A060 · Sequence-Conditioned Generation of Genome-Targeting Integrases with a Genomic Foundation Model

Presenter: Tanggis Bohnuud — Basecamp Research

Authors: Geraldene Munsamy, Gavin Ayres, Carla Greco, Keith Kam, Gus Minto-Cowcher, John St. John, Tanggis Bohnuud, Matthew Bakalar, William Chow, Robert Pecoraro, Marcelo D.T. Torres, Aaron Kollasch, Marcus Leung, Hassan Sirelkhatim, Francesco Farina, Connor McGinnis, Srijani Sridhar, Daniel Anderson, Francesco Oteri, Ali Taghibakhshi, Jeremie Dona, Tyler Shimko, Cedric Steenbeke, Alexandros Papadopoulos, Malcolm Krolick, Fabian Spoendlin, Purba Gupta, Sandeep Kumar, Anne Bara, Jared Wilbur, Noelia Ferruz, Timur Rvachov, Fangping Wan, Hanqun Cao, Hyun-Su Lee, Japan Mehta, Raphael Chaleil, Valerio Pereno, Sid Potti, Chris Emerson, Roy Tal Dew, Kevin K Yang, Eric Nguyen, Neha Tadimeti, Jillian F. Banfield, Alicia Frame, Emma Bolton, David Ruau, Rory Kelleher, Anthony Costa, Kimberley Powell, Cesar de la Fuente-Nunez, Glen-Oliver Gowers, Oliver Vince, Jonathan Finn, Philipp Lorenz

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Designing proteins to recognize user-specified genomic sequences requires learning the complex relationship between protein sequence, DNA recognition, and catalytic function. Here, we formulate genome-targeted integrase design as a conditional sequence-generation problem and use EDEN, a 28-billion-parameter genomic foundation model trained on 9.7 trillion nucleotide tokens from diverse metagenomes (BaseData), to generate de novo large serine recombinases (LSRs). LSRs and their cognate attachment sites mined from BaseData provide naturally occurring paired protein–DNA sequences from which to learn the evolutionary constraints linking recombinase sequence to target specificity. We condition EDEN on only 30 nucleotides representing a desired attB target site and generate corresponding LSR protein sequences. To test generalization to unseen sequence specifications, we prompted EDEN with human genomic sequences absent from the training data, spanning ten disease-associated loci and four potential safe-harbor sites. EDEN generated multiple experimentally active recombinases for every tested genomic locus, achieving a 63.2% functional hit rate across diverse DNA prompts. High-performing generated proteins diverged substantially from their parental sequences, reaching as low as 52% sequence identity while retaining biochemical activity. Generated LSRs also translated to cellular function: 50% of tested designs were active in human cells, including targeted integration of CAR constructs in primary human T cells. These results demonstrate that genomic foundation models can learn conditional relationships between short DNA sequences and the proteins that recognize them, enabling functional protein generation directly from user-specified genomic targets.