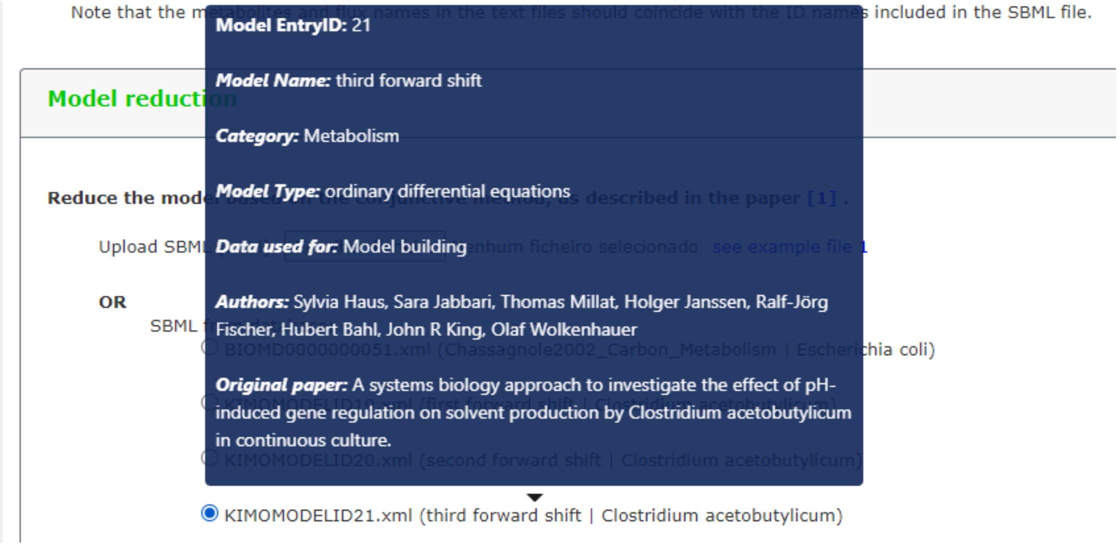


This section provides useful information about the provenance of the data

**Manuscript title** – title of the original publication where the data are published or a general title if not published yet.
**PubMed ID** - the PubMed ID related to this manuscript.
**Journal** - the journal name of the manuscript.
**Year** - the year of publication.
**Authors** - the authors of the manuscript.
**Affiliations** - the affilation(s) of the first author.
**Keywords** - general keywords that characterize the data.
**Full text article** - article file where the data are described.
**Project name** - the comprehensive name (acronym) of the project (if the data is part of a general project).

| | |
|---|---|
| Manuscript title | |
| PubMed ID | |
| Journal | |
| Year | 2002 |

Figure 5.13: The new tooltip describing the Data Entry ID metadata topics.



Note that the m... s included in the SBML file.

Model reducti...

Reduce the mod... s described in the paper [1] .

Upload SBML ... nhum ficheiro selecionado see example file 1

OR

SBML ...

**Model EntryID:** 21

***Model Name:*** third forward shift

***Category:*** Metabolism

***Model Type:*** ordinary differential equations

***Data used for:*** Model building

***Authors:*** Sylvia Haus, Sara Jabbari, Thomas Millat, Holger Janssen, Ralf-Jörg Fischer, Hubert Bahl, John R King, Olaf Wolkenhauer

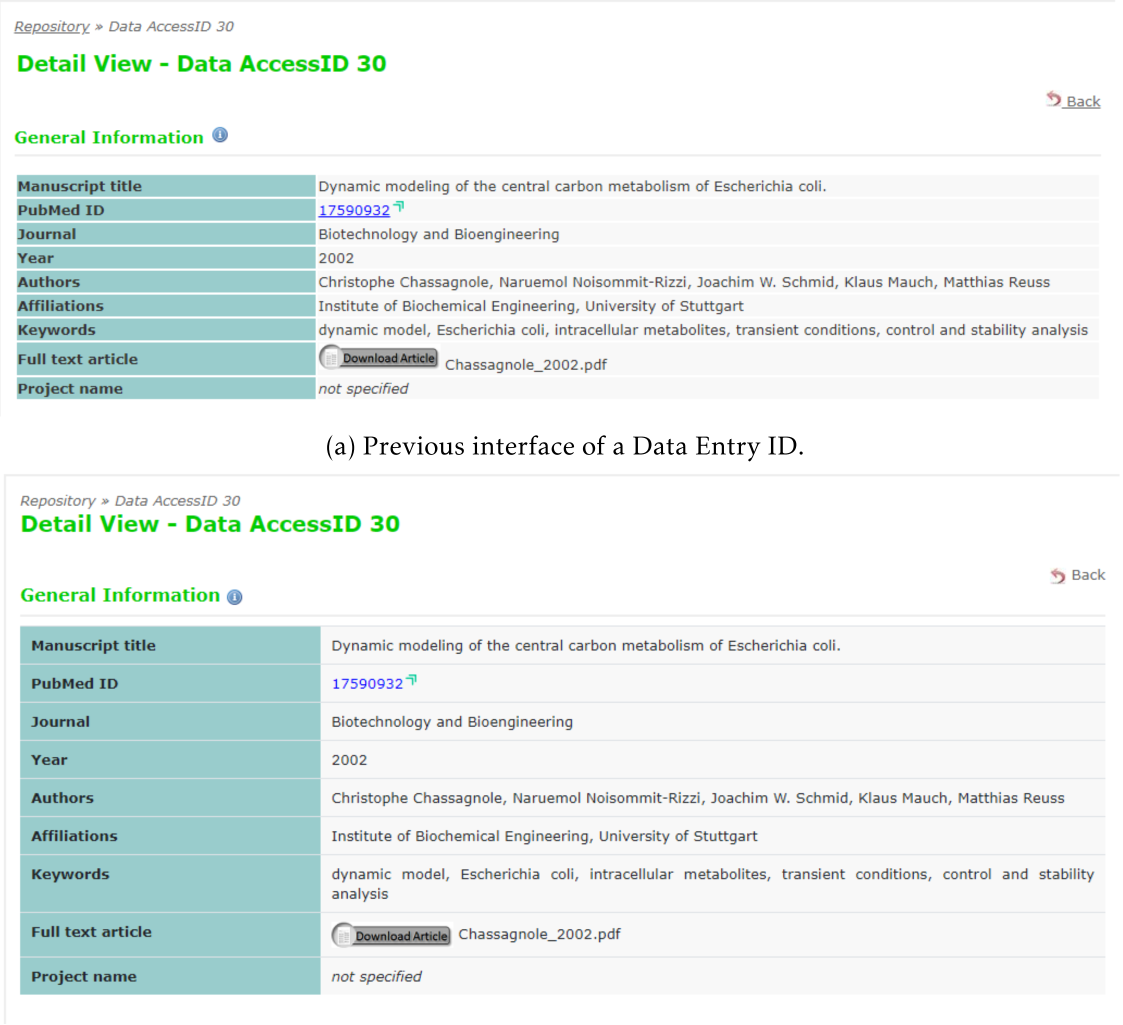
***Original paper:*** A systems biology approach to investigate the effect of pH-induced gene regulation on solvent production by Clostridium acetobutylicum in continuous culture.

○ BIOMD0000000051.xml (Chassagnole2002_Carbon_Metabolism | Escherichia coli)

○ KIMOMODELID20.xml (second forward shift | Clostridium acetobutylicum)

● KIMOMODELID21.xml (third forward shift | Clostridium acetobutylicum)

Figure 5.14: A new tooltip that depicts a Model Entry ID in the tools section.



Repository » Data AccessID 30

Detail View - Data AccessID 30

Back

General Information

| | |
|---|---|
| Manuscript title | Dynamic modeling of the central carbon metabolism of Escherichia coli. |
| PubMed ID | 17590932 |
| Journal | Biotechnology and Bioengineering |
| Year | 2002 |
| Authors | Christophe Chassagnole, Naruemol Noisommit-Rizzi, Joachim W. Schmid, Klaus Mauch, Matthias Reuss |
| Affiliations | Institute of Biochemical Engineering, University of Stuttgart |
| Keywords | dynamic model, Escherichia coli, intracellular metabolites, transient conditions, control and stability analysis |
| Full text article | Download Article Chassagnole_2002.pdf |
| Project name | *not specified* |

(a) Previous interface of a Data Entry ID.

Repository » Data AccessID 30

Detail View - Data AccessID 30

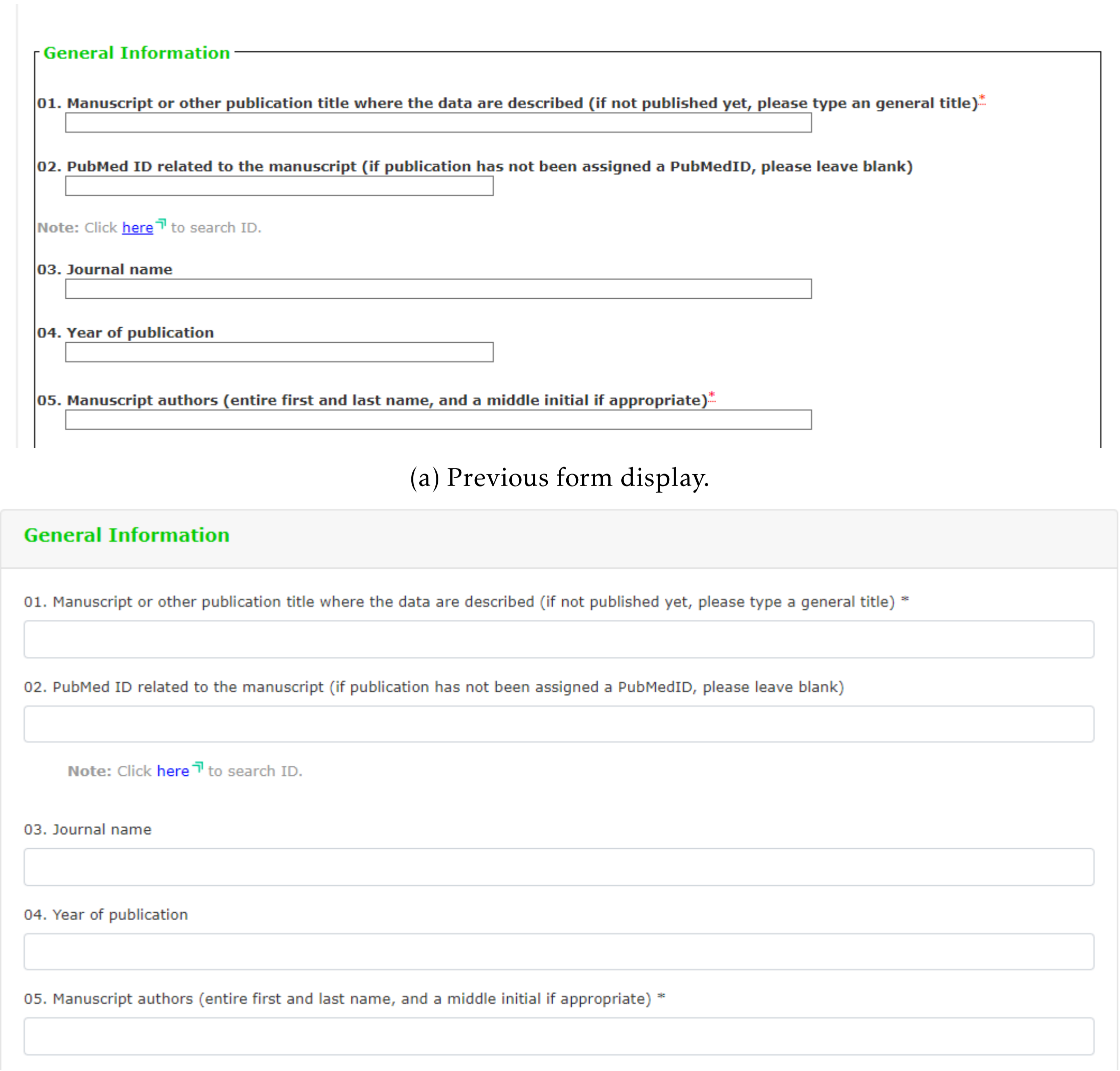
Back

General Information

| | |
|---|---|
| Manuscript title | Dynamic modeling of the central carbon metabolism of Escherichia coli. |
| PubMed ID | 17590932 |
| Journal | Biotechnology and Bioengineering |
| Year | 2002 |
| Authors | Christophe Chassagnole, Naruemol Noisommit-Rizzi, Joachim W. Schmid, Klaus Mauch, Matthias Reuss |
| Affiliations | Institute of Biochemical Engineering, University of Stuttgart |
| Keywords | dynamic model, Escherichia coli, intracellular metabolites, transient conditions, control and stability analysis |
| Full text article | Download Article Chassagnole_2002.pdf |
| Project name | *not specified* |

(b) The new visuals, offering a better read of the table information.

Figure 5.15: The interface differences in a Data Entry ID page.



General Information

01. Manuscript or other publication title where the data are described (if not published yet, please type an general title)*

02. PubMed ID related to the manuscript (if publication has not been assigned a PubMedID, please leave blank)

Note: Click here to search ID.

03. Journal name

04. Year of publication

05. Manuscript authors (entire first and last name, and a middle initial if appropriate)*

(a) Previous form display.

General Information

01. Manuscript or other publication title where the data are described (if not published yet, please type a general title) *

02. PubMed ID related to the manuscript (if publication has not been assigned a PubMedID, please leave blank)

Note: Click here to search ID.

03. Journal name

04. Year of publication

05. Manuscript authors (entire first and last name, and a middle initial if appropriate) *

(b) The new form visual, created with Bootstrap.

Figure 5.16: The differences between the forms, in this particular case the Data Entry ID submission form.

submitting, now the user only needs to fill the template with the actual data, and the metadata submitted in the online form is appended to that Excel file. In this way, the excess amount of time required to the user to fill the same information twice disappears, and contributes to make sharing more attractive. In Appendix A it is depicted the new dataset submitting process.

### 5.2.8 KiMoSys 2.0 Architecture

The KiMoSys' 2.0 web platform view for an unregistered user is composed by 7 tabs:

- HOME - A brief description of the repository with its goals and features is displayed in this tab.

- REPOSITORY - Here it is possible to find the main repository table with the filter panel, as well as the data submission options.
- TOOLS - Where various features concerning model construction and simulation are available.
- STATISTICS - Only introduced in KiMoSys 2.0, this tab displays several plots depicting the data present in the repository.
- DOCUMENTATION - On this page, the user can find information regarding the web platform usage, including a user manual, a FAQ, and the proper recognition to the website contributors and funders.
- LINKS - Multiple external links to kinetic simulation and parameter estimation tools, standards and exchange formats, and similar databases, such as the previously mentioned BRENDA and SABIO-RK.
- CONTACT US - Here a Google Form is provided, where the users can send their suggestions, comments, questions, or complaints.

For a registered user another tab is present, named MY REPOSITORY, where the users can find their submitted data.

### 5.2.9 reCAPTCHA Results

Google reCAPTCHA has already helped to block some registrations. These are believed to be done by bots. Fig 5.17 depicts the number of approved and unapproved requests to create new accounts in KiMoSys 2.0.

## 5.3 Advantages of KiMoSys 2.0

KiMoSys 1.0 already had new features: the ability to associate kinetic models to datasets that helped to build them. Adding to that, KiMoSys is not a single-omics type database such as PRIDE or CeCaFDB. The introduction of the RDF format also enhances KiMoSys 2.0 abilities concerning machine-reading formats.

Although other repositories already have some of the features implemented in the course of this project, like data visualization, model simulation interfaces, more appealing user interfaces, a data persistent identifier system, or diverse downloadable formats, KiMoSys 2.0 comprises all these characteristics. Adding these to the features already present in KiMoSys 1.0, which includes a tool that helps the construction of kinetic models, KiMoSys 2.0 now has an extensive set of functionalities.

Considering this, KiMoSys 2.0 is a well-rounded repository, aggregating important capabilities not only in the kinetic models' field but also in experimental data.

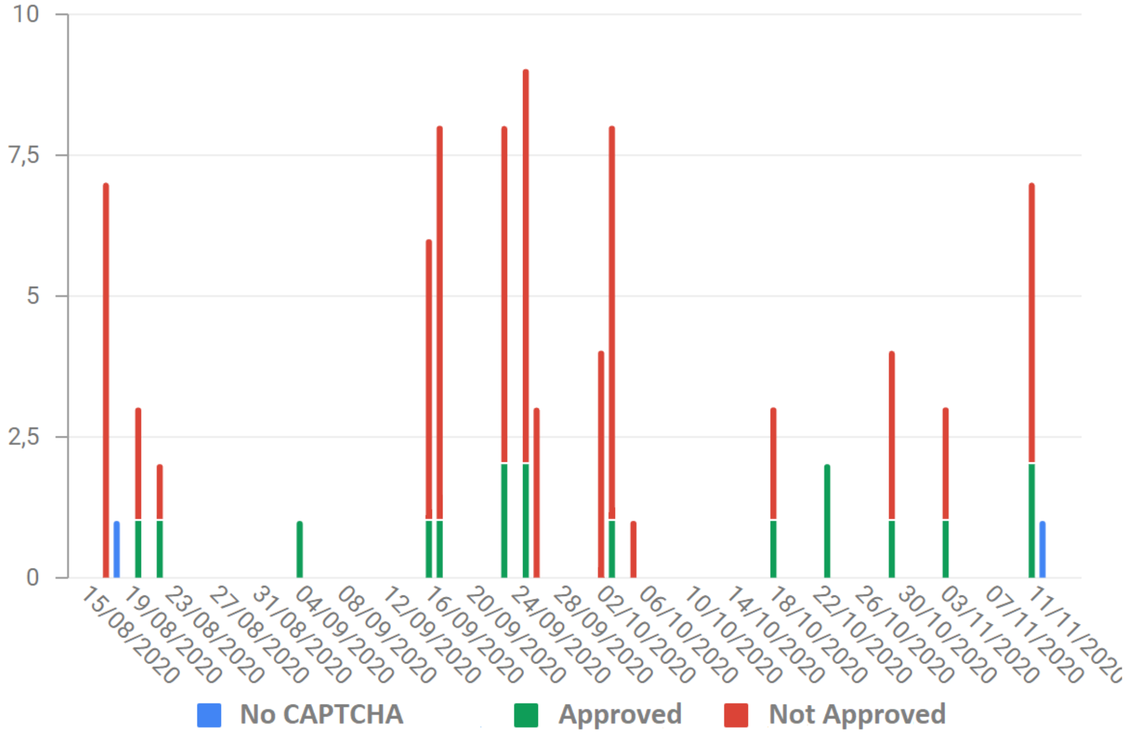


Figure 5.17: The Google reCAPTCHA results, with the approved and not approved requests to create new accounts.

## 5.4 Validation

Given the short time in which the repository is online, it is not yet easy to assess the adhesion of new users and entries produced by the new version of KiMoSys. Nevertheless, a paper was successfully submitted to DATABASE - The Journal of Biological Databases and Curation (https://academic.oup.com/database), and its acceptance validates somehow the quality of the new KiMoSys features and suggests its wider use in the future.

# Chapter 6

# Conclusion

In this chapter the achievements of the thesis project and the scopes for future directions are discussed.

## 6.1 Main Conclusions

Public repositories are gaining importance every year, and data sharing has never been as central as it is today. KiMoSys aims to provide a web-accessible data management platform, offering published experimental data, which is connected to the corresponding associated kinetic models. The ability to browse the datasets that helped to validate the kinetic models is a relevant feature. Adding to this, the options to deal with these models, such as Model Reduction or Kinetic Equations Translation, conferred a relevant position in the modeling area to this platform, that lead to some important recommendations by respected scientific journals.

In this work a new version of KiMoSys was developed, and more functionalities regarding the modeling part were added, being now possible to make two types of simulations: time-course, and steady-state. The introduction of data visualization and preview are important to enhance the user experience, allowing for a better perception of the data present in KiMoSys. Concerning this topic, the new statistics page is very informative about the repository data. The introduction of data citations and new downloadable formats facilitate data usage. The interface update gave KiMoSys a modern look, which has great importance nowadays, as websites give more and more relevance to User Interfaces and also User Experience. The addition of a Unique Persistent Identifier System granted KiMoSys the Findable principle, making it fully adherent to the FAIR principles. The improved submission tool can now spare some time to the users, encouraging them to interact more with the website.

All these new functionalities confer more data FAIRness to KiMoSys, being now even more unique in its capabilities. The introduction of popular libraries such as Bootstrap, eases the future upgrades, as these tools receive constant maintenance and have a large community of users. In the same way, the introduction of COPASI facilitates the implementation of new features for kinetic models in the future. The next big step is to introduce more datasets to the repository, in order have a wider community using KiMoSys.

## 6.2 Future Directions

This section addresses possible implementations to introduce in the future, and the measure of the results produced in this thesis.

### 6.2.1 Insert new Datasets/Models

Having more data in the repository can potentially attract more users to KiMoSys. The inclusion of an improved submission tool that allows users to deposit data in an easier way was already made according to this line of thinking. Right now KiMoSys has around 100 entries in total (counting Data + Model Entry IDs), which is a relatively small number when comparing to other repositories described in Chapter 3.

There are two ways to introduce new data to the repository: through collecting data manually from literature or using data mining algorithms. Data present in KiMoSys was introduced using the first option, with permission from the authors. By using data mining algorithms to extract data from other repositories, it would also be necessary to ask for their authorization. These algorithms would have to adapt and adjust the extracted data to fit KiMoSys' structure. As other repositories may follow different conventions in terms of metadata vocabulary or arrangement of the datasets files, a curation process would possibly need to be done by the system administrators.

### 6.2.2 Extend COPASI functionalities

The COPASI integration is already done through its Python language bindings, making it possible to do steady-state and time-course simulations. But COPASI has many more abilities concerning kinetic simulation, so it would be interesting to add more of its functionalities. The python-copasi package main flaw is the lack of documentation, which hinders the task of building custom scripts for different simulations. On the other hand, this package receives constant updates, and has a user forum where questions are answered promptly.

### 6.2.3 Measurement of Results

The new version of the website has been online for a small amount of time, so it has not yet been possible to measure the impact of the new features. In a few months it will be important to check the new interactions with the repository, such as new data submissions, new registrations, and the website affluence, to properly measure the results.

## 6.3 Scientific Contribution

During the development of this work, one scientific article was written. Chapters 1, 2, 4 and 5 are based on the publication: "KiMoSys 2.0: an upgraded database for submitting, storing and accessing experimental data for kinetic modeling" (https://doi.org/10.1093/database/baaa093), adapted with changes to the format of the thesis.

The goal of this paper is to introduce the improvements and new features of KiMoSys' latest version to the scientific community. The acceptance of the paper authenticates these new capabilities, and confer KiMoSys a relevant space among repositories for experimental data and kinetic modeling.

# Bibliography

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (2016).

[2] ScientificData team. *Data Policies*. [Online; accessed 29-January-2020]. URL: https://www.nature.com/sdata/policies/data-policies#repo-criteria.

[3] PLOS team. *PLOS Criteria for Recommended Data Repositories*. [Online; accessed 29-January-2020]. URL: https://blogs.plos.org/everyone/2018/03/01/criteria-for-recommended-data-repositories.

[4] S.-A. Sansone, P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, and M. Thurston. "FAIRsharing as a community approach to standards, repositories and policies." In: *Nature biotechnology* 37.4 (2019), pp. 358–367.

[5] R. S. Costa, A. Veríssimo, and S. Vinga. "Ki MoSys: a web-based repository of experimental data for KInetic MOdels of biological SYStems." In: *BMC systems biology* 8.1 (2014), p. 85.

[6] Scientific Data Team. *Recommended Data Repositories | Scientific Data*. [Online; accessed 12-February-2020]. URL: https://www.nature.com/sdata/policies/repositories.

[7] C. Machado, R. S. Costa, M. Rocha, I. Rocha, B. Tidor, and E. C. Ferreira. "Model transformation of metabolic networks using a Petri net based framework." In: *CEUR Workshop Proceedings*. Vol. 827. 2010, pp. 101–115.

[8] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver, et al. "BioModels: ten-year anniversary." In: *Nucleic acids research* 43.D1 (2014), pp. D542–D548.

[9] P. McQuilton, S.-A. Sansone, H. Cousijn, M. Cannon, W. M. Chan, I. Carnevale, I. Cranston, S. C. Edmunds, N. Everitt, E. Ganley, and et al. *FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter*. 2020. DOI: 10.17605/OSF.IO/N9QJ7. URL: osf.io/n9qj7.

[10] S. G. Villas-Bôas, S. Mas, M. Åkesson, J. Smedsgaard, and J. Nielsen. "Mass spectrometry in metabolome analysis." In: *Mass spectrometry reviews* 24.5 (2005), pp. 613–646.

[11] M. R. Antoniewicz. "A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications." In: *Metabolic Engineering* (2020).

[12] M. Tyers and M. Mann. "From genomics to proteomics." In: *Nature* 422.6928 (2003), pp. 193–197.

[13] W. Timp and G. Timp. "Beyond mass spectrometry, the next step in proteomics." In: *Science Advances* 6.2 (2020), eaax8978.

[14] C. F. Taylor, D. Field, S.-A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P.-A. Binz, M. Bogue, T. Booth, et al. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." In: *Nature biotechnology* 26.8 (2008), pp. 889–896.

[15] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." In: *Nature genetics* 29.4 (2001), p. 365.

[16] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, et al. "Toward interoperable bioscience data." In: *Nature genetics* 44.2 (2012), p. 121.

[17] N. Le Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, et al. "The systems biology graphical notation." In: *Nature biotechnology* 27.8 (2009), p. 735.

[18] T. Czauderna, C. Klukas, and F. Schreiber. "Editing, validating and translating of SBGN maps." In: *Bioinformatics* 26.18 (2010), pp. 2340–2341.

[19] Dong, Xiaoxi and Luo, Weijun. *SBGNview: Pathway based omics data integration, visualization and analysis*. [Online; accessed 03-February-2020]. URL: https://www.bioconductor.org/packages/devel/bioc/vignettes/SBGNview/inst/doc/SBGNview.Vignette.html.

[20] Bergmann, Frank and König, Matthias. *SBGN_ML*. [Online; accessed 05-February-2020]. URL: https://github.com/sbgn/sbgn/wiki/SBGN_ML.

[21] S. Husen, Z. de Wilde, A. de Waard, and H. Cousijn. "Recommended versus certified repositories: mind the gap." In: *Available at SSRN 3020994* (2017).

[22] D. J. Clarke, L. Wang, A. Jones, M. L. Wojciechowicz, D. Torre, K. M. Jagodnik, S. L. Jenkins, P. McQuilton, Z. Flamholz, M. C. Silverstein, et al. "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources." In: *Cell systems* 9.5 (2019), pp. 417–421.

[23] C. S. Burns, A. Lana, and J. M. Budd. "Institutional repositories: Exploration of costs and value." In: *D-Lib Magazine* 19.1/2 (2013).

[24] L. Lyon. "Dealing with data: Roles, rights, responsibilities and relationships consultancy report." In: (2007).

[25] R. C. Amorim, J. A. Castro, J. R. Da Silva, and C. Ribeiro. "A comparison of research data management platforms: architecture, flexible metadata and interoperability." In: *Universal Access in the Information Society* 16.4 (2017), pp. 851–862.

[26] J. F. Sowa. "Ontology, metadata, and semiotics." In: *International conference on conceptual structures*. Springer. 2000, pp. 55–81.

[27] C. Willis, J. Greenberg, and H. White. "Analysis and synthesis of metadata goals for scientific data." In: *Journal of the American Society for Information Science and Technology* 63.8 (2012), pp. 1505–1520.

[28] E. Duval, W. Hodgins, S. Sutton, and S. L. Weibel. "Metadata principles and practicalities." In: *D-lib Magazine* 8.4 (2002), pp. 1082–9873.

[29] S.-T. Liaw, C. Pearce, H. Liyanage, G. S. Cheah-Liaw, and S. De Lusignan. "An integrated organisation-wide data quality management and information governance framework: theoretical underpinnings." In: *Journal of Innovation in Health Informatics* 21.4 (2014), pp. 199–206.

[30] F. Dreher, T. Kreitler, C. Hardt, A. Kamburov, R. Yildirimman, K. Schellander, H. Lehrach, B. M. Lange, and R. Herwig. "DIPSBC-data integration platform for systems biology collaborations." In: *BMC bioinformatics* 13.1 (2012), p. 85.

[31] H. E. Plesser. "Reproducibility vs. replicability: a brief history of a confused terminology." In: *Frontiers in neuroinformatics* 11 (2018), p. 76.

[32] U. Wittig, R. Kania, M. Golebiewski, M. Rey, L. Shi, L. Jong, E. Algaa, A. Weidemann, H. Sauer-Danzwith, S. Mir, et al. "SABIO-RK—database for biochemical reaction kinetics." In: *Nucleic acids research* 40.D1 (2011), pp. D790–D796.

[33] D. Machado, R. S. Costa, M. Rocha, I. Rocha, B. Tidor, and E. C. Ferreira. "A critical review on modelling formalisms and simulation tools in computational biosystems." In: *International Work-Conference on Artificial Neural Networks*. Springer. 2009, pp. 1063–1070.

[34] E. Stalidzans, A. Seiman, K. Peebo, V. Komasilovs, and A. Pentjuss. "Model-based metabolism design: constraints for kinetic and stoichiometric models." In: *Biochemical Society Transactions* 46.2 (2018), pp. 261–267.

[35] R. S. Costa, A. Hartmann, and S. Vinga. "Kinetic modeling of cell metabolism for microbial production." In: *Journal of biotechnology* 219 (2016), pp. 126–141.

[36] J. L. Reed and B. Ø. Palsson. "Thirteen years of building constraint-based in silico models of Escherichia coli." In: *Journal of bacteriology* 185.9 (2003), pp. 2692–2699.

[37] R. S. Costa, D. Machado, I. Rocha, and E. C. Ferreira. "Critical perspective on the consequences of the limited availability of kinetic data in metabolic dynamic modelling." In: *IET systems biology* 5.3 (2011), pp. 157–163.

[38] R. Alves, E. Vilaprinyo, B. Hernández-Bermejo, and A. Sorribas. "Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways." In: *Biotechnology and Genetic Engineering Reviews* 25.1 (2008), pp. 1–40.

[39] D. Visser and J. J. Heijnen. "Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics." In: *Metabolic engineering* 5.3 (2003), pp. 164–176.

[40] J. J. Heijnen. "Approximative kinetic formats used in metabolic network modeling." In: *Biotechnology and bioengineering* 91.5 (2005), pp. 534–545.

[41] E. J. Clement, B. J. Wysocki, G. A. Soliman, T. A. Wysocki, and P. H. Davis. "Dynamic Modeling and Stochastic Simulation of Metabolic Networks." In: *BioRxiv* (2018), p. 336677.

[42] U. Kaplan, M. Türkay, L Biegler, and B. Karasözen. "Modeling and simulation of metabolic networks for estimation of biomass accumulation parameters." In: *Discrete applied mathematics* 157.10 (2009), pp. 2483–2493.

[43] E. Vasilakou, D. Machado, A. Theorell, I. Rocha, K. Nöh, M. Oldiges, and S. A. Wahl. "Current state and challenges for dynamic metabolic modeling." In: *Current opinion in microbiology* 33 (2016), pp. 97–104.

[44] P. A. Saa and L. K. Nielsen. "Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks." In: *Biotechnology advances* 35.8 (2017), pp. 981–1003.

[45] SABIO-RK team. *Statistics - Sabio-RK*. [Online; accessed 13-November-2020]. URL: http://sabio.h-its.org/layouts/content/statistic.gsp.

[46] I. Schomburg, A. Chang, and D. Schomburg. "BRENDA, enzyme data and metabolic information." In: *Nucleic acids research* 30.1 (2002), pp. 47–49.

[47] Z. Zhang, T. Shen, B. Rui, W. Zhou, X. Zhou, C. Shang, C. Xin, X. Liu, G. Li, J. Jiang, C. Li, R. Li, M. Han, S. You, G. Yu, Y. Yi, H. Wen, Z. Liu, and X. Xie. "CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13C-fluxomics." In: *Nucleic Acids Research* 43.D1 (Nov. 2014), pp. D549–D557. ISSN: 0305-1048. DOI: 10.1093/nar/gku1137. eprint: https://academic.oup.com/nar/article-pdf/43/D1/D549/17436700/gku1137.pdf. URL: https://doi.org/10.1093/nar/gku1137.

[48] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. "MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data." In: *Nucleic Acids Research* 41.D1 (Oct. 2012), pp. D781–D786. ISSN: 0305-1048. DOI: 10.1093/nar/gks1004. eprint: https://academic.oup.com/nar/article-pdf/41/D1/D781/18785371/gks1004.pdf. URL: https://doi.org/10.1093/nar/gks1004.

[49] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, and R. Aebersold. "The peptideatlas project." In: *Nucleic acids research* 34.suppl_1 (2006), pp. D655–D658.

[50] P. Jones, R. G. Côté, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob, and R. Apweiler. "PRIDE: a public repository of protein and peptide identifications for the proteomics community." In: *Nucleic Acids Research* 34.suppl_1 (Jan. 2006), pp. D659–D663. ISSN: 0305-1048. DOI: 10.1093/nar/gkj138. eprint: https://academic.oup.com/nar/article-pdf/34/suppl_1/D659/3926173/gkj138.pdf. URL: https://doi.org/10.1093/nar/gkj138.

[51] E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D. S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R. L. Moritz, J. J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, and J. A. Vizcaíno. "The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition." In: *Nucleic Acids Research* 45.D1 (Oct. 2016), pp. D1100–D1106. ISSN: 0305-1048. DOI: 10.1093/nar/gkw936. eprint: https://academic.oup.com/nar/article-pdf/45/D1/D1100/8847220/gkw936.pdf. URL: https://doi.org/10.1093/nar/gkw936.

[52] BioModels Team. *Path2Models project page*. [Online; accessed 13-February-2020]. URL: https://www.ebi.ac.uk/biomodels/path2models.

[53] B. G. Olivier and J. L. Snoep. "Web-based kinetic modelling using JWS Online." In: *Bioinformatics* 20.13 (2004), pp. 2143–2144.

[54] T. Yu, C. M. Lloyd, D. P. Nickerson, M. T. Cooling, A. K. Miller, A. Garny, J. R. Terkildsen, J. Lawson, R. D. Britten, P. J. Hunter, and P. M. F. Nielsen. "The Physiome Model Repository 2." In: *Bioinformatics* 27.5 (Jan. 2011), pp. 743–744. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq723. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/5/743/5518220/btq723.pdf. URL: https://doi.org/10.1093/bioinformatics/btq723.

[55] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure. "myExperiment: a repository and social network for the sharing of bioinformatics workflows." In: *Nucleic Acids Research* 38.suppl_2 (May 2010), W677–W682. ISSN: 0305-1048. DOI:

10.1093/nar/gkq429. eprint: https://academic.oup.com/nar/article-pdf/38/suppl_2/W677/16773278/gkq429.pdf. URL: https://doi.org/10.1093/nar/gkq429.

[56] J. Singh et al. "FigShare." In: *Journal of Pharmacology and Pharmacotherapeutics* 2.2 (2011), p. 138.

[57] H. White, S. Carrier, A. Thompson, J. Greenberg, and R. Scherle. "The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment." In: *Dublin Core Conference*. 2008, pp. 157–162.

[58] C. Durinx, J. McEntyre, R. Appel, R. Apweiler, M. Barlow, N. Blomberg, C. Cook, E. Gasteiger, J.-H. Kim, R. Lopez, et al. "Identifying ELIXIR core data resources." In: *F1000Research* 5 (2016).

[59] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. "COPASI—a complex pathway simulator." In: *Bioinformatics* 22.24 (2006), pp. 3067–3074.

[60] D. T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions." In: *Journal of computational physics* 22.4 (1976), pp. 403–434.

[61] COPASI team. *COPASI: Support/Features*. [Online; accessed 30-January-2020]. URL: http://copasi.org/Support/Features/.

[62] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano. "CellDesigner 3.5: a versatile modeling tool for biochemical networks." In: *Proceedings of the IEEE* 96.8 (2008), pp. 1254–1265.

[63] E. T. Somogyi, J.-M. Bouteiller, J. A. Glazier, M. König, J. K. Medley, M. H. Swat, and H. M. Sauro. "libRoadRunner: a high performance SBML simulation and analysis library." In: *Bioinformatics* 31.20 (2015), pp. 3315–3321.

[64] H. Sauro. "33 JARNAC: a system for interactive metabolic analysis." In: (2000).

[65] SQLite Team. *What Is SQLite?* [Online; accessed 10-August-2020]. URL: https://www.sqlite.org/index.html.

[66] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck. "ChEBI in 2016: Improved services and an expanding collection of metabolites." In: *Nucleic acids research* 44.D1 (2016), pp. D1214–D1219.

[67] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes." In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[68] U. Consortium. "UniProt: a worldwide hub of protein knowledge." In: *Nucleic acids research* 47.D1 (2019), pp. D506–D515.

[69] Ruby on Rails Team. *Active Storage Overview*. [Online; accessed 11-August-2020]. URL: https://guides.rubyonrails.org/active_storage_overview.html.

[70] Burns, Mike. *Deprecating Paperclip*. [Online; accessed 11-August-2020]. URL: https://thoughtbot.com/blog/closing-the-trombone.

[71] Oliver, Chris. *How to Migrate from Paperclip to Rails ActiveStorage (Example) | GoRails - GoRails*. 2018. URL: https://gorails.com/episodes/migrate-from-paperclip-to-rails-active-storage.

[72] Burns, Mike. *How to Migrate from Paperclip to Rails ActiveStorage (Example) | GoRails - GoRails*. 2018. URL: https://github.com/thoughtbot/paperclip/blob/master/MIGRATING.md.

[73] Negreiros, Leonardo. *Migrating from Paperclip to ActiveStorage*. 2018. URL: https://blog.codeminer42.com/migrating-from-paperclip-to-activestorage-b37ef187fb17/.

[74] Custer, Kevin. *Migrate a Rails Project from Paperclip to ActiveStorage*. 2018. URL: https://www.kevin-custer.com/blog/migrate-a-rails-project-from-paperclip-to-active-storage/.

[75] Confreaks. *RailsConf 2019 - How to migrate to Active Storage without losing your mind by Colleen Schnettler*. 2019. URL: https://www.youtube.com/watch?v=tZ_WNUytO9o.

[76] DataTables team. *DataTables | Table plug-in for jQuery*. [Online; accessed 11-August-2020]. URL: https://datatables.net.

[77] B. Russell and A. N. Whitehead. *Principia mathematica to* 56*. Vol. 2. Cambridge University Press Cambridge, UK, 1997.

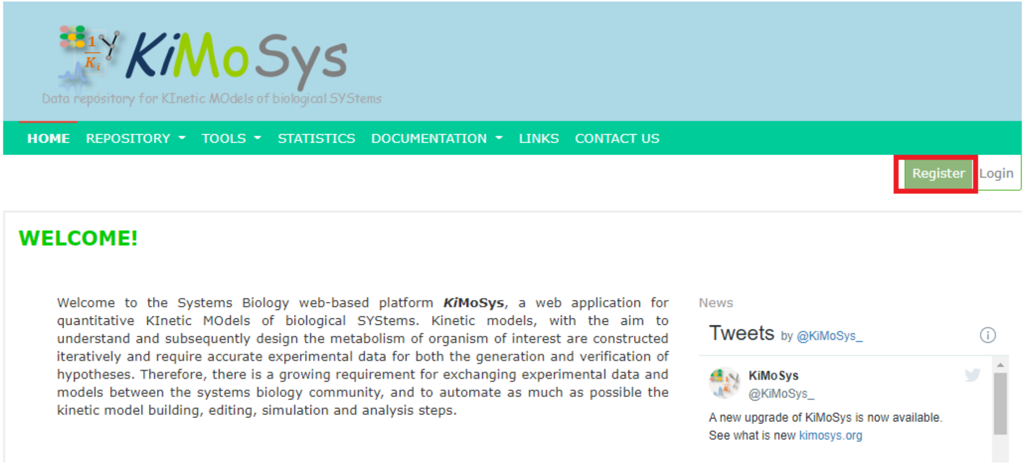
[78] S.-M. Fendt and U. Sauer. “Transcriptional regulation of respiration in yeast metabolizing differently repressive carbon substrates.” In: *BMC systems biology* 4.1 (2010), p. 12.

# Appendix A

# Example of Submitting a Dataset

In this appendix it is described step by step how to submit a new dataset into the repository. For this purpose, the experimental data from Fendt et al. [78], will be used.

Before starting the actual submitting process, it is necessary to create an account first. From the KiMoSys home page (`https://kimosys.org`), the user should click in the "Register" button.



WELCOME!

Welcome to the Systems Biology web-based platform ***KiMoSys***, a web application for quantitative KInetic MOdels of biological SYStems. Kinetic models, with the aim to understand and subsequently design the metabolism of organism of interest are constructed iteratively and require accurate experimental data for both the generation and verification of hypotheses. Therefore, there is a growing requirement for exchanging experimental data and models between the systems biology community, and to automate as much as possible the kinetic model building, editing, simulation and analysis steps.

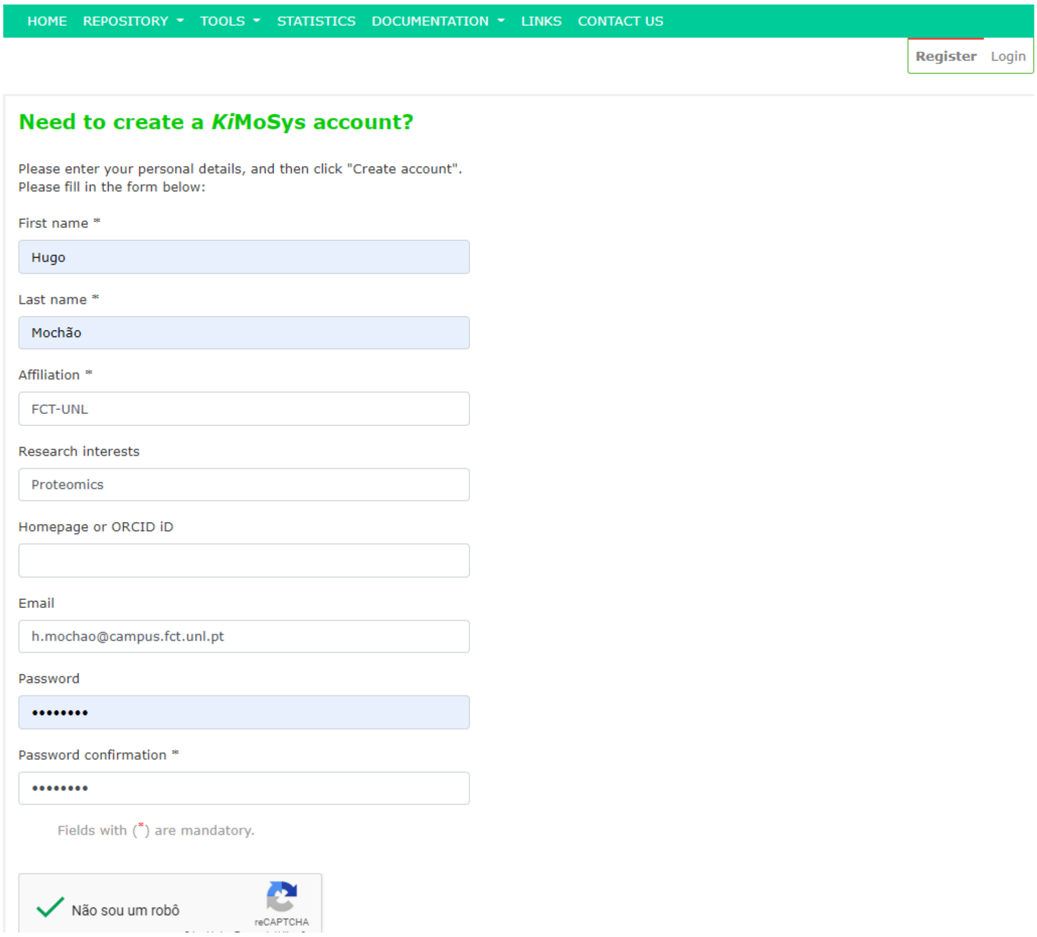
Tweets by @KiMoSys_

KiMoSys
@KiMoSys_

A new upgrade of KiMoSys is now available. See what is new kimosys.org

Figure A.1: Screenshot of the KiMoSys' main page.

Then, the user has to complete the registration form with his data (`https://kimosys.org/users/sign_up`).



Need to create a *Ki*MoSys account?

Please enter your personal details, and then click "Create account".
Please fill in the form below:

First name *

Hugo

Last name *

Mochão

Affiliation *

FCT-UNL

Research interests

Proteomics

Homepage or ORCID iD

Email

h.mochao@campus.fct.unl.pt

Password

••••••••

Password confirmation *

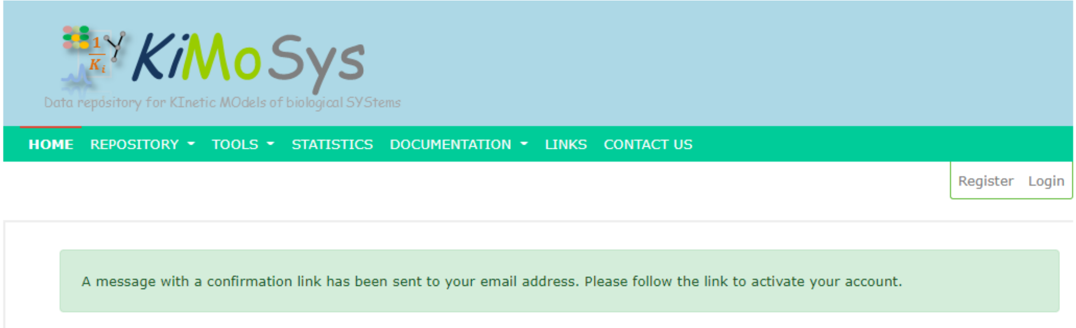
••••••••

Fields with (*) are mandatory.

✓ Não sou um robô

Figure A.2: View of the registration form.

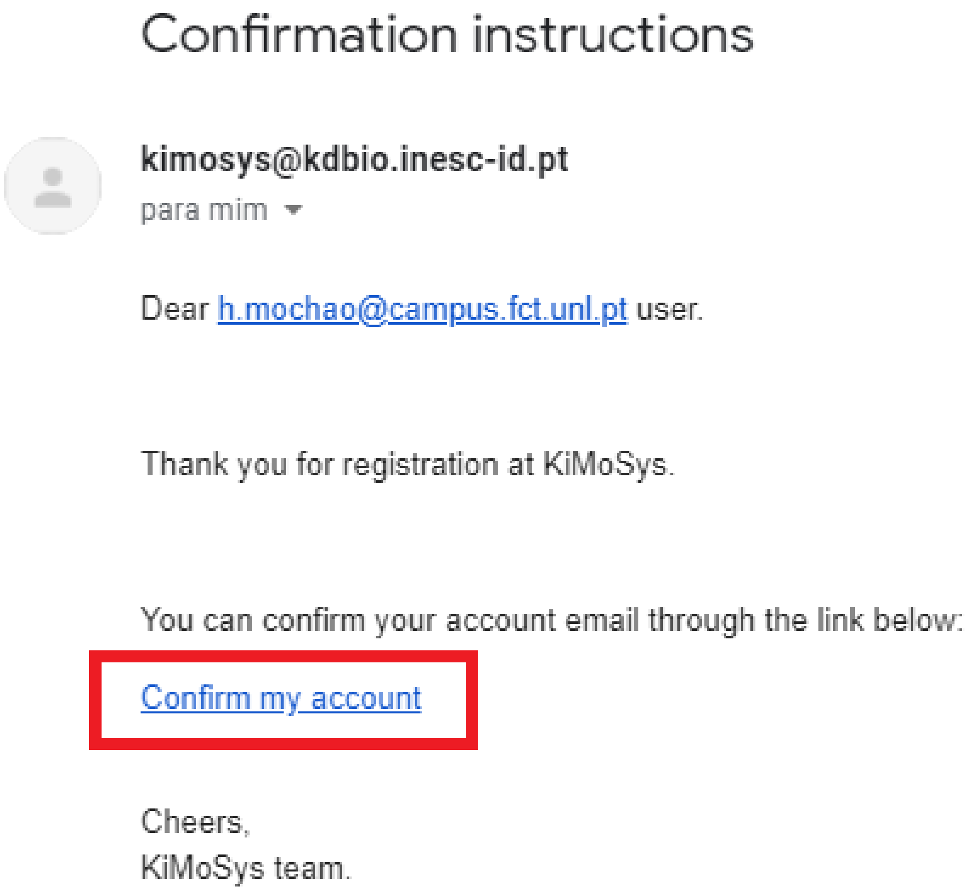
After the submission, the user is redirected to the main KiMoSys page (`https://kimosys.org`), and a confirmation email is sent.



A message with a confirmation link has been sent to your email address. Please follow the link to activate your account.

Figure A.3: Snapshot of the redirected view after the successful registration, were a message informs the user that a confirmation link has been sent via email.

Then, it is provided a confirmation link in the email, that redirects the user to the login page (https://kimosys.org/users/sign_in).



Confirmation instructions

kimosys@kdbio.inesc-id.pt

para mim

Dear h.mochao@campus.fct.unl.pt user.

Thank you for registration at KiMoSys.

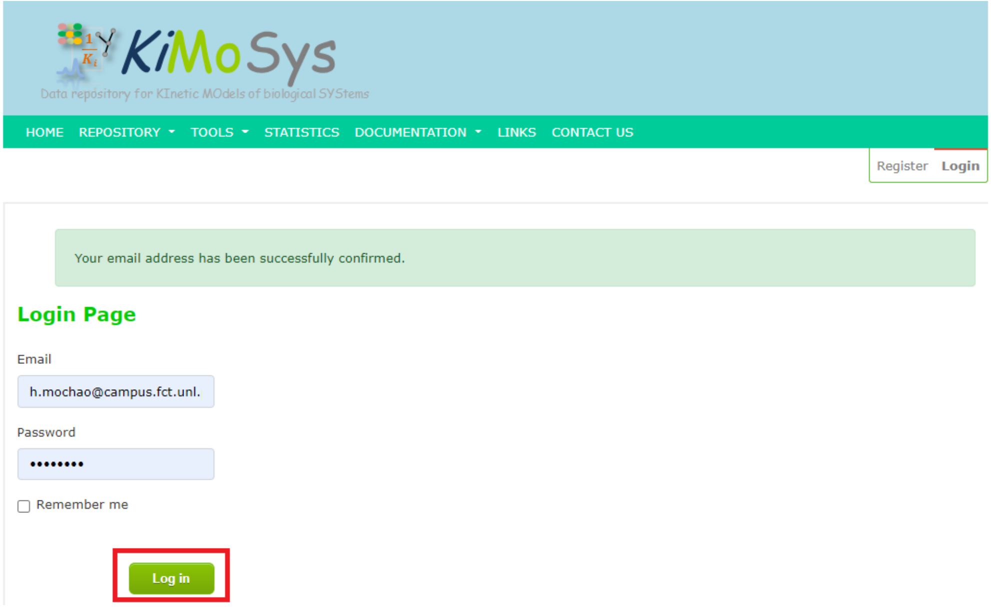
You can confirm your account email through the link below:

Confirm my account

Cheers,
KiMoSys team.

Figure A.4: Screenshot of the email that contains the confirmation instructions.

After clicking on the link, the user is redirected to the login page, where he can use the credentials previously submitted in the registration.


KiMoSys
Data repository for KInetic MOdels of biological SYStems
HOME REPOSITORY TOOLS STATISTICS DOCUMENTATION LINKS CONTACT US
Register Login
Log in

Your email address has been successfully confirmed.

Login Page

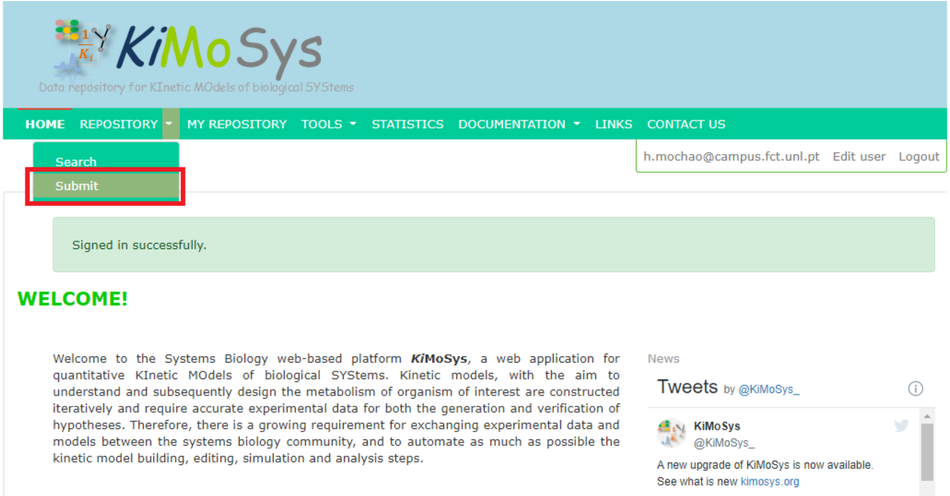
Email: h.mochao@campus.fct.unl.

Password: ••••••••

□ Remember me

Figure A.5: Snapshot of the login page.

Now logged in and in the KiMoSys home page again (`https://kimosys.org`), the user should click in the arrow next to the "REPOSITORY" tab, and click on "Submit".


KiMoSys
Data repository for KInetic MOdels of biological SYStems
HOME REPOSITORY MY REPOSITORY TOOLS STATISTICS DOCUMENTATION LINKS CONTACT US
Search
Submit
h.mochao@campus.fct.unl.pt Edit user Logout
News
Tweets by @KiMoSys_

Signed in successfully.

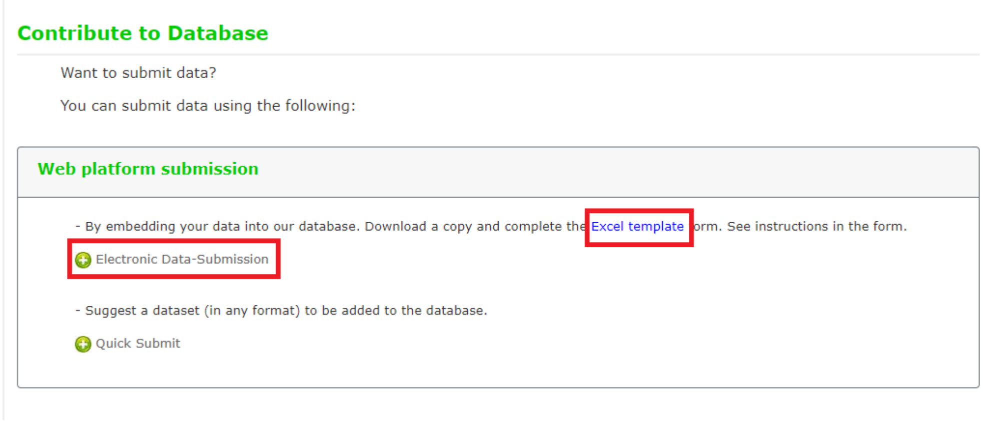
WELCOME!

Welcome to the Systems Biology web-based platform ***KiMoSys***, a web application for quantitative KInetic MOdels of biological SYStems. Kinetic models, with the aim to understand and subsequently design the metabolism of organism of interest are constructed iteratively and require accurate experimental data for both the generation and verification of hypotheses. Therefore, there is a growing requirement for exchanging experimental data and models between the systems biology community, and to automate as much as possible the kinetic model building, editing, simulation and analysis steps.

KiMoSys @KiMoSys_
A new upgrade of KiMoSys is now available. See what is new kimosys.org

Figure A.6: Screenshot of the KiMoSys home page, with the options to navigate to the data submission area.

The user lands in the web platform submission interface view (https://kimosys.org/repository#contribute). Before going to the actual "Electronic Data-Submission" form, it is necessary to download the "Excel templates" of the repository.



Contribute to Database

Want to submit data?

You can submit data using the following:

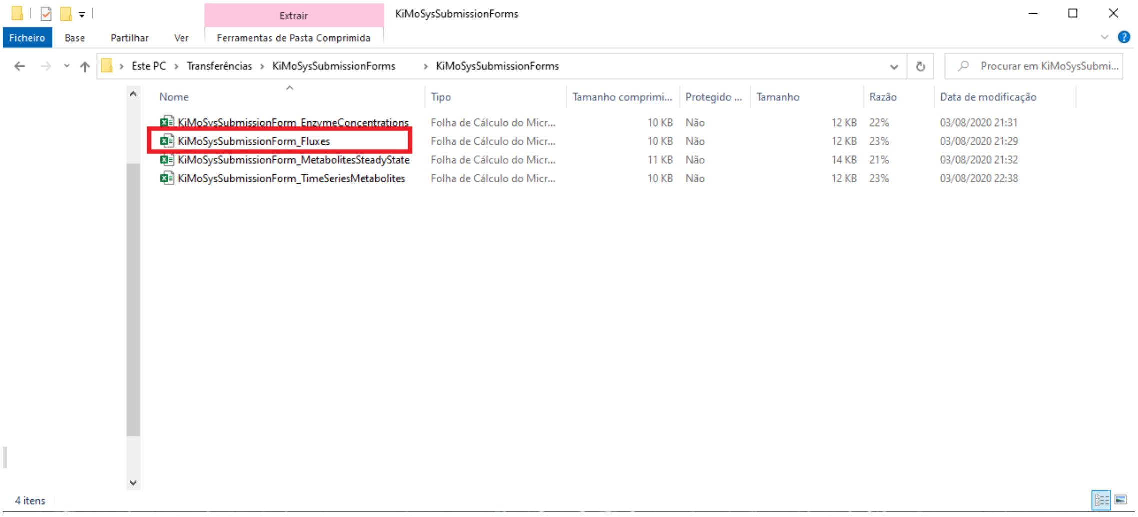
Web platform submission

- By embedding your data into our database. Download a copy and complete the Excel template form. See instructions in the form.

- Suggest a dataset (in any format) to be added to the database.

Figure A.7: View of the Web platform submission interface.

There are four data types in KiMoSys, hence the four templates. In this case, since it is a flux dataset, the user extracts and opens the "KiMoSysSubmissionForm_Fluxes" file.

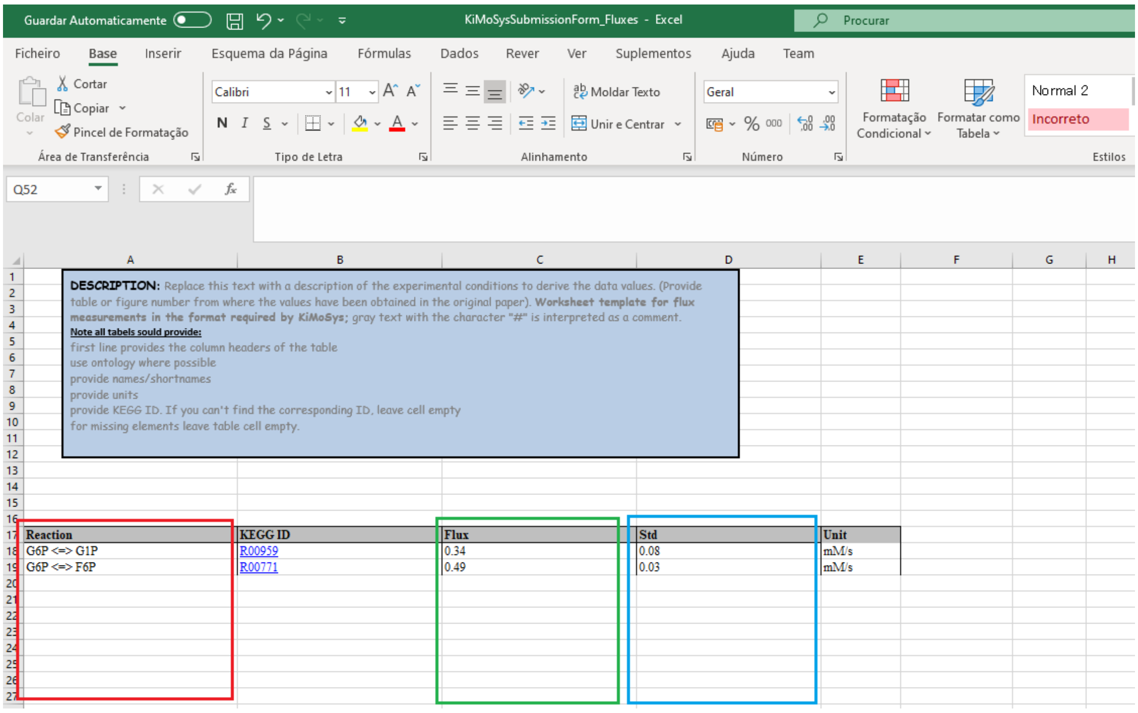


Extrair
KiMoSysSubmissionForms
Ficheiro
Base
Partilhar
Ver
Ferramentas de Pasta Comprimida
Este PC › Transferências › KiMoSysSubmissionForms › KiMoSysSubmissionForms
Procurar em KiMoSysSubmi...
4 itens

| Nome | Tipo | Tamanho comprimi... | Protegido ... | Tamanho | Razão | Data de modificação |
|---|---|---|---|---|---|---|
| KiMoSysSubmissionForm_EnzymeConcentrations | Folha de Cálculo do Micr... | 10 KB | Não | 12 KB | 22% | 03/08/2020 21:31 |
| KiMoSysSubmissionForm_Fluxes | Folha de Cálculo do Micr... | 10 KB | Não | 12 KB | 23% | 03/08/2020 21:29 |
| KiMoSysSubmissionForm_MetabolitesSteadyState | Folha de Cálculo do Micr... | 11 KB | Não | 14 KB | 21% | 03/08/2020 21:32 |
| KiMoSysSubmissionForm_TimeSeriesMetabolites | Folha de Cálculo do Micr... | 10 KB | Não | 12 KB | 23% | 03/08/2020 22:38 |

Figure A.8: Snapshot of the downloaded file containing all the Excel templates.

This file has some instructions of how it should be filled, and two rows serving as an example.



**DESCRIPTION:** Replace this text with a description of the experimental conditions to derive the data values. (Provide table or figure number from where the values have been obtained in the original paper). **Worksheet template for flux measurements in the format required by KiMoSys;** gray text with the character "#" is interpreted as a comment.

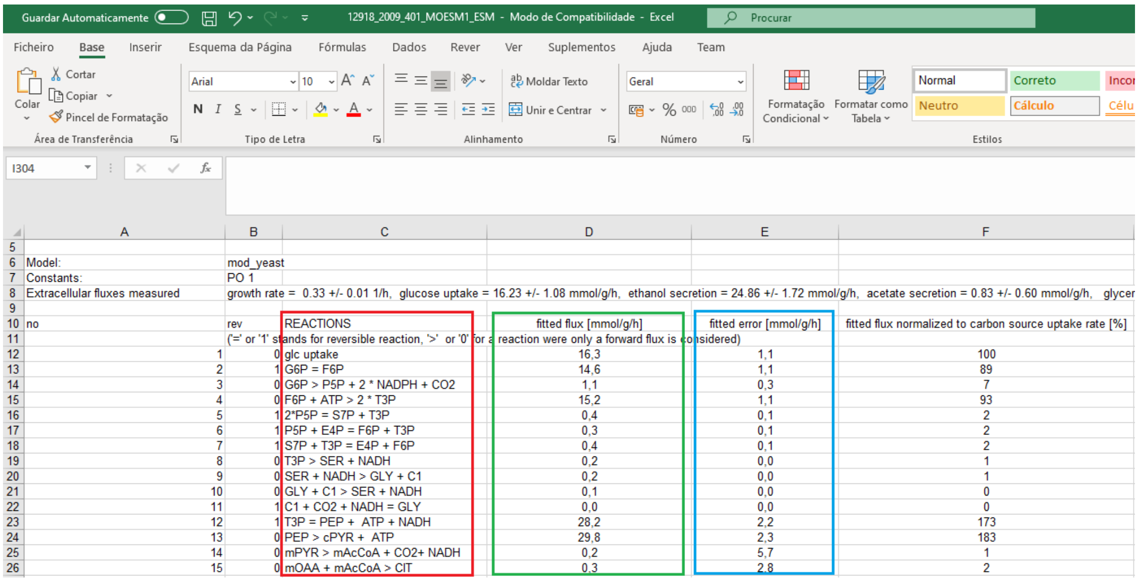
**Note all tabels sould provide:**

first line provides the column headers of the table
use ontology where possible
provide names/shortnames
provide units
provide KEGG ID. If you can't find the corresponding ID, leave cell empty
for missing elements leave table cell empty.

| Reaction | KEGG ID | Flux | Std | Unit |
|---|---|---|---|---|
| G6P <=> G1P | R00959 | 0.34 | 0.08 | mM/s |
| G6P <=> F6P | R00771 | 0.49 | 0.03 | mM/s |

Figure A.9: The Flux Measurements template.

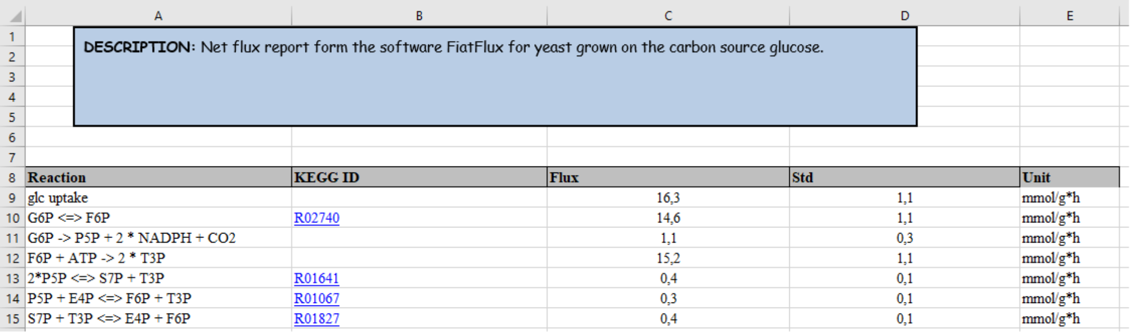
Now, it is necessary to adapt the experimental data file to fit the KiMoSys template. The columns "fitted flux" and "fitted error" correspond to KiMoSys' "Flux" and "Std", respectively, while the "REACTION" is the same. So, these values are copied. However, the KEGG IDs are not provided, so they need to be inserted manually. At https://www.genome.jp/kegg/reaction/ it is possible to search for the reactions' KEGG IDs.



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 5 | | | | | | |
| 6 | Model: | mod_yeast | | | | |
| 7 | Constants: | PO 1 | | | | |
| 8 | Extracellular fluxes measured | growth rate = 0.33 +/- 0.01 1/h, glucose uptake = 16.23 +/- 1.08 mmol/g/h, ethanol secretion = 24.86 +/- 1.72 mmol/g/h, acetate secretion = 0.83 +/- 0.60 mmol/g/h, glycer | | | | |
| 9 | | | | | | |
| 10 | no | rev | REACTIONS | fitted flux [mmol/g/h] | fitted error [mmol/g/h] | fitted flux normalized to carbon source uptake rate [%] |
| 11 | | ('=' or '1' stands for reversible reaction, '>' or '0' for a reaction were only a forward flux is considered) | | | | |
| 12 | 1 | 0 | glc uptake | 16,3 | 1,1 | 100 |
| 13 | 2 | 1 | G6P = F6P | 14,6 | 1,1 | 89 |
| 14 | 3 | 0 | G6P > P5P + 2 * NADPH + CO2 | 1,1 | 0,3 | 7 |
| 15 | 4 | 0 | F6P + ATP > 2 * T3P | 15,2 | 1,1 | 93 |
| 16 | 5 | 1 | 2*P5P = S7P + T3P | 0,4 | 0,1 | 2 |
| 17 | 6 | 1 | P5P + E4P = F6P + T3P | 0,3 | 0,1 | 2 |
| 18 | 7 | 1 | S7P + T3P = E4P + F6P | 0,4 | 0,1 | 2 |
| 19 | 8 | 0 | T3P > SER + NADH | 0,2 | 0,0 | 1 |
| 20 | 9 | 0 | SER + NADH > GLY + C1 | 0,2 | 0,0 | 1 |
| 21 | 10 | 0 | GLY + C1 > SER + NADH | 0,1 | 0,0 | 0 |
| 22 | 11 | 1 | C1 + CO2 + NADH = GLY | 0,0 | 0,0 | 0 |
| 23 | 12 | 1 | T3P = PEP + ATP + NADH | 28,2 | 2,2 | 173 |
| 24 | 13 | 0 | PEP > cPYR + ATP | 29,8 | 2,3 | 183 |
| 25 | 14 | 0 | mPYR > mAcCoA + CO2+ NADH | 0,2 | 5,7 | 1 |
| 26 | 15 | 0 | mOAA + mAcCoA > CIT | 0,3 | 2,8 | 2 |

Figure A.10: The file with the results produced by Fendt et al..

After the introduction of the KEGG IDs and the fluxes values, the first table should look like fig A.11. This dataset has 4 tables/tabs, so the process is repeated until all the tables are correctly filled.

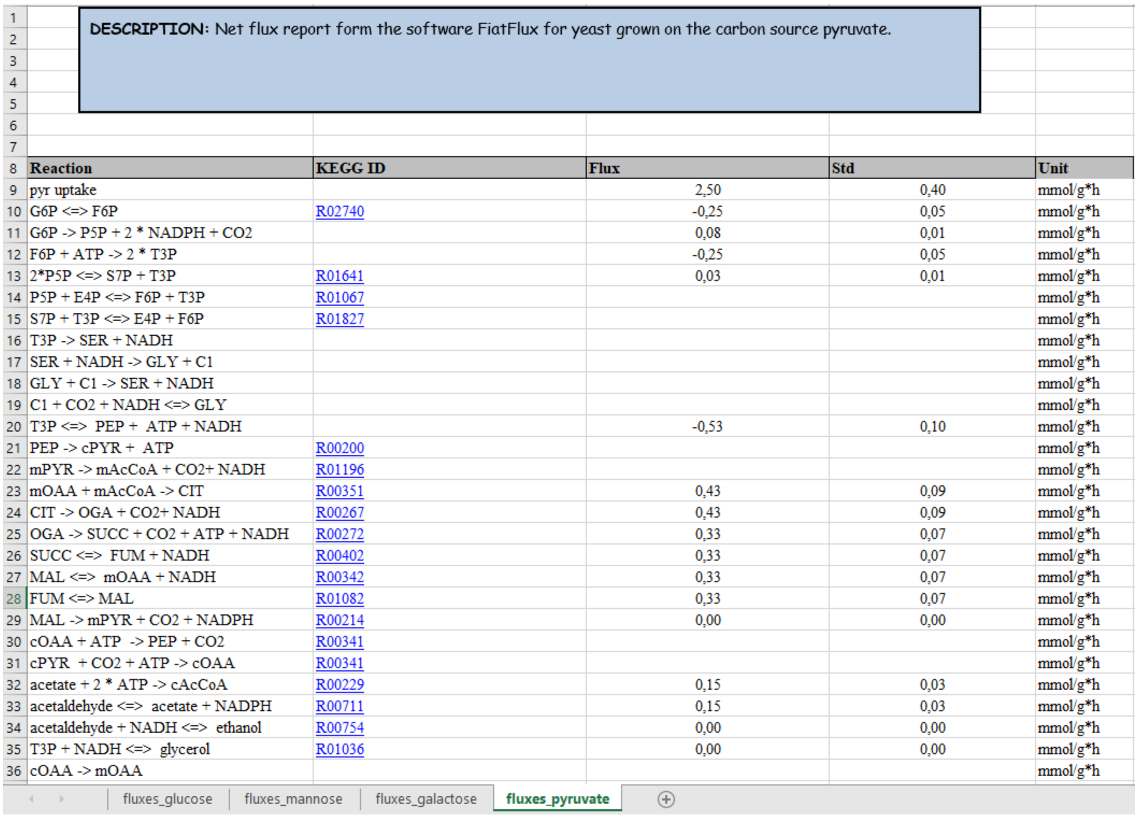


**DESCRIPTION:** Net flux report form the software FiatFlux for yeast grown on the carbon source glucose.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 8 | Reaction | KEGG ID | Flux | Std | Unit |
| 9 | glc uptake | | 16,3 | 1,1 | mmol/g*h |
| 10 | G6P <=> F6P | R02740 | 14,6 | 1,1 | mmol/g*h |
| 11 | G6P -> P5P + 2 * NADPH + CO2 | | 1,1 | 0,3 | mmol/g*h |
| 12 | F6P + ATP -> 2 * T3P | | 15,2 | 1,1 | mmol/g*h |
| 13 | 2*P5P <=> S7P + T3P | R01641 | 0,4 | 0,1 | mmol/g*h |
| 14 | P5P + E4P <=> F6P + T3P | R01067 | 0,3 | 0,1 | mmol/g*h |
| 15 | S7P + T3P <=> E4P + F6P | R01827 | 0,4 | 0,1 | mmol/g*h |

Figure A.11: Snapshot of the KiMoSys template with the first table completed.

After finishing completing the last table, the user clicks in the fig A.7 link "Electronic Data Submission".

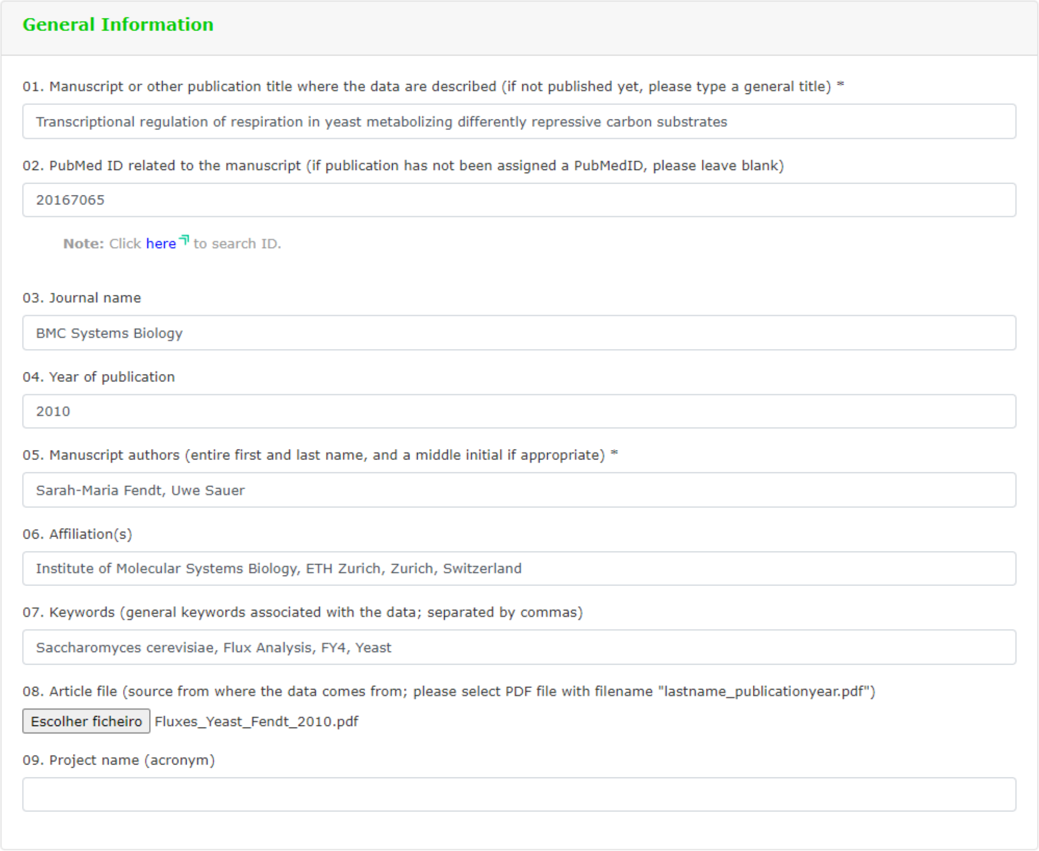


**DESCRIPTION:** Net flux report form the software FiatFlux for yeast grown on the carbon source pyruvate.

| Reaction | KEGG ID | Flux | Std | Unit |
|---|---|---|---|---|
| pyr uptake | | 2,50 | 0,40 | mmol/g*h |
| G6P <=> F6P | R02740 | -0,25 | 0,05 | mmol/g*h |
| G6P -> P5P + 2 * NADPH + CO2 | | 0,08 | 0,01 | mmol/g*h |
| F6P + ATP -> 2 * T3P | | -0,25 | 0,05 | mmol/g*h |
| 2*P5P <=> S7P + T3P | R01641 | 0,03 | 0,01 | mmol/g*h |
| P5P + E4P <=> F6P + T3P | R01067 | | | mmol/g*h |
| S7P + T3P <=> E4P + F6P | R01827 | | | mmol/g*h |
| T3P -> SER + NADH | | | | mmol/g*h |
| SER + NADH -> GLY + C1 | | | | mmol/g*h |
| GLY + C1 -> SER + NADH | | | | mmol/g*h |
| C1 + CO2 + NADH <=> GLY | | | | mmol/g*h |
| T3P <=> PEP + ATP + NADH | | -0,53 | 0,10 | mmol/g*h |
| PEP -> cPYR + ATP | R00200 | | | mmol/g*h |
| mPYR -> mAcCoA + CO2+ NADH | R01196 | | | mmol/g*h |
| mOAA + mAcCoA -> CIT | R00351 | 0,43 | 0,09 | mmol/g*h |
| CIT -> OGA + CO2+ NADH | R00267 | 0,43 | 0,09 | mmol/g*h |
| OGA -> SUCC + CO2 + ATP + NADH | R00272 | 0,33 | 0,07 | mmol/g*h |
| SUCC <=> FUM + NADH | R00402 | 0,33 | 0,07 | mmol/g*h |
| MAL <=> mOAA + NADH | R00342 | 0,33 | 0,07 | mmol/g*h |
| FUM <=> MAL | R01082 | 0,33 | 0,07 | mmol/g*h |
| MAL -> mPYR + CO2 + NADPH | R00214 | 0,00 | 0,00 | mmol/g*h |
| cOAA + ATP -> PEP + CO2 | R00341 | | | mmol/g*h |
| cPYR + CO2 + ATP -> cOAA | R00341 | | | mmol/g*h |
| acetate + 2 * ATP -> cAcCoA | R00229 | 0,15 | 0,03 | mmol/g*h |
| acetaldehyde <=> acetate + NADPH | R00711 | 0,15 | 0,03 | mmol/g*h |
| acetaldehyde + NADH <=> ethanol | R00754 | 0,00 | 0,00 | mmol/g*h |
| T3P + NADH <=> glycerol | R01036 | 0,00 | 0,00 | mmol/g*h |
| cOAA -> mOAA | | | | mmol/g*h |

Figure A.12: Screenshot of the KiMoSys template now completely filled.

Now in the web submission form (https://kimosys.org/repository/new). Here is the first part of the submission, the "General Information". The form fields are filled according to the original paper information. Some of the fields are not specified or not applicable, so they are left blank.



General Information

01. Manuscript or other publication title where the data are described (if not published yet, please type a general title) *

Transcriptional regulation of respiration in yeast metabolizing differently repressive carbon substrates

02. PubMed ID related to the manuscript (if publication has not been assigned a PubMedID, please leave blank)

20167065

Note: Click here to search ID.

03. Journal name

BMC Systems Biology

04. Year of publication

2010

05. Manuscript authors (entire first and last name, and a middle initial if appropriate) *

Sarah-Maria Fendt, Uwe Sauer

06. Affiliation(s)

Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

07. Keywords (general keywords associated with the data; separated by commas)

Saccharomyces cerevisiae, Flux Analysis, FY4, Yeast

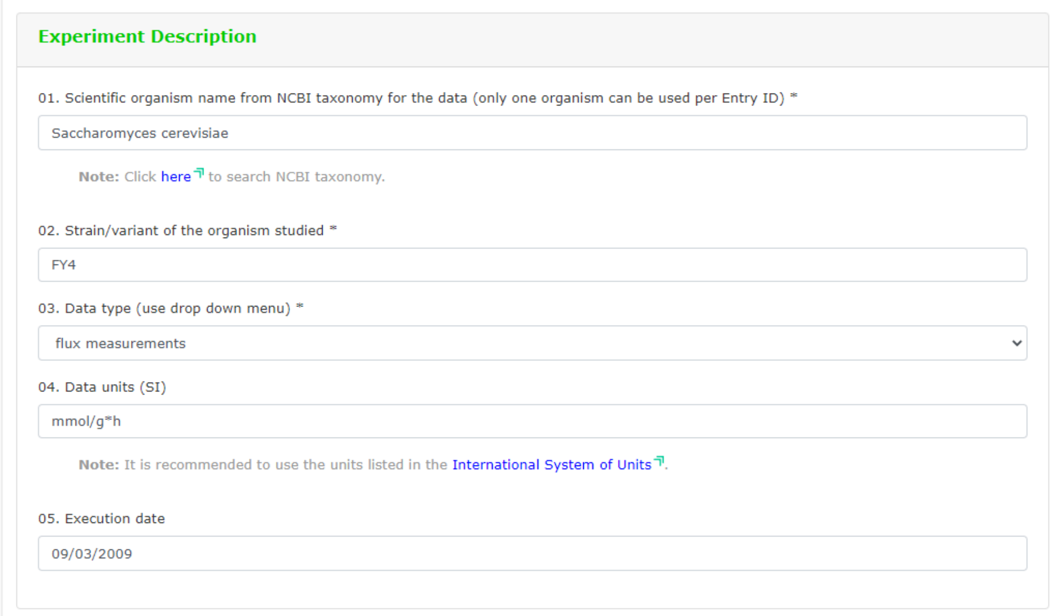
08. Article file (source from where the data comes from; please select PDF file with filename "lastname_publicationyear.pdf")

Escolher ficheiro Fluxes_Yeast_Fendt_2010.pdf

09. Project name (acronym)

Figure A.13: View of the first part of the web submission form.

Then the user advances to the "Experiment Description" part and completes it.



Experiment Description

01. Scientific organism name from NCBI taxonomy for the data (only one organism can be used per Entry ID) *

Saccharomyces cerevisiae

**Note:** Click here to search NCBI taxonomy.

02. Strain/variant of the organism studied *

FY4

03. Data type (use drop down menu) *

flux measurements

04. Data units (SI)

mmol/g*h

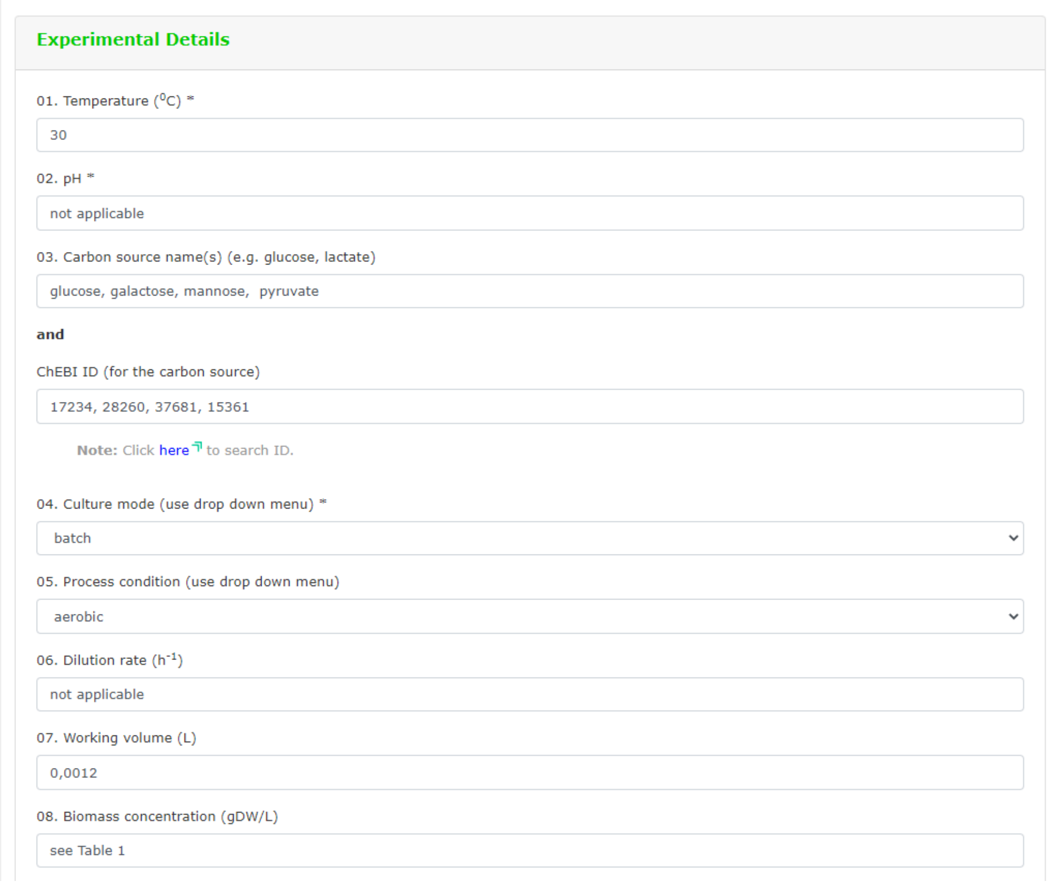
**Note:** It is recommended to use the units listed in the International System of Units.

05. Execution date

09/03/2009

Figure A.14: Snapshot of the second part of the form.

Finally, the "Experimental Details" are filled.



Experimental Details

01. Temperature (°C) *

30

02. pH *

not applicable

03. Carbon source name(s) (e.g. glucose, lactate)

glucose, galactose, mannose, pyruvate

**and**

ChEBI ID (for the carbon source)

17234, 28260, 37681, 15361

**Note:** Click here to search ID.

04. Culture mode (use drop down menu) *

batch

05. Process condition (use drop down menu)

aerobic

06. Dilution rate ($h^{-1}$)

not applicable

07. Working volume (L)

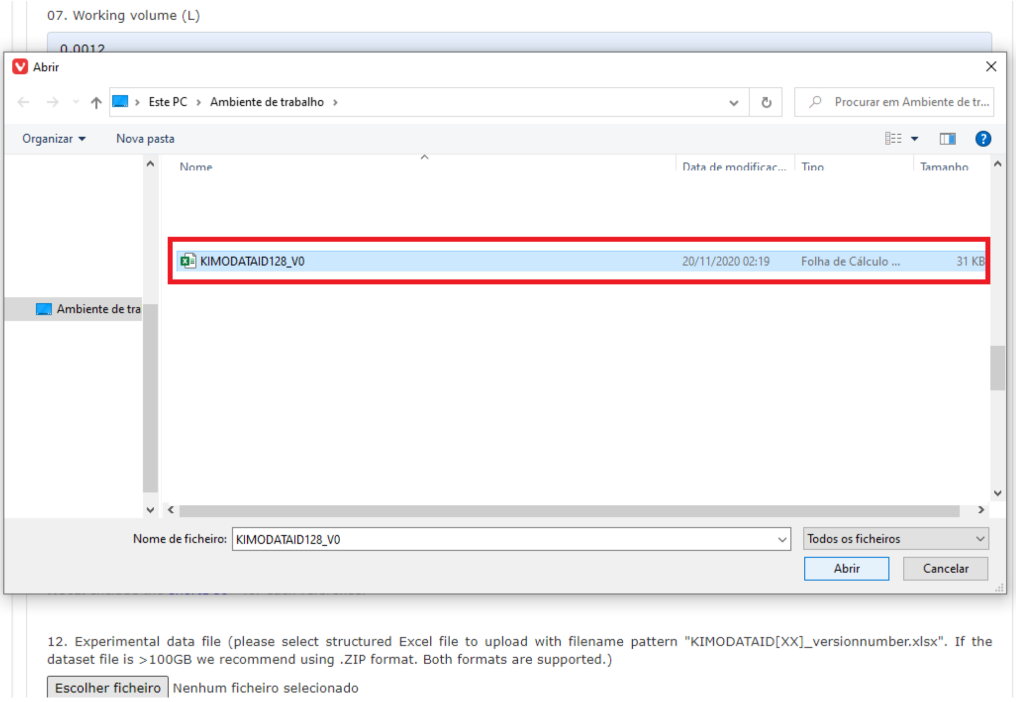
0,0012

08. Biomass concentration (gDW/L)

see Table 1

Figure A.15: Screenshot of the last fields to be filled, corresponding to the "Experimental Details".

In this submission part, the dataset file is required. The user chooses the template file previously filled.



07. Working volume (L)

0.0012

Abrir
Este PC › Ambiente de trabalho ›
Procurar em Ambiente de tr...
Organizar ▾
Nova pasta
Ambiente de tra

| Nome | Data de modificac... | Tipo | Tamanho |
|---|---|---|---|
| KIMODATAID128_V0 | 20/11/2020 02:19 | Folha de Cálculo ... | 31 KB |

Nome de ficheiro: KIMODATAID128_V0

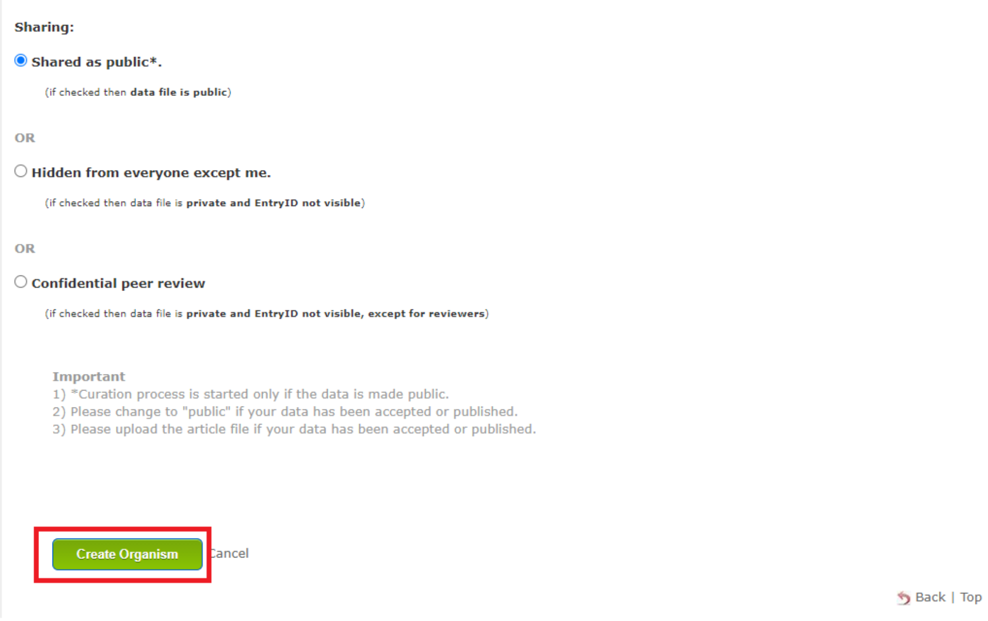
Todos os ficheiros

12. Experimental data file (please select structured Excel file to upload with filename pattern "KIMODATAID[XX]_versionnumber.xlsx". If the dataset file is >100GB we recommend using .ZIP format. Both formats are supported.)

Escolher ficheiro Nenhum ficheiro selecionado

Figure A.16: View of the user choosing the dataset file to submit.

The final step is to submit. It is chosen the option "Shared as public" , which means that the new Data Entry will be publicly available.



**Sharing:**

● **Shared as public*.**

(if checked then **data file is public**)

OR

○ **Hidden from everyone except me.**

(if checked then data file is **private and EntryID not visible**)

OR

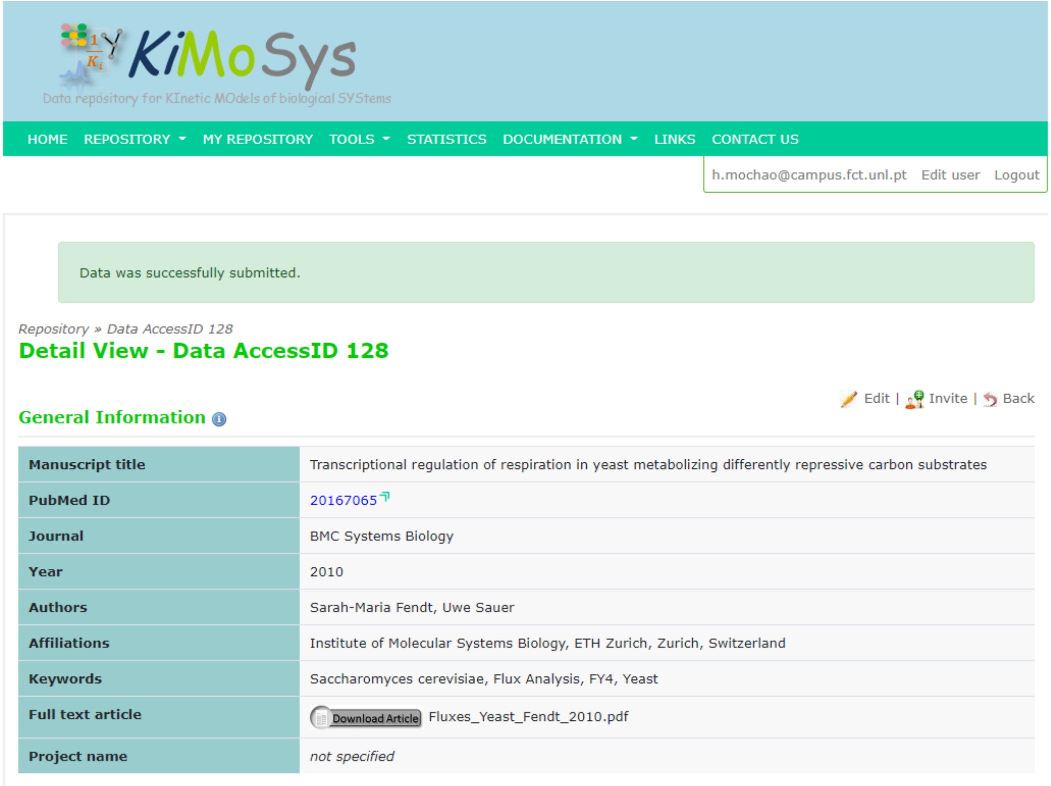
○ **Confidential peer review**

(if checked then data file is **private and EntryID not visible, except for reviewers**)

**Important**
1) *Curation process is started only if the data is made public.
2) Please change to "public" if your data has been accepted or published.
3) Please upload the article file if your data has been accepted or published.

Figure A.17: Snapshot of the submission final step.

The data is submitted. Now it is possible to see the file, and navigate through it using the new KiMoSys functionalities (https://kimosys.org/repository/128).



Data was successfully submitted.

Repository » Data AccessID 128

Detail View - Data AccessID 128

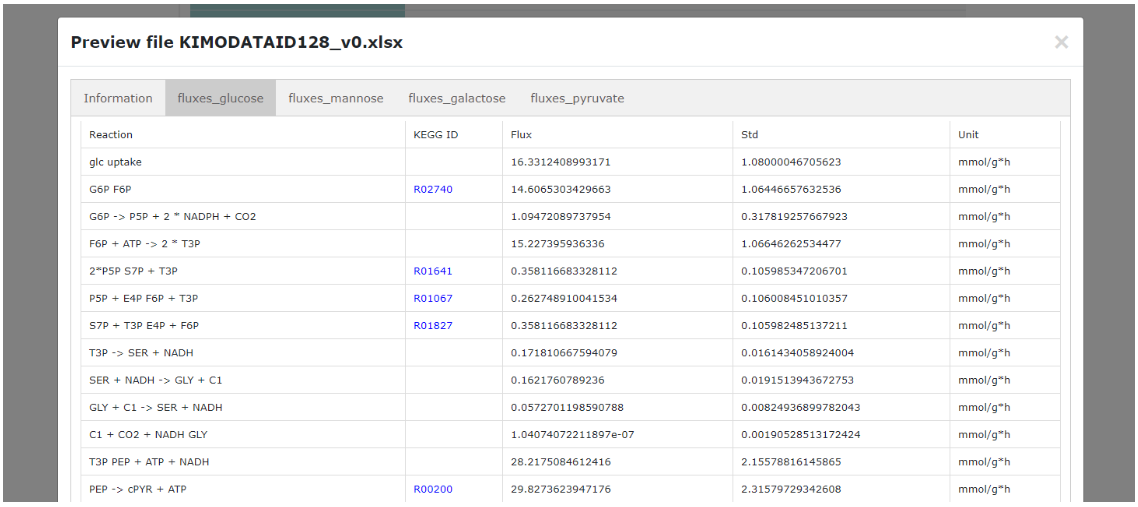
Edit | Invite | Back

General Information

| | |
|---|---|
| Manuscript title | Transcriptional regulation of respiration in yeast metabolizing differently repressive carbon substrates |
| PubMed ID | 20167065 |
| Journal | BMC Systems Biology |
| Year | 2010 |
| Authors | Sarah-Maria Fendt, Uwe Sauer |
| Affiliations | Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland |
| Keywords | Saccharomyces cerevisiae, Flux Analysis, FY4, Yeast |
| Full text article | Download Article Fluxes_Yeast_Fendt_2010.pdf |
| Project name | *not specified* |

Figure A.18: Screenshot of the new Data Entry ID 128.

Finally, the user checks the preview of the file and confirms that the data is correct. The "Information" tab was automatically added with the information filled by the user in the web submission form.



Preview file KIMODATAID128_v0.xlsx

| Reaction | KEGG ID | Flux | Std | Unit |
|---|---|---|---|---|
| glc uptake | | 16.3312408993171 | 1.08000046705623 | mmol/g*h |
| G6P F6P | R02740 | 14.6065303429663 | 1.06446657632536 | mmol/g*h |
| G6P -> P5P + 2 * NADPH + CO2 | | 1.09472089737954 | 0.317819257667923 | mmol/g*h |
| F6P + ATP -> 2 * T3P | | 15.227395936336 | 1.06646262534477 | mmol/g*h |
| 2*P5P S7P + T3P | R01641 | 0.358116683328112 | 0.105985347206701 | mmol/g*h |
| P5P + E4P F6P + T3P | R01067 | 0.262748910041534 | 0.106008451010357 | mmol/g*h |
| S7P + T3P E4P + F6P | R01827 | 0.358116683328112 | 0.105982485137211 | mmol/g*h |
| T3P -> SER + NADH | | 0.171810667594079 | 0.0161434058924004 | mmol/g*h |
| SER + NADH -> GLY + C1 | | 0.1621760789236 | 0.0191513943672753 | mmol/g*h |
| GLY + C1 -> SER + NADH | | 0.0572701198590788 | 0.00824936899782043 | mmol/g*h |
| C1 + CO2 + NADH GLY | | 1.04074072211897e-07 | 0.00190528513172424 | mmol/g*h |
| T3P PEP + ATP + NADH | | 28.2175084612416 | 2.15578816145865 | mmol/g*h |
| PEP -> cPYR + ATP | R00200 | 29.8273623947176 | 2.31579729342608 | mmol/g*h |

Figure A.19: Snapshot of the dataset file preview for Data Entry 128.

Now the dataset will be reviewed by the admin, and assigned a DOI.