# Prediction of terpene synthase activity using self-supervised deep learning

Samuchevich Raman1,2,3, Tereza Čalounová1,4, Roman Bushuiev1,5, Adéla Tajovská1, Josef Šivic2,

Pluskal Tomáš1

1 Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences

2 Czech Institute of Informatics, Robotics and Cybernetics (CIIRC CTU)

3 Faculty of Chemical Technology, University of Chemistry and Technology

*4 Faculty of Science, Charles University in Prague*

*5 Faculty of Information Technology, Czech Technical University in Prague*

Terpene synthases (TPSs) are enzymes responsible for the biosynthesis of the largest class of natural products, including widely used flavors, fragrances, and first-line medicines. The amount of available TPS protein sequences is increasing exponentially due to rapid advances in high-throughput sequencing. However, characterizing the function of each TPS requires challenging and time-consuming experiments as well as significant domain expertise. To help overcome these challenges, we employ self-supervised deep learning models to understand the sequence space of TPSs and make predictions regarding the substrates and reactions catalyzed by yet-uncharacterized TPS enzymes. Our model can now outperform established bioinformatic methods based on hidden Markov models for the prediction of TPS substrates. To facilitate the development of novel predictive methods for TPS characterization, we also assembled a curated database of ~1,500 characterized TPS-catalyzed reactions, which is currently the largest such dataset available.