High-performance prediction of riboswitch–ligand interactions with LLM-guided active learning

Sofia Mun[1], Sunggi An[1], and Hyun Uk Kim[1,*]

[1]*Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea*
Corresponding author: ehukim@kaist.ac.kr

Riboswitches, cis-regulatory RNA elements, control gene expression by binding to specific small-molecule ligands. They are central to cellular metabolism and have emerged as powerful tools in synthetic biology. Yet, despite their importance, the full spectrum of riboswitch–ligand interactions is still unclear. Limited datasets and the absence of robust predictive models have hindered progress in this area. To address this challenge, we introduce a machine learning framework designed to predict ligand-binding affinity from both sequence and secondary-structure features. At the core of our approach lies large language model (LLM)-guided active learning, which combines raw RNA sequence information with expert-defined structural rules. This strategy allows the model to refine its predictions iteratively and leverage prior biological knowledge. The results are promising: the model achieves a mean absolute error (MAE) below that of earlier methods and attains 76% accuracy in distinguishing strong from weak binders. Beyond benchmarking gains, the framework provides a scalable platform that could accelerate the design of synthetic riboswitches and extend naturally to other RNA–ligand systems.