A035 · Cherimoya: Lightweight modeling of genomic modalities enables organism-wide analyses

Presenter: Jacob Schreiber — UMass Chan Medical School

Authors: Christian Ramirez, Achsah Marlene Aruva, Zhiping Weng, Jacob Schreiber

Session: Day 1 · Thu Oct 1, 2026 · 4:15–5:15 PM

Accurate prediction of genomic function across diverse modalities and cell types has become central to modern regulatory genomics, but state-of-the-art models often achieve their performance by growing in size and complexity in a manner that limits their practical ability to tackle genome- and organism-scale analyses. We introduce Cherimoya, a compact neural network that uses an order of magnitude fewer parameters than existing state-of-the-art models while achieving higher accuracy on both observational and variant effect predictions and running up to 10x faster. Cherimoya supports prediction across a variety of genomic modalities with state-of-the-art performance, enabling scalable in silico experimentation. By drastically reducing compute and memory requirements, Cherimoya makes large-scale genomic design practical, including tasks such as designing cell type-specific enhancers across dozens to hundreds of related cell types. Finally, Cherimoya is fully integrated into modern agentic systems, having used such systems to optimize model architecture and hyperparameters, and providing documentation and skills for such systems to easily use Cherimoya in their analyses.

Cherimoya: Lightweight modeling of genomic modalities enables organism-wide analyses

Deep learning models that map genomic sequence to biochemical readouts have become foundational tools for studying regulatory genomics. These models have enabled the *de novo* discovery of regulatory motifs, characterization of higher-order rules governing cis-regulatory syntax, and prioritization of causal non-coding variants underlying complex traits and rare diseases. An aspirational goal of the field is to move beyond prediction on, and interpretation of, endogenous sequences toward design by engineering synthetic regulatory elements with prescribed cell type-specific activity. Achieving this goal requires models that are both accurate and fast enough to be evaluated potentially millions of times during the iterative design process.

We introduce Cherimoya, a compact, efficient neural network architecture that builds on the sequence-to-profile framework of ChromBPNet but introduces changes that dramatically reduce parameter count while also improving predictive performance. The core innovation is the Cheri block, an adaptation of the ConvNeXt block for noisy genomic signals. Each Cheri Block is a dilated depthwise convolution, layer norm, projection to a higher-dimensional feature space, GeLU activation, and a projection back, with a residual connection modulated by a fixed layer scaling. To achieve higher throughput, the dilated depthwise convolution and subsequent layer normalization are fused into a custom GPU kernel, providing a ~4x speedup over a naive PyTorch implementation. Cherimoya also includes a fused megakernel of the entire Cheri Block for inference-only tasks that increases speed by another 2x at fp16 and 3x at fp32.

Cherimoya establishes a new performance frontier for genomic sequence-to-function (S2F) modeling. With only 32 filters in each layer – ~1.25% of ChromBPNet's total parameter count – Cherimoya achieves comparable predictive performance on chromatin accessibility benchmarks. Increasing to 96 filters (<5% of ChromBPNet's parameters) yields a substantial improvement in accuracy over ChromBPNet. In terms of wall-clock inference time, the 96-filter model runs 7.4x faster than ChromBPNet (10.2x faster at half precision), and the 32-filter variant is 14.2x (18.9x) faster.

We evaluate Cherimoya on a comprehensive benchmark of 1,125 chromatin accessibility models. Using matched training protocols, Cherimoya outperforms ChromBPNet on profile and count prediction metrics, with an average Pearson correlation of 0.741 (0.715) and 0.811 (0.797) on ATAC- and DNase-seq log count prediction respectively (versus ChromBPNet), and similar improvements on profile Pearson (0.661 vs 0.647 on ATAC-seq, 0.375 vs 0.349 on DNase-seq). For variant effect prediction, we benchmark against chromatin accessibility quantitative trait loci identified in lymphoblastoid cell lines of diverse ancestry. Across six African populations, Cherimoya achieves an average signed Pearson r of 0.693 compared to 0.686 for ChromBPNet, demonstrating that the efficiency gains do not come at the cost of biological fidelity.

Cherimoya demonstrates that large parameter budgets are not necessary to achieve state-of-the-art S2F prediction. By reducing inference cost by over an order of magnitude, Cherimoya enables agentic systems to train and use these models easily when investigating hypotheses, performing variant effect, or designing DNA – even in complicated or whole-organism settings. We have provided agentic skills files with progressive disclosure and comprehensive documentation to guide these systems to detailed instructions on their usage, and anticipate this will be a core feature.