[BIOINFO 2025]

ParaEM : Sequence-based paratope predictor using Expectation-Maximization with CDR prior

Taegyu Kim[1], and Sun Kim[1,2,3,4,*]

[1] *Interdisciplinary Program in Artificial Intelligence, Seoul National University*
[2] *Department of Computer Science and Engineering, Seoul National University*
[3] *Interdisciplinary Program in Bioinformatics, Seoul National University*
[4] *AIGENDRUG Co., Ltd.*
** Corresponding author: sunkim.bioinfo@snu.ac.kr*

Accurate identification of antibody paratopes—the residues that bind antigen—is critical for antibody engineering and vaccine design, yet existing structure-based predictors depend on high confidence 3D structures that are unavailable for most sequences. Here, we introduce a purely sequence-based paratope predictor that integrates pretrained ESM3 embeddings with trainable CDR region vectors in an EM driven cross-attention framework to infer latent antibody antigen alignments. Our model achieves state-of-the-art performance at predicting paratope on three different benchmark datasets. Ablation studies demonstrate that the latent-alignment EM algorithm is the primary driver of performance improvements, and that CDR priors serve as a powerful initialization-accelerating EM convergence and yielding additional improvements. By eliminating reliance on structural inputs and offering interpretable residue-level alignments, our approach scales to millions of antibody sequences and accelerates rational therapeutic development.