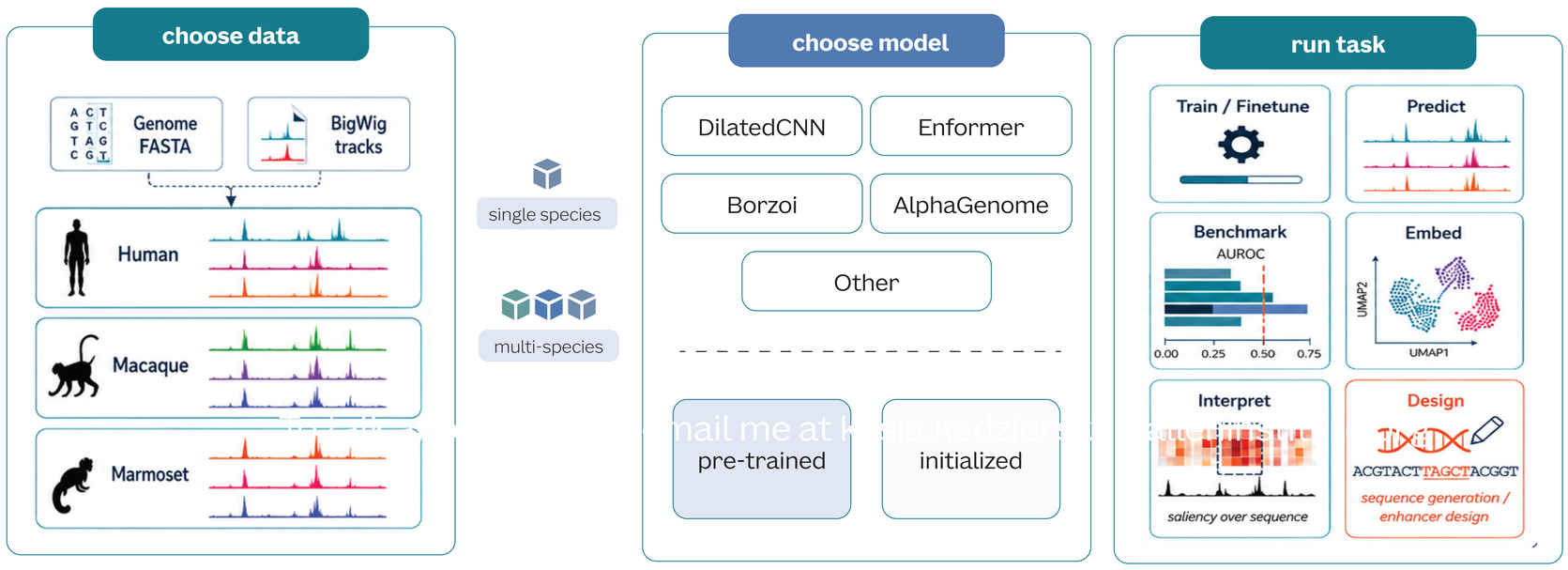
Benchmarking sequence-to-function models for cell-type-specificity

Kasia Kedzierska Alan DenAdel Nelson Johansen Grace Huynh **Allen Institute, Seattle, WA, USA**

Enhancer Designer

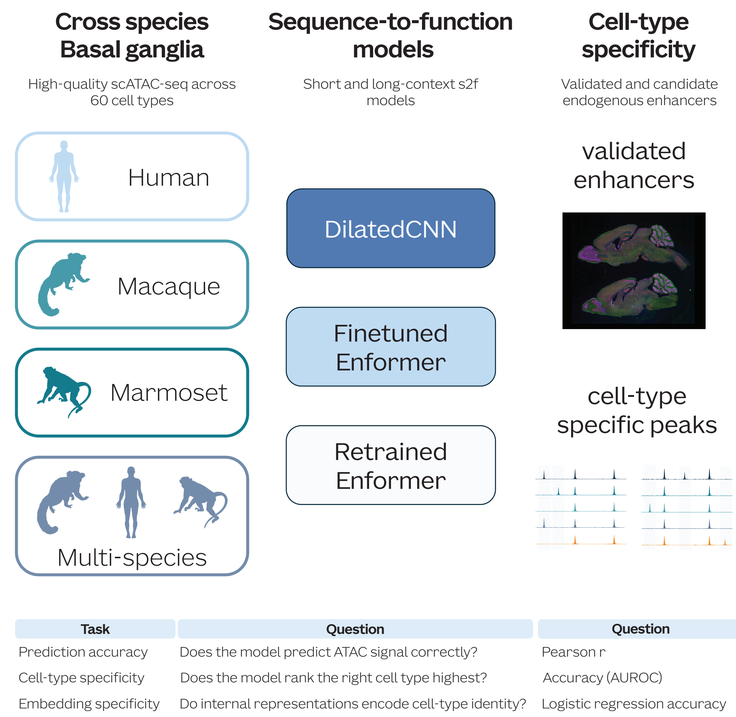
Modular sequence-to-function modeling and sequence design framework



Zarr PyTorch Lightning Weights & Biases Hugging Face

Easily swap data, model, and task via config Hydra YAML

Experimental setup

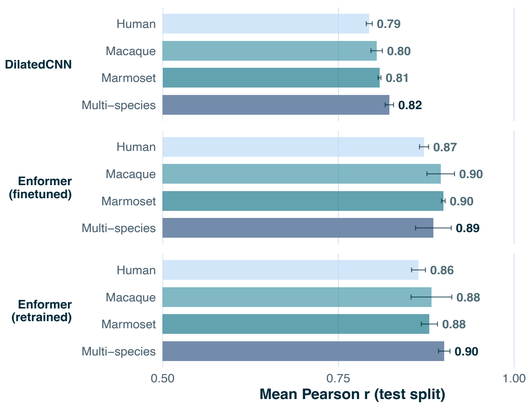


Task	Question	Question
Prediction accuracy	Does the model predict ATAC signal correctly?	Pearson r
Cell-type specificity	Does the model rank the right cell type highest?	Accuracy (AUROC)
Embedding specificity	Do internal representations encode cell-type identity?	Logistic regression accuracy

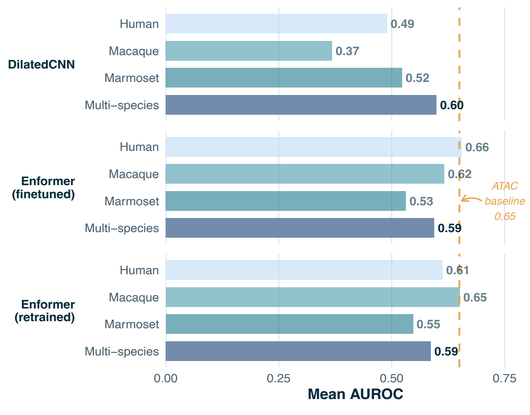
Predicting accessibility ≠ discriminating cell types.

Cell-type specificity requires direct evaluation.

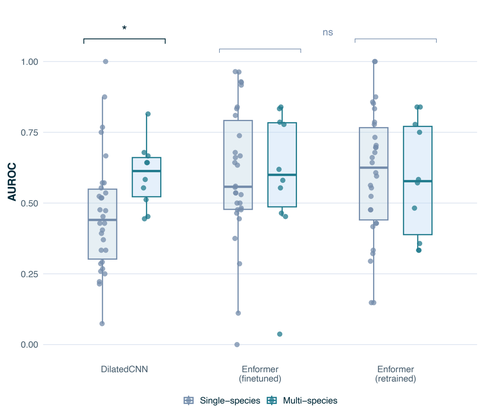
Models accurately predict chromatin accessibility...



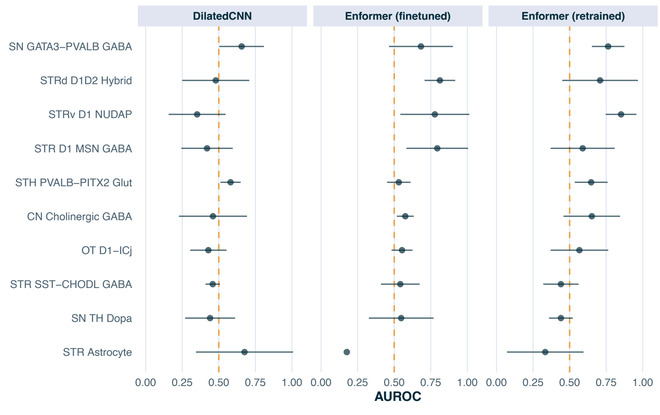
...but cell-type discrimination barely exceeds ATAC alone.



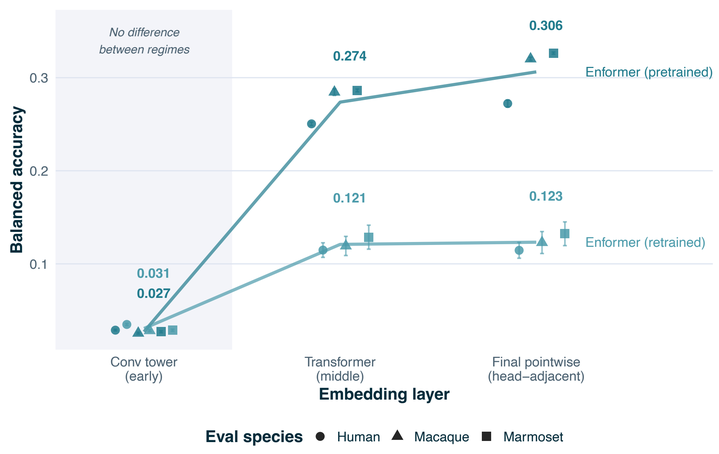
DilatedCNN benefits from a multi-species training regime.



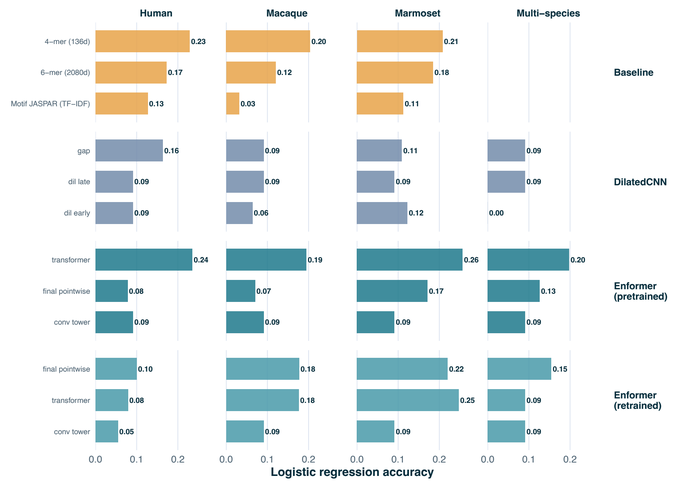
The performance of models is highly variable across cell types.



Enformer's pretrained representations encode cell-type identity that retraining on brain data struggles to recover.



Pretrained Enformer embeddings outperform BG-trained representations in cell-type classification.



Key takeaways

Accessibility ≠ cell-type specificity. Most models achieve Pearson r > 0.8, yet their best cell-type AUROC (~0.66) only modestly exceeds the ATAC signal proportion baseline (~0.60).

Multi-species training helps DilatedCNN but not Enformer. 60 closely related brain cell types may be too few to benefit a large transformer without stronger regularization.

Broad pretraining is not replaceable by task-specific retraining. Enformer's pretraining on diverse tissues and assays across ENCODE captures cell-type distinctions that fine-tuning on brain data alone cannot recover.

Next steps

Whole-brain models: Scale to whole-brain cell atlas and incorporate epigenomically diverse tissues to provide contrastive signal for cell-type discrimination.

Optimize training: Explore multi-species training regimes that balance cross-species alignment with per-species cell-type resolution, and disentangle conservation signal from data augmentation effects.

Experiment with design choices: Benchmark additional architectures and training strategies, including gradual unfreezing and context-length-efficient model variants.

To talk about science, email me at kasia.kedzierska@alleninstitute.org

Scan & learn more

kasia.codes/EnhancerDesigner

BG scATAC-seq data: Johansen NJ, Fu Y et al. *Cross-species consensus atlas of the primate basal ganglia.* bioRxiv 2025. doi:10.64898/2025.12.15.694496

Benchmark data: Wirthlin ME et al. *A Cross-Species Enhancer-AAV Toolkit for Targeting the Basal Ganglia.* bioRxiv 2026. doi:10.64898/2026.02.23.706695

BICCN challenge: Johansen NJ, Kempynck N et al. *Evaluating methods for predicting cell-type-specific enhancers in the mammalian cortex.* Cell Genomics 5:100879 (2025). doi:10.1016/j.xgen.2025.100879

Enformer model: Avsec Ž et al. *Effective gene expression prediction from sequence by integrating long-range interactions.* Nat Methods 18:1196 (2021). doi:10.1038/s41592-021-01252-x

dCNN model: Kempynck N, De Winter S et al. *CREsted: modeling cell-type-specific enhancers across tissues and species.* Nat Methods (2026). doi:10.1038/s41592-026-03057-2

allen institute/

This research was supported by the Allen Institute, founded by Jody Allen – chair and co-founder of Allen Family Philanthropies, and the late Paul G. Allen – investor, philanthropist, and co-founder of Microsoft. We gratefully acknowledge their vision and generosity, which make this work possible.