ModelSEED: Assembling a biological thermodynamic database

Andrew Philip Freiburger1, \*, Sam Seaver2, Chris Henry2, Ethan Chan1, and Heather L. Buckley 1,3

1Civil Engineering and 3 Chemistry Departments at the University of Victoria

2Argonne National Laboratory

\* Author for contact [andrewfreiburger@gmail.com](mailto:andrewfreiburger@gmail.com);

**Abstract**

**Summary:** Computational biology is predicated upon biochemical databases. The Thermodynamics of Enzyme-Catalyzed Reactions database (TECRDB) by NIST is a renowned database of thermodynamic reactions, however, the website has become neglected and consequently the data has become inaccessible to biological modelers. We present work that salvaged, reformatted, and contributed the TECRDB data to the ModelSEED database, where the data will be actively curated and readily accessible to biological modelers. The salvaging, reformatting, and contributing methods were completed through Python scripts, which are provided with the associated documentation in our GitHub repository (<https://github.com/freiburgermsu/Biochemical-databases>).

Systems biology and bioinformatics are test theoretical and experimental biology. Whole cell models (Karr and Goldberg, 2020), for example, critique theoretical schema of interacting biological systems whilst also predicting experimental results by perturbing parameters. These systems-level models are predicated upon biochemical data and databases (Bhat and Balaji, 2020; Zvi *et al.*, 2008; Arakawa *et al.*, 2003). Database are most prevalently online, since the internet reaches a maximum audience and optimally allows real-time updates. Numerous website databases, however, have become neglected as webmasters leave their duties, and consequently many databases have lost usefulness to contemporary modelers.

The TECRDB (Goldberg *et al.*, 2004) is such a database. The TECRDB impressively amalgamates of literature review and database assemblage to create a singular source of biochemical data for 417 enzymatic reactions. The database website, however, has lost critical features that were detailed in the original publication. The website, for example, has disabled the download function, which renders the data practically non-existent for biological modelers who are unlikely to manually transcribe the data. The website has further lost the cited background description of the database, which may allow researchers to understand the data and the associated methods that were used to acquire and\or approximate the data. The symbol key for the sub- and super-scripts, which must adhere to IUPAC and IUBMB standards, are also absent from the database. The querying methods are moreover dissimilar from the published methods, where the current methods are rudimentary. The database furthermore contained 25 broken pages, where the thermodynamic data tables were unpresented.

We recognized the neglected yet valuable data of the TECRDB and elected to refresh the data. This work both reintroduces an important thermodynamics resource to the systems biology community and may ideally inspire bioinformatitians to consider salvaging neglected thermodynamic datasets and contributing the data to a standard resource for the community (Dogucu and Çetinkaya-Rundel, 2020). We used myriad Python scripts, which are published in our GitHub repository (<https://github.com/freiburgermsu/Biochemical-databases>), to web scrape, parse, organize and reformat, and then contribute the TECRDB data to the ModelSEED database, which is an actively curated database of biochemical reactions.

The TECRDB was first web scraped through the BeautifulSoup library of Python. Script iteratively searched the URLs for each functioning database page through the Requests module. The webpages were subsequently parsed the table elements through the <tr> and <th> HTML tags via the BeautifulSoup LXML interpreter, and the acquired data was cleaned of spaces and non-ASCII characters (Massimino, 2016). The data from scraping each webpage were compiled into a single table, where columns of similar data like equilibrium constants were all combined.

The organized TECRDB data was reformatted for the ModelSEED database (Seaver *et al.*, 2021). The ModelSEED database accepts community contributions of data to create a centralized and actively currated resource for computational biologists. The contribution of TECRDB data required that the TECRDB reactions be paired with the ModelSEED reactions, was accomplished via the Pandas and JSON modules. The first attempt of matching the TECRDB reactions occurred through matching the chemical equations after each involved compound was matched between the TERCRDB and ModelSEED database. The compounds were matched principally by name, although, the myriad common and IUPAC synonyms for every chemical prevented high success of compound matching; thus, extensive manual investigation was required to matched previously unmatched compounds. The second attempt of matching reactions, where at least one compound failed to matched, were matched through the KEGG IDs that the TECRDB ascribed to every one of its reactions. The data from the TECRDB may then be contributed to the matched reaction entry in the ModelSEED database. Reactions and comopunds that were undescribed by the ModelSEED database yet were verified as being chemically coherent were introduced to the database after manual vetting.

The salvaging of neglected data will become increasingly necessary as website databases age. Bioinformatitians must resist the bystander effect and proactively repurpose old databases that have fallen from popular use, yet may support modelers in a new and accessible form. The ModelSEED database offers a highly accessible resource of centralized and curated biochemical data, which is an invaluable resource for biological modelers. The ModelSEED database is eager to receive community contributions of data, for which a step-wise diagram is provided in our GitHub repository, and to become an inclusive resource that expedites computational research. The TECRDB, despite malfunctions and shortcomings of the website per se, may find renewed value through expression in the ModelSEED database.

# **References**

Arakawa,K. *et al.* (2003) G-language genome analysis environment: A workbench for nucleotide sequence data mining. *Bioinformatics*, **19**, 305–306.

Bhat,N.G. and Balaji,S. (2020) Whole-Cell Modeling and Simulation: A Brief Survey. *New Gener. Comput.*, **38**, 259–281.

Dogucu,M. and Çetinkaya-Rundel,M. (2020) Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *J. Stat. Educ.*

Goldberg,R.N. *et al.* (2004) Thermodynamics of enzyme-catalyzed reactions - A database for quantitative biochemistry. *Bioinformatics*, **20**, 2874–2877.

Karr,J. and Goldberg,A. (2020) An introduction to whole-cell modeling.

Massimino,B. (2016) Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data. *J. Bus. Logist.*, **37**, 34–42.

Seaver,S.M.D. *et al.* (2021) The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.*, **49**, D575–D588.

Zvi,A. *et al.* (2008) Whole genome identification of Mycobacterium tuberculosisvaccine candidates by comprehensive data mining and bioinformatic analyses. *BMC Med. Genomics*, **1**, 1–25.