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

Yair Schiff

Chia-Hsiang Kao

Aaron Gokaslan

Tri Dao

Albert Gu

Volodymyr Kuleshov

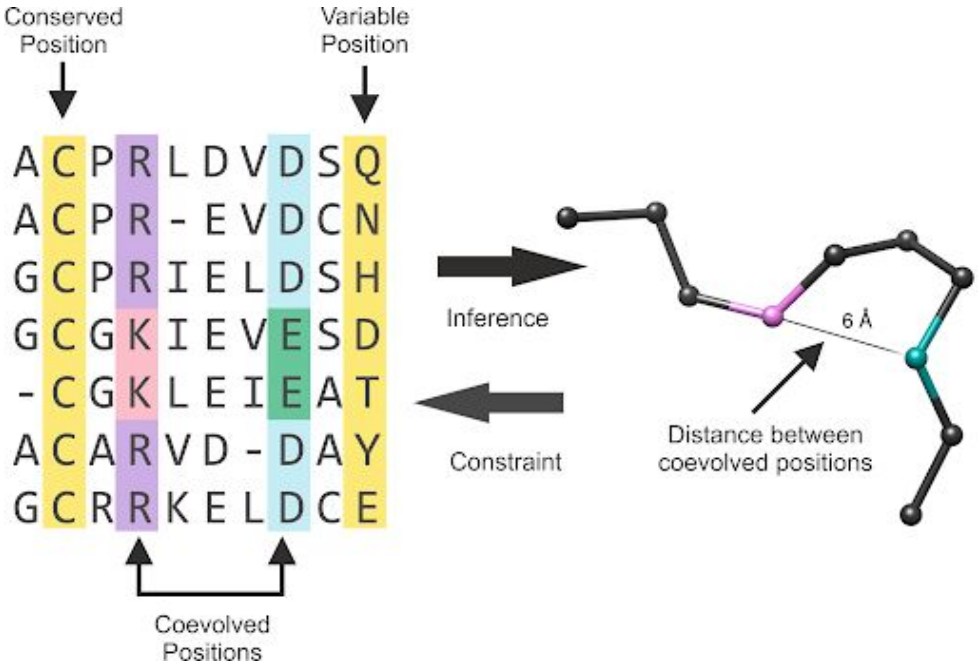
ICML International Conference On Machine Learning

July 2024

Motivation

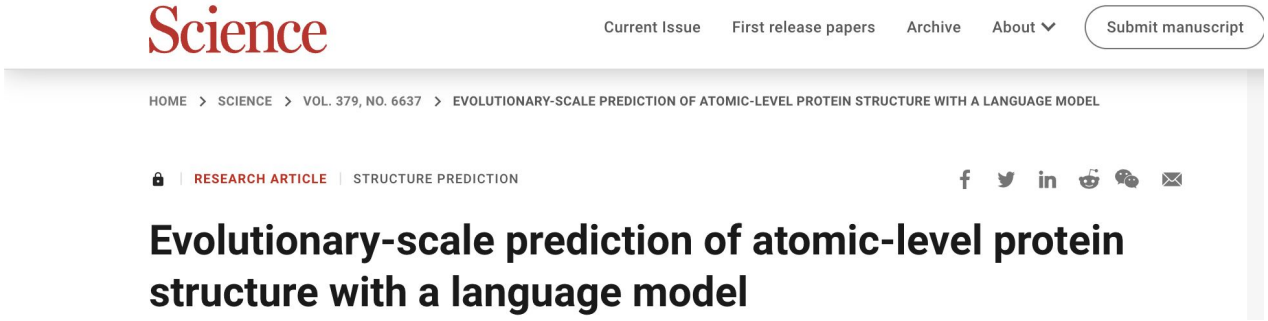
Why Biological Foundation Models?

A generative model over sequences can learn conservation and evolutionary pressure.



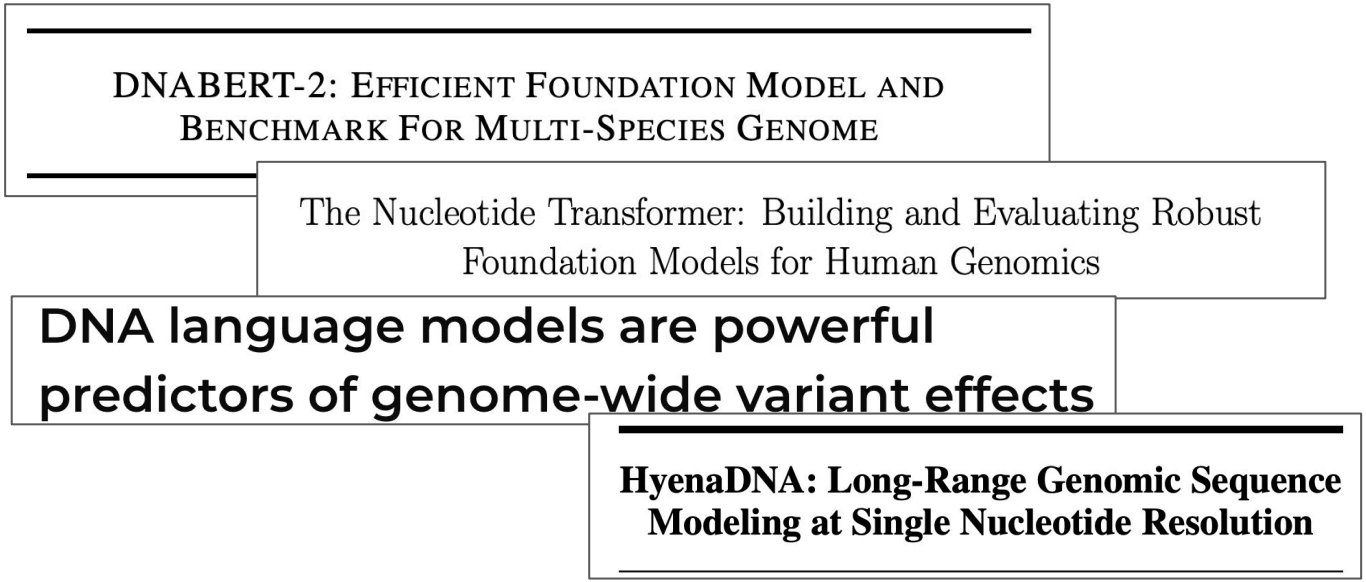
These models can improve downstream tasks such as variant effect prediction, protein folding, etc.

Existing Applications of Foundation Models in Biology



Why DNA Foundation Models?

- Extend to non-coding and regulatory regions of the genome
- Solve tasks that protein models cannot solve, e.g., gene annotation



Challenges

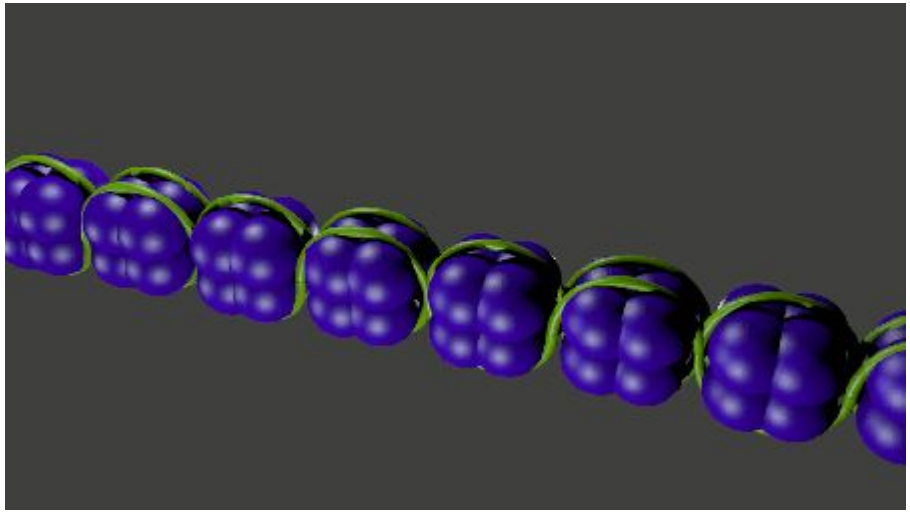
Long-range interactions

Bi directionality

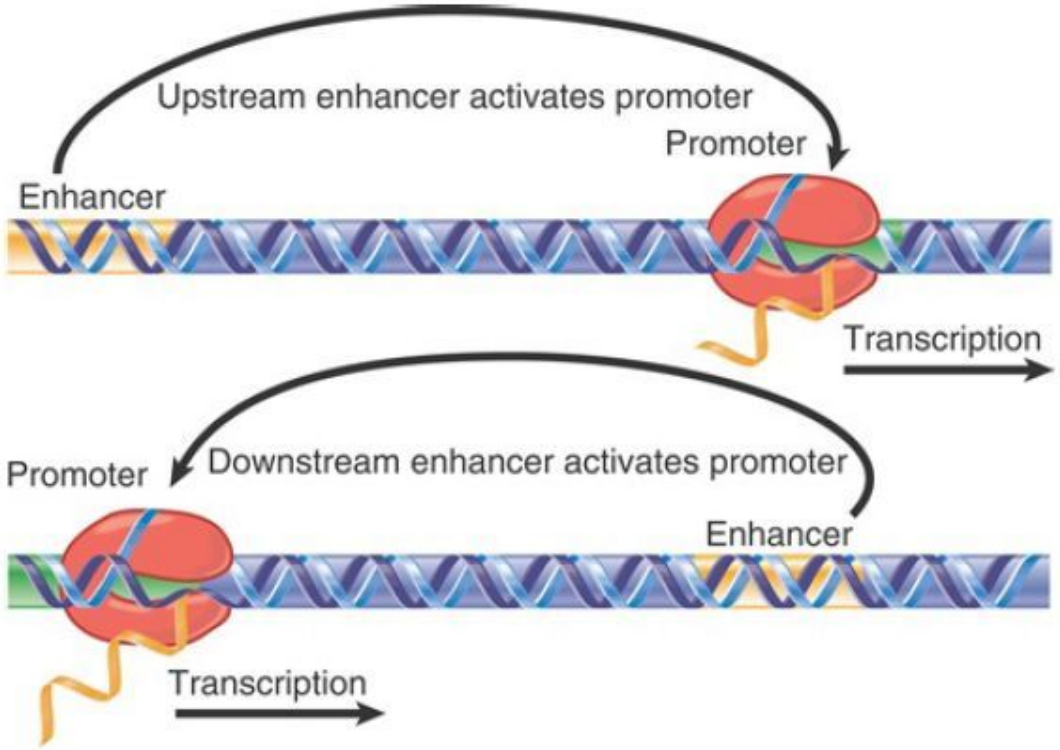
Reverse complement strands

Distal interactions

- Unlike proteomics, genomics requires modeling distal interactions (up to 1M base pairs)



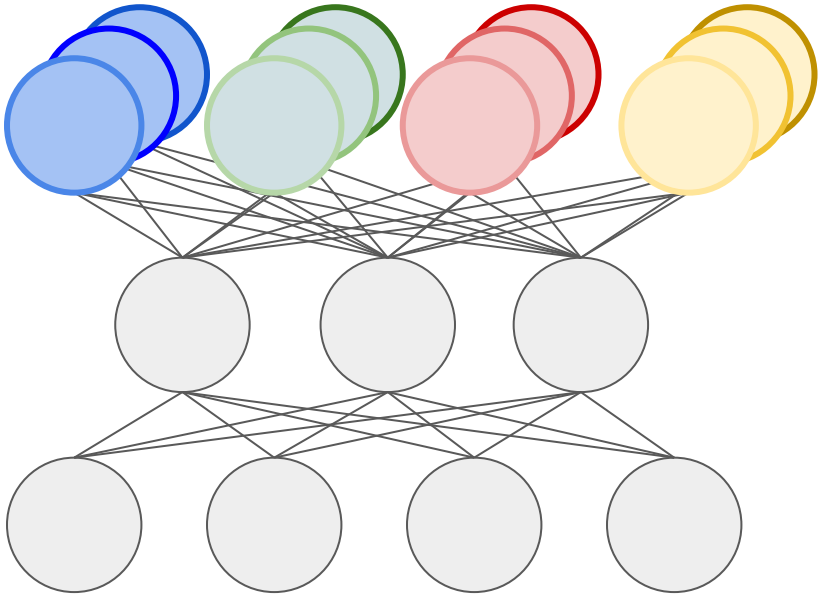
Causal models insufficient to model DNA



Reverse complement (RC) DNA strands contain equivalent information

✅ Models that respect this symmetry have been found to yield improved performance (Zhou et. al, 2021; Mallet et. al, 2021)

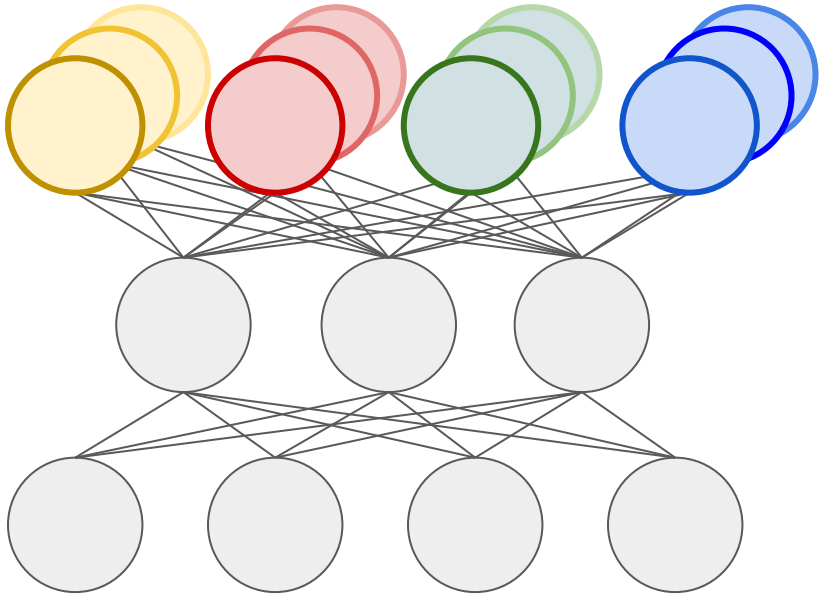
Equivariance: Models commute with RC operation



CATTGT

Equivariance: Models commute with RC operation

RC input corresponds to RC output



ACAATG

CATTGT

Reverse Complement sequences

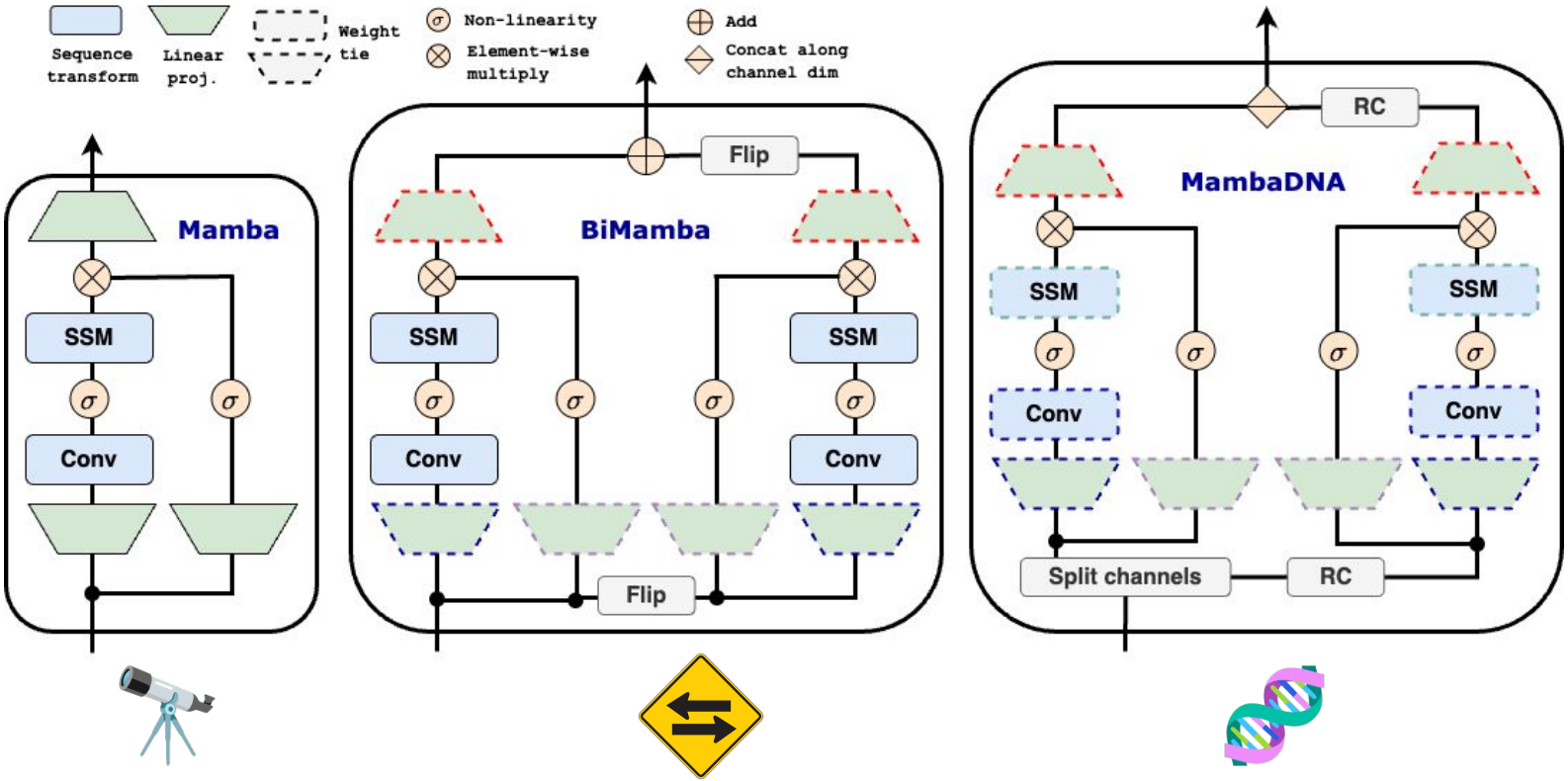
Caduceus

Caduceus highlights

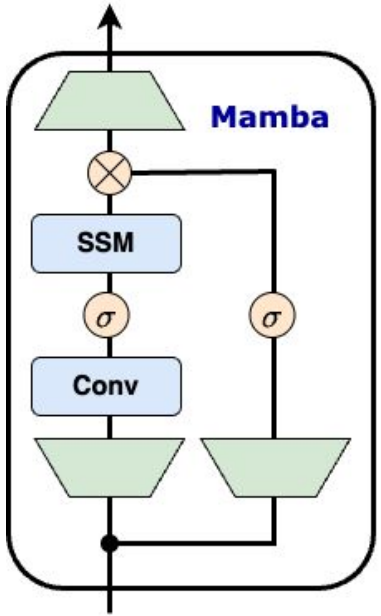
- ✅ Memory-efficient, bi-directional extension of Mamba
- ✅ RC-equivariant language modeling
- ✅ Improved performance over much larger Transformers-based and comparably-sized SSM-based models

	Long Range	Bi-dir.	RC Equivariant
DNA BERT	❌	✅	❌
Nucleotide Transformer	❌	✅	❌
GPN	❌	✅	❌
HyenaDNA	✅	❌	❌
Mamba	✅	❌	❌
Caduceus	✅	✅	✅

Building towards Caduceus



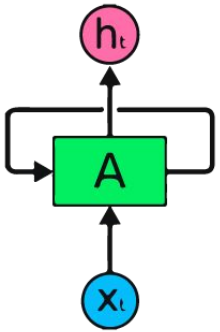
Building towards Caduceus: **Long-range**



Leveraging the original Mamba module takes advantage of the improved **long-range** sequence modeling of this architecture

Two dominant approaches to sequence modeling

Transformers



Recurrent Neural Networks

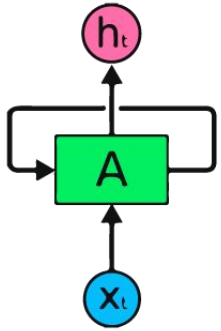
Transformers

✅ Interactions between all elements

✅ Efficient training

❌ Fixed context size

❌ $O(n^2)$ scaling in sequence length



Recurrent Neural Networks

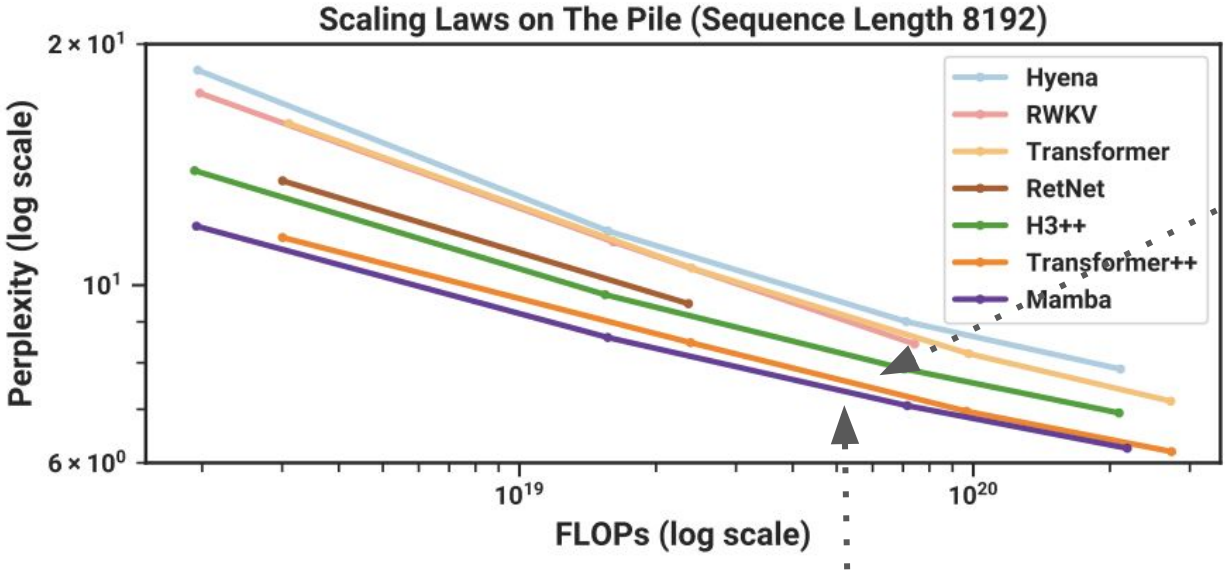
- ✅ ‘Infinite’ context width
- ✅ Linear scaling at inference
- ❌ Fixed-dimensional hidden representations
- ❌ Vanishing / exploding gradients
- ❌ Slow to train

Mamba (and friends)

New hardware-aware architectures targeting large language models

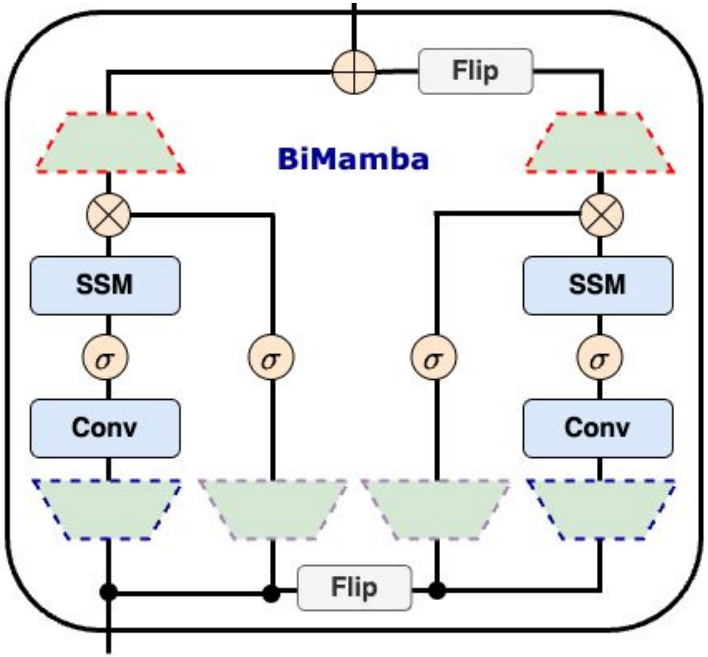
- Mamba (Gu and Dao 2023)
- S5 (Smith et al. 2022)
- Based (Arora et al. 2024)
- Griffin (De et al. 2024)
- GLA (Yang et al. 2023)
- RetNet (Sun et al. 2023)

Why is this important now?



Mamba (Gu and Dao 2023)

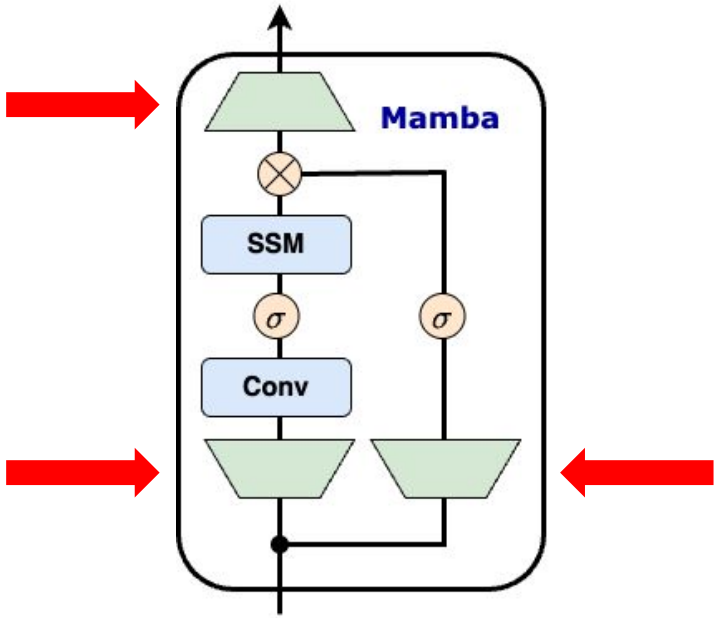
Building towards Caduceus: **Bi-directional**



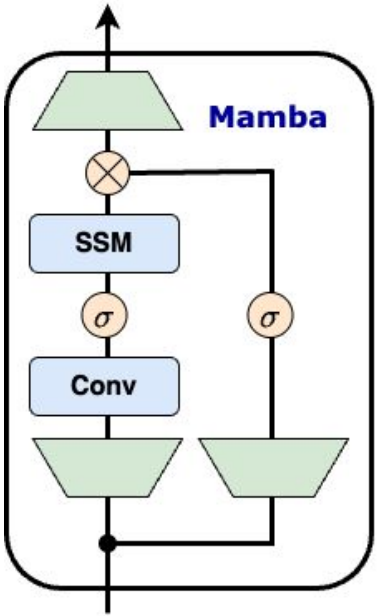
Sharing weights for forward and backward projections enables memory-efficient **bi-directional** sequence modeling

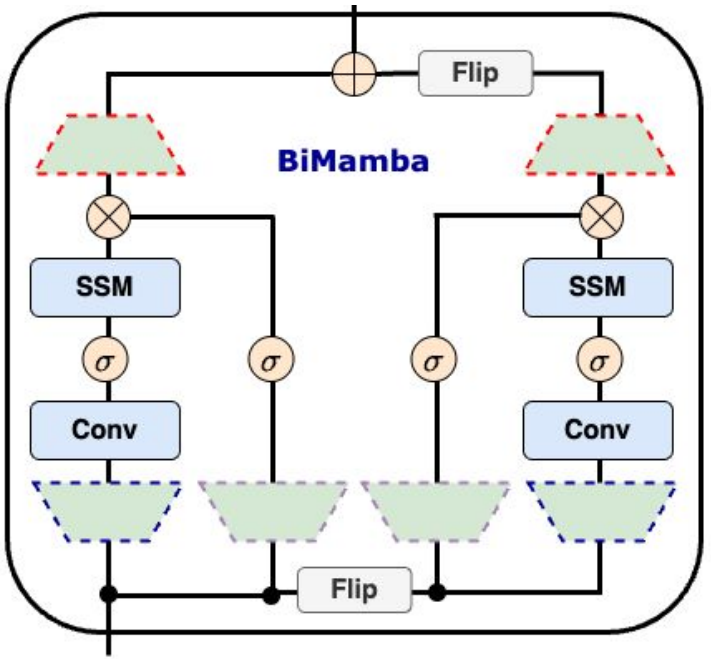
“Strategic” Weight-tieing

Majority of the parameters in a Mamba block come from linear projections

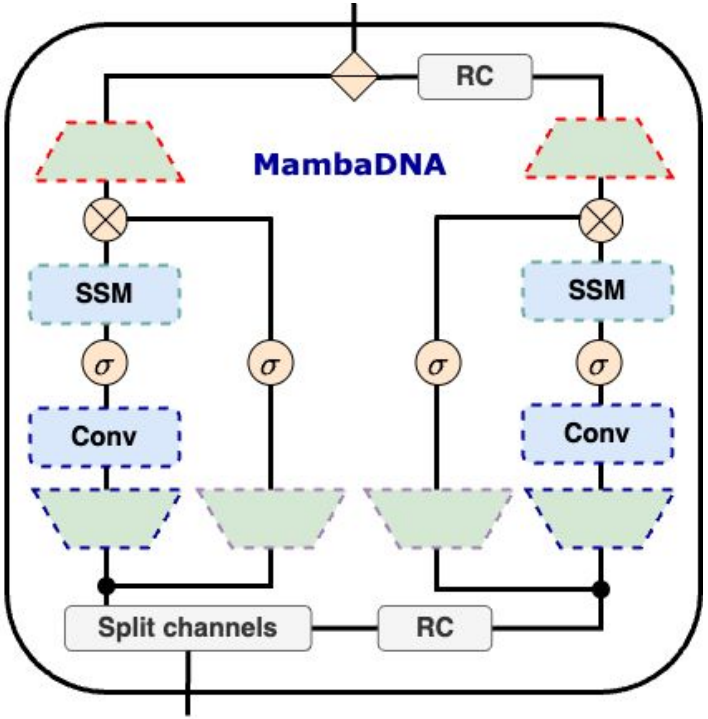


“Strategic” Weight-tieing



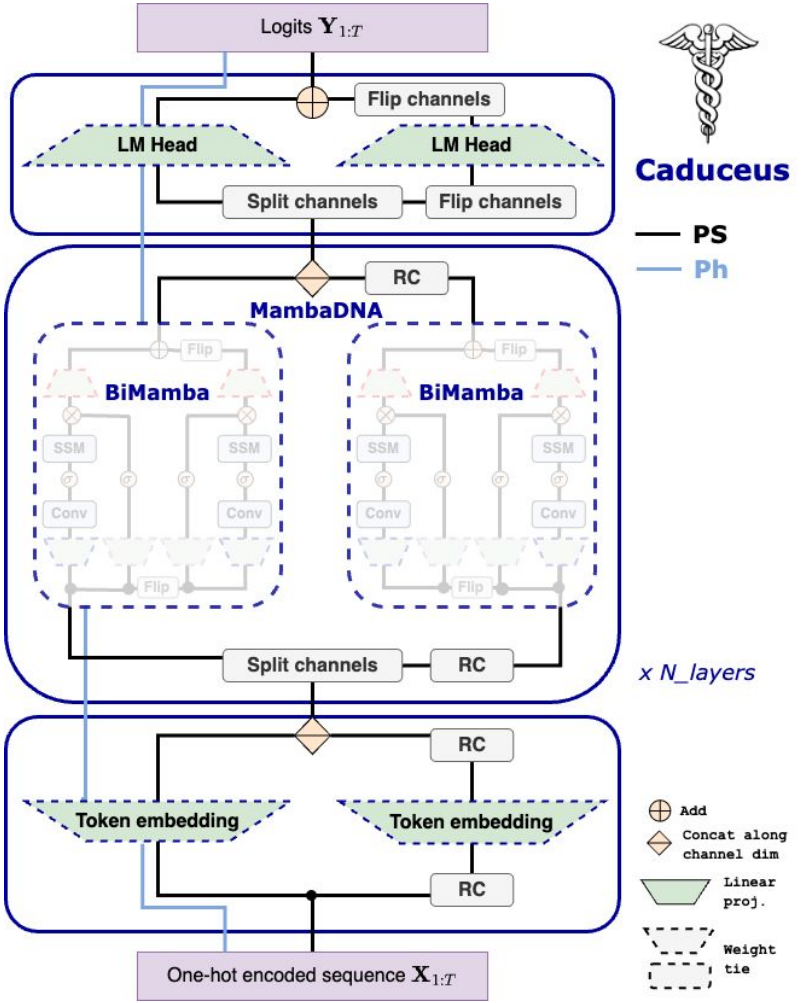


Building towards Caduceus: **RC equivariant**



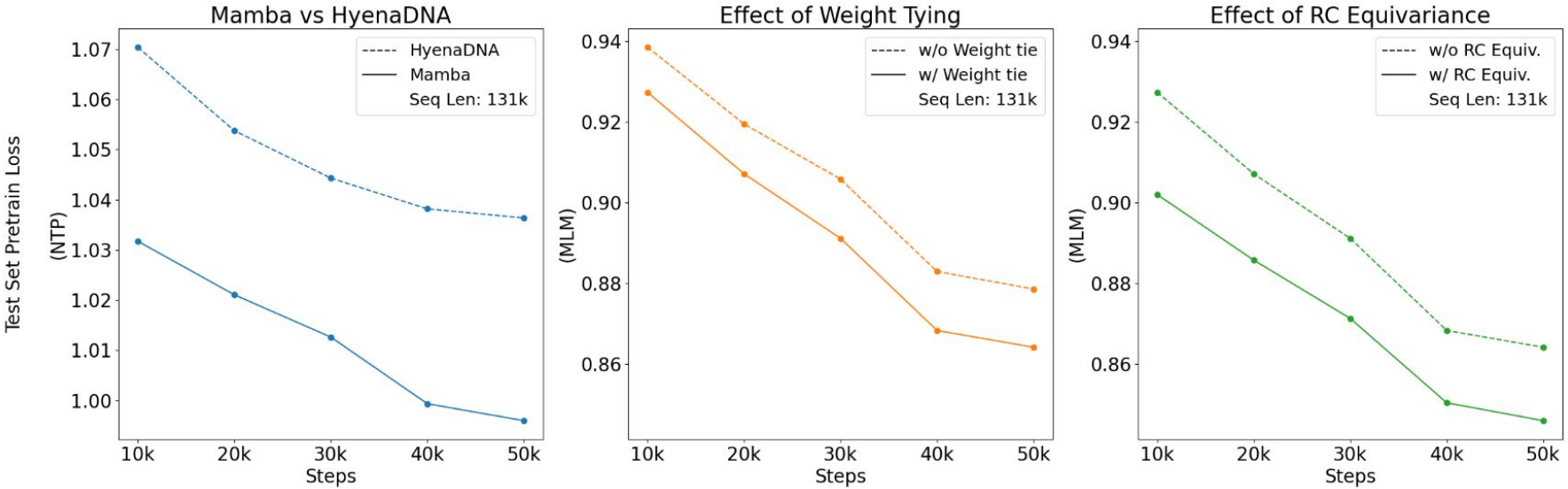
Using parameter-sharing and running modules on sequences and their RC versions enables **RC-equivariance**

Putting it all together



Experiments

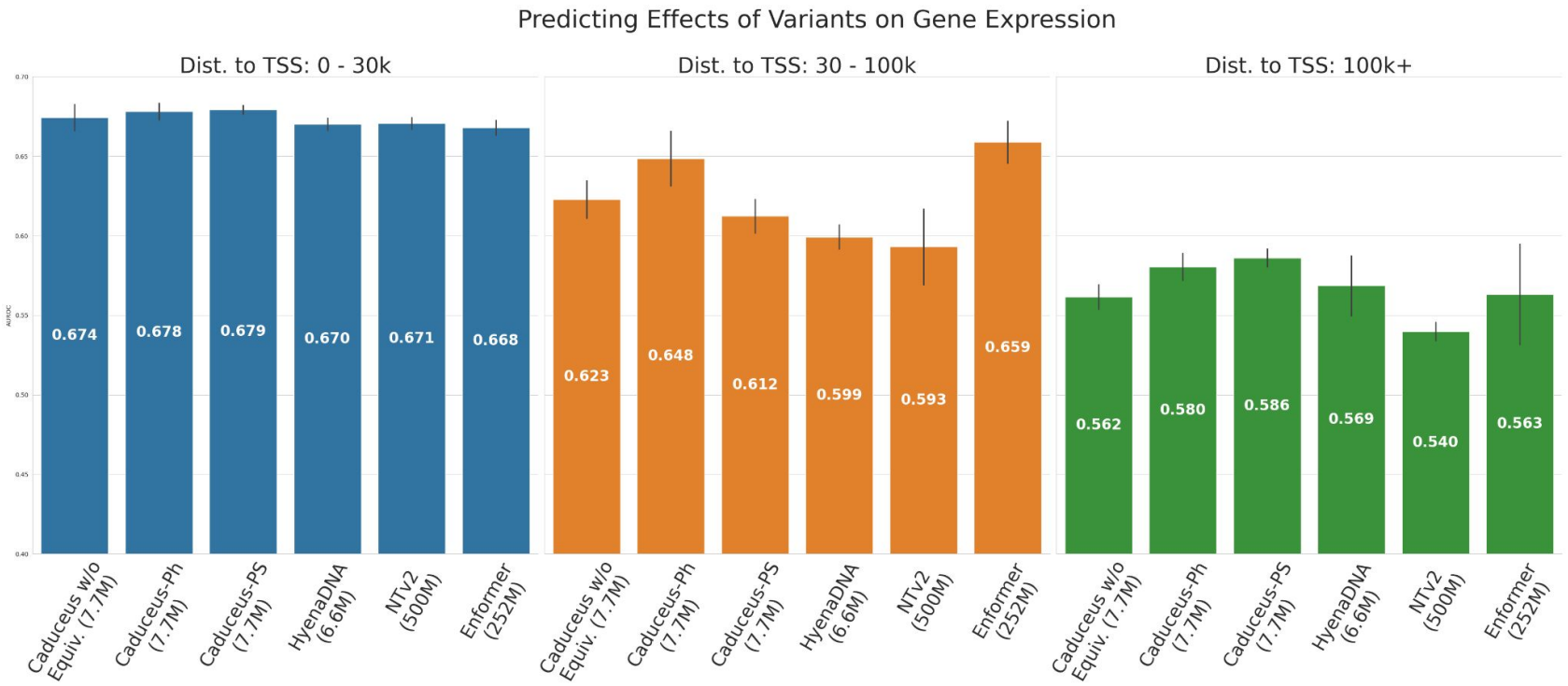
Improvements in pre-training loss



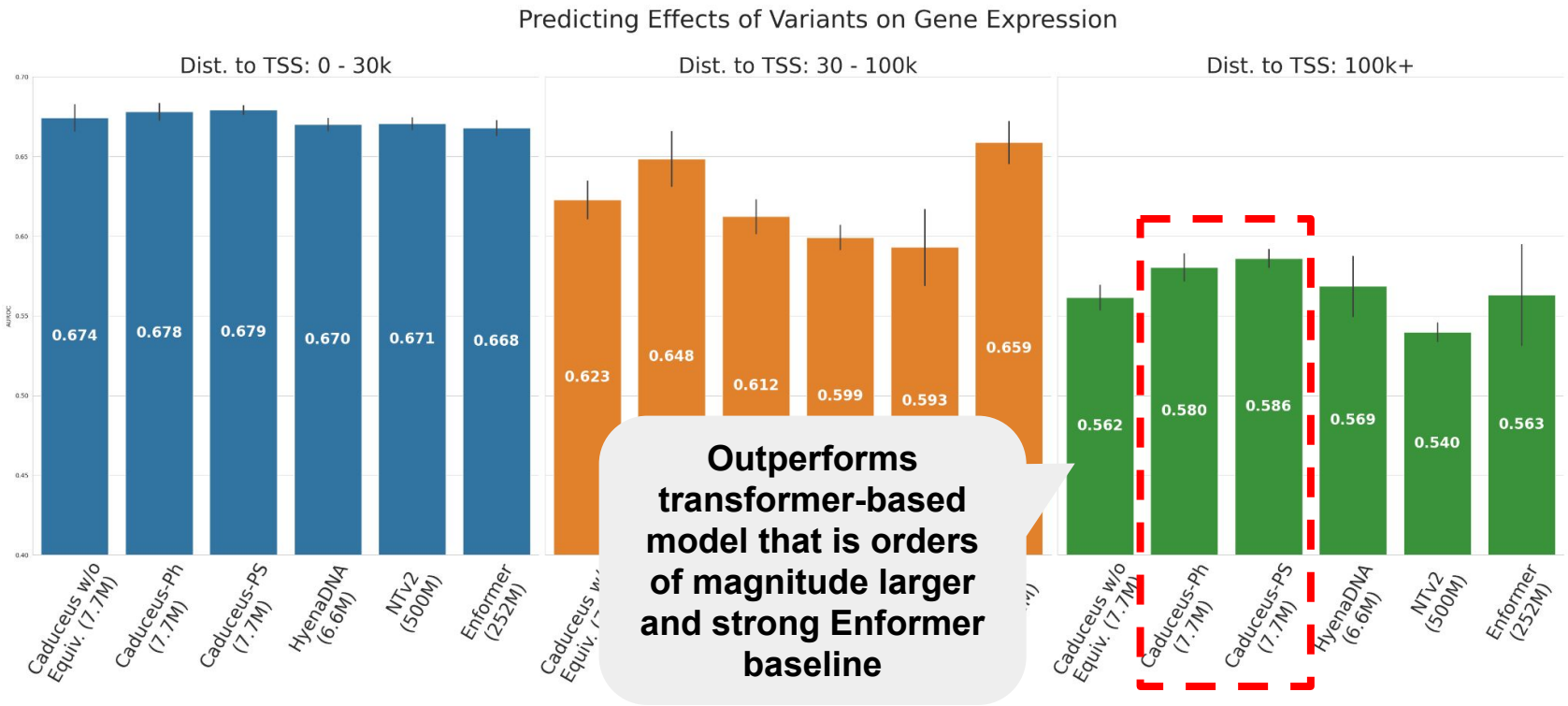
Competitive on standard benchmarks even against much larger models

	> 100M Param. Models			< 2M Param. Models		
	Enformer (252M)	DNABERT-2 (117M)	NT-v2 (500M)	HyenaDNA (1.6M)	Caduceus-Ph (1.8M)	Caduceus-PS (1.8M)
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

Improved eQTL causal SNP prediction with Caduceus



Improved eQTL causal SNP prediction with Caduceus



Conclusion

- ✅ Introduced bi-directional and RC equivariant extensions of Mamba
- ✅ Proposed Caduceus, a novel DNA foundation model
- ✅ Improved performance over much larger Transformers-based and comparably-sized SSM-based models

Thank you!

Thu 25 Jul 11:30 a.m. – 1 p.m. CEST
Hall C 4-9 #314

https://arxiv.org/abs/2403.03234

https://github.com/kuleshov-group/caduceus