

# Biochemical Data Integration: Standardization of Parameter Units and Normalization of Chemical Compound Names in SABIO-RK

Martin Golebiewski<sup>1,\*</sup>, Jasmin Šarić<sup>1,2</sup>, Henriette Engelken<sup>1</sup>, Andreas Weidemann<sup>1</sup>,  
Renate Kania<sup>1</sup>, Olga Krebs<sup>1</sup>, Saqib Mir<sup>1</sup>, Ulrike Wittig<sup>1</sup>, Isabel Rojas<sup>1</sup>

1. Scientific Databases and Visualization, EML Research gGmbH, Heidelberg, Germany

2. Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

\*Email: [martin.golebiewski@eml-r.villa-bosch.de](mailto:martin.golebiewski@eml-r.villa-bosch.de)

## Background

SABIO-RK (<http://sabio.villa-bosch.de/SABIORK>), a database system we have developed to provide experimental data, offers information about biochemical reactions and their corresponding kinetics [1]. The database is populated by merging information about reactions and pathways mainly derived from public databases with the corresponding kinetic data manually extracted from literature. The integration of the collected data from various sources is indispensable to make it comparable and homogenous, e.g. for the set-up of biochemical models.

The heterogeneity of the data, such as the usage of synonymous or aberrant notations of compounds and enzymes, different units of kinetic parameters or missing information about assay proceedings and experimental conditions, causes several obstacles for consistent data integration. Standardization methods can be designed to increase the consistency of the data. With this aim, we have developed two algorithms that accomplish standardizations which are necessary for the integration of data from various sources.

## Standardization of Parameter Units

The first algorithm standardizes parameter values describing comparable measurements to defined base units. This is an important task that increases the comparability of parameter values and also makes them accessible for simulation platforms that currently are unable to make correct calculations based on diverse parameter units.

This transformation proceeds in two steps (Figure 1): Each parameter unit, composed of unit kinds that are compliant with SBML (Systems Biology Mark-Up Language [2]), is scaled to a base level. In a second step, all parameters of the same type are transformed to uniform composite units. In the next version of the SABIO-RK user interface we will release the unit transformation for data export, resulting in SBML files that contain standardized parameters with uniform units.

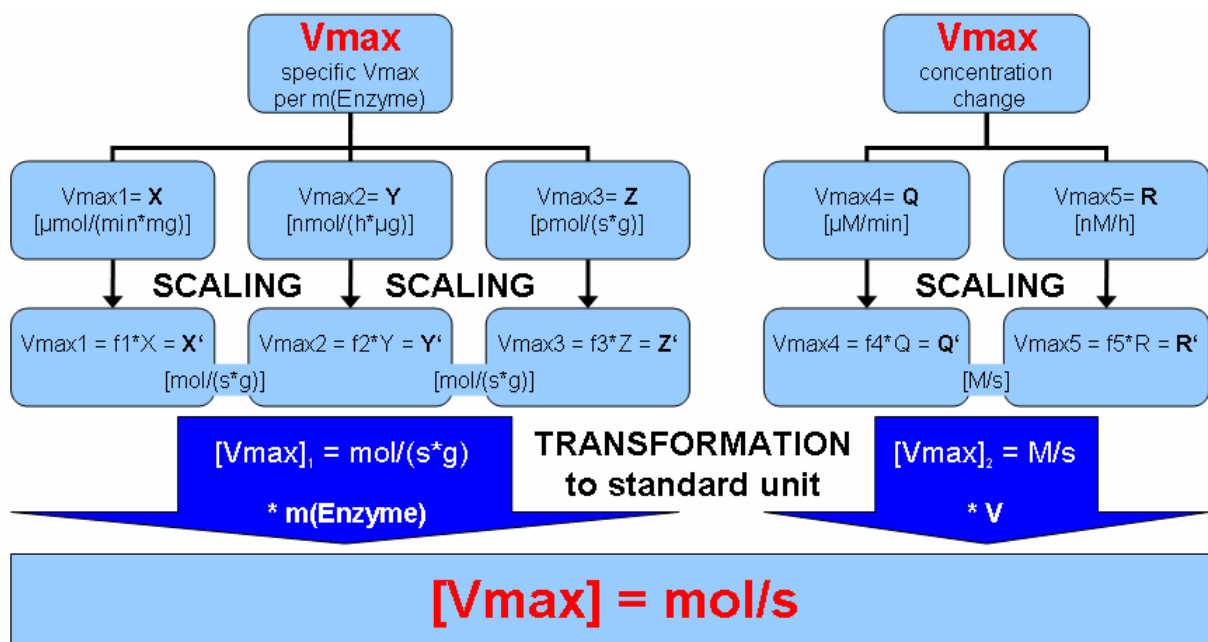


Figure 1: The standardization of parameters results in uniform units for each parameter type. This is an example for the concerted standardization of parameters describing the maximum velocity ( $V_{max}$ ) of a reaction in different units.

### Normalization of Chemical Compound Names

The second algorithm was implemented to facilitate the bundling of corresponding data referring to identical compounds. Since chemical compounds can have many different names - trivial, as well as systematic names - the identification of a chemical compound solely based on its name is a delicate task. However, this identification is crucial for the integration of biochemical data, as many publications exclusively describe a compound by its name. The methods we apply are based on natural language processing (NLP) and focus on the systematic normalization and subsequent comparison of chemical compound names. By normalizing different name variations the tool is able to match synonymous notations and, in many cases, to map to the corresponding systematic name as recommended by the International Union of Pure and Applied Chemistry (<http://www.iupac.org/>).

The identification of synonyms will be combined with our approach to construct chemical structures from chemical compound names (CHEMorph [3]). CHEMorph analyzes systematic names and yields a semantic representation. This representation is translated into a chemical structure (SMILES string) and classified by functional groups. The combination of CHEMorph with our methods for the normalization and matching of synonyms will constitute a platform for the unambiguous identification of compounds based on their various names (Figure 2).

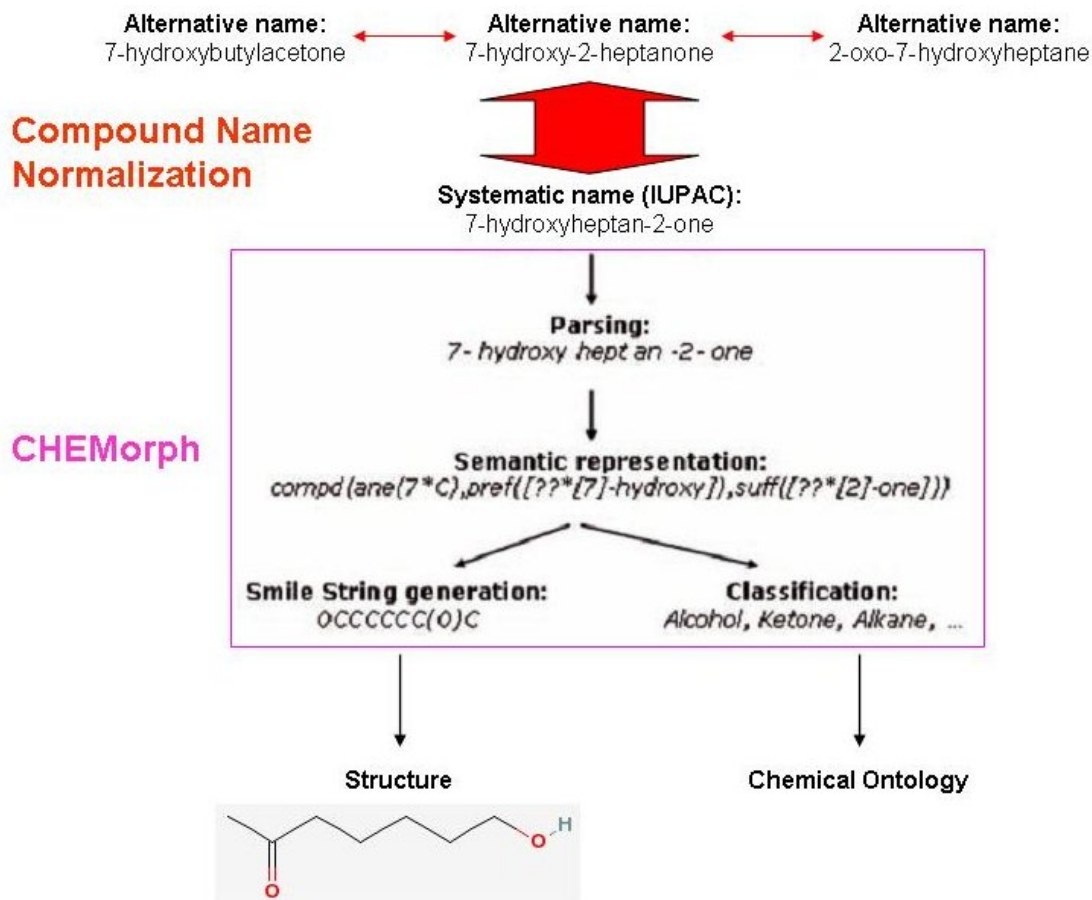


Figure 2: The combination of normalization methods merging different synonyms that describe the same chemical compound with a semantic analysis of the corresponding systematic name help to unambiguously identify and classify a compound.

## References

- [1] Wittig U, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Anstein S, Saric J, Rojas I (2006):  
SABIO-RK: Integration and Curation of Reaction Kinetics Data  
In proceedings of the 3rd International workshop on Data Integration in the Life Sciences 2006 (DILS'06). Lecture Notes in Computer Science, 4075: 94-103.
- [2] Hucka M, et al. (2003):  
The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. Bioinformatics, 19: 524-531.
- [3] Kremer G, Anstein S, Reyle U (2006):  
Analysing and Classifying Names of Chemical Compounds with CHEMorph.  
In Sophia Ananiadou and Juliane Fluck, (editors) Proceedings of the Second International Symposium on Semantic Mining in Biomedicine: 37-43.  
JULIE Lab, Friedrich-Schiller-Universität Jena, Germany.