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

Advanced Biomedicine Seminar

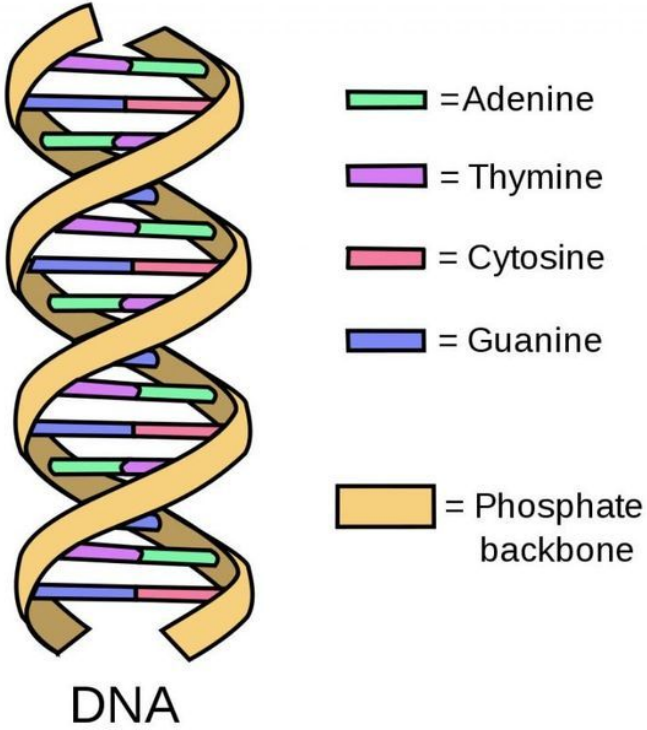
28/11/2024 - John Shahla

Roadmap

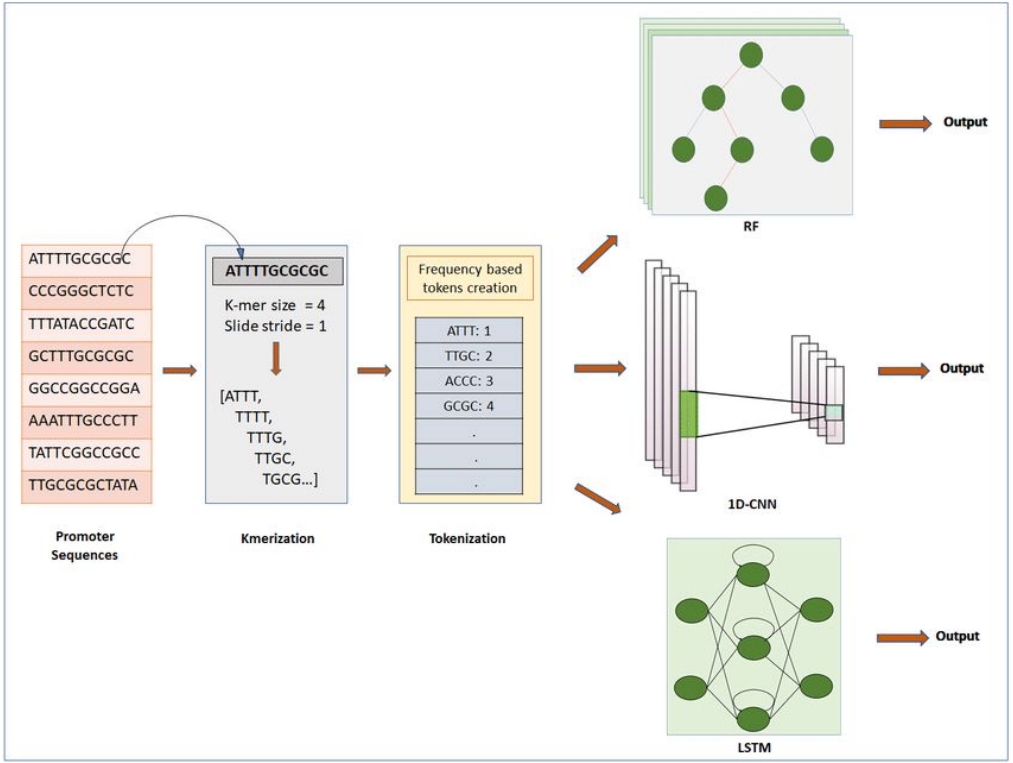
- Definitions
- Motivation
- Models
- Results

Definitions

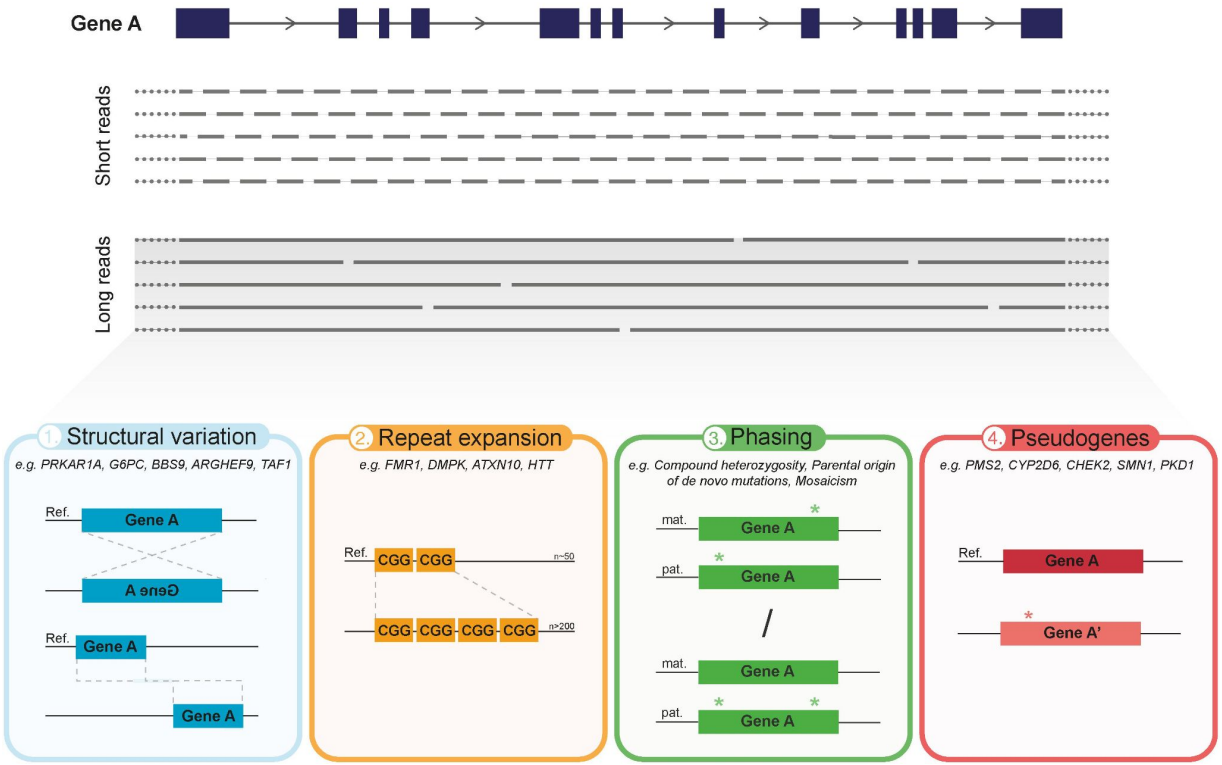
DNA Structure



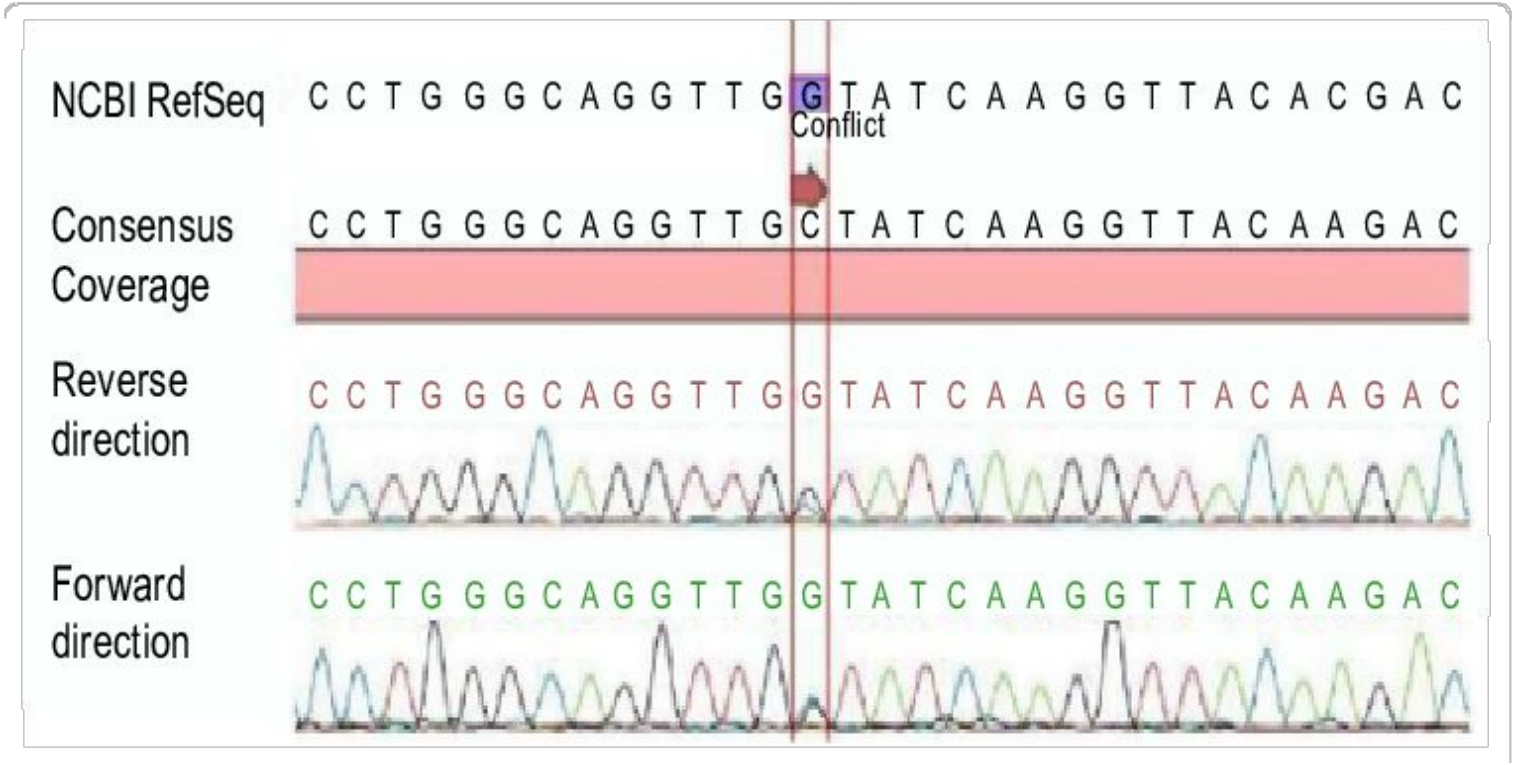
DNA Sequence Modelling



Long-Range DNA Sequencing



Bi-Directional Sequencing



Reverse Complementarity (RC)

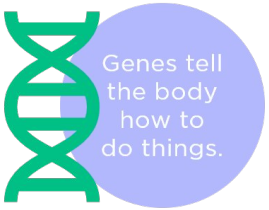
Writing sequences

- Written 5’-3’
  - ATGGGTAGCGGTCATGATAC
- Complement
  - TACCCATCGCCAGTACTATG
- Reverse (inverse)
  - CATAGTACTGGCGATGGGTA
- Reverse complement
  - GTATCATGACCGCTACCCAT

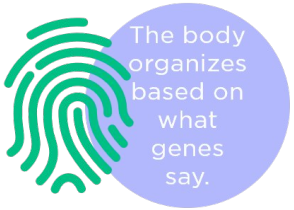
Upstream and downstream

Phenotypes

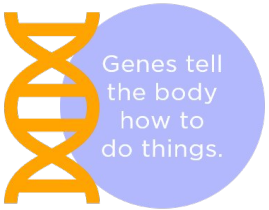
Genotype: DNA



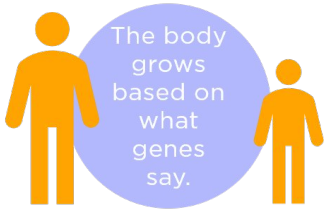
Phenotype: Thumbprint



Genotype: DNA



Phenotype: Height



Caduceus use

Variant effect prediction: a task to detect whether a genetic mutation influences a phenotype

Motivation

- Bi-Directionality
- Reverse Complementarity
- Long-Range Dependencies
- Limitations of Existing Models

Models

Mamba

- Processes DNA in long-range sequences.
- Recognizes reverse complement strands.

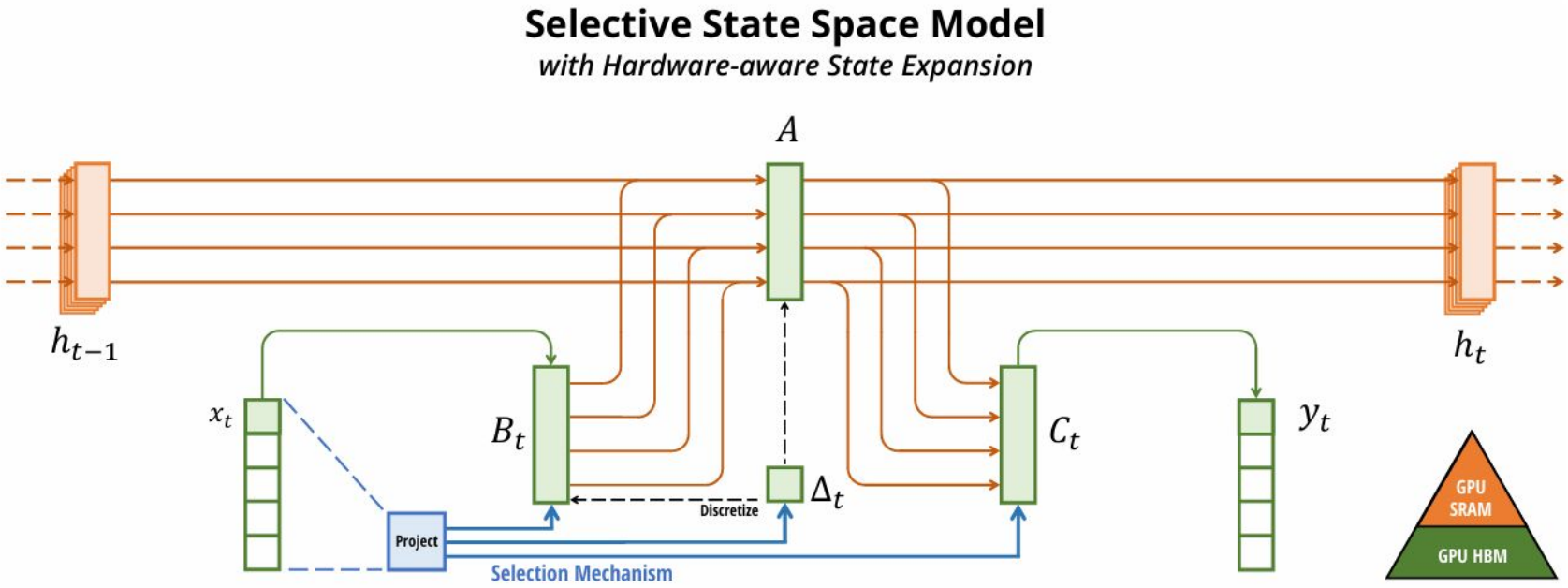
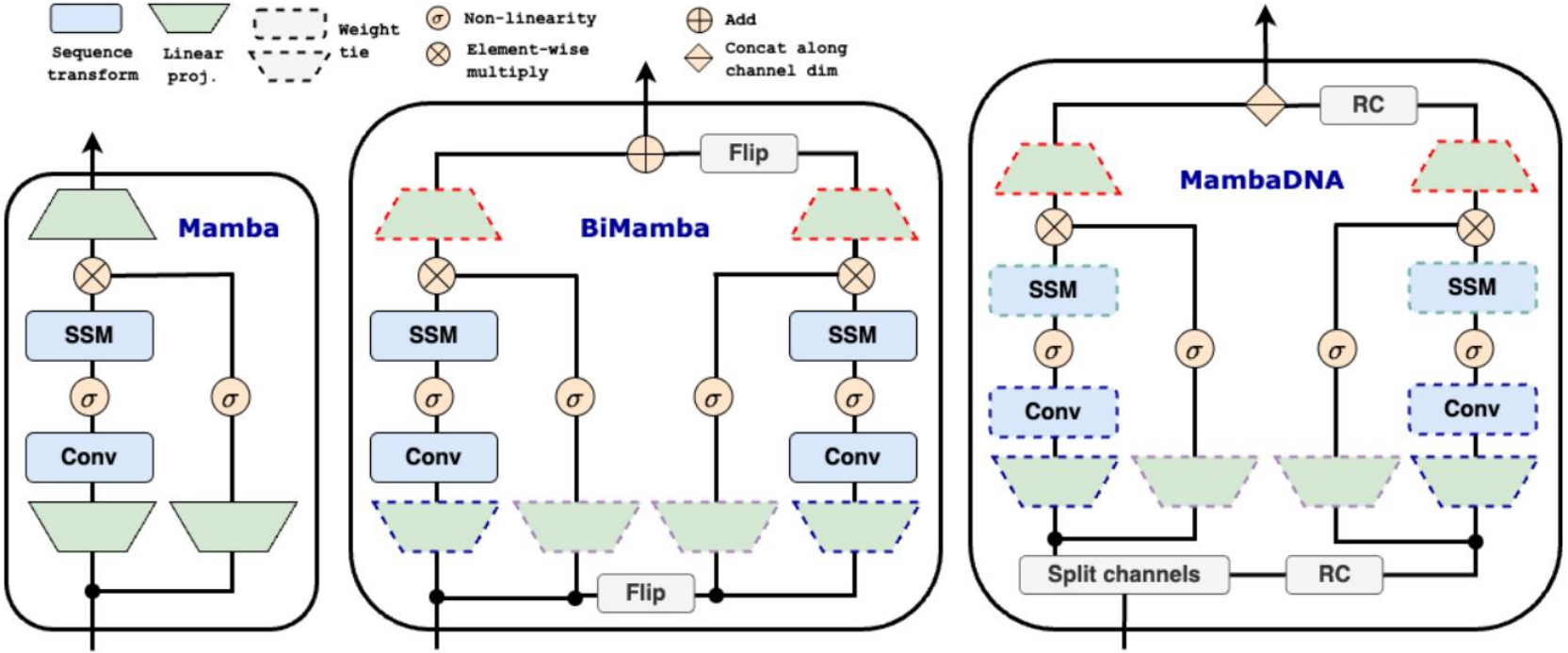
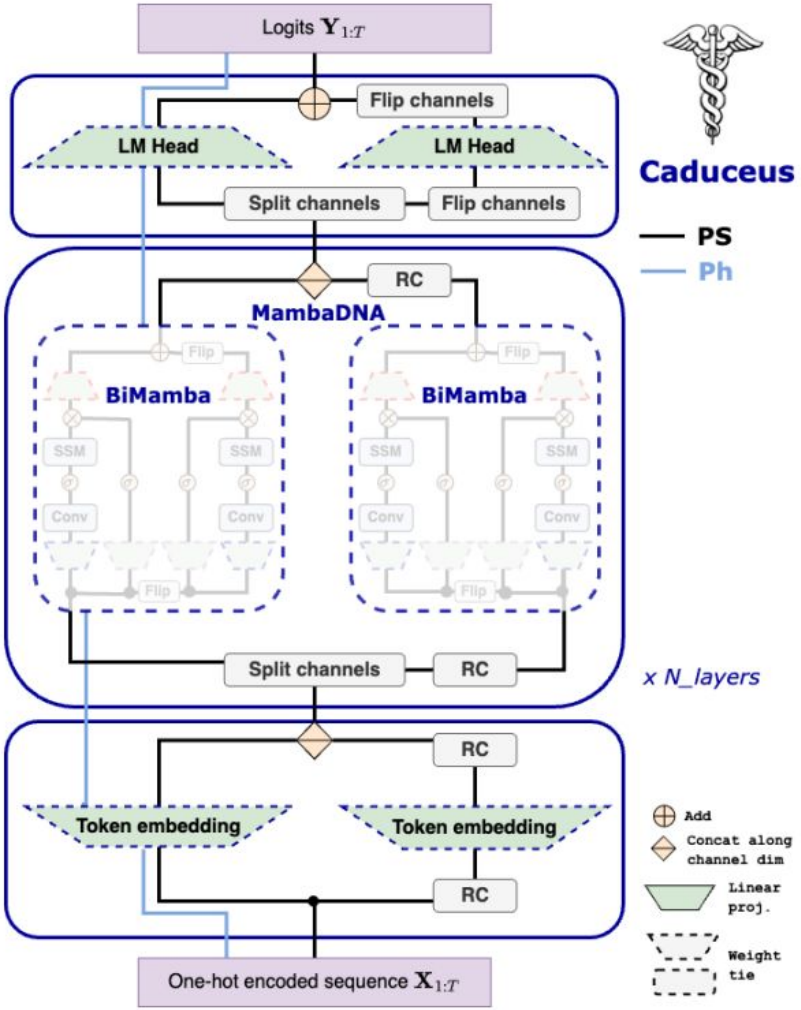


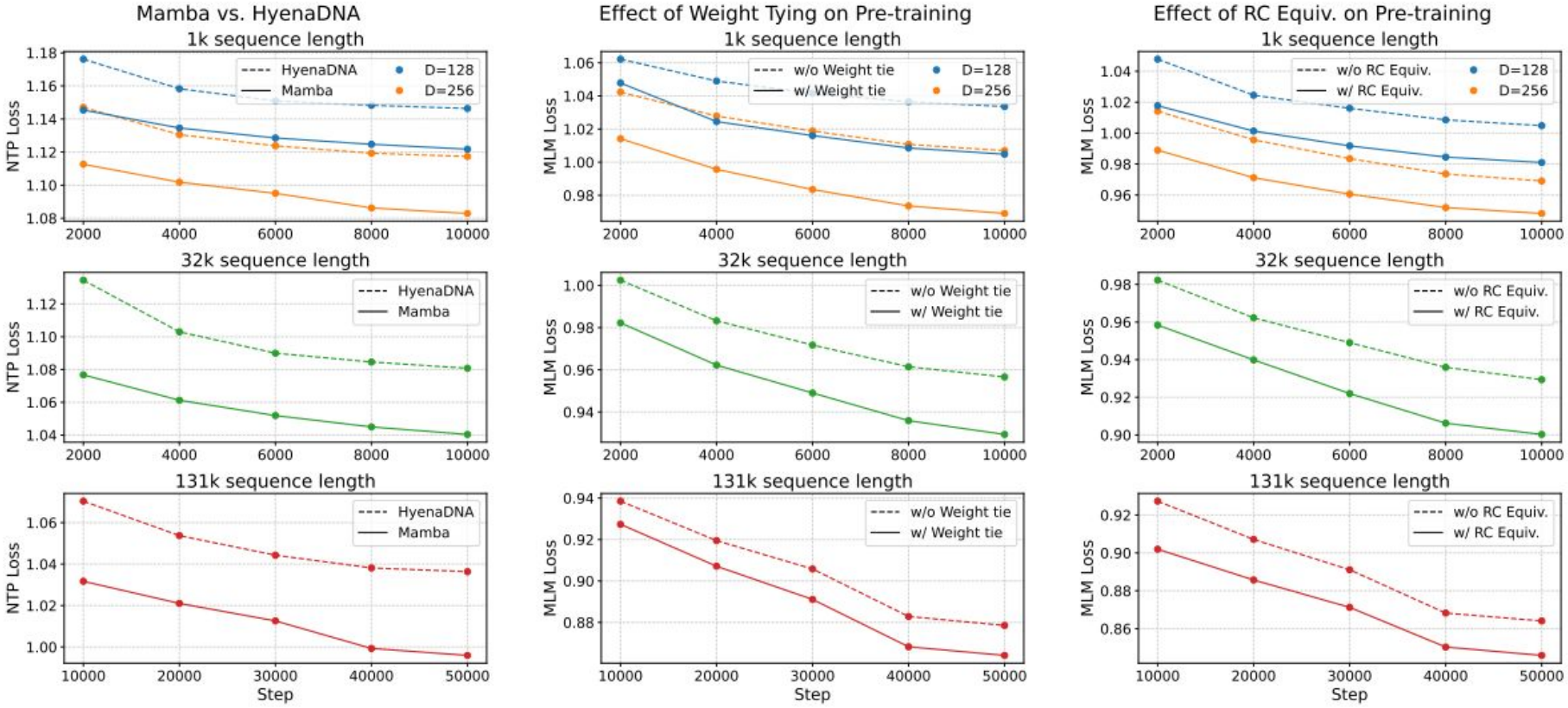
Figure 1: (**Overview**.) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the (Δ, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$) parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

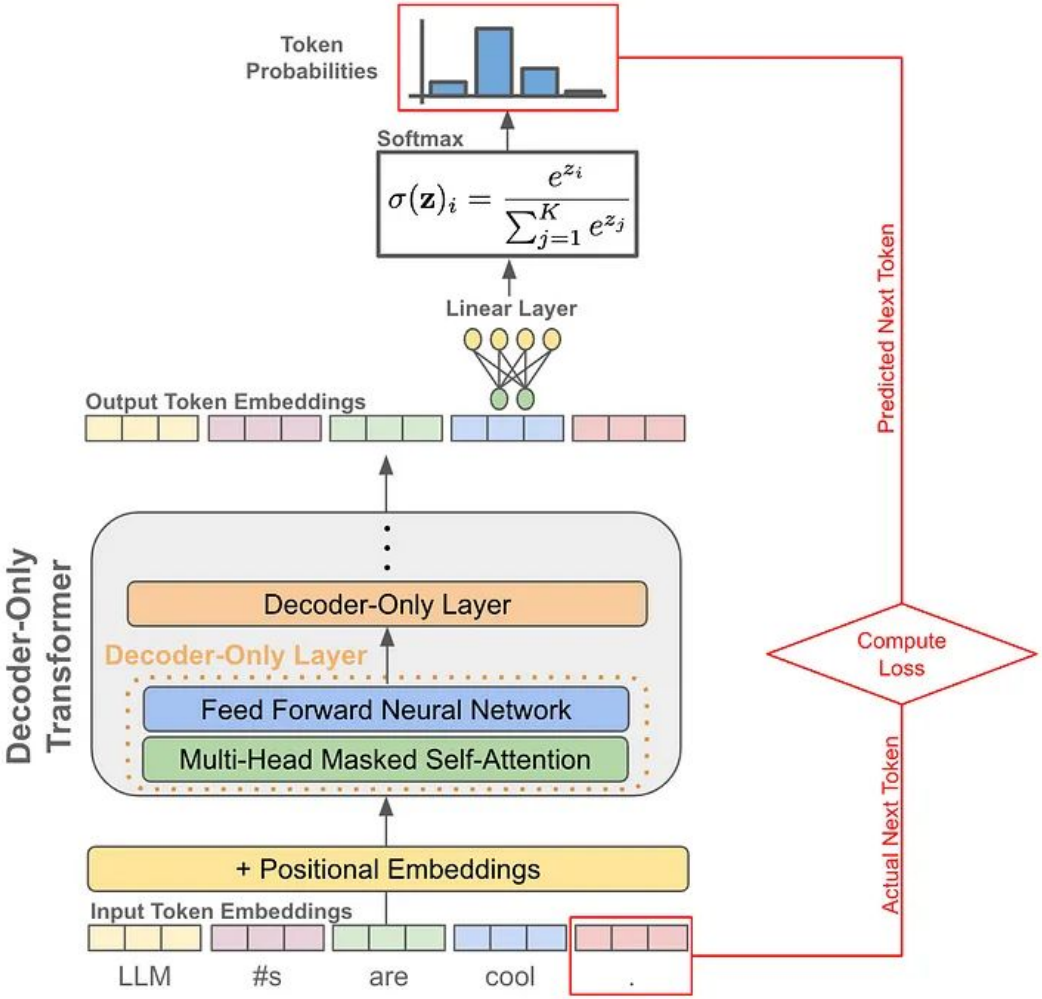

Sequence transform
Linear proj.
Weight tie
Non-linearity
Element-wise multiply
Add
Concat along channel dim
Mamba
SSM
Conv
BiMamba
Flip
MambaDNA
RC
Split channels

Caduceus

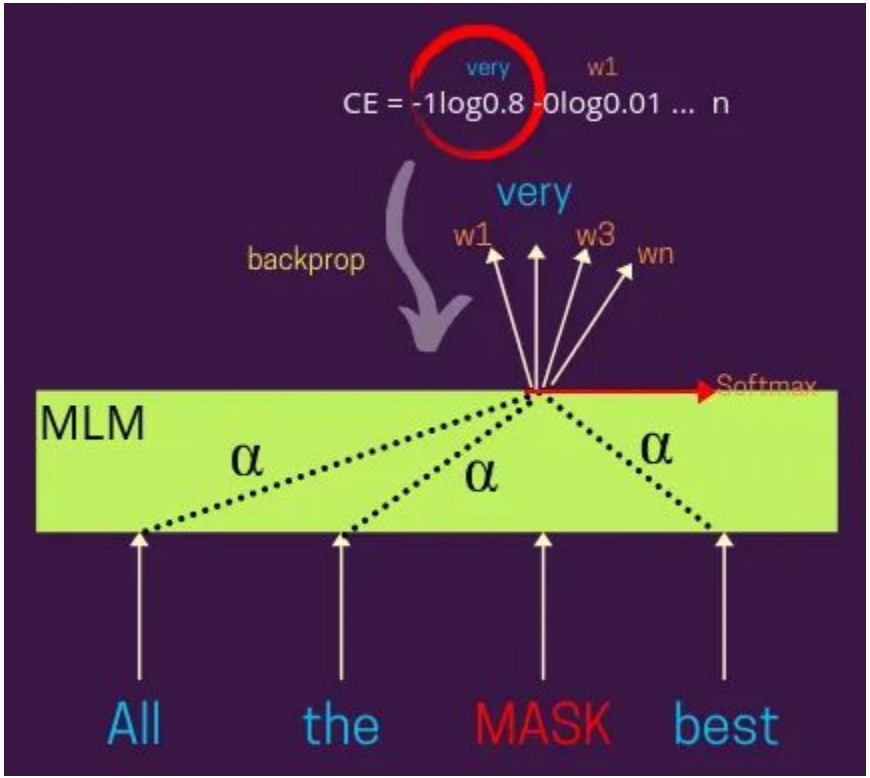

Logits $\mathbf{Y}_{1:T}$
Flip channels
LM Head
LM Head
Split channels
Flip channels
RC
MambaDNA
Flip
BiMamba
SSM
SSM
Conv
Conv
Flip
Flip
BiMamba
SSM
SSM
Conv
Conv
Flip
Split channels
RC
x N_layers
RC
Token embedding
Token embedding
RC
One-hot encoded sequence $\mathbf{X}_{1:T}$
Caduceus
PS
Ph
Add
Concat along channel dim
Linear proj.
Weight tie

Results


Mamba vs. HyenaDNA
1k sequence length
32k sequence length
131k sequence length
NTP Loss
HyenaDNA
Mamba
D=128
D=256
Step
(a)
Effect of Weight Tying on Pre-training
1k sequence length
32k sequence length
131k sequence length
MLM Loss
w/o Weight tie
w/ Weight tie
D=128
D=256
Step
(b)
Effect of RC Equiv. on Pre-training
1k sequence length
32k sequence length
131k sequence length
MLM Loss
w/o RC Equiv.
w/ RC Equiv.
D=128
D=256
Step
(c)


Token Probabilities
Softmax
$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
Linear Layer
Output Token Embeddings
Predicted Next Token
Decoder-Only Transformer
Decoder-Only Layer
Decoder-Only Layer
Feed Forward Neural Network
Multi-Head Masked Self-Attention
Compute Loss
+ Positional Embeddings
Input Token Embeddings
Actual Next Token
LLM
#s
are
cool
.


very
MLM
All
the
MASK
best


very
w1
CE = -1log0.8 -0log0.01 ... n
very
backprop
w1
w3
wn
Softmax
MLM
α
α
α
All
the
MASK
best

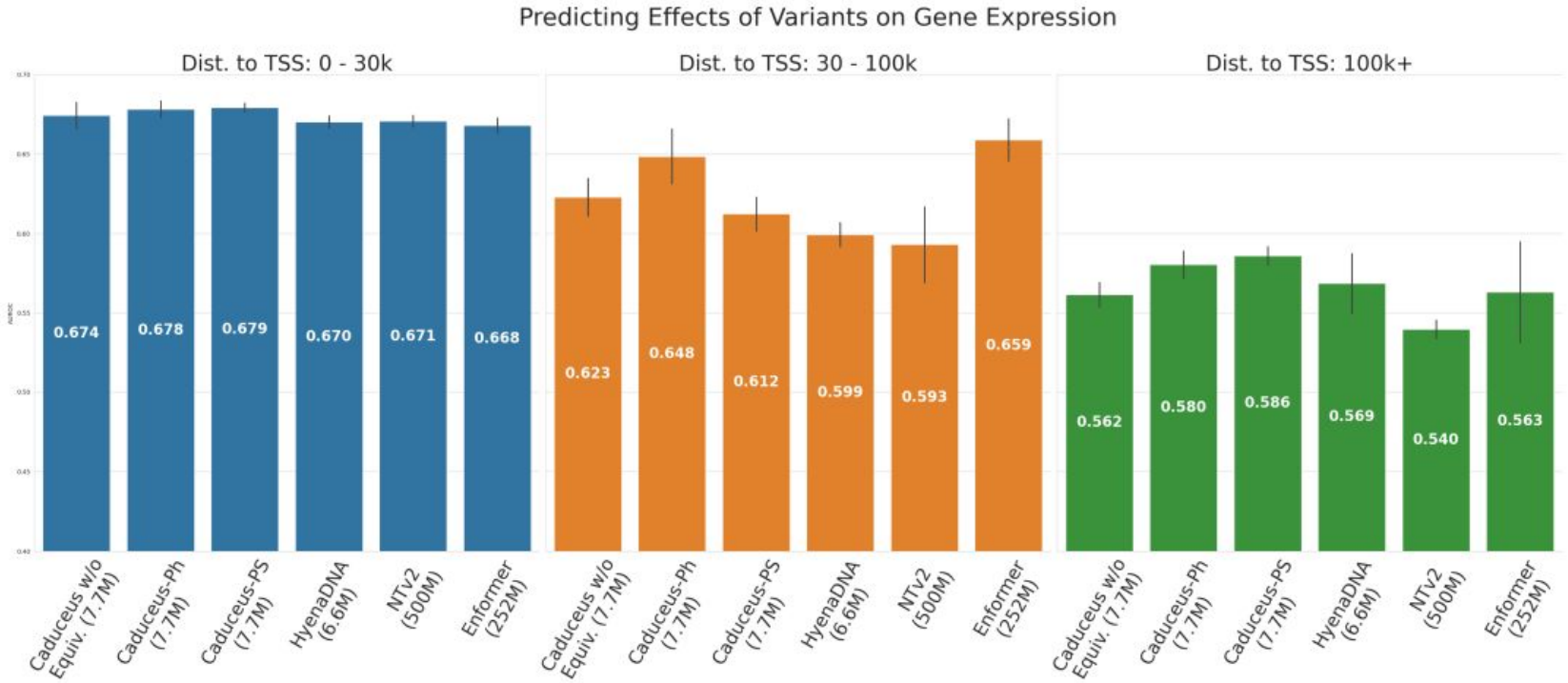
Genomics Benchmarks

	CNN (264K)	HyenaDNA (436K)	Mamba (468K)	Caduceus w/o Equiv. (470K)	Caduceus-Ph (470K)	Caduceus-PS (470K)
Mouse Enhancers	0.715 ±0.087	*0.780* ±0.025	0.743 ±0.054	0.770 ±0.058	0.754 ±0.074	**0.793** ±0.058
Coding vs. Intergenomic	0.892 ±0.008	0.904 ±0.005	0.904 ±0.004	0.908 ±0.003	**0.915** ±0.003	*0.910* ±0.003
Human vs. Worm	0.942 ±0.002	0.964 ±0.002	0.967 ±0.002	*0.970* ±0.003	**0.973** ±0.001	0.968 ±0.002
Human Enhancers Cohn	0.702 ±0.021	0.729 ±0.014	0.732 ±0.029	0.741 ±0.008	**0.747** ±0.004	*0.745* ±0.007
Human Enhancer Ensembl	0.744 ±0.122	0.849 ±0.006	0.862 ±0.008	0.883 ±0.002	*0.893* ±0.008	**0.900** ±0.006
Human Regulatory	0.872 ±0.005	0.869 ±0.012	0.814 ±0.211	0.871 ±0.007	*0.872* ±0.011	**0.873** ±0.007
Human OCR Ensembl	0.698 ±0.013	0.783 ±0.007	0.815 ±0.002	0.818 ±0.003	**0.828** ±0.006	*0.818* ±0.006
Human NonTATA Promoters	0.861 ±0.009	0.944 ±0.002	0.933 ±0.007	0.933 ±0.006	**0.946** ±0.007	*0.945* ±0.010

Nucleotide Transformer Tasks

	> 100M Param. Models			< 2M Param. Models		
	Enformer (252M)	DNABERT-2 (117M)	NT-v2 (500M)	HyenaDNA (1.6M)	Caduceus-Ph (1.9M)	Caduceus-PS (1.9M)
Histone Markers						
H3	0.719±0.048	0.785±0.033	0.784±0.047	0.779±0.037	**0.815**±0.048	*0.799*±0.029
H3k14ac	0.288±0.077	0.516±0.028	0.551±0.021	*0.612*±0.065	**0.631**±0.026	0.541±0.212
H3k36me3	0.344±0.055	0.591±0.020	**0.625**±0.013	*0.613*±0.041	0.601±0.129	0.609±0.109
H3k4me1	0.291±0.061	0.511±0.028	**0.550**±0.021	0.512±0.024	*0.523*±0.039	0.488±0.102
H3k4me2	0.211±0.069	0.336±0.040	0.319±0.045	*0.455*±0.095	**0.487**±0.170	0.388±0.101
H3k4me3	0.158±0.072	0.352±0.077	0.410±0.033	**0.549**±0.056	*0.544*±0.045	0.440±0.202
H3k79me3	0.496±0.042	0.613±0.030	0.626±0.026	0.672±0.048	**0.697**±0.077	*0.676*±0.026
H3K9ac	0.420±0.063	0.542±0.029	0.562±0.040	0.581±0.061	**0.622**±0.030	*0.604*±0.048
H4	0.732±0.076	0.796±0.027	*0.799*±0.025	0.763±0.044	**0.811**±0.022	0.789±0.020
H4ac	0.273±0.063	0.463±0.041	0.495±0.032	*0.564*±0.038	**0.621**±0.054	0.525±0.240
Regulatory Annotation						
Enhancer	0.451±0.108	0.516±0.098	**0.548**±0.144	0.517±0.117	*0.546*±0.073	0.491±0.066
Enhancer types	0.309±0.134	0.423±0.051	*0.424*±0.132	0.386±0.185	**0.439**±0.054	0.416±0.095
Promoter: All	0.954±0.006	*0.971*±0.006	**0.976**±0.006	0.960±0.005	0.970±0.004	0.967±0.004
NonTATA	0.955±0.010	*0.972*±0.005	**0.976**±0.005	0.959±0.008	0.969±0.011	0.968±0.006
TATA	*0.960*±0.023	0.955±0.021	**0.966**±0.013	0.944±0.040	0.953±0.016	0.957±0.015
Splice Site Annotation						
All	0.848±0.019	0.939±0.009	**0.983**±0.008	*0.956*±0.011	0.940±0.027	0.927±0.021
Acceptor	0.914±0.028	*0.975*±0.006	**0.981**±0.011	0.958±0.010	0.937±0.033	0.936±0.077
Donor	0.906±0.027	*0.963*±0.006	**0.985**±0.022	0.949±0.024	0.948±0.025	0.874±0.289

Gene Expression


Predicting Effects of Variants on Gene Expression
Dist. to TSS: 0 - 30k
Dist. to TSS: 30 - 100k
Dist. to TSS: 100k+
0.674
0.678
0.679
0.670
0.671
0.668
0.623
0.648
0.612
0.599
0.593
0.659
0.562
0.580
0.586
0.569
0.540
0.563
Caduceus w/o Equiv. (7.7M)
Caduceus-Ph (7.7M)
Caduceus-PS (7.7M)
HyenaDNA (6.6M)
NTv2 (500M)
Enformer (252M)

Any Questions?

Thank you