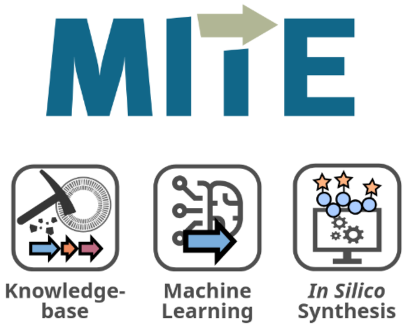
Master Thesis Project

# Machine Learning on Tailoring Enzymes: Harnessing MITE DB

| | |
|---|---|
| **Supervisor** | Ass.-Prof. **Mitja M. Zdouc**. ZdoucLab.org<br>Division of Pharmacognosy, University of Vienna. |
| **Topic** | Machine learning, data science, enzymology |
| **Duration** | 6 Months (full-time), **computational work only** |
| **Language** | English/German |

## Description

Many medicines are derived from natural products (NPs), small molecules produced by a variety of organisms. A specific class of proteins called tailoring enzymes are responsible for converting the nascent metabolites into their mature and often bioactive forms. Therefore, knowledge about tailoring enzymes and their reactions is essential in understanding NP biosynthesis and performing tasks such as bioactivity prediction. To this purpose, we have recently developed **MITE**, the **M**inimum **I**nformation about a **T**ailoring **E**nzyme database [1] (https://mite.lisc.univie.ac.at/). MITE connects tailoring enzymes with their reaction- and substrate-specificity in a machine-readable format, and provides rich metadata on reaction types, co-factors, and biosynthetic context. MITE cross-references with databases such as MIBiG and Rhea and is already used as a reference database by the popular genome mining tool antiSMASH.



In this project, you will investigate the use of the MITE database as a **resource for machine learning**. You will primarily focus on supervised learning in tasks such as **enzyme classification, reaction prediction, and substrate recognition.** Your work will include data extraction and augmentation, feature engineering, and workflow development. You will evaluate different algorithms for their performance and prepare them for follow-up case studies on tailoring enzyme classes. By providing workflows, training and test datasets, you will support the adoption of MITE as a benchmark database in enzyme-related machine learning tasks. Eventually, your work will serve as a foundation for better prioritization algorithms that speed up NP discovery.

## Learning outcomes

At the end of this Master thesis, you will be able to:

- **Categorize** machine learning algorithms and **apply** them in protein-related predictive tasks.
- **Prepare** datasets by data augmentation and feature engineering.
- **Test** and interpret model results and interpret their generalizability.
- **Present** findings in word and text and **understand** the manuscript writing process.

## How to apply

Please contact Ass.-Prof. Mitja Zdouc via E-Mail (mitja.zdouc@univie.ac.at). Include your CV and a brief summary of the progress of your studies, your research interest, and why you would like to work on this project. Indicate any prior knowledge related to the topic, as well as your preferred working style, supervision expectations, and preferred starting date. Applicants will be invited for a brief interview where the project will be discussed in detail.

**References:**
1 Rutz et al. 2025 Nucl Acids Res DOI: 10.1093/nar/gkaf969