Structure of EnzymeML documents

1. Content of the SBML file

1.1. `model` tag

In SBML, the `model` tag forms the container of all defined lists. The name of the experiment is saved in the `name` attribute. The metadata of the experiment file is saved as annotation in Dublin Core RDF format and contains information about the author of the file, the creation date, and the modification dates of the file. The information about the COMBINE Archive[1] and files included therein is saved as MIRIAM annotation. Information on the methods used in the different experiments is stored in the `reaction` tag.

1.2. Units

A strict definition of units is crucial to correctly describe the result of experiments and to ensure comparability and reproducibility. SBML defines SI units by the `listOfUnitDefinitions` tag, from which the experimentalist can choose. A new unit is defined by base units that can have a `scale`, a `multiplier`, and an `exponent`. Throughout the document, this unit is referred to by its `id`.

1.3 Compartments

The reactions take place in different environments, which are described in SBML with the `compartment` tag in the `listOfCompartments` and are annotated by MIRIAM. The `compartment` tag has attributes which describe whether it is `constant`, as well as its `size` and `units`. In EnzymeML, a compartment is a test tube, and each compartment describes a different experimental condition, such as initial substrate concentrations used in the reaction. Throughout the document, the compartment is referred to by its `id`, and its `name` describes the compartment in a human-readable form.

1.4 Species

In SBML, the `species` element defines an entity that is considered indistinguishable from each other, may participate in reactions, and is located in a specific compartment. The species element also defines attributes such as the `sboTerm` attribute (**Table S2**) which specifies the role of the species in the reaction, the `compartment` where the species is located, and the initial concentration or amount of the species. The latter can be described using the distribution package of SBML (*http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/distrib*) to store value

distributions and ranges. The species name is stored in the `name` attribute, while further information about the species is stored in the respective annotations. The compartment description also includes other species such as buffer or additives, resulting in a comprehensive documentation of the reaction conditions.

1.5 Reaction

The reactions of the experiment are represented by the reaction element in the `listOfReactions` tag. They include the `listOfReactants`, the `listOfProducts`, and the `listOfModifiers` tags to represent the chemical equation of the experiment and their respective stoichiometry. Cascade reactions can be represented by different reaction elements in the same reaction vessel (defined via `compartment`) connected by the same species `id`s. Each reaction element is described by an `id`, `name`, and the information whether the reaction is `reversible`. The `listOfModifiers` tag gives information about the catalyst of the reaction and other species that interact with the reaction. The `kineticLaw` tag of SBML is used to describe a kinetic model fitted to the experimental data, including the mathematical model as well as parameters such as k_{cat}. The EnzymeML annotation `enzymeml:data` of the `listOfReactions` tag and the `reaction` tag specify the data stored in the CSV formatted experimental data file and links them to the respective `id`s of the species. The annotation `enzymeml:data` consists of three lists, beginning with `listOfFormats`, where columns for each CSV file found in the second list `listOfFiles` are described. Entries in `listOfMeasurements` denote which file belongs to a measurement which further links experiment metadata found in the EnzymeML document to the data.

1.6 Kinetic models

The `kineticLaw` tag describes the kinetic modelling process, where the estimated parameters are stored as `localParameters` and the model in MathML format, resulting in a comprehensive, reproducible description of the modelling procedure. The MIRIAM annotation provides further information about the model and the modelling method. Additional information can be stored as plain text. The EnzymeML annotation also provides information about the data used for the parameter estimation.

2 Experimental data file

The experimental data on substrate or product concentration as a function of time is stored in the tabular CSV format. Each column is separated by a comma, and each row is separated by a new line. This simple file format allows the user to easily write and read the file with a spreadsheet program such as Excel or a text editor. Because it is machine readable, it can also be generated and analyzed by a program.

The format of the experimental data file is described by the EnzymeML annotation in the experiment file and defines the content of the columns (time or concentration). Each concentration column refers to a species and a unit. Each experimental data file can contain multiple experiments, each experiment multiple replica and the `measurement id` and the `replica` attribute, respectively, are used to identify the begin and end row of a measurement or a replica.

3 Annotations

3.1 EnzymeML

EnzymeML annotations contain additional information that is not part of SBML or MIRIAM, such as pH value, temperature, and pressure (**Figure S1**). The annotations of a species element depend on its type. Small molecules are described by their IUPAC name,[2] their SMILES (simplified molecular-input line-entry system) code, [3] or their InChI (International Chemical Identifier) code,[4] proteins are described by their amino acid sequence. EnzymeML annotations in the `listOfReactions` (**Figure S2**) and `reaction` tags control the automated use of the data stored in the experimental data files, such as format, arrangement of data, and replica. This leads to a list of measurements, which can be later used for kinetic modelling. Furthermore, all datasets are assigned to their corresponding experiments, thus enabling reuse of data. A complete specification as XML Schema is available at GitHub (*https://github.com/EnzymeML/PyEnzyme/blob/main/xsd/EnzymeML.xsd*).

3.2 MIRIAM

The MIRIAM annotations[5] are stored in the RDF format (*http://www.w3.org/TR/rdf-schema/)* and structured in a subject-predicate-object format. The subject is usually the `metaid` attribute of the annotated element, while the predicate is defined using one of the BioModels Qualifiers (**Figure S3**). The object is an RDF Bag container with a list of elements, that include as a resource attribute a link to *http://identifiers.org* that links to the ontology resource. The output

format of the link can be requested as an RDF document, which is machine-readable and can be used by a program to retrieve further information about the annotated elements.

The RDF format is used to describe metadata. In the subject, which is represented by the `Description` element, the `metaid` is linked to the attribute `about` and the '#' sign followed by the `metaid`. The predicate determines how the object is represented by the subject. The subject in EnzymeML is any SBML element defined with `metaid`. The predicates are listed in the MIRIAM annotation or in other namespaces.