[BIOINFO 2025]

# Predicting substrate specificity of acyltransferase domains in polyketide synthases using machine learning

Joon Young Kwon[1], Byeongsub Lee[2], Jihi Yeom[1], Byung Tae Lee[1], and Hyun Uk Kim[1,2,*]

*[1] Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea*

*[2] Graduate School of Engineering Biology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea*

**Corresponding author: ehukim@kaist.ac.kr*

Polyketides are an important family of natural products that include many useful medicines, such as antibiotics, anticancer agents, and immunosuppressants. They are synthesized by large modular enzymes called polyketide synthases (PKSs), with the acyltransferase (AT) domain playing a key role by selecting which extender unit—such as malonyl-CoA or methylmalonyl-CoA—will be added next. This decision strongly influences the chemical diversity and biological activity of polyketides. While some sequence motifs have been linked to AT specificity, predicting substrate preference directly from sequence remains challenging. To predict AT substrate specificity, we constructed a curated dataset by extracting AT domain sequences from the MiBIG database and validating substrate assignments through the primary literature. Using this dataset, we developed a multi-class classification model based on the pretrained protein language model, Evolutionary Scale Modeling (ESM)-2, keeping early layers frozen while fine-tuning later layers to capture substrate-specific sequence features efficiently. Beyond robust classification across diverse AT families, the model can predict the substrate specificity of uncharacterized AT domains from newly sequenced PKSs, provide interpretable substrate probability distributions through a softmax output layer, and perform in silico mutational analysis using a beam search strategy to suggest amino acid substitutions likely to alter substrate preference. Our results demonstrate how integrating protein language models with biosynthetic domain knowledge can accelerate the interpretation and engineering of complex enzymatic systems, offering a computational route to the discovery of novel polyketide-based therapeutics.