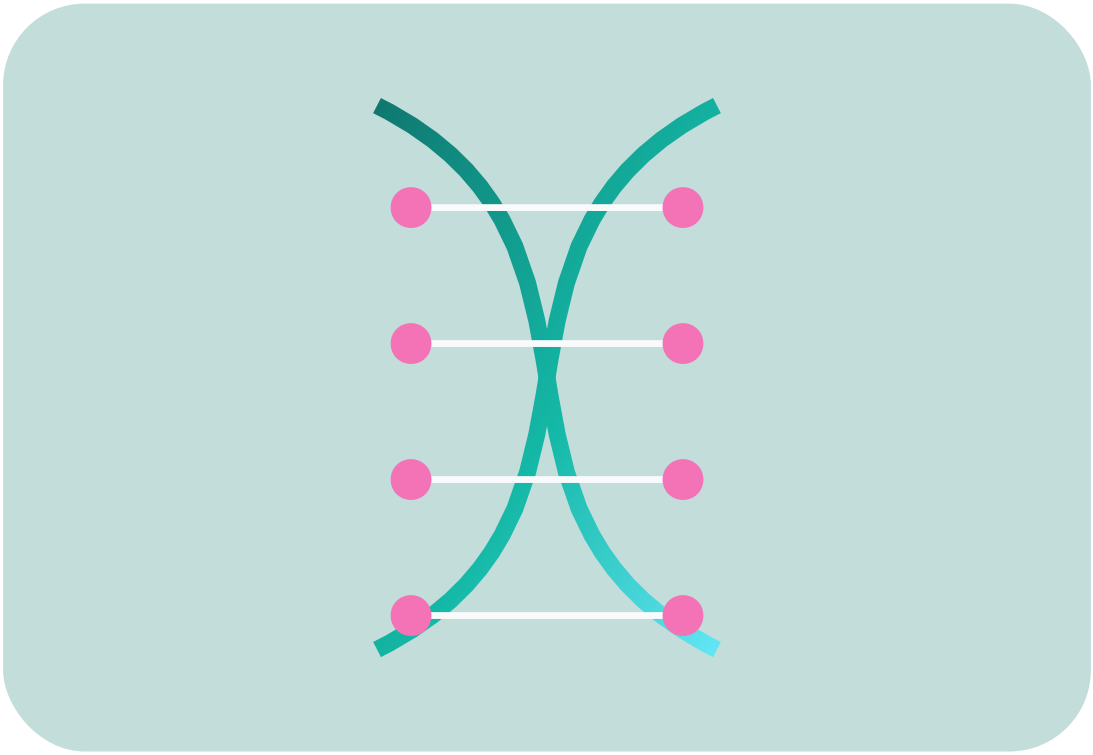
Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling




```
<h3>Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence
Modeling · Concept Sketch</h3>
<p>Genome-scale signal aggregation framing PRS vs. foundation
model granularity.</p>
```

Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling

Authors: Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, Volodymyr Kuleshov
Year: 2024
Venue: 41st International Conference on Machine Learning (ICML), PMLR 235

---

1. Classification

- **Domain Category:**
  - **Genomics FM.** The paper develops long-range sequence models specifically for DNA and evaluates them on genomics tasks such as regulatory element classification and variant effect prediction.
- **FM Usage Type:**
  - **Core FM development.** The main contribution is a new family of DNA foundation models (Caduceus) and underlying architectural blocks (BiMamba and MambaDNA), together with pretraining and fine-tuning strategies.
- **Key Modalities:**
  - Single-modality DNA sequence (human reference genome; nucleotide-level modeling).

---

2. Executive Summary

This paper introduces Caduceus, a family of long-range DNA language models designed to capture two key biological symmetries: bi-directional context and reverse complement (RC) equivariance of DNA. Standard sequence models either struggle with very long genomic contexts or ignore that DNA's two strands carry equivalent information in opposite directions, which wastes data and harms generalization. Building on the Mamba structured state space model, the authors propose BiMamba for efficient bi-directional sequence modeling and MambaDNA for RC-equivariant processing, then assemble these into Caduceus variants that operate as DNA foundation models trained with masked language modeling. They pretrain on the human genome and fine-tune on a wide range of genomics benchmarks, showing that Caduceus matches or outperforms both HyenaDNA (another long-range SSM-based model) and much larger Transformer-based models such as Nucleotide Transformer v2 and Enformer, especially for tasks requiring long-range sequence context and symmetry handling. A highlight is variant effect prediction, where Caduceus competes with or exceeds 10× larger attention-based models at long genomic distances from genes. For a new grad student, this paper is a concrete example of how to incorporate biological inductive biases (symmetries) into modern foundation-model-style architectures to unlock better performance at manageable scale.

---

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - How to build foundation-scale models that understand long DNA sequences, especially non-coding regions that regulate gene expression.
  - Specifically, the model should:
    - Use **bi-directional context**: regulatory effects can depend on bases both upstream and downstream of a position.
    - Be **reverse complement (RC) aware**: the two strands of DNA encode the same information in opposite directions (A?T, C?G), and biological assays may sequence either strand.

        - Capture **long-range interactions**: variants up to hundreds of kilobases (or more) away from a gene can affect expression.
    - Downstream, they particularly care about **variant effect prediction (VEP)**—predicting whether a single nucleotide polymorphism (SNP) causally affects gene expression.
- **Why this is hard**
    - **Long-range dependencies:**
        - Transformers scale quadratically with sequence length, making it costly to model context lengths of 100k–1M base pairs.
        - Biological regulatory effects can span these very long distances, so truncating context harms performance.
    - **Bi-directionality:**
        - Many language models are causal (left-to-right) and only see past context; for DNA, both "past" and "future" bases matter symmetrically.
    - **Reverse complement symmetry:**
        - A DNA sequence and its reverse complement should yield equivalent predictions for most tasks.
        - Standard models treat RC pairs as unrelated sequences, requiring explicit data augmentation and still not enforcing strict symmetry.
    - **Data and modeling complexity:**
        - Non-coding DNA is vast and noisy; many relevant signals are subtle (e.g., conservation, motif patterns).
        - Tokenization choices (e.g., k-mers) can make small sequence changes look large in token space.

---

4. Data and Modalities

- **Datasets used**
    - **Pretraining:**
        - Human reference genome **HG38 / GRCh37** (from the Genome Reference Consortium), split into ~34,021 segments extended to length (2^{20} ,048,576) base pairs.
        - Total scale ≈ **35 billion nucleotide tokens**.
    - **Downstream benchmarks:**
        - **Genomics Benchmarks** suite (Grešová et al., 2023): eight regulatory element classification tasks (e.g., mouse enhancers, human enhancers, open chromatin regions, promoters).
        - **Nucleotide Transformer benchmark** (Dalla-Torre et al., 2023): 18 datasets, including histone mark prediction, enhancer and promoter annotations, and splice site prediction.
        - **Variant effect prediction (VEP) dataset** derived from Enformer (Avsec et al., 2021) and Trop et al. (2023), with SNPs labeled as causal vs non-causal for gene expression using SuSiE fine-mapping.
- **Modalities**
    - Single modality: **DNA sequence** at single-nucleotide resolution (A/C/G/T).
    - Outputs are task-specific labels (e.g., regulatory class, histone mark presence, variant effect).
- **Preprocessing / representation**
    - **Character-level tokenization**: each nucleotide (A, C, G, T) is a token, avoiding k-mer tokenization.
        - Motive: k-mers make small base changes cause large token changes, which complicates learning; single-nucleotide tokens preserve locality.

- For downstream tasks:
  - Sequences are trimmed or padded to task-specific lengths (e.g., 200–2,000 bps for Genomics Benchmarks).
  - Final hidden states are pooled (e.g., mean pooling over positions) to obtain fixed-size embeddings.
- For VEP:
  - They extract features from windows centered at SNP positions (e.g., 1,536 bp windows for SSM-based models) and use these embeddings with an external classifier (SVM).

---

5. Model / Foundation Model

- **Model Type**
  - Based on **Structured State Space Models (SSMs)**—specifically the **Mamba** architecture (Gu & Dao, 2023), which models sequences via linear state-space dynamics plus input-dependent "selection" mechanisms.
  - Mamba operates in **linear time in sequence length**, enabling long contexts (up to hundreds of thousands of base pairs) without quadratic cost.
- **Is it a new FM or an existing one?**
  - The paper **extends Mamba** with new architectural components tailored to DNA:
    - **BiMamba**: a bi-directional variant that processes sequences in both forward and reversed order with shared parameters.
    - **MambaDNA**: an RC-equivariant block that enforces reverse complement symmetry.
  - These blocks are used to build **Caduceus**, a **new family of DNA foundation models** pretrained with a masked language modeling objective.
- **Key components and innovations**

Component	Role / Idea
Mamba	Base SSM block combining selective state space layers with a gated MLP; originally causal (uni-directional).
BiMamba	Applies Mamba to the sequence and its reversed version, then flips and adds outputs; uses shared projections to keep parameter count low while modeling bi-directional context.
MambaDNA	RC-equivariant module: splits channels, applies Mamba (or BiMamba) to forward and reverse-complement sequences with shared parameters, and recombines to guarantee RC equivariance.
Caduceus-PS	Fully RC-equivariant LM via parameter sharing: RC-equivariant embeddings, MambaDNA blocks, and LM head; predictions for RC inputs transform appropriately.
Caduceus-Ph	Uses BiMamba with RC data augmentation during pretraining and **post-hoc conjoining** at inference (averaging predictions on forward and RC sequences) to enforce RC invariance downstream.

- **Architectural details (intuitive)**
  - **BiMamba:**
    - Take a sequence (X) and a reversed copy.
    - Run the Mamba block on both with **shared in/out projections**.
    - Flip the reversed output back along the sequence length and add it to the forward output.
    - This yields a representation that incorporates information from both directions without doubling parameters.
  - **MambaDNA:**
    - Split the channel dimension into two halves.
    - Apply Mamba (or BiMamba) to the forward half and to the **reverse complement** of the other half, with shared parameters.
    - Apply the RC transform again to the reversed output and **concatenate** the two halves back, resulting in a representation that is provably RC-equivariant.
  - **Caduceus-PS:**
    - Uses RC-equivariant embeddings, stacks of MambaDNA blocks with BiMamba inside, and an RC-equivariant LM head that combines channel-flipped outputs.
    - Enforces RC-equivariant behavior already during pretraining, so RC data augmentation is not needed.
  - **Caduceus-Ph:**
    - Uses BiMamba without built-in RC equivariance in the LM.
    - Relies on **RC data augmentation** during pretraining and **post-hoc conjoining** (averaging outputs for forward and RC inputs) at downstream inference to enforce RC invariance.

- **Training setup**
  - **Pretraining objective:**
    - **Masked Language Modeling (MLM)**, similar to BERT:
      - Mask 15% of tokens; among these, 80% replaced with [MASK], 10% replaced with random bases, 10% unchanged.
    - Causal next-token prediction is used for some baseline comparisons (e.g., Mamba vs HyenaDNA), but Caduceus FM training is primarily MLM-based and bi-directional.
  - **Pretraining data & scale:**
    - Human reference genome with long sequence segments (1k, 32k, 131k bps), keeping number of tokens per batch approximately constant across lengths.
    - Several model variants with different depths and hidden dimensions; small (~hundreds of thousands to a few million parameters) compared to Transformers like Nucleotide Transformer and Enformer.
  - **Optimization details (high level):**
    - ADAM optimizer, cosine learning rate decay, learning rate around (8 ^{-3}) for Mamba-based models.
    - RC data augmentation used for non-RC-equivariant models (including HyenaDNA and Caduceus-Ph during pretraining).
  - **Fine-tuning:**
    - For downstream tasks, they pool final hidden states (often mean pooling), then train task-specific heads.
    - Hyperparameters (learning rate, batch size) are tuned per task using cross-validation (5-fold for Genomics Benchmarks, 10-fold for Nucleotide Transformer tasks).

6. Multimodal / Integration Aspects (If Applicable)

This paper **does not perform multimodal integration**: it focuses on a single modality (DNA sequence) and does not combine genomics with other data types such as expression, imaging, or clinical variables.

- **Relation to integration and the baseline plan**
  - Within the broader **Integration Baseline Plan**, Caduceus informs the **"Genetics embedding hygiene and attribution"** principle:
    - It emphasizes **reverse-complement handling (RC equivariance or RC averaging)** as a core inductive bias for DNA encoders.
    - It uses deterministic, nucleotide-level tokenization rather than k-mers, aligning with the plan's caution about unstable tokenization for variant-level analyses.
  - If Caduceus embeddings are later integrated with other modalities (e.g., brain imaging, clinical phenotypes), the plan would advocate:
    - **Preserving modality-specific representations** (i.e., using Caduceus as a frozen or carefully fine-tuned encoder, then fusing at a later stage).
    - Applying the robustness practices (standardization, residualization, consistent CV, calibrated metrics) discussed in the plan.

7. Experiments and Results

- **Pretraining analyses**
  - **Mamba vs HyenaDNA (next-token prediction):**
    - On human-genome pretraining with different sequence lengths (1k, 32k, 131k bps), **Mamba achieves lower cross-entropy loss** than HyenaDNA at similar model sizes.
    - Mamba also appears more robust to higher learning rates, supporting its use as the core building block.
  - **Effect of BiMamba's parameter sharing:**
    - BiMamba with **projection weight tying** enables deeper bi-directional models for the same parameter budget.
    - These deeper, tied models achieve **better MLM loss** than naive bi-directional Mamba (without weight tying, shallower models).
  - **Effect of RC equivariance:**
    - RC-equivariant language modeling (as in Caduceus-PS) leads to **improved MLM loss** across sequence lengths compared to non-equvariant models.
    - This suggests that encoding RC symmetry directly into the model improves pretraining quality, not just downstream metrics.
- **Genomics Benchmarks (8 classification tasks)**
  - Baselines:
    - CNN trained from scratch, HyenaDNA, non-RC-equvariant Mamba and Caduceus backbones.
  - Key findings:
    - Across all eight tasks, **Caduceus variants attain the best or near-best accuracy**.

        - **Caduceus-Ph** often provides the strongest performance, sometimes slightly surpassing Caduceus-PS and non-equvariant Caduceus.
            - Example trends (qualitative):
                - Mouse enhancers: Caduceus models outperform CNN and HyenaDNA, with Caduceus-PS performing best among Caduceus variants.
                - Human enhancer and promoter tasks: Caduceus models typically exceed Mamba-only and HyenaDNA baselines.
        - The results show that combining long-range SSMs with bi-directionality and RC handling is consistently beneficial on regulatory sequence tasks.
- **Nucleotide Transformer benchmark (18 tasks)**
    - Baselines:
        - Large Transformer-based models: Enformer (~252M parameters), DNABERT-2 (~117M), Nucleotide Transformer v2 (~500M).
        - HyenaDNA (~1.6M parameters).
    - Metrics: MCC for histone marks/enhancers, F1 for promoters and splice annotation, accuracy for "splice sites all".
    - Key findings:
        - **Caduceus-Ph and Caduceus-PS (~1.9M parameters)** perform **competitively with much larger Transformers**, particularly on histone marks and regulatory annotation tasks.
        - Caduceus models generally **outperform HyenaDNA** on most histone and regulatory tasks.
        - HyenaDNA remains strong on some splice site annotation tasks, where Caduceus performance is more mixed.
        - Overall, Caduceus demonstrates that carefully designed SSM-based FMs with biological inductive biases can rival or outperform huge Transformers on many genomics tasks at a fraction of the parameter count.
- **Variant Effect Prediction (VEP) on gene expression**
    - Setup:
        - Use embeddings from Caduceus, HyenaDNA, Nucleotide Transformer v2, and Enformer.
        - For each SNP, extract embeddings from a window centered at the SNP (e.g., 1,536 bp window for SSM models, shorter effective windows for Transformer baselines), optionally concatenated with tissue information.
        - Train an **SVM with RBF kernel** to classify whether the SNP is causal for gene expression, stratifying by distance to the nearest transcription start site (TSS): short (0–30k bp), medium (30–100k bp), long (>100k bp).
    - Key findings:
        - **Caduceus models consistently outperform HyenaDNA** across distance buckets.
        - **Caduceus-PS** often **matches or exceeds Nucleotide Transformer v2 (500M params)**, especially at **long ranges**.
        - For SNPs >100k bp from TSS, **Caduceus even surpasses Enformer**, which is a strong, supervised baseline with large receptive fields.
        - These trends demonstrate that Caduceus' long-range, bi-directional, RC-aware representations are especially powerful when long genomic contexts matter most.

---

8. Strengths, Limitations, and Open Questions

- **Strengths**
  - **Biologically grounded inductive biases:**
    - Incorporates **bi-directionality** and **RC equivariance** directly into the architecture, aligning model invariances with DNA's physical properties.
  - **Long-context efficiency:**
    - Uses SSM-based Mamba blocks to handle sequences up to ~131k bps and beyond with **linear-time scaling**, enabling realistic genomic context sizes.
  - **Strong empirical performance:**
    - Outperforms or matches both other SSM-based models (HyenaDNA) and much larger Transformers (Nucleotide Transformer, Enformer) on many benchmarks.
    - Particularly strong on **long-range variant effect prediction**, a high-impact biological task.
  - **Clear architectural story:**
    - Breaks down innovations into modular components (BiMamba, MambaDNA, Caduceus-PS/Ph), making it easier to reuse or extend in other settings.
- **Limitations**
  - **Single-modality focus:**
    - Only models DNA sequence; does not integrate other relevant data (RNA expression, chromatin contact maps, clinical covariates).
  - **Task coverage:**
    - Evaluations focus on classification tasks and VEP; less attention to generative uses (e.g., sequence design) or interpretability tools (e.g., motif discovery pipelines).
  - **Complexity of RC handling:**
    - While the theory is clean, implementing and debugging RC-equivariant models and post-hoc conjoining pipelines may be non-trivial in practice.
  - **Reliance on external classifiers for VEP:**
    - The VEP evaluation uses SVMs on top of embeddings, which might not fully exploit the capacity of the representations compared to end-to-end fine-tuning.
- **Open Questions and Future Directions:**
  - **Multimodal integration:**
    - How do Caduceus embeddings combine with other modalities (e.g., expression, epigenomics, imaging) in late-fusion or contrastive frameworks, as suggested in the integration baseline plan?
  - **Interpretability:**
    - Can the RC-equivariant and bi-directional structure be exploited for clearer attribution of variants, motif discovery, or mechanistic interpretation?
  - **Scaling laws:**
    - How do performance and sample efficiency scale with model size and context length for Mamba-based DNA FMs compared to Transformers?
  - **End-to-end VEP:**
    - Does training end-to-end predictors on top of Caduceus (instead of SVMs) further improve performance and robustness?

    - **Generalization across species and assays:**
        - How well do Caduceus models transfer to other species, tissue types, or new assay types (e.g., single-cell, new histone marks)?

---

9. Context and Broader Impact

- **Position in the FM landscape**
    - In the **genomics FM** space, Caduceus is part of a movement from Transformer-based models (DNABERT, Nucleotide Transformer, Enformer) toward **efficient long-range architectures** like HyenaDNA and SSM-based FMs.
    - Conceptually, you can think of Caduceus as **"like BERT but for DNA, built on Mamba instead of a Transformer, and explicitly aware of DNA's symmetries"**.
    - It complements work on large language models in biology (e.g., protein LMs) by focusing on non-coding genomic sequence and variant effects.
- **Relation to well-known ideas**
    - **Bi-directionality** echoes BERT and ELMo, which improved NLP by considering both left and right context during pretraining.
    - **Equivariance** connects to broader ML trends of embedding symmetry (e.g., rotational equivariance in vision) into architectures to improve data efficiency and generalization.
    - The use of SSMs like Mamba fits into a growing ecosystem of non-attention-based, long-context sequence learners.
- **Connection to the integration baseline plan**
    - The integration baseline plan explicitly cites **Caduceus** as a reference for **genetics embedding hygiene**:
        - RC handling (e.g., RC averaging, post-hoc conjoining) and stable tokenization are central to trustworthy variant-level modeling.
    - For multimodal or clinical integration projects, Caduceus-style encoders can serve as **frozen or lightly fine-tuned DNA feature extractors**, feeding into late-fusion models or contrastive frameworks as described in the plan.
    - The evaluation discipline in this paper (clear splits, cross-validation, AUROC/AUPRC reporting) is aligned with the plan's emphasis on robust, statistically careful evaluation.

---

10. Key Takeaways (Bullet Summary)

- **Problem**
    - Long-range DNA sequence modeling needs models that can handle **very long contexts**, use **bi-directional information**, and respect **reverse complement symmetry**.
    - Existing models either struggle with scaling (Transformers) or ignore key biological invariances (many sequence models).
- **Method / Model**
    - The authors introduce **BiMamba**, a parameter-efficient bi-directional extension of the Mamba SSM, and **MambaDNA**, an RC-equivariant module built around Mamba/BiMamba.

    - They combine these components into **Caduceus**, a family of **DNA foundation models** trained with masked language modeling on the human genome.
    - Two main variants, **Caduceus-PS** (RC-equivariant via parameter sharing) and **Caduceus-Ph** (post-hoc RC conjoining), provide different trade-offs between strict equivariance and practical performance.
- **Results**
    - Mamba-based models achieve **better pretraining loss** than HyenaDNA and benefit from weight tying and RC equivariance.
    - On **Genomics Benchmarks**, Caduceus variants consistently outperform CNN, HyenaDNA, and non-RC Mamba baselines across eight regulatory tasks.
    - On the **Nucleotide Transformer benchmark**, Caduceus (~1.9M params) matches or surpasses much larger Transformer models on many histone and regulatory tasks while beating HyenaDNA on most of them.
    - For **variant effect prediction**, Caduceus models outperform HyenaDNA and **Caduceus-PS exceeds the performance of a 500M-parameter Nucleotide Transformer and even Enformer at long distances to TSS**.
- **Why it matters**
    - Caduceus shows that carefully designed, symmetry-aware, long-range SSM-based architectures can function as powerful **genomic foundation models** at modest scale.
    - For a grad student, this paper is a template for combining **domain knowledge (DNA symmetries)** with **modern sequence modeling techniques** to achieve strong performance on biologically important tasks like variant effect prediction.

Generated via custom pipeline · 2025-11-18