STANFORD
COMPUTER SCIENCE

Stanford MEDICINE | Department of Genetics

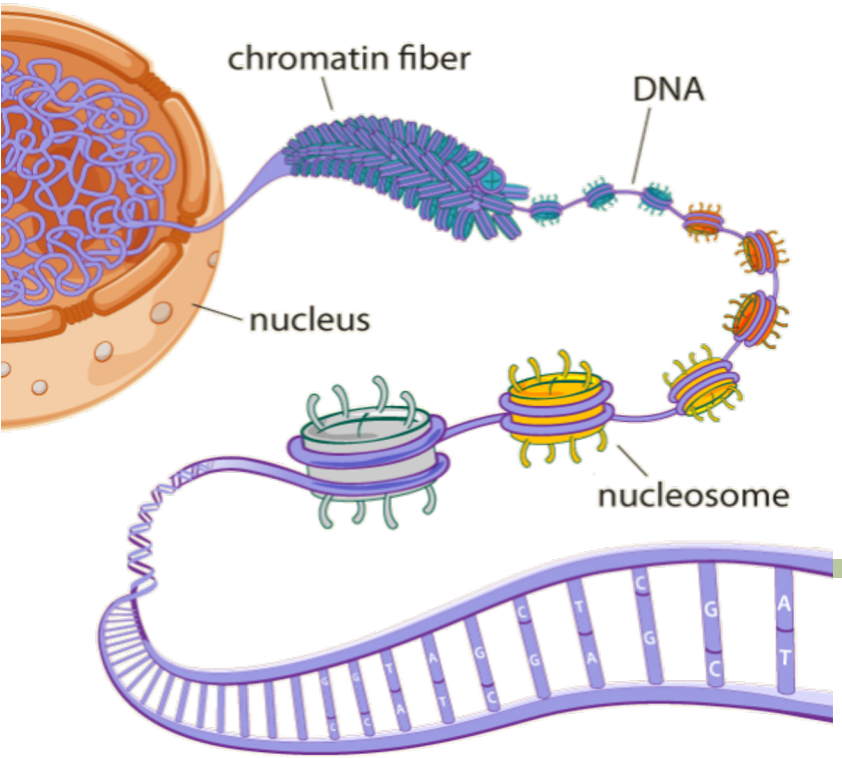
Machine learning for genomic discovery

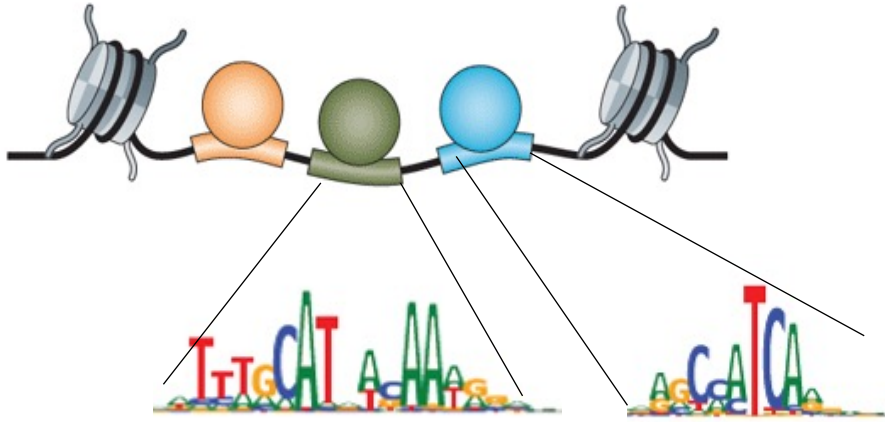
Anshul Kundaje

Twitter:@anshulkundaje

Website: http://anshul.kundaje.net

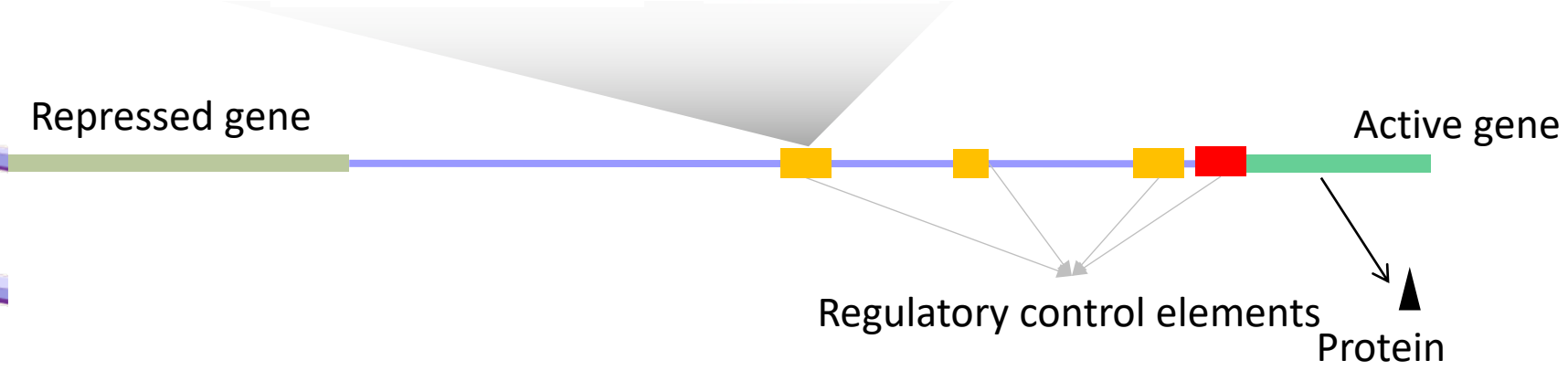
Decoding regulatory DNA





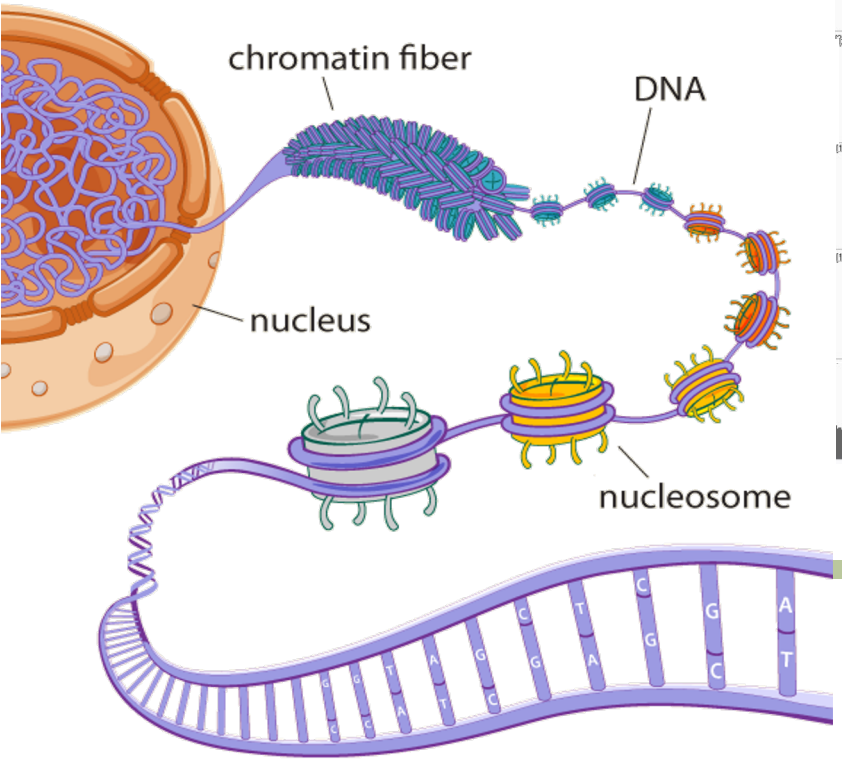
Motif syntax: rules of
- Composition
- Affinity
- Arrangement
- Spacing
- Orientation

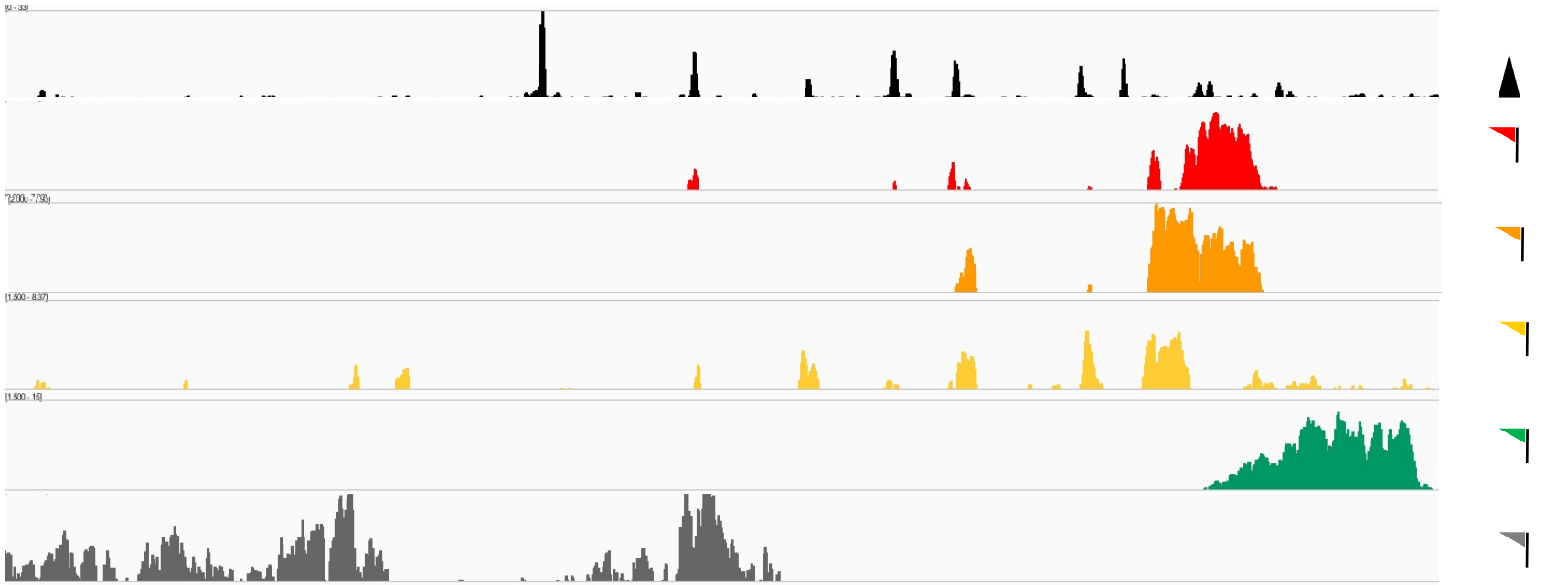
=> cooperativity

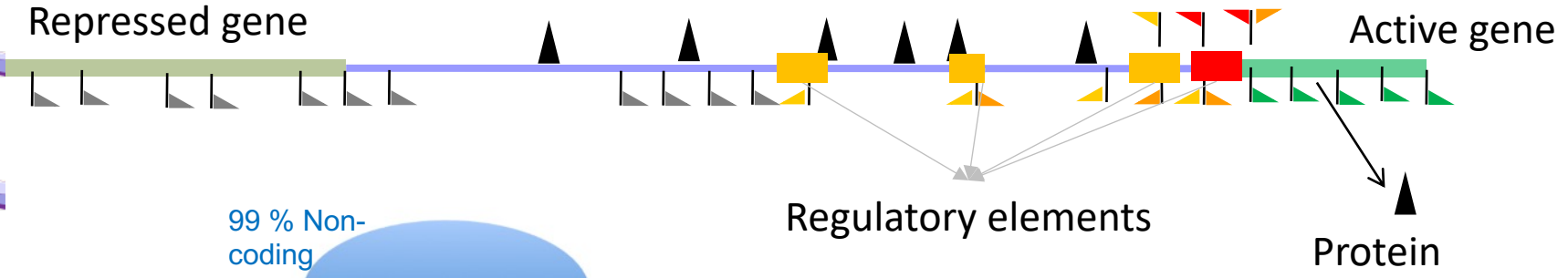


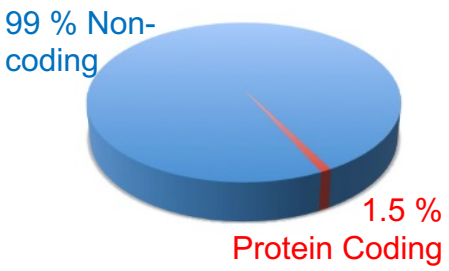
https://www.broadinstitute.org/news/1504

Mapping biochemical markers of regulatory activity

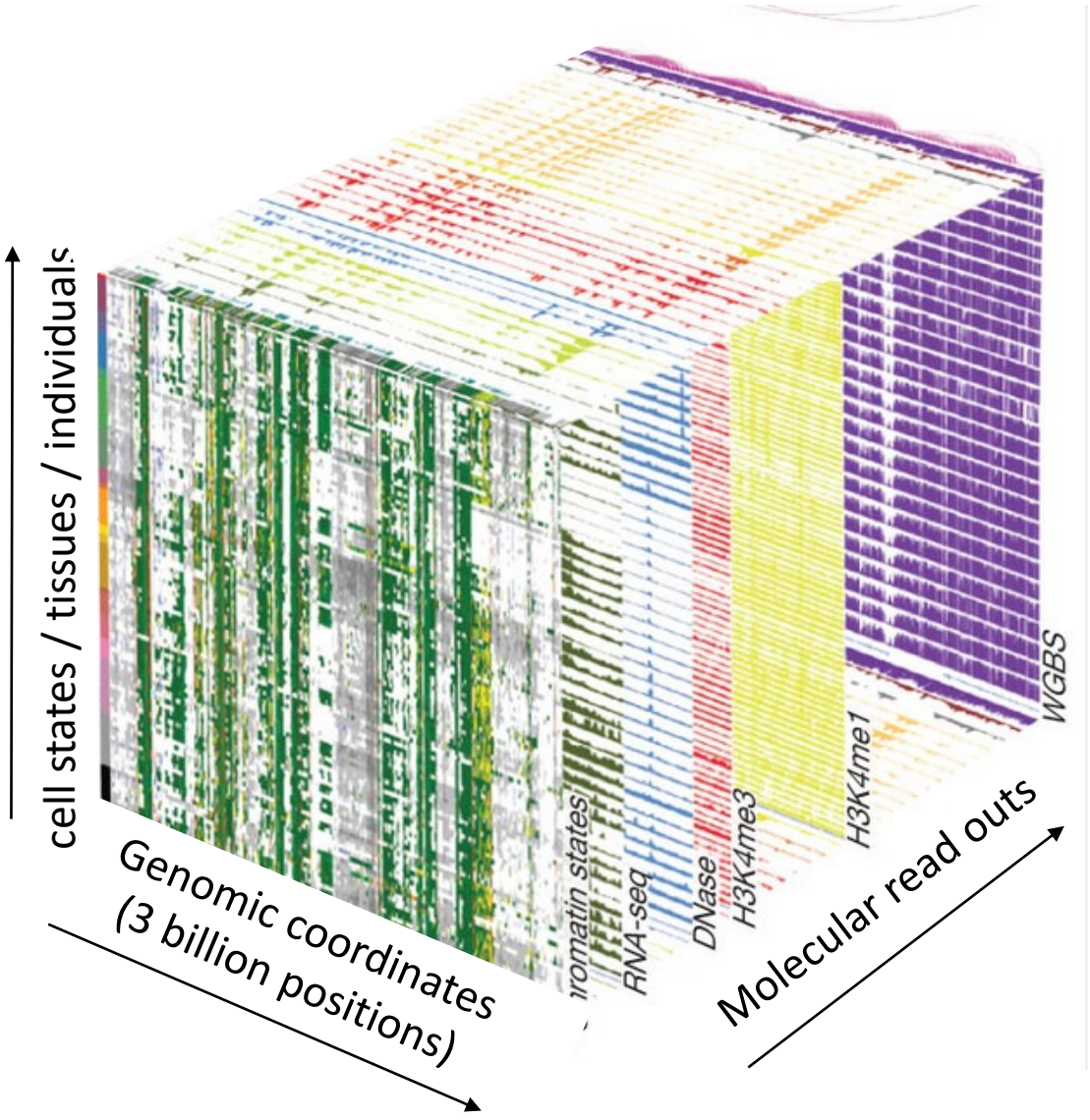


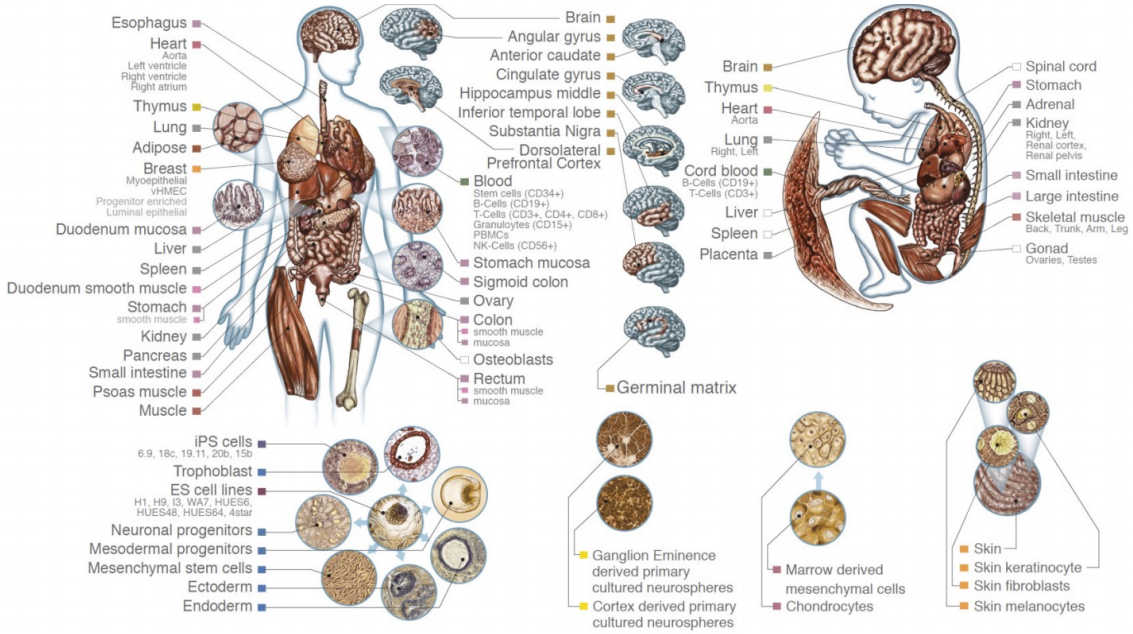






https://www.broadinstitute.org/news/1504





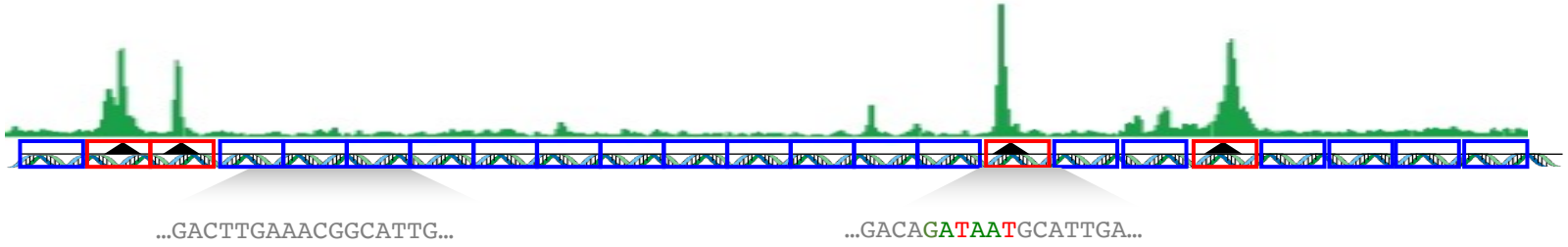
100s of Cell-Types/Tissues

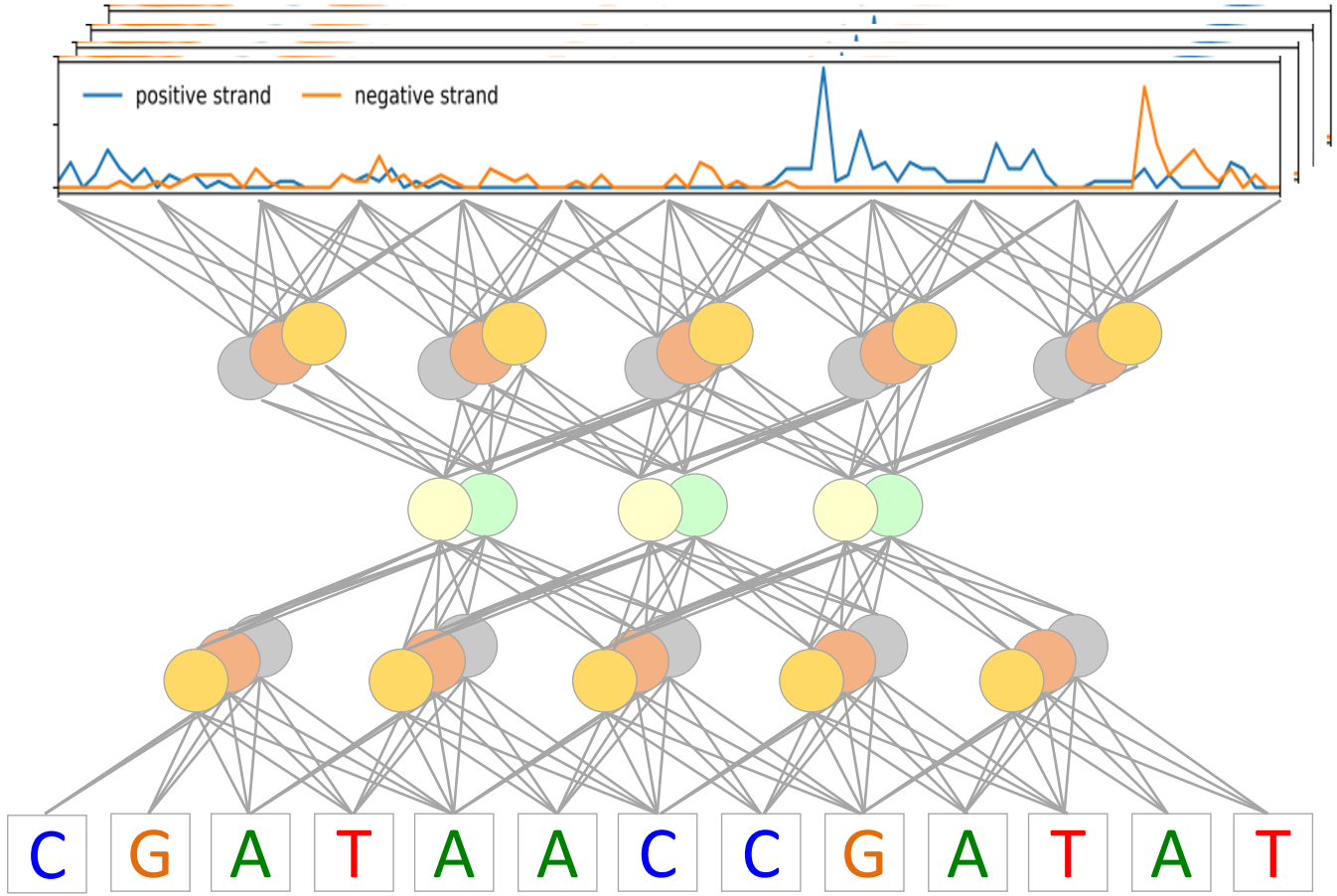
Machine learning models

Decoding regulatory DNA sequence

Predicting functional genetic variant & mutations

BPNet: Mapping DNA sequence to base-pair resolution profiles





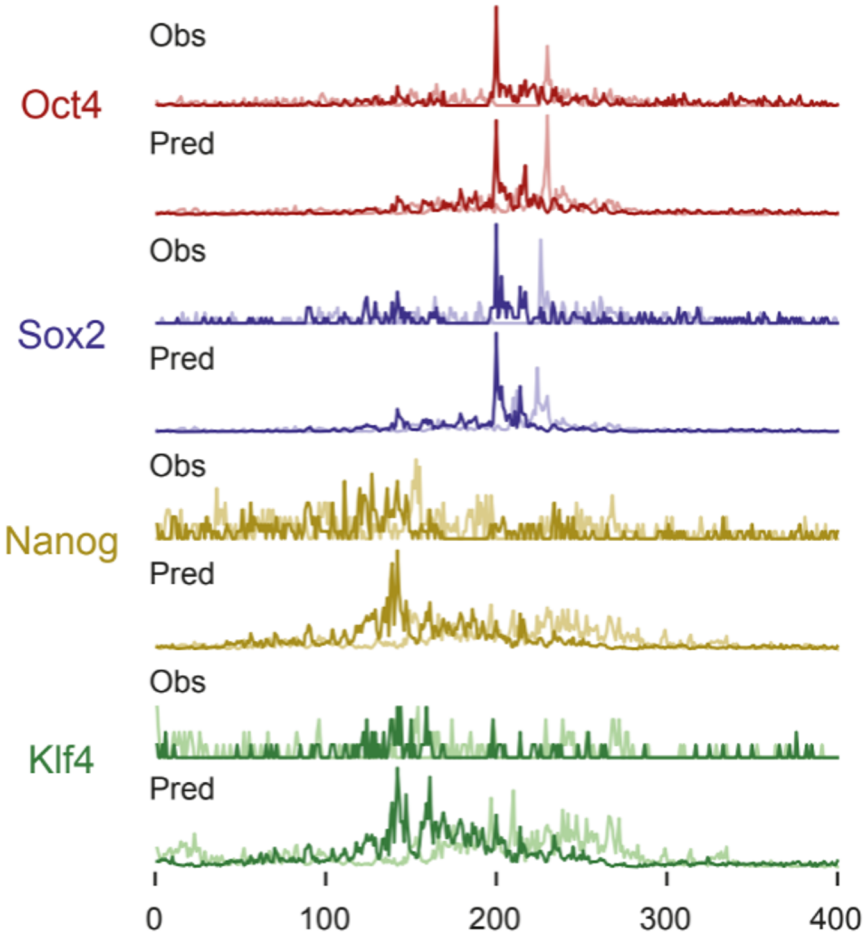
Ziga Avsec

Julia Zeitlinger

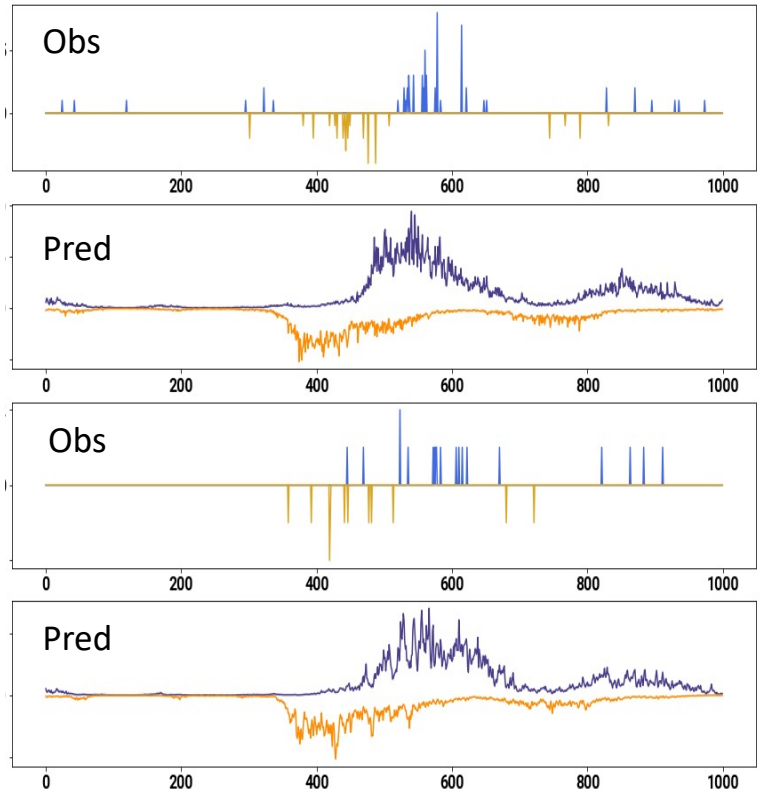
Avsec et al. 2021 Nature Genetics

BPNet predicts reg. profiles from sequence with unprecedented accuracy

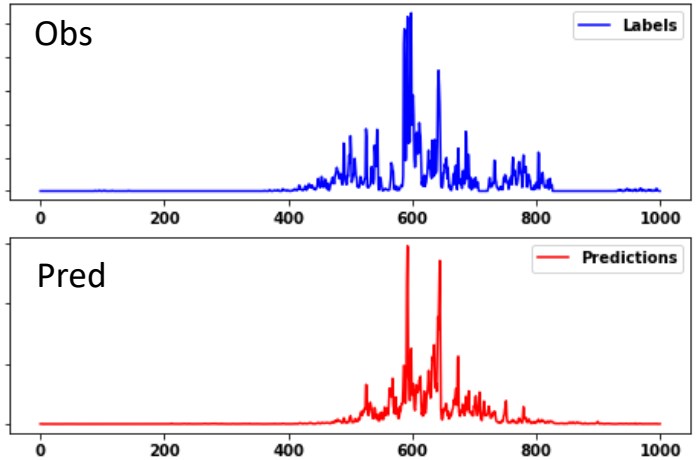
TF ChIP-exo



TF ChIP-seq

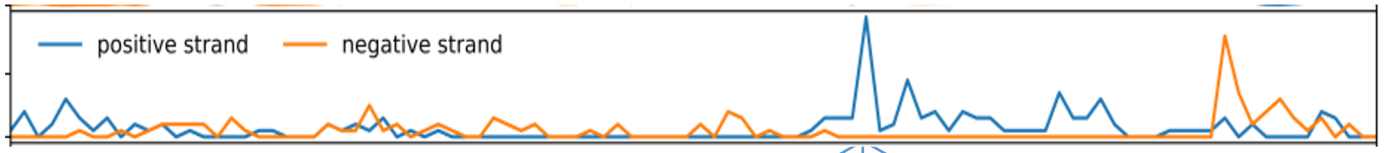


DNase-seq / ATAC-seq /
pseudo-bulk scATAC-seq



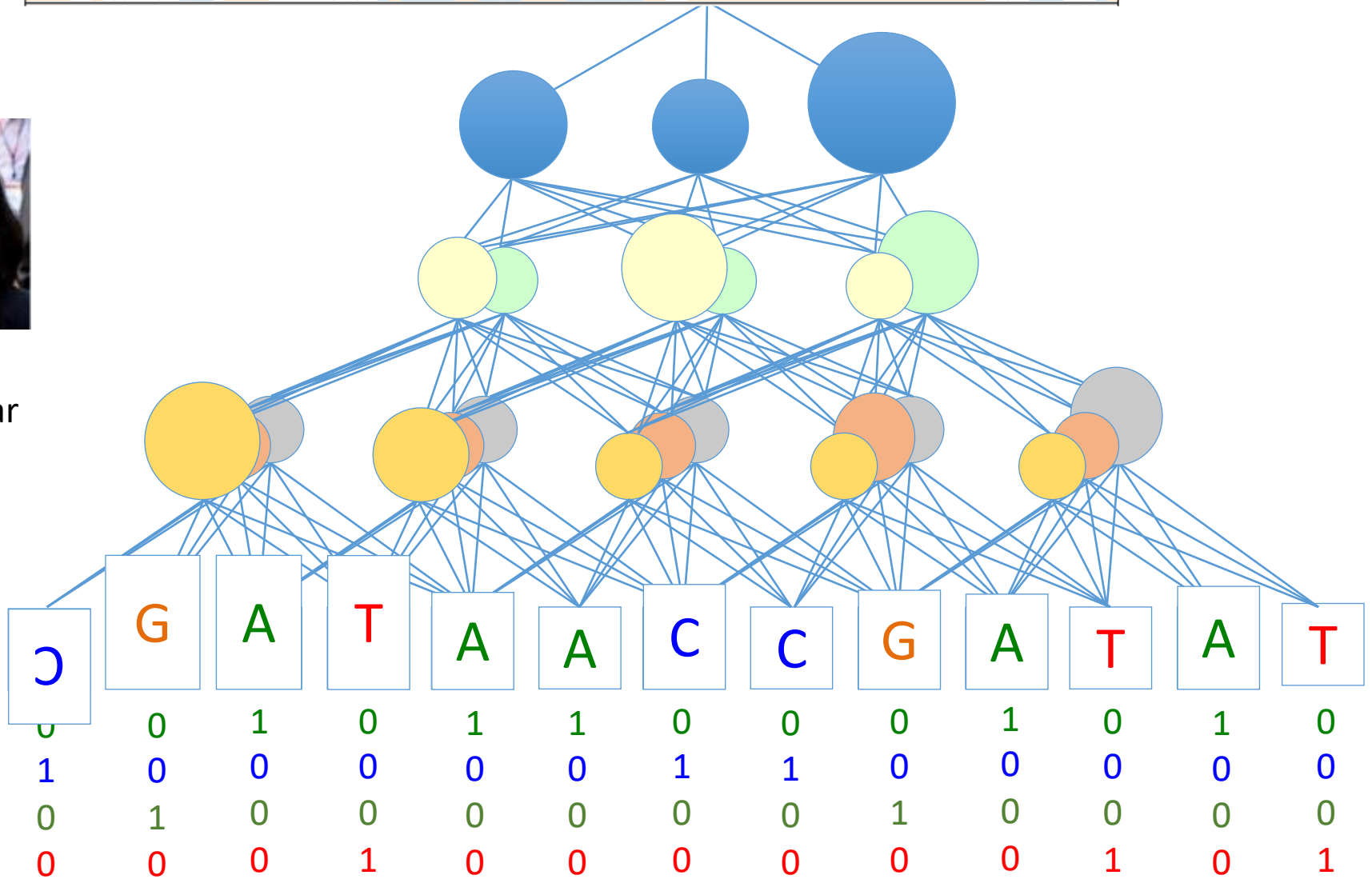
Opening up the blackbox

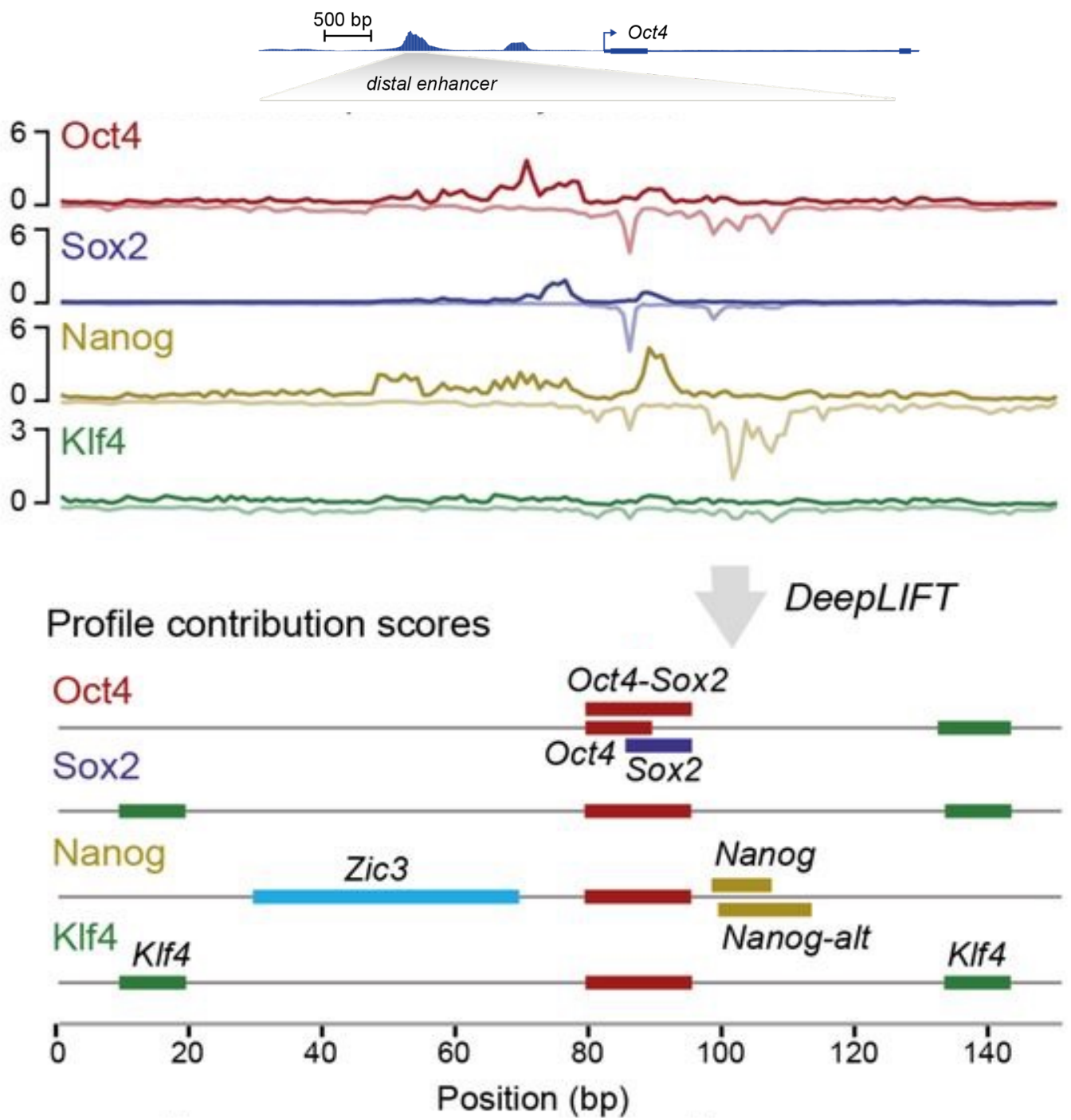
DeepLIFT: Inferring predictive nucleotides at individual binding events



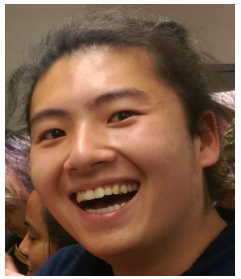
Shrikumar et al. ICML 2017
Lundberg et al. NeurIPS 2017

Avanti Shrikumar





Avanti Shrikumar

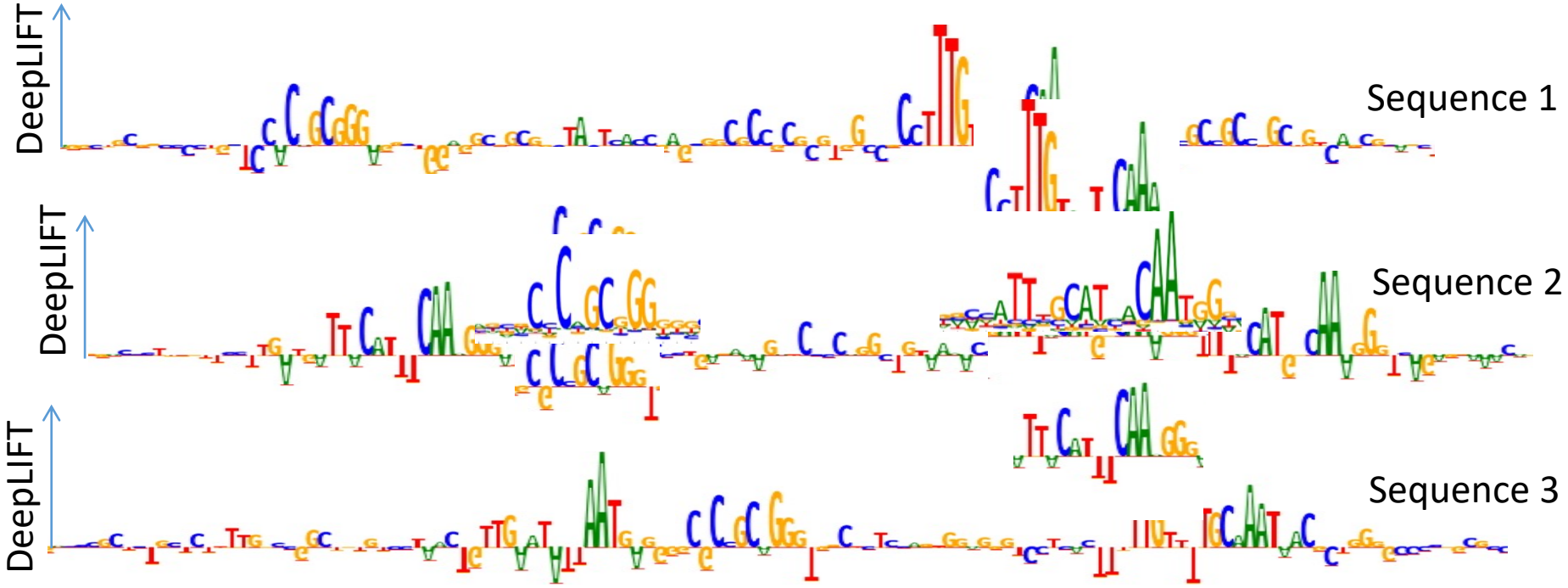


Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

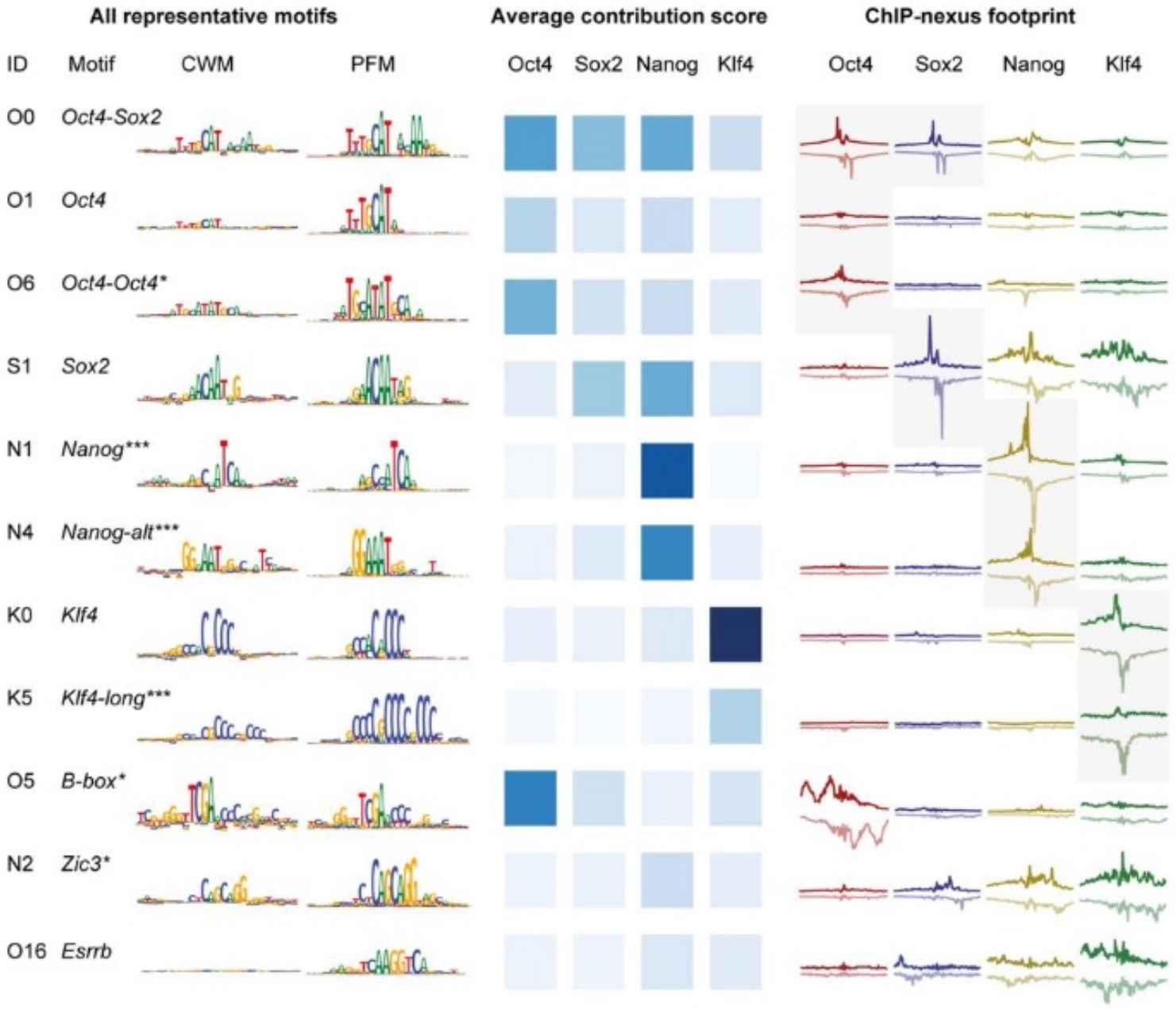
Insight: conv. filter contributions are integrated at the nucleotide level



Shrikumar et al. 2018, arxiv

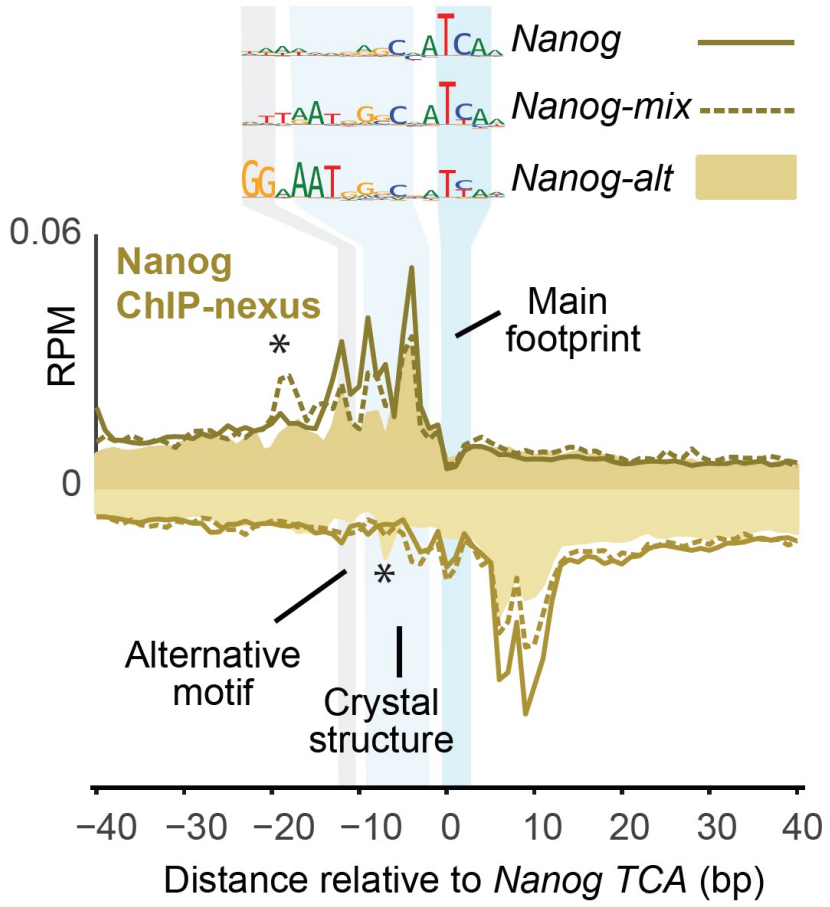
CODE: *https://github.com/kundajelab/tfmodisco*

Complex repertoire of motifs due to cooperative binding

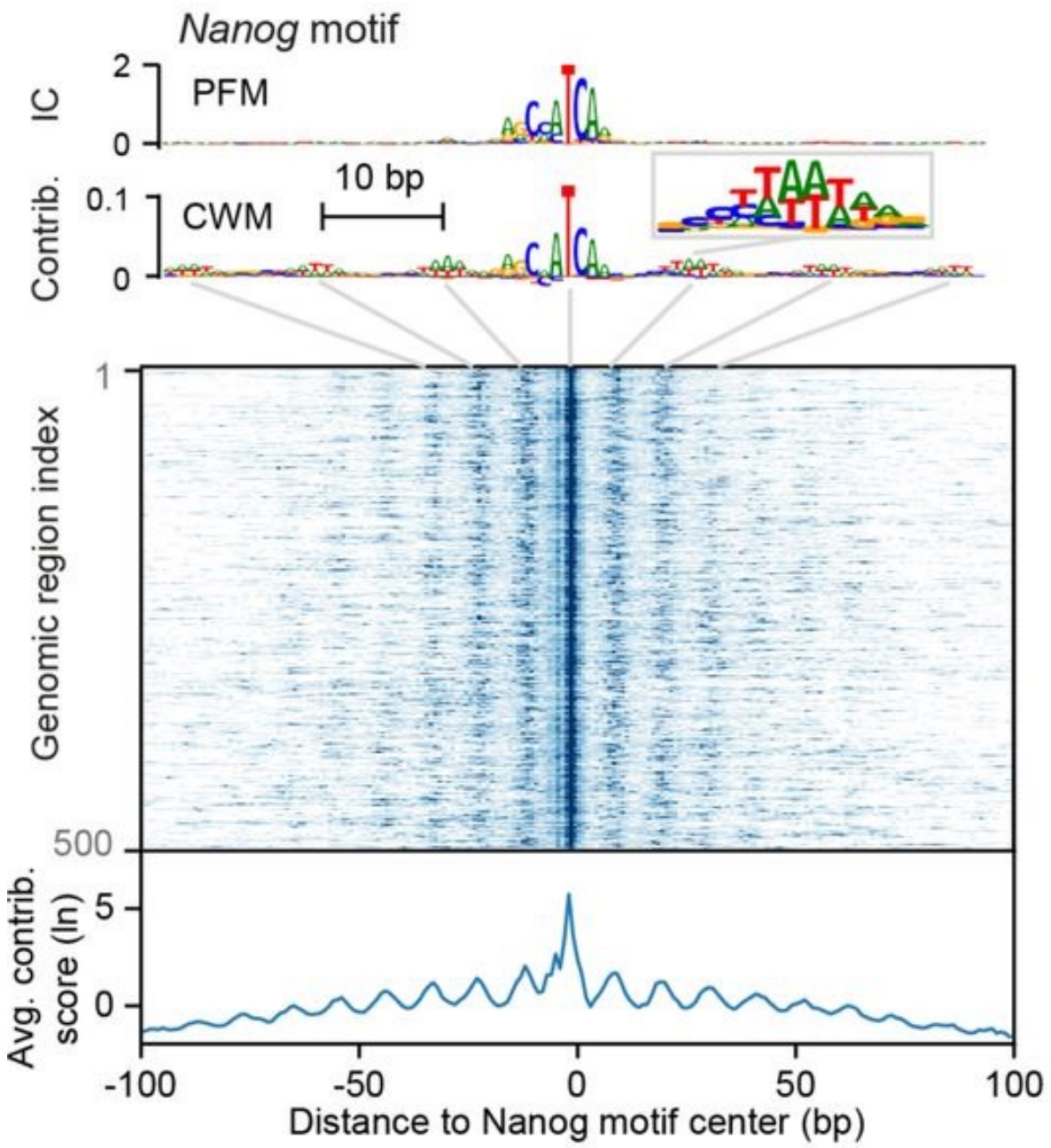


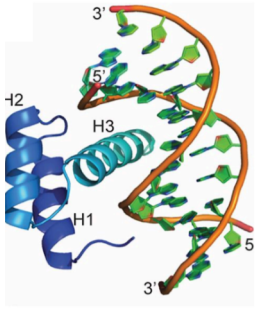
50 motifs for 4 TFs!

Subtle differences in Nanog motifs

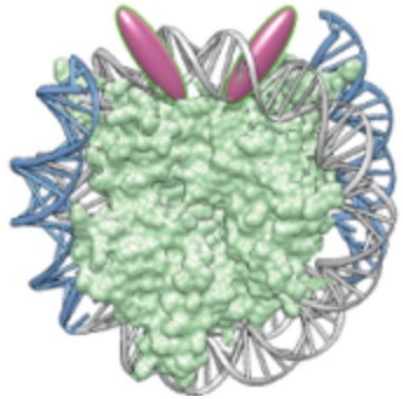


Subtle low affinity patterns with helical periodicity flanking Nanog motif



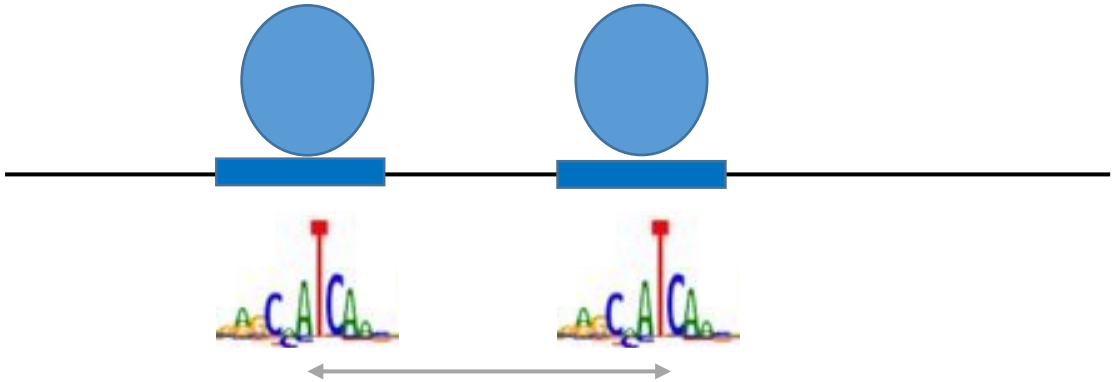


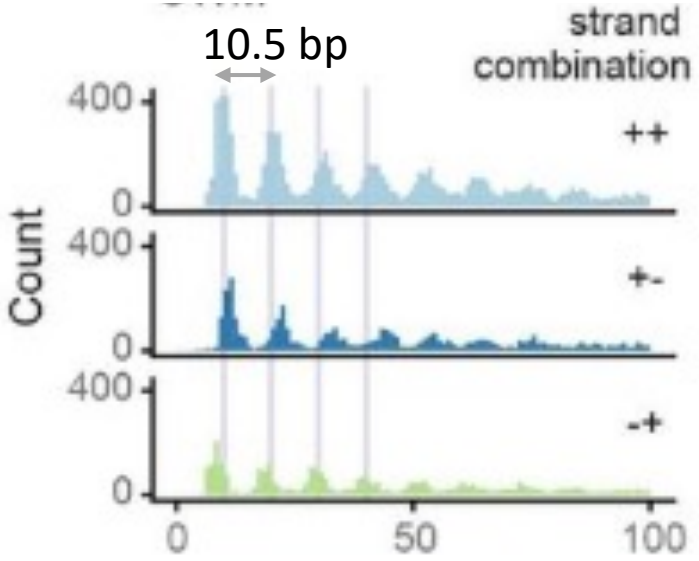
Nanog homeodomain
Hayakshi et al. PNAS 2015



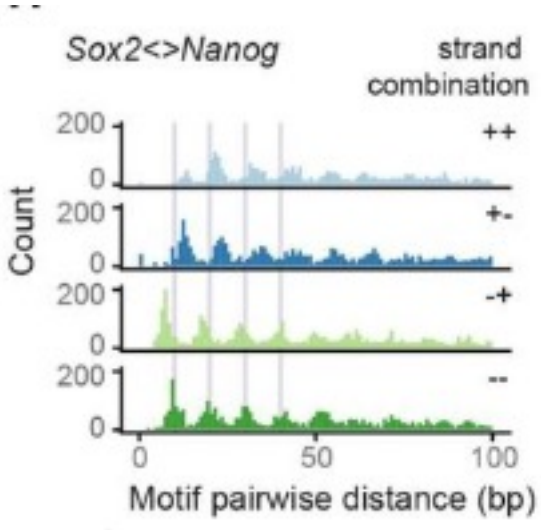
10 bp periodic binding of homeobox TFs to nucleosome DNA
from recent *in vitro* NCAP-SELEX data
(Zhu et al. Nature 2018)

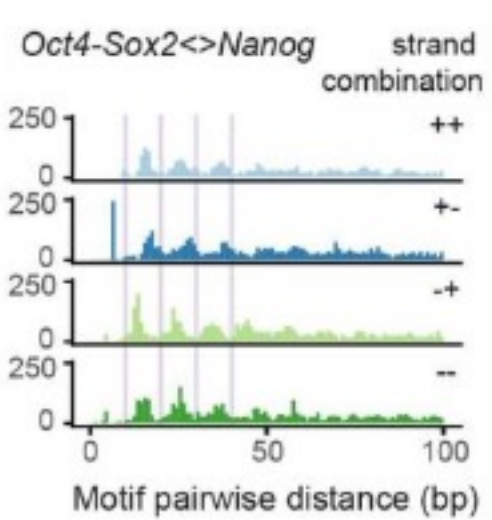
Soft syntax: helical spacing preference between Nanog motifs in the genome

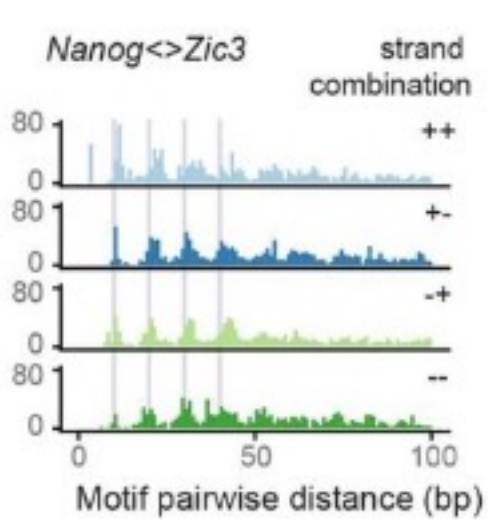




Motif pairwise distance

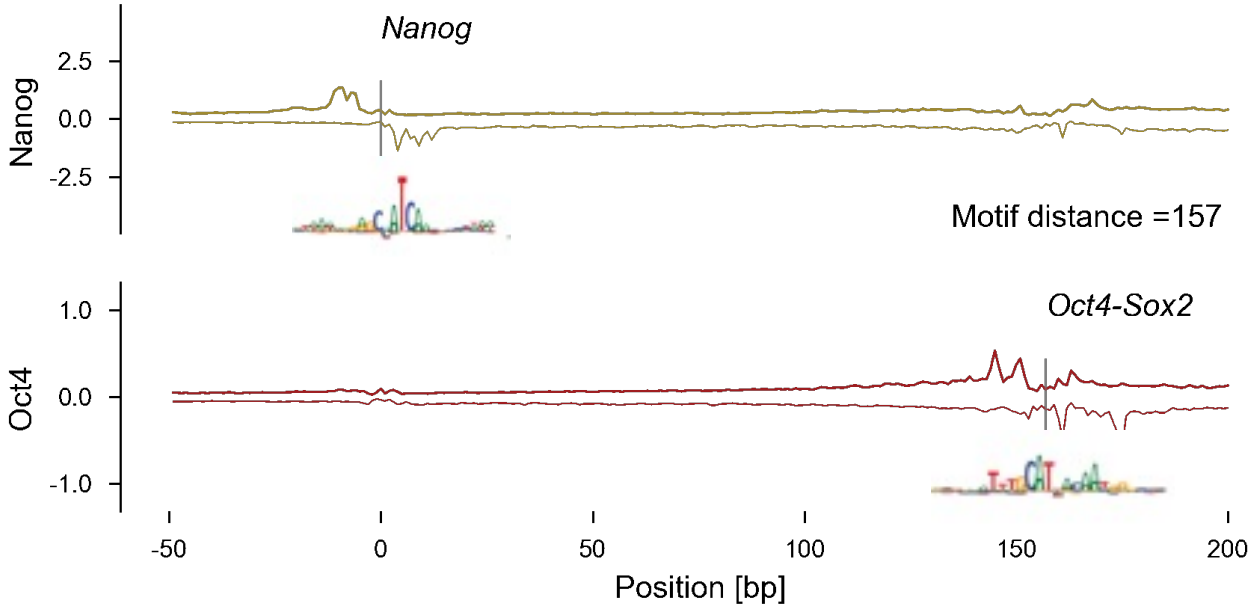


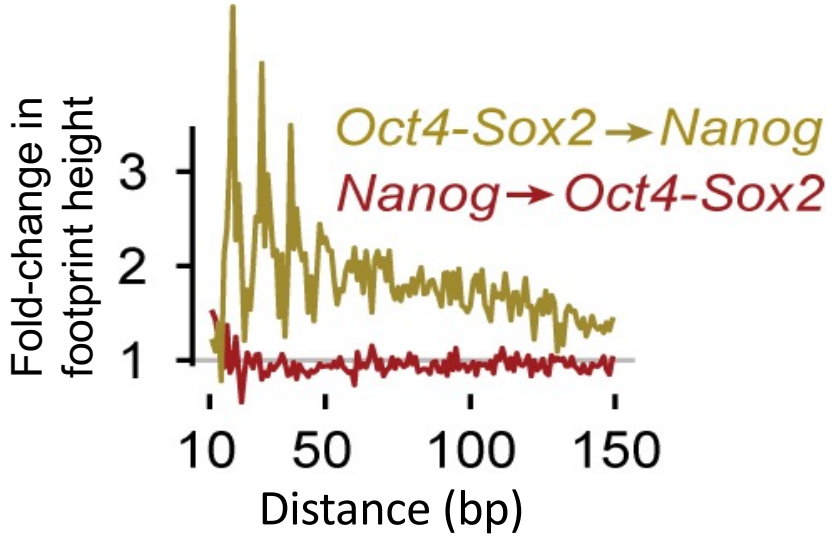




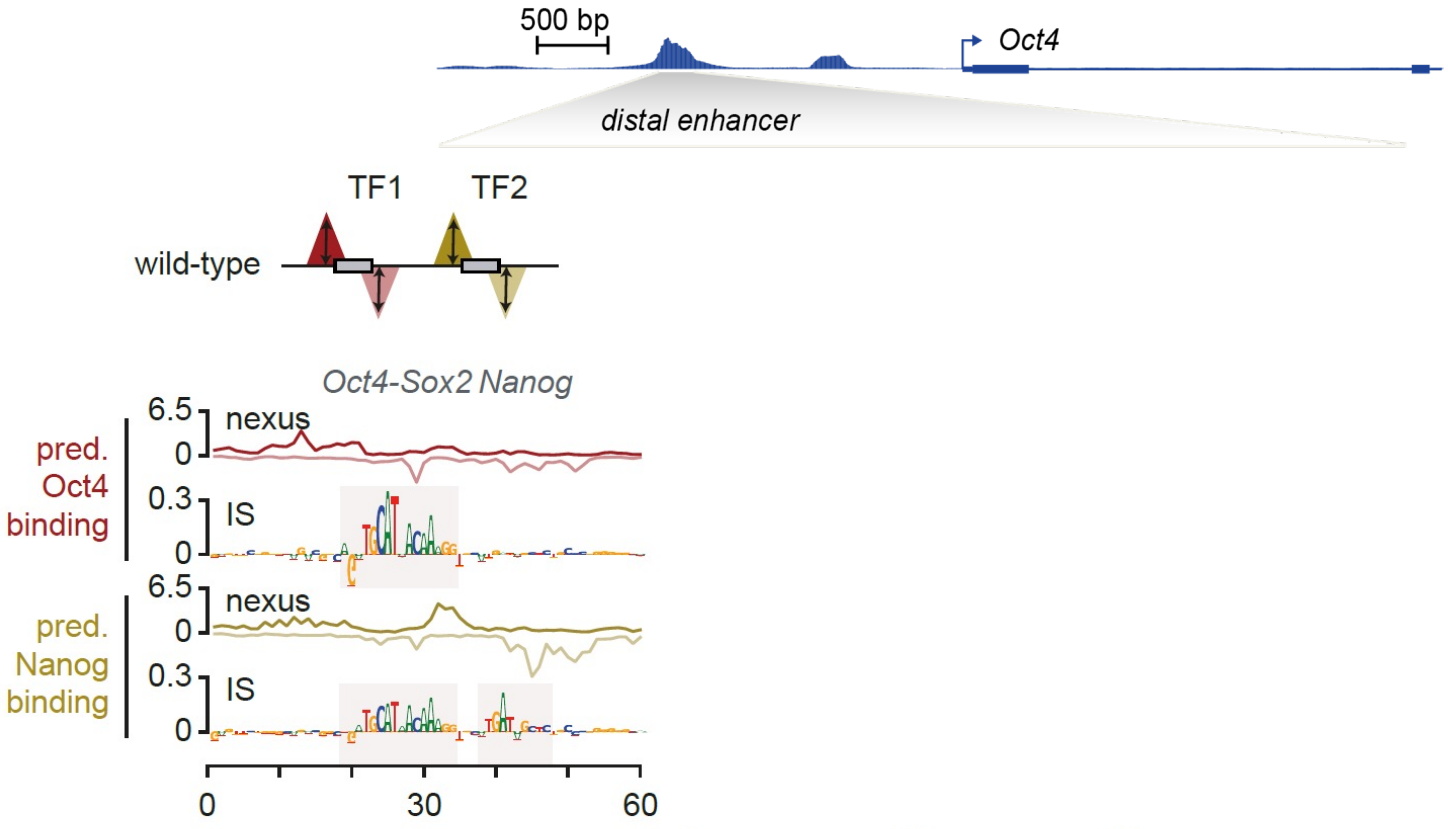
Using the model as an “oracle” to perform large-scale *in-silico* experiments

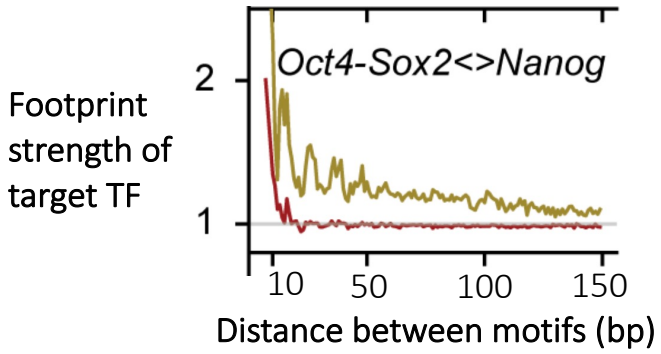
Deciphering syntax dependent TF cooperativity with synthetic designed sequences



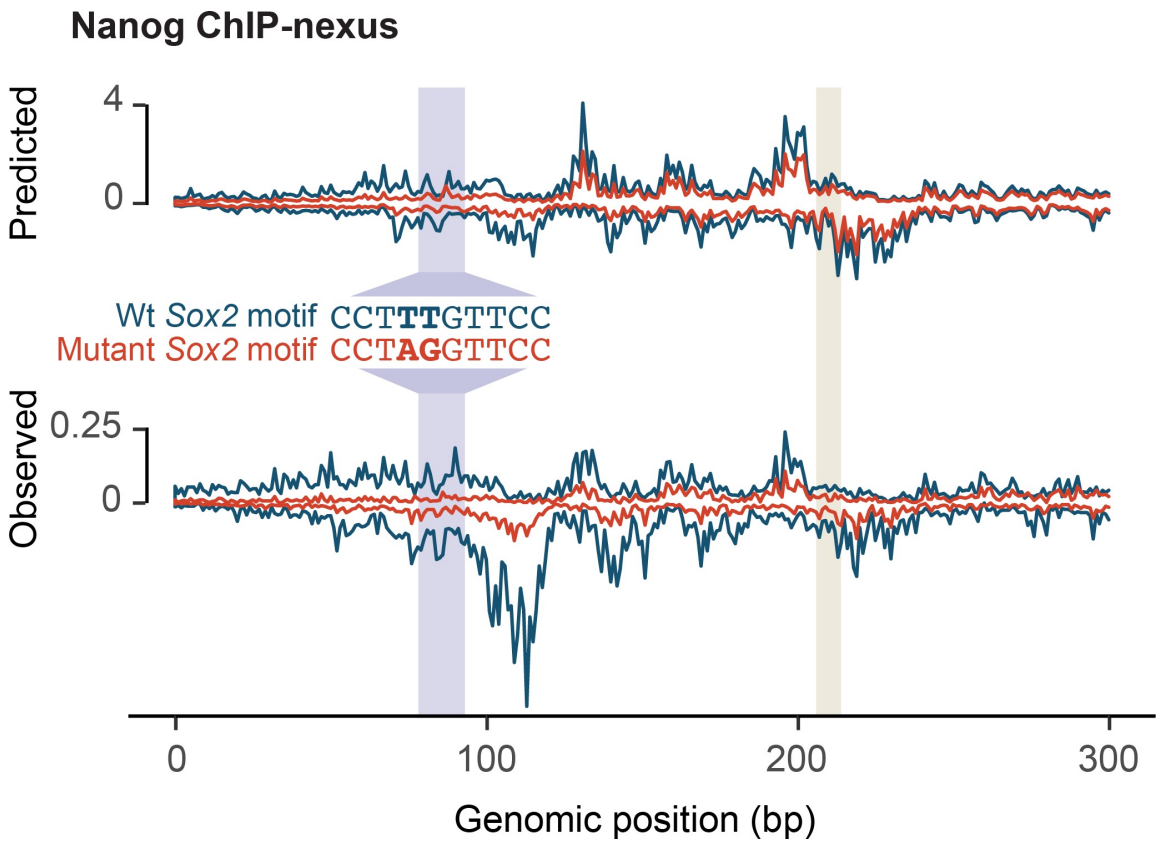


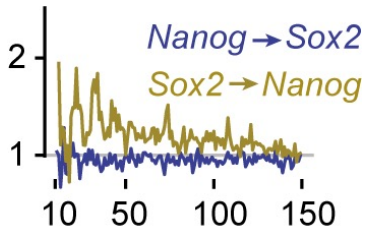
Deciphering syntax dependent TF cooperativity with *in-silico* genome editing

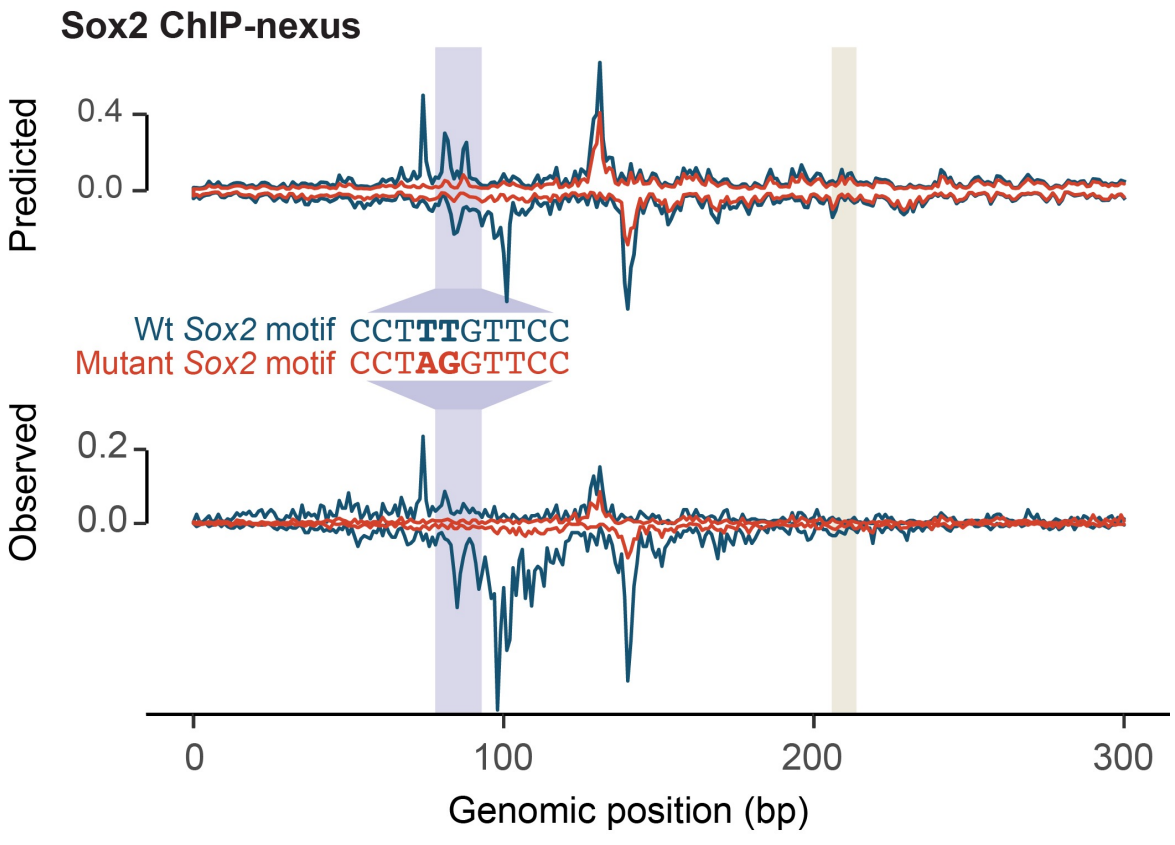


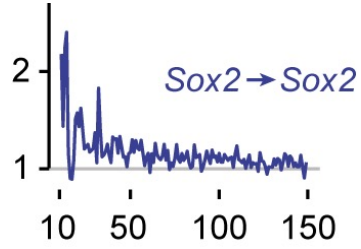


Using the model to design CRISPR experiments to validate discoveries







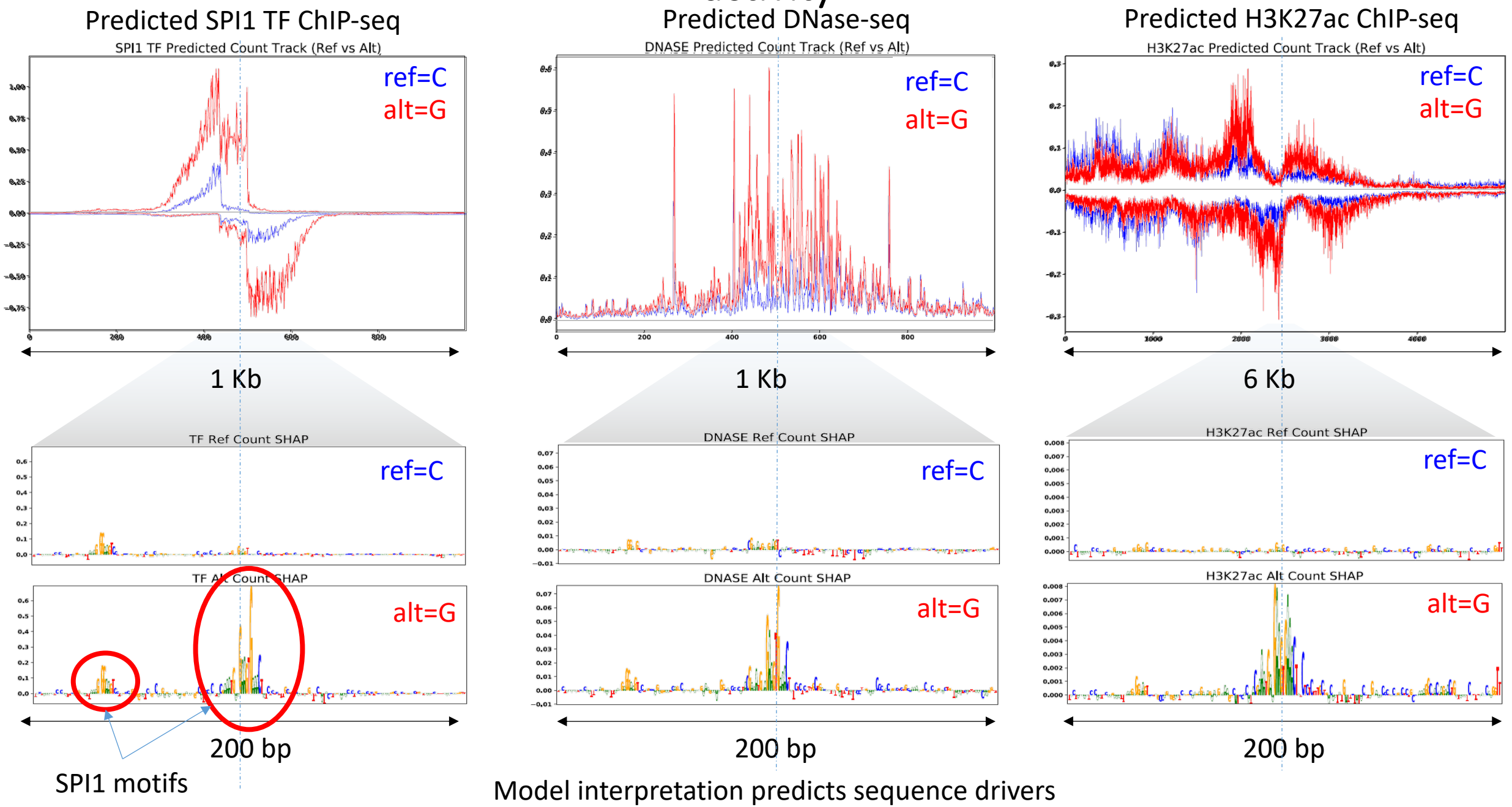


Julia Zeitlinger

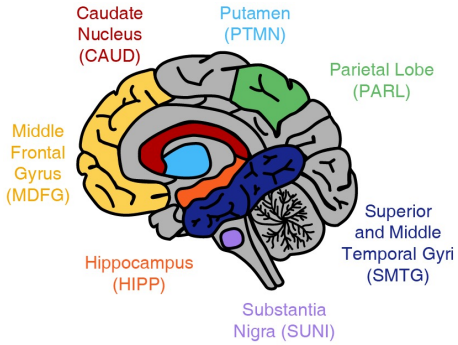
Sabrina Krueger, Melanie Weilert

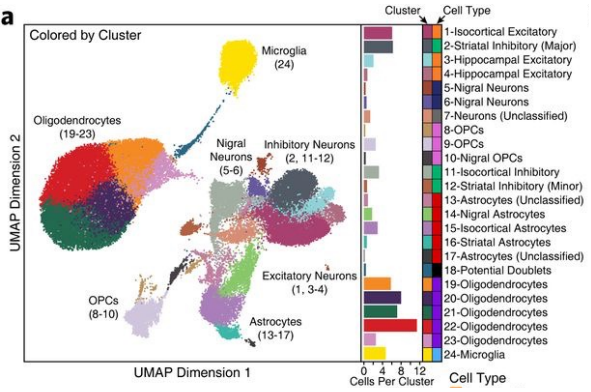
Model-driven prioritization of functional genetic variation

Predicting and interpreting variants influencing multiple layers of regulatory activity

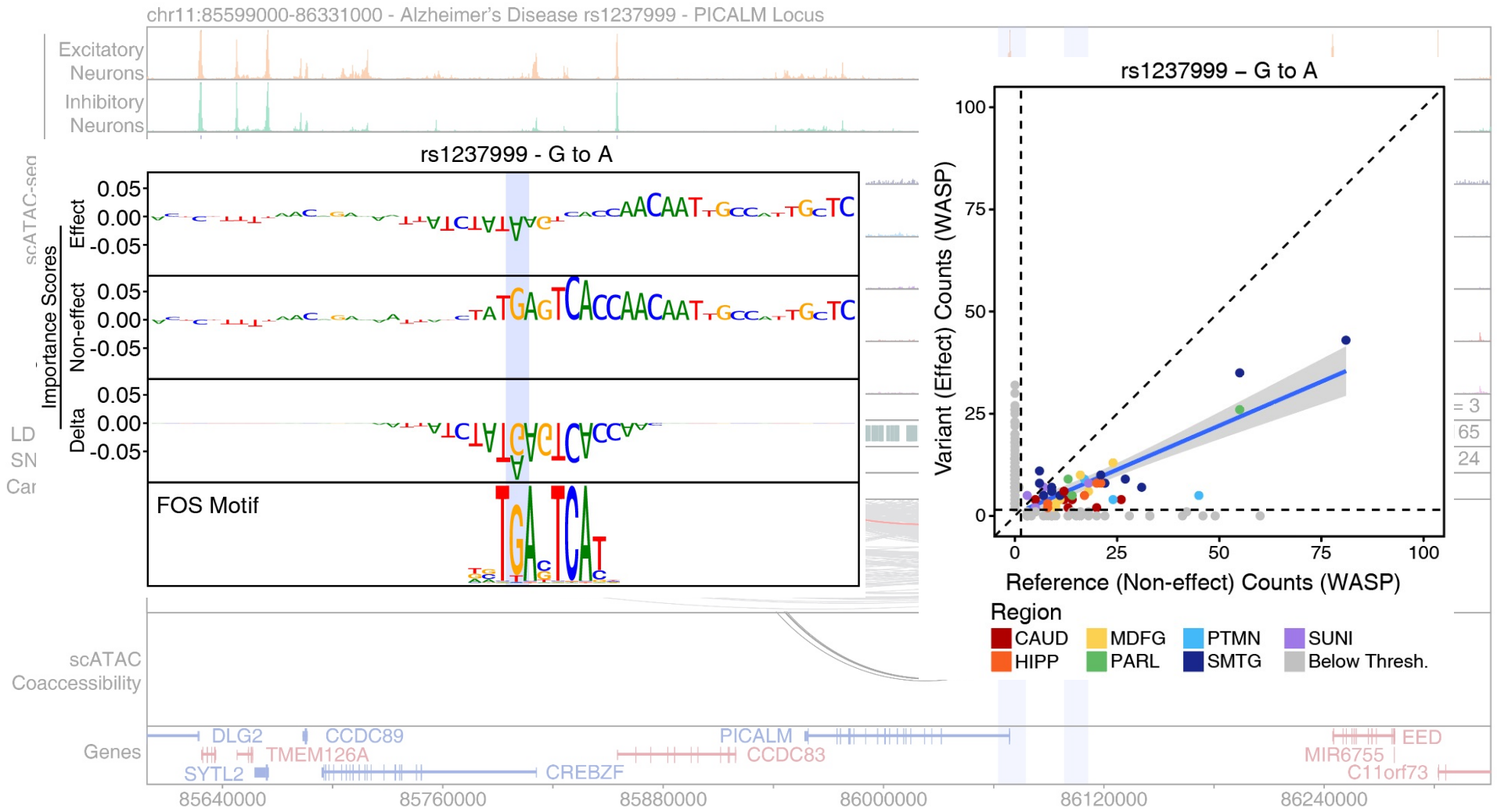


Prioritizing putative causal variants in disease-associated loci





Corces et al. 2020, Nature Genetics

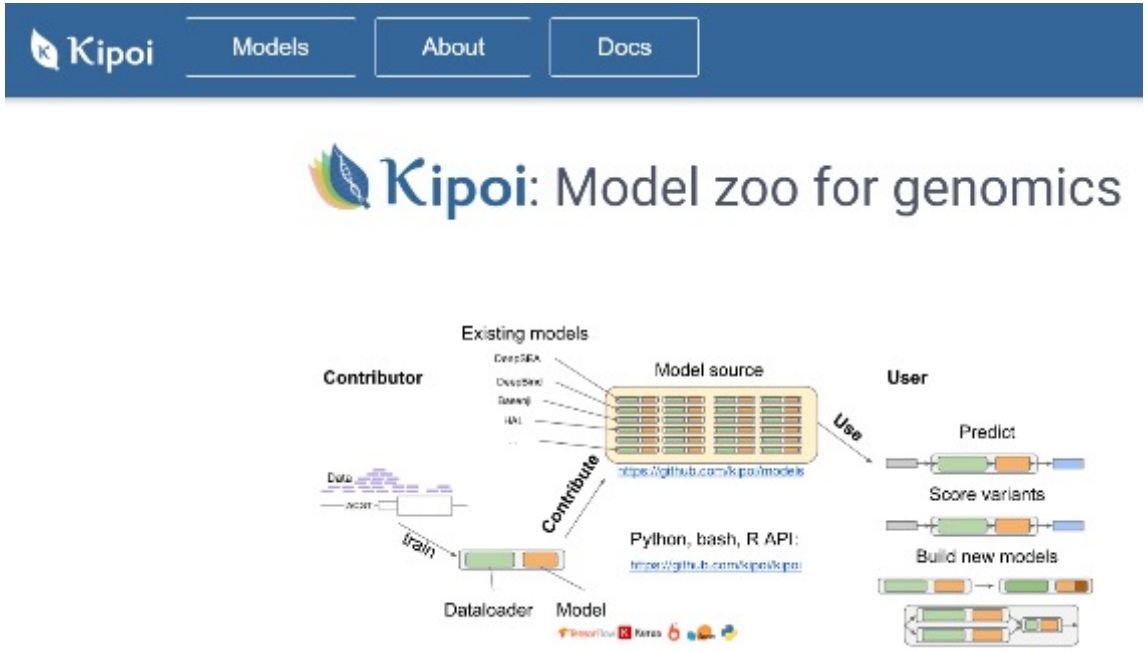


rs1237999 in the PICALM locus for Alzheimer's disease GWAS disrupts an oligodendrocyte-specific FOS enhancer

Summary & Outlook

- Predictive blackbox models + interpretation frameworks
  - Prediction, de-noising & Imputation
  - Biological discovery of causal phenomena
  - Hypothesis generation and optimized experimental design
- Important to be transparent about the limits, blind spots, biases & pitfalls of each model
- What do we need?
  - Large-scale, harmonized ML-ready observational and perturbational data
  - Decentralized, scalable, affordable compute resources
  - Unified ecosystem: Compute ⇔ Data Portals ⇔ Model Zoos ⇔ Literature Mining
  - New user-interfaces to models for interactive discovery, search and design
  - Incentivizes collaborative efforts & diverse contributions

Democratizing ML for genomics: http://kipoi.org/



Kipoi (pronounce: kípi; from the Greek κήποι: gardens) is an API and a repository of ready-to-use trained models for regulatory genomics. It currently contains 1709 different models, covering canonical predictive tasks in transcriptional and post-transcriptional gene regulation. Kipoi's API is implemented as a python package (github.com/kipoi/kipoi) and it is also accessible from the command line or R.

Numbers

# of models: 1709

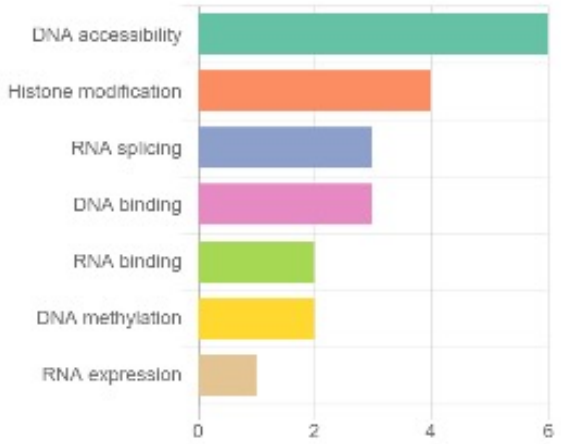
# of model groups: 16

# of contributors: 6

# of model groups supporting postprocessing:

- **Variant effect prediction:** 11/16

Model groups by tag



- Easy installation of dependencies
- Few lines of code to use models to predict
- Exactly reproduce analyses
- Trivial to compare models
- Retrain models
- Fine tune models
- Combine models
- Contribute models

Avsec et al. 2019 Nature Biotech

STANFORD COMPUTER SCIENCE

Kundaje lab

STANFORD SCHOOL of MEDICINE | Genetics


Daniel Kim (BMI)


Kelly Cochran (CS)


Soumya Kundu (CS)


Surag Nair (CS)


Maxim Zaslavsky (CS)


Vivek Ramalingam (Postdoc)


Caleb Lareau (Postdoc)


Akshay Balsubramani (Postdoc)


Georgi Marinov (Postdoc)


Alex Tseng (CS)

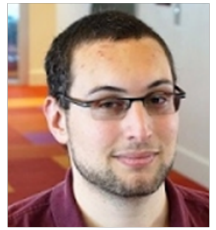

Amr Alexandari (CS)


Abhimanyu Banerjee (Physics)


Laksshman Sundaram (CS)


Anusri Pampari (CS)


Kristy Mualim (Bioinformatician)

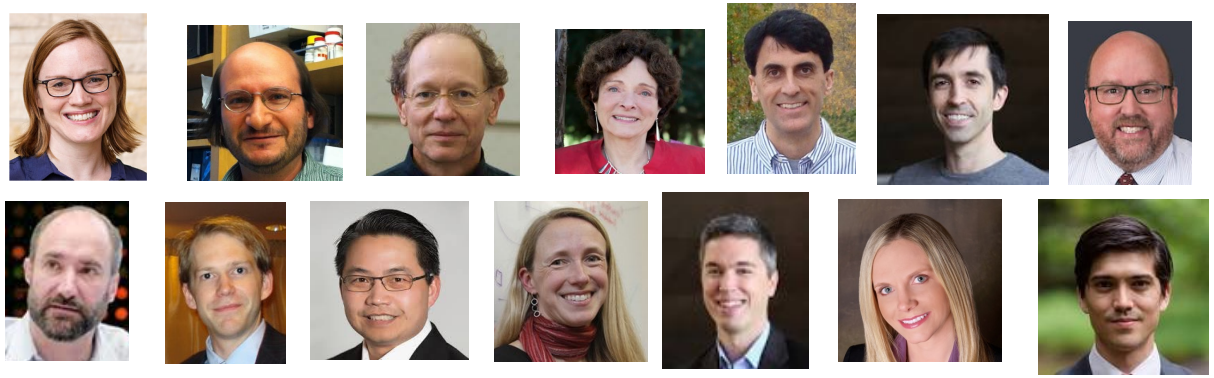

Jacob Schreiber (Postdoc)


Mahfuza Sharmin (Postdoc)


Eran Kotler (Postdoc)


Zahoor Zafrulla (ML engineer)

Collaborators



Funding

1R01HG009674

1U01HG009431

1U24HG009446

R01ES02500902

1DP2OD022870