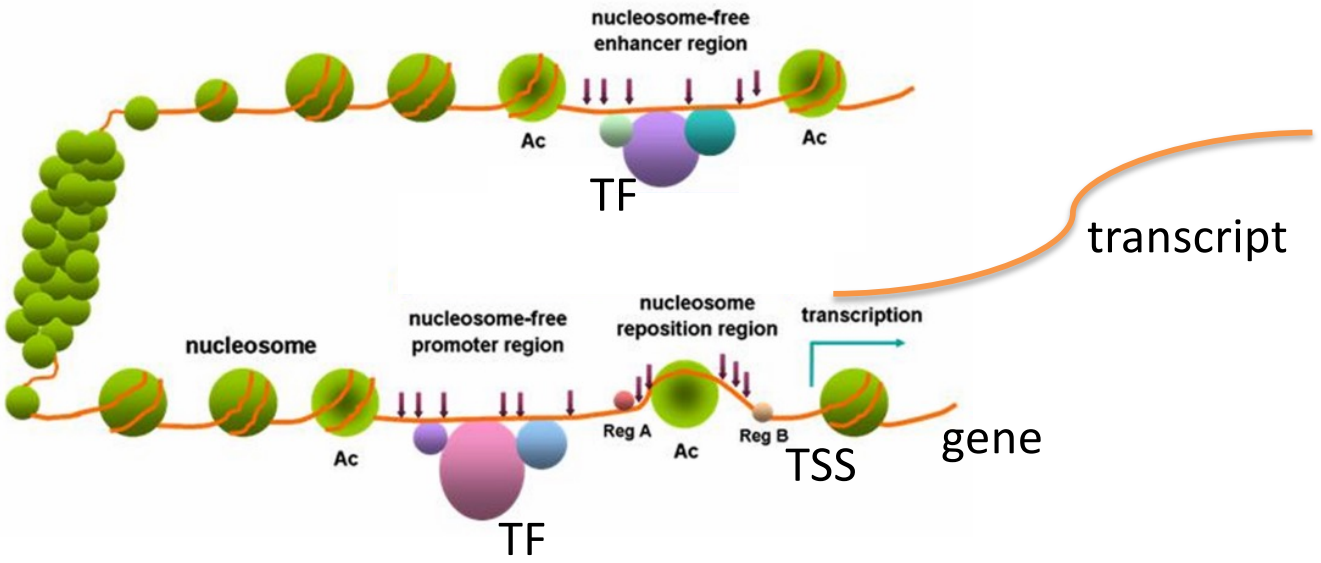
The 3D genome and predictive models of gene regulation

Christina Leslie
Computational and Systems Biology Program
Memorial Sloan Kettering Cancer Center
http://cbio.mskcc.org/leslielab

Deciphering gene regulation

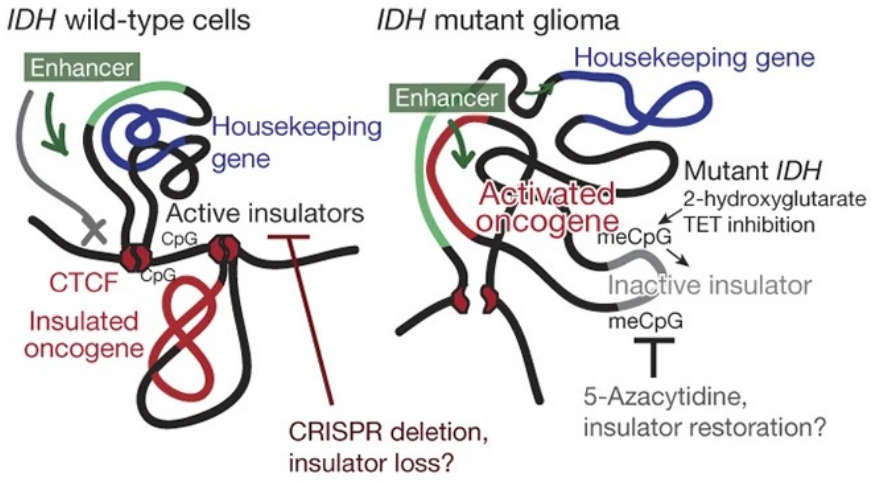
- Transcriptional regulation coordinated by transcription factors (TFs) binding to promoter and enhancer elements
- Distal enhancers may be >1Mb from promoters, physically interact via chromatin looping
- 1D epigenomic data (chromatin accessibility, histone marks) map candidate enhancer elements but not their connectivity



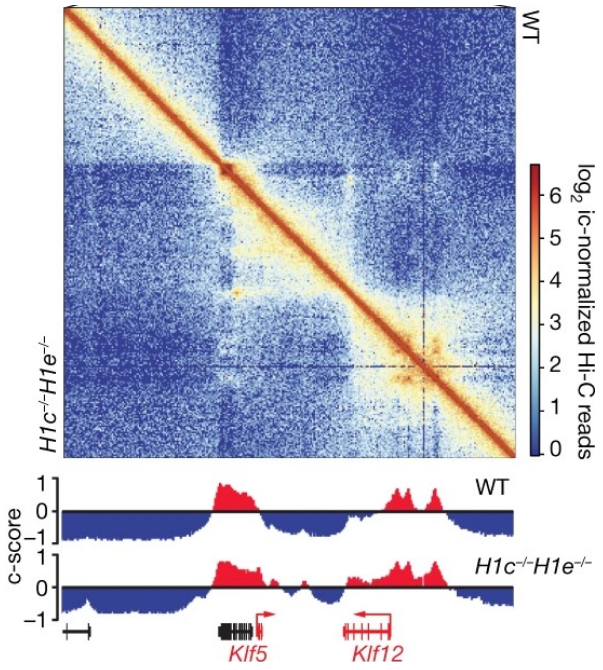
- Advances in **chromosome conformation capture** assays (e.g. Hi-C, HiChIP) can now resolve 3D genomic interactions relevant to gene regulation

3D dysregulation in cancer cells

- Somatic alteration or epigenetic disruption of 3D organization can lead to oncogene activation, tumorigenic processes
  - E.g. hypermethylation can lead to loss of insulation of *PDGFRA* in *IDH* mutant gliomas
  - E.g. loss of H1 linker histone leads to chromatin “decompaction” in germinal center B cells



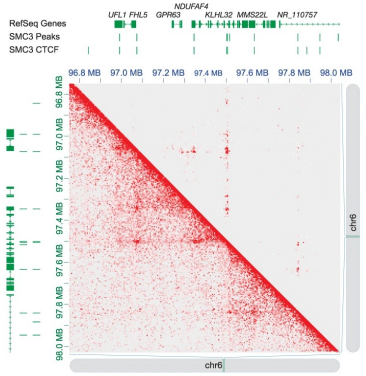
Flavahan et al., *Nature* 2016



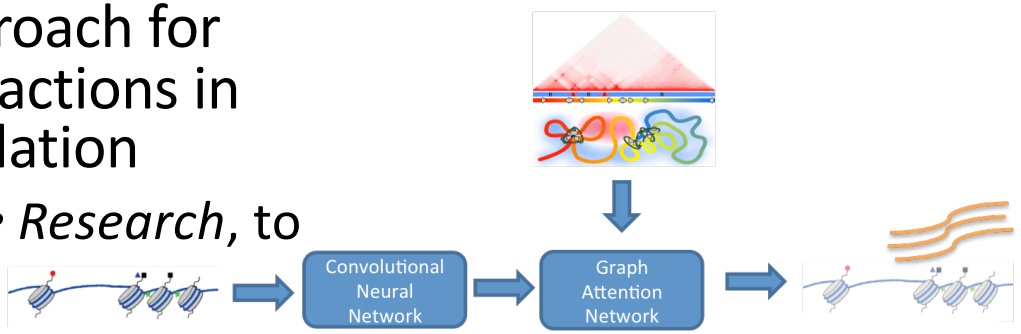
Yusufova et al., *Nature* 2021

Outline

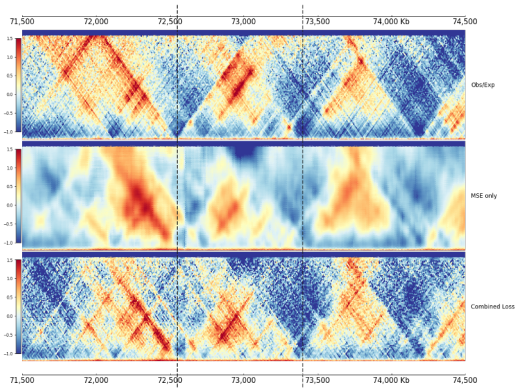
- HiC-DC+: a statistical framework for calling significant and differential interactions in Hi-C and HiChIP
  - Sahin et al., *Nat Commun* 2021



- GraphReg: a deep learning approach for incorporating 3D genomic interactions in predictive models of gene regulation
  - Karbalayghareh et al., *Genome Research*, to appear



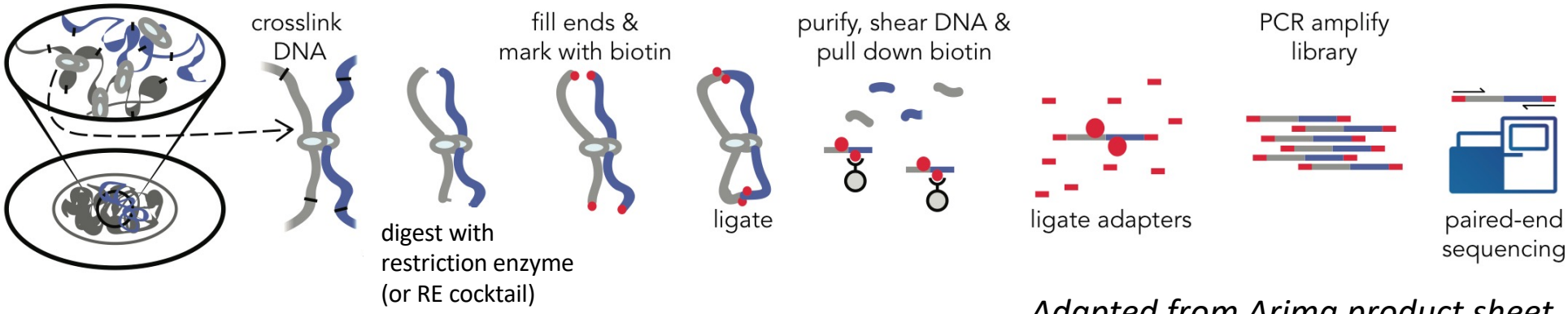
- Epiphany: a deep learning model to predict the 3D contact matrix from 1D epigenomic data
  - Yang, Das, et al., *bioRxiv* 2021



Mapping the 3D genome and calling 3D interactions

Mapping the 3D genome

- Hi-C, chromosome conformation capture
  - Capture 3D interactions: crosslink DNA (now in situ), restriction enzyme digest, proximity ligation, pull down, paired-end sequencing

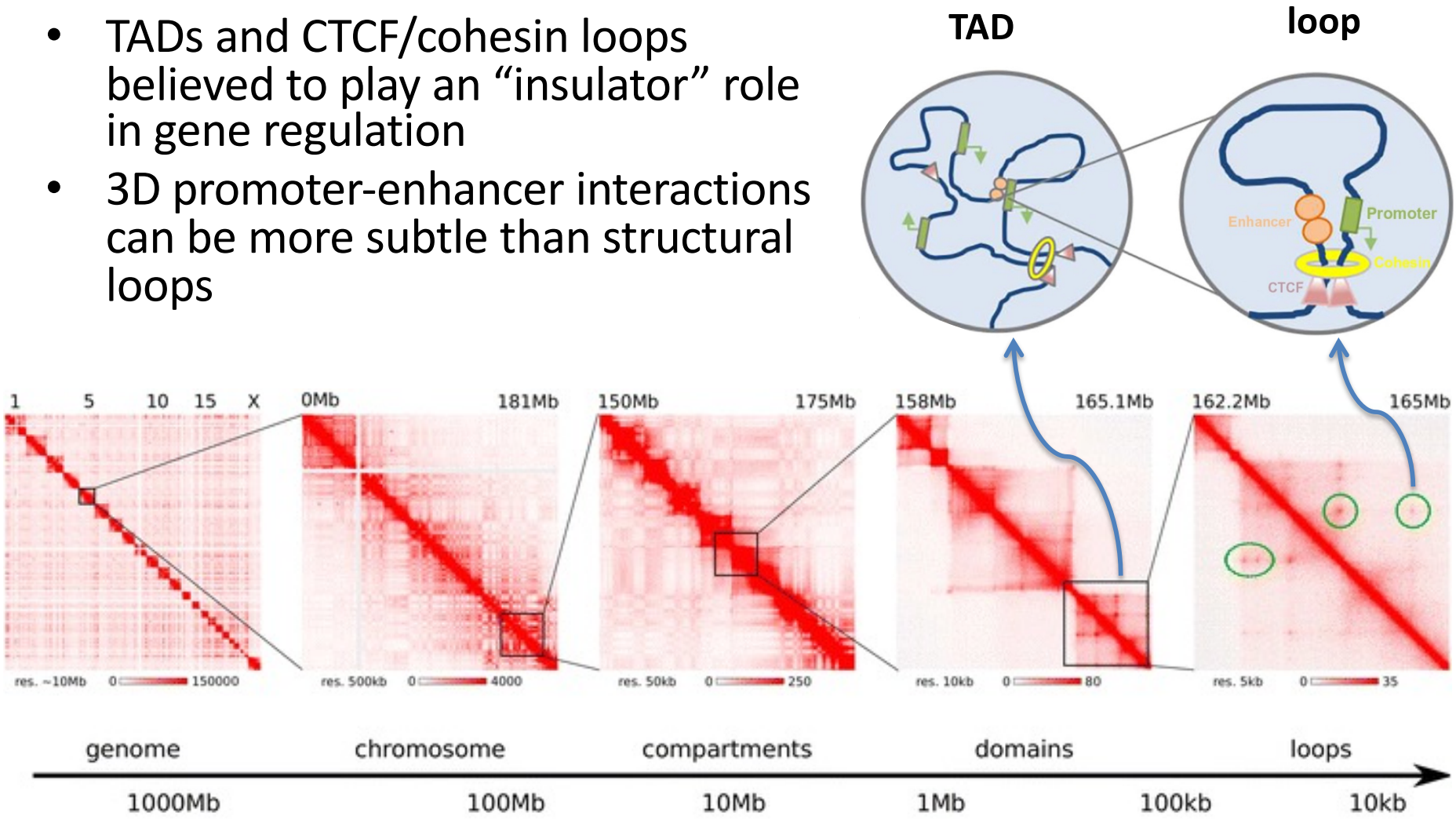


Adapted from Arima product sheet

  - Read pair = "contact"; build contact matrix for input cell population: C_{ij}= #paired end reads with anchors in bin_i and bin_j

Hierarchical folding of chromatin

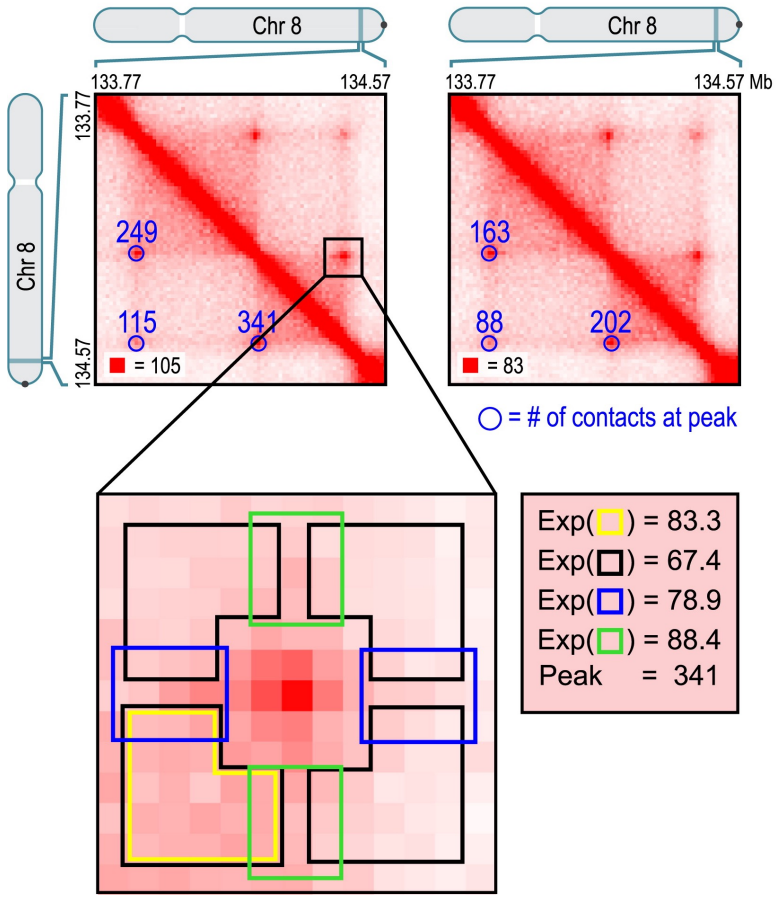
- TADs and CTCF/cohesin loops believed to play an “insulator” role in gene regulation
- 3D promoter-enhancer interactions can be more subtle than structural loops



Adapted from Wright et al., 2019

Calling 3D loops vs “interactions”

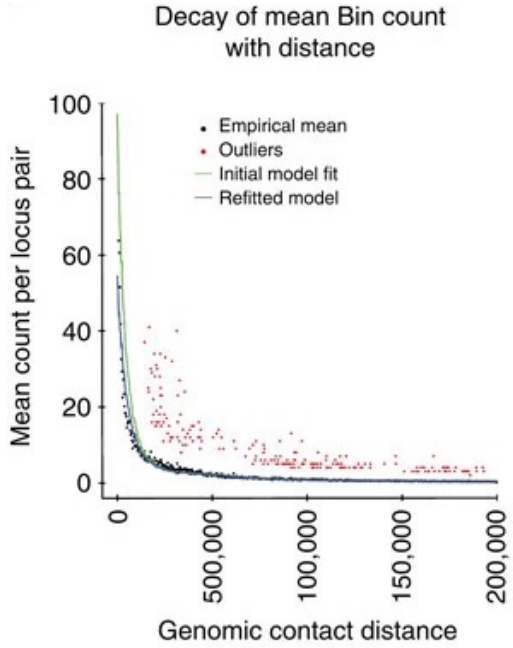
- Typical Hi-C loop callers treat the contact map like an image
  - Find pixels that are brighter than surrounding pixels
  - May use normalization, smoothing to improve signal-to-noise
  - Generally no good estimate of statistical significance
  - Conservative, calls structural “loops”
- Need more sensitive approach to find 3D “interactions” like promoter-enhancer contacts

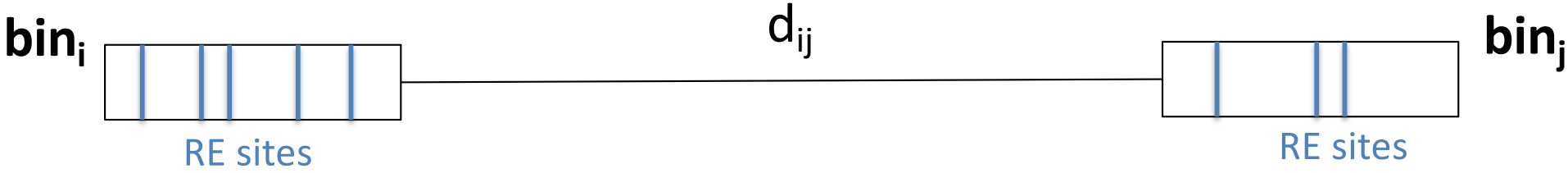


Rao et al., *Cell* 2014

Methods matter: HiC-DC+

- "Hi-C direct caller": use read counts from raw contact matrix directly, without normalization
  - Estimate background model (expected read count) directly from data using negative binomial regression
  - Covariates: genomic distance (spline fit), mappability, effective bin size (related to restricting enzyme density), GC content
  - Assign *P* value (or *Z*-score) to interactions



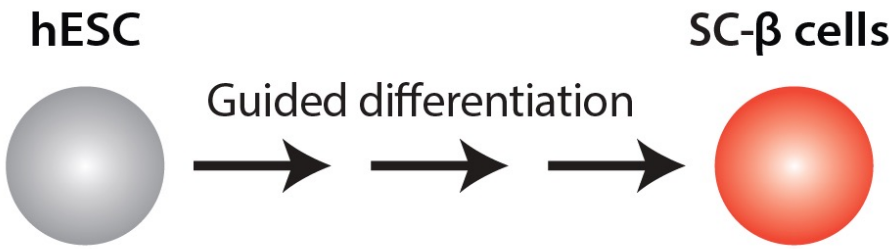


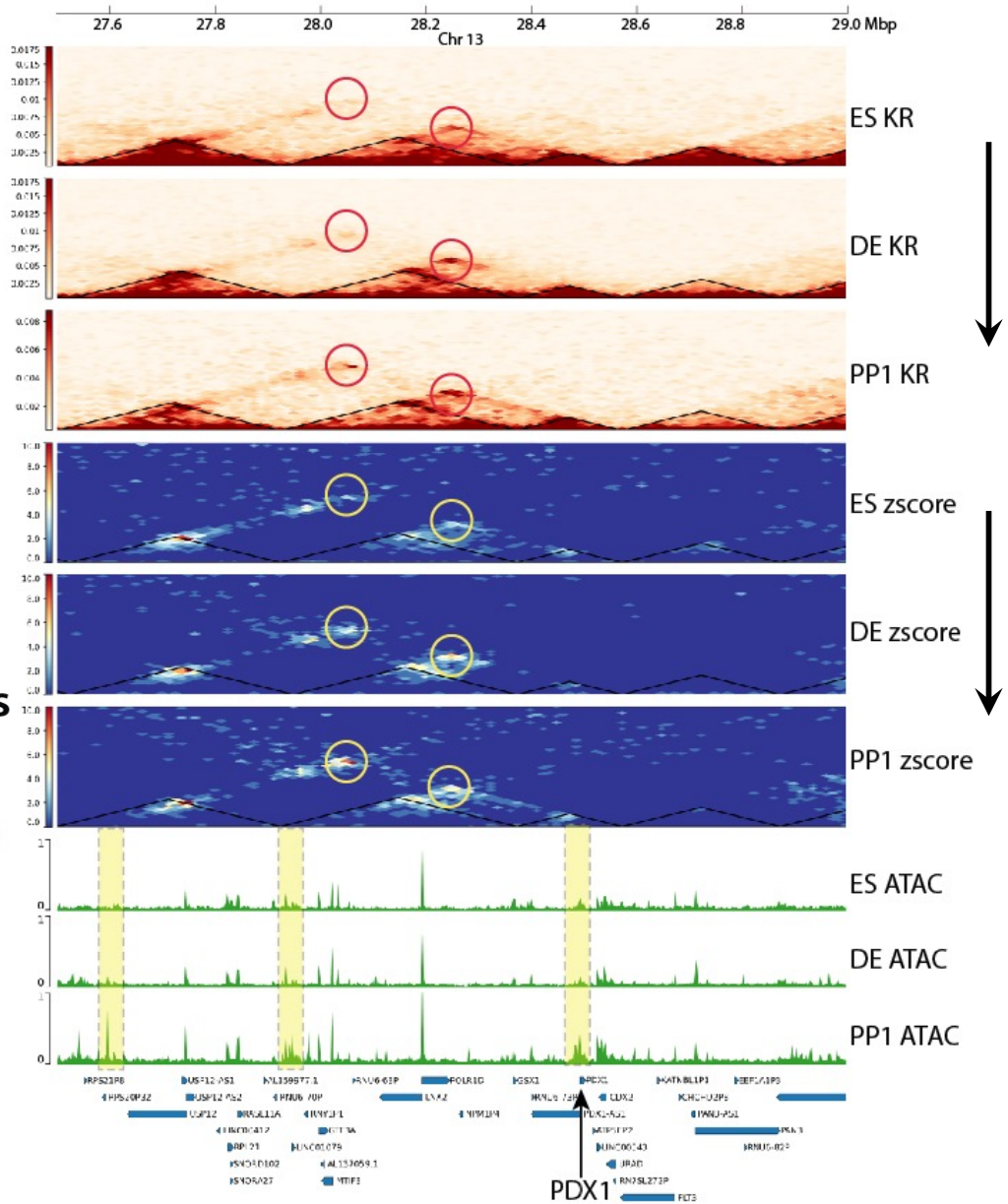
- HiC-DC+: Efficient code, extends to HiChIP, *differential* interactions between cell types

Carty et al., Nat Commun 2017;
Sahin et al., Nat Commun 2021

Methods matter: HiC-DC+

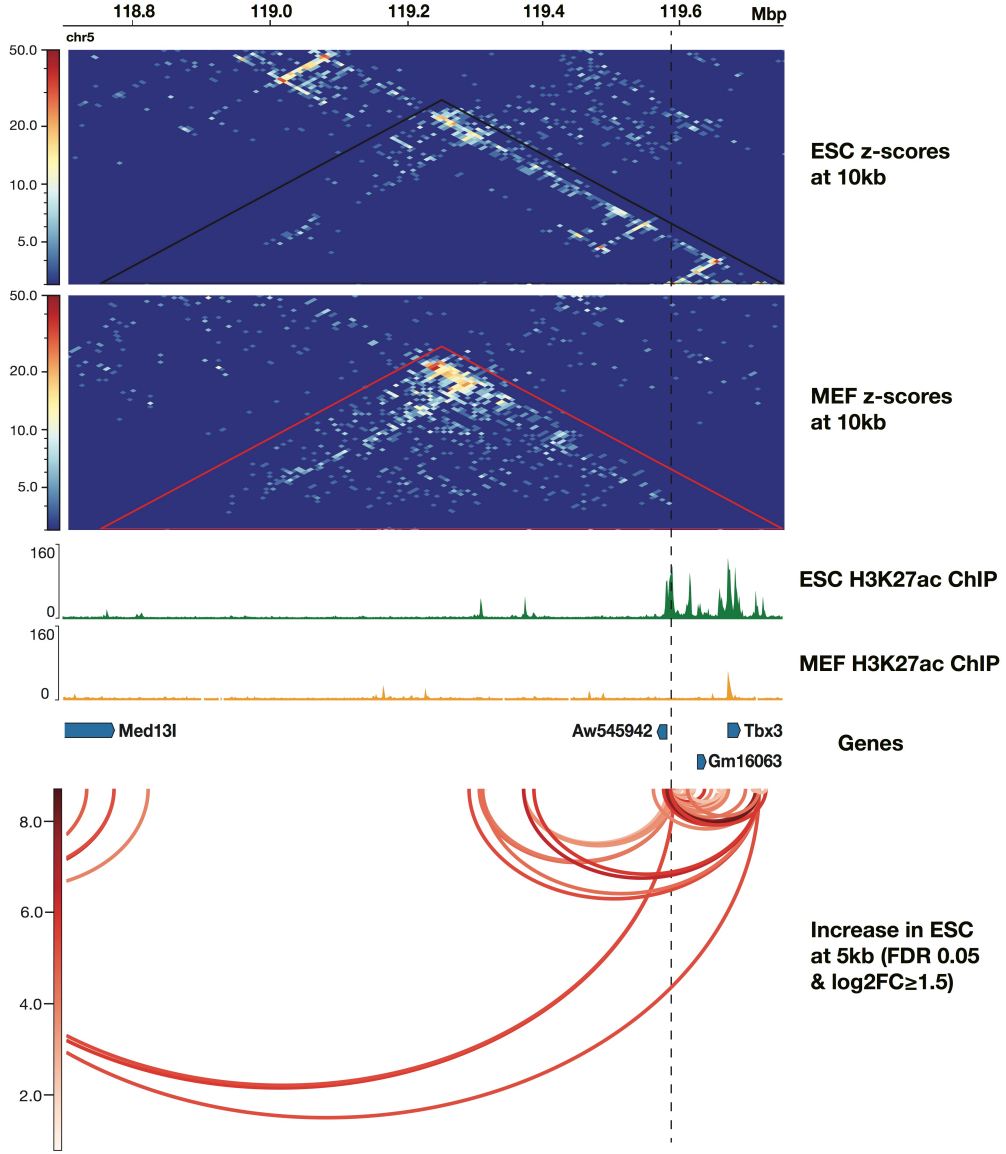
- Gain of promoter-enhancer for developmental gene *PDX1* in guided pancreatic differentiation
  - With Danwei Huangfu and Eftychia Apostolou (as 4D Nucleome project)





HiC-DC+ analysis of HiChIP

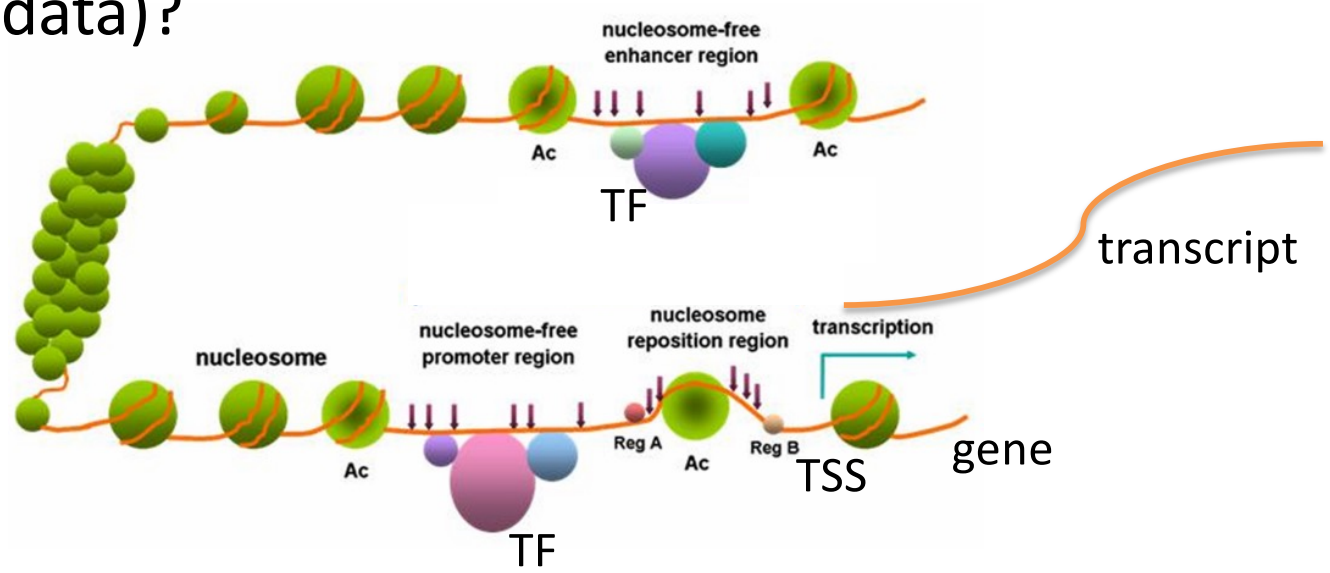
- HiChIP: Hi-C contact library followed by chromatin IP for protein/histone mark of interest
  - H3K27ac HiChIP: regulatory interactions, i.e. promoter-enhancer, enhancer-enhancer, etc.
- E.g. “enhancer hub” via differential analysis between mouse ESC and MEFs
  - Data from Effie Apostolou lab (cf: Di Giammartino et al., 2019*)*



Graph neural networks for predictive models of gene regulation

Predictive models of gene regulation

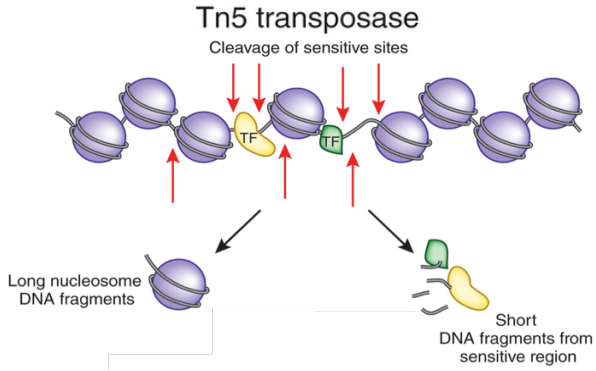
- Can we learn to predict gene expression levels from 1D (e.g. DNA sequence, epigenomic signals) and 3D (physical interaction data)?

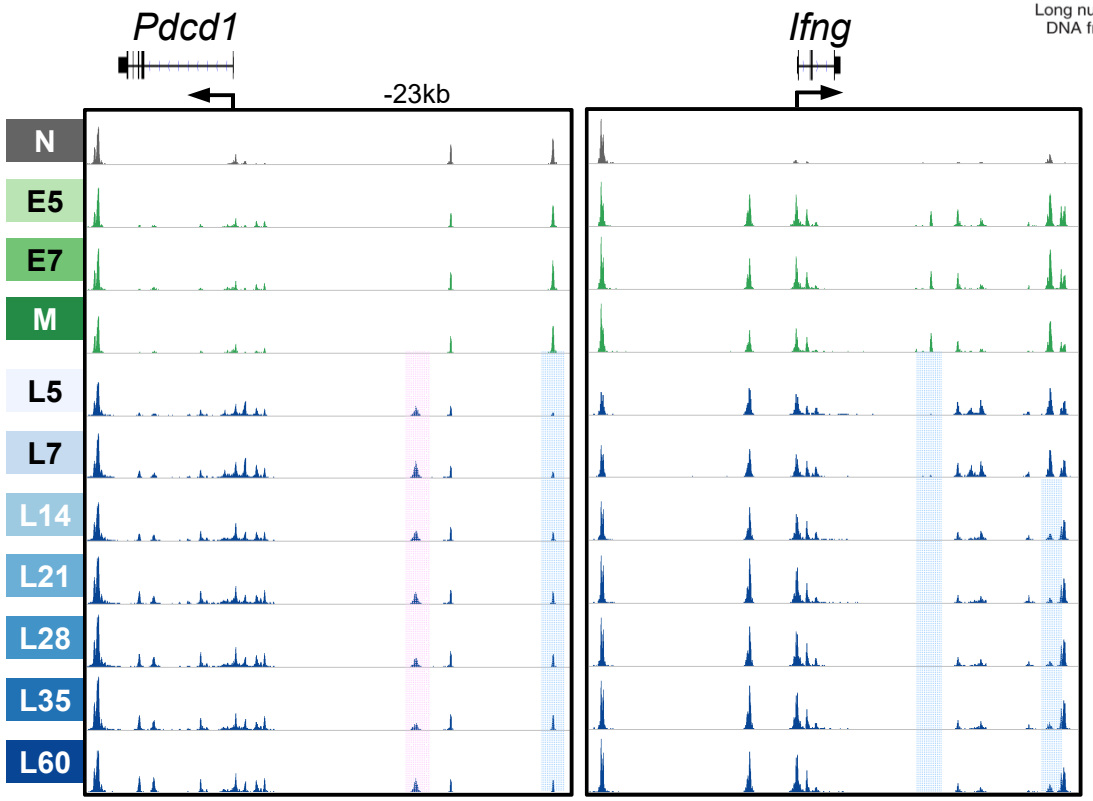


- If we could learn an accurate predictive model and *interpret* the model, we could:
  - Identify functional enhancers and TF regulators of genes
  - Predict how gene expression would change under perturbations

Epigenomic data encodes regulatory information

- E.g. chromatin accessibility (ATAC-seq) maps local regulatory elements and encodes global differentiation state





Functional CD8 T cells

Tumor-specific dysfunctional CD8 T cells

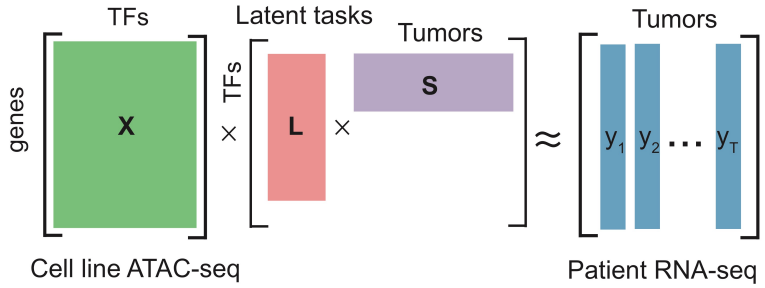
low *Pdcd1*, high *Ifng* expression

gain of *Pdcd1*, loss of *Ifng* expression

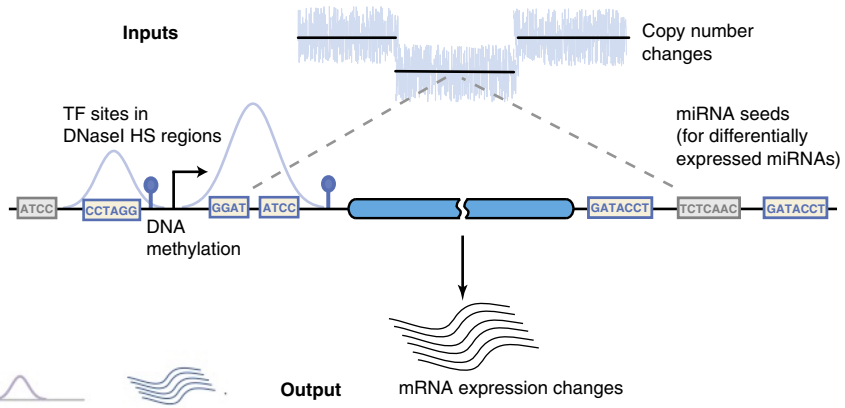
Philip et al., Nature 2017

Predictive gene regulatory models

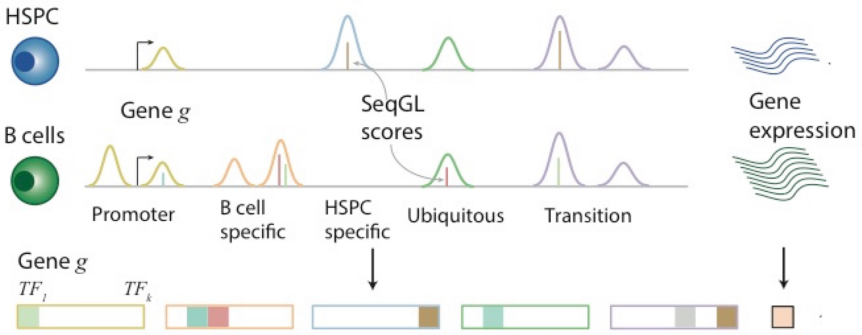
- Previous GRMs predict gene expression (or fold change) from DNA sequence and accessibility/activity of regulatory elements *in order to decipher gene regulation*



Osmanbeyoglu et al., Nat Commun 2019



Setty et al., Mol Syst Biol 2012

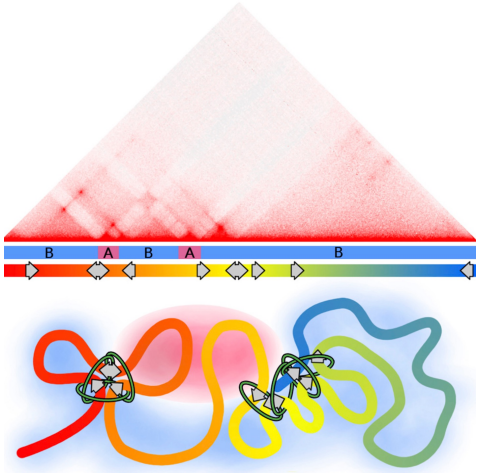


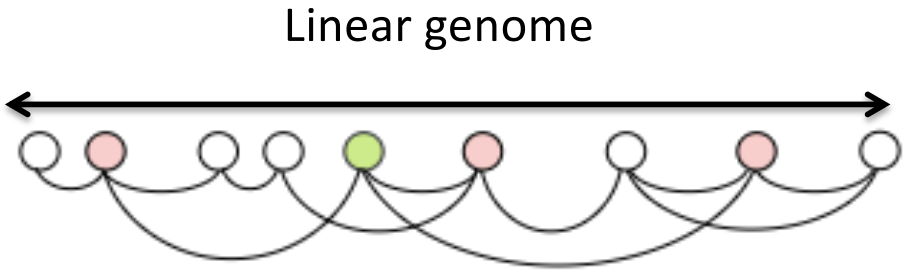
Gonzalez, Setty* et al., Nat Genet 2015*

- Missing information: *connectivity* of promoter and enhancers
- Idea: use 3D interaction data in graph neural network GRMs

GraphReg: graph neural networks for gene regulatory models

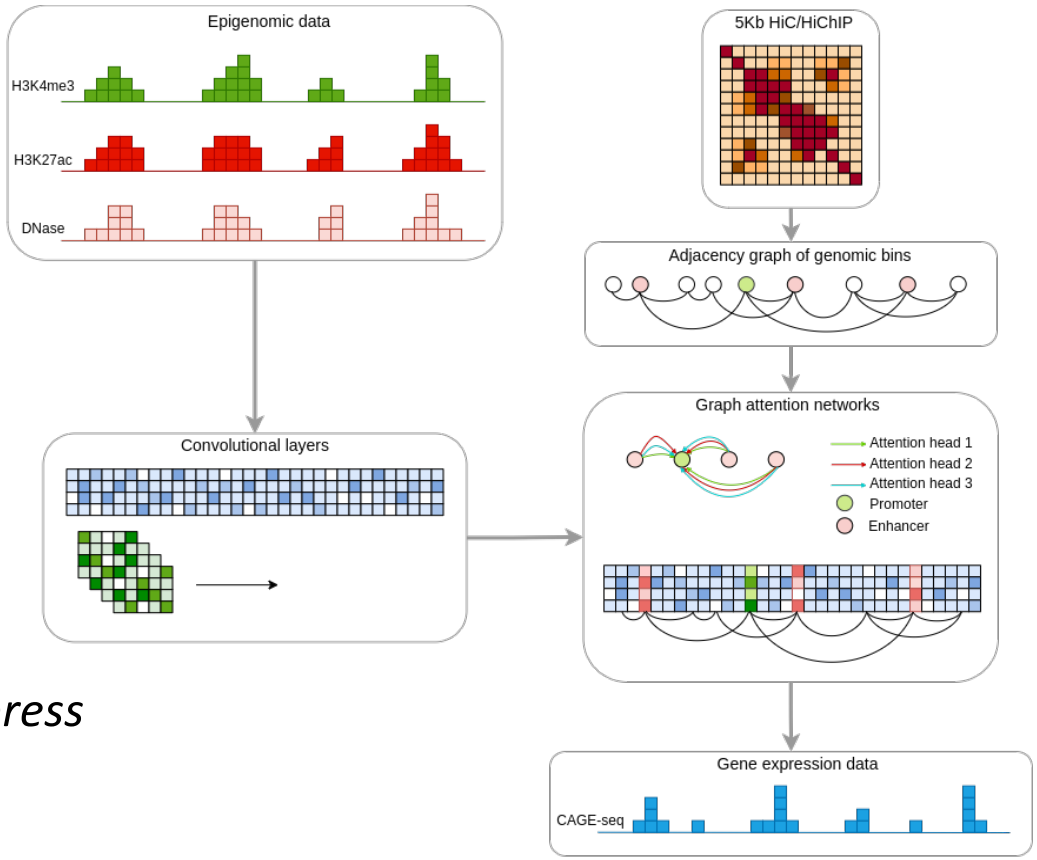
- Idea: use Hi-C/HiChIP to encode long-range chromatin interactions as a graph, propagate information information via *graph neural networks* (GNNs)
- Nodes of graph = genomic bins, edges = 3D genomic interactions
- Input features: epigenomic data or DNA sequence
- Output: gene expression (at node)
- Compare to (dilated) *convolutional neural networks* (CNNs), use only 1D data





Epigenome-based GraphReg

1D input: chromatin accessibility and histone modifications data



3D input: regulatory chromatin interactions (H3K27ac HiChIP)

Karbalayghareh et al., Genome Research, in press

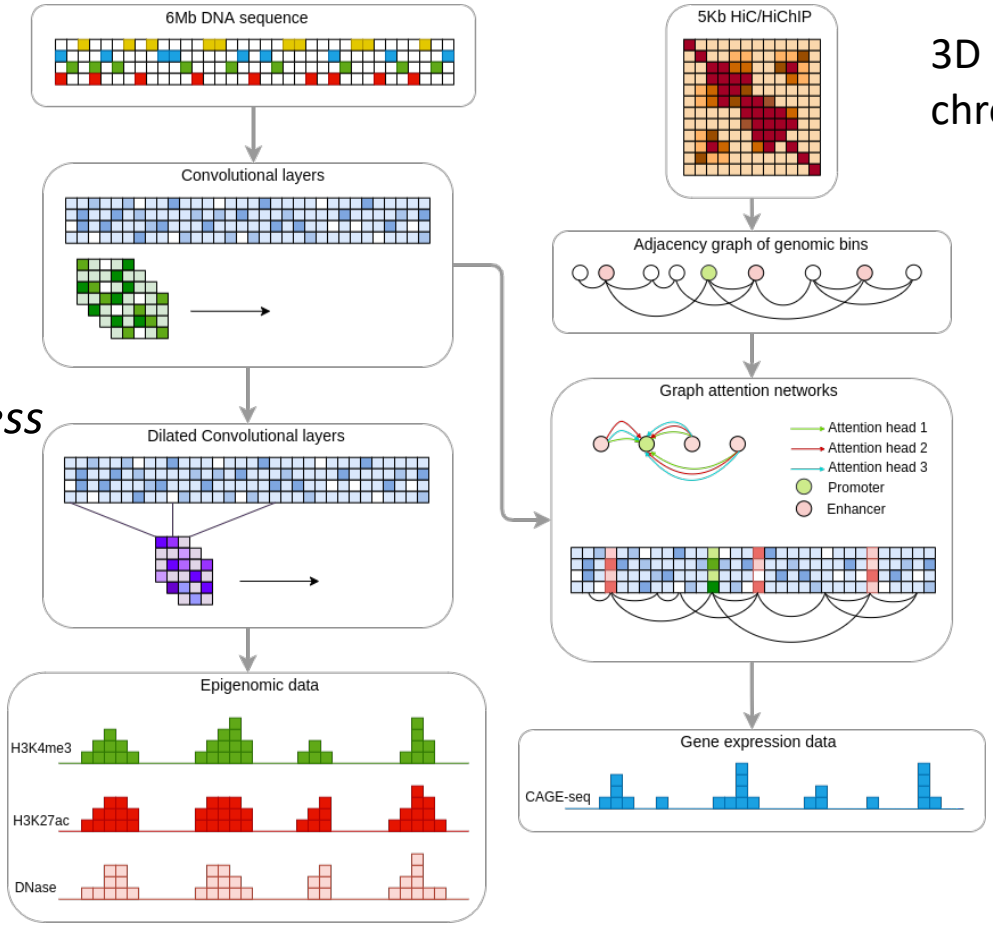
Output: CAGE-seq (gene expression at TSS)

- Predict gene expression from *activity* and *connectivity* of regulatory elements
- "Cell-type-agnostic": can generalize to a new cell type given cell-type specific 1D and 3D inputs

Sequence-based GraphReg

1D input: DNA sequence

Karbalayghareh et al., Genome Research, in press



3D input: regulatory chromatin interactions

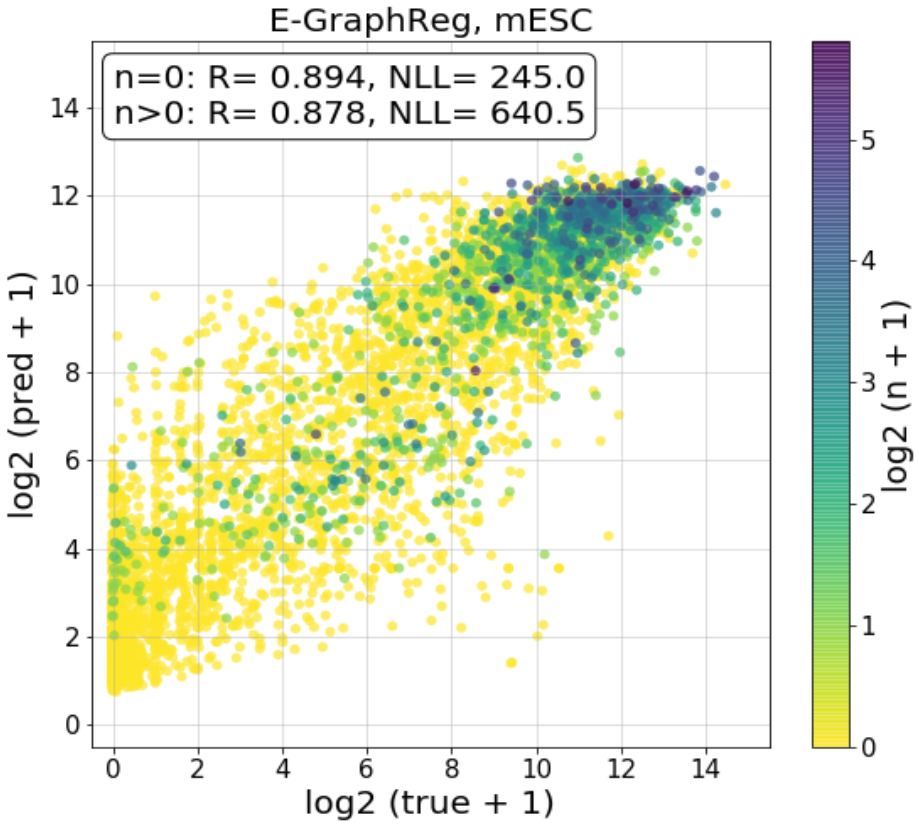
Output: accessibility, histone marks

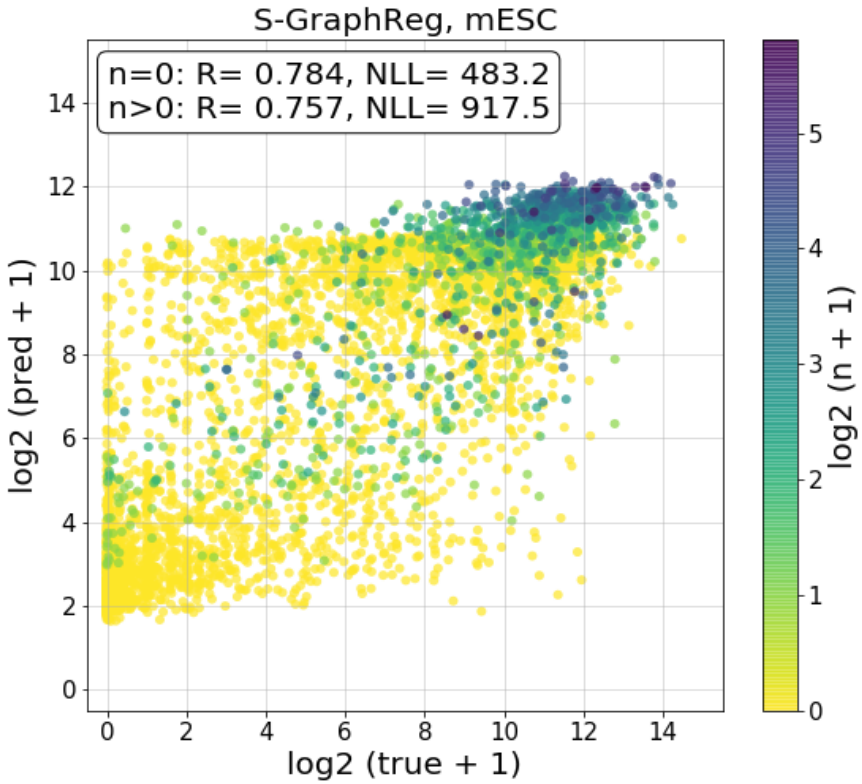
Output: CAGE-seq

- Predict expression and 1D epigenomic signals from genomic DNA sequence + 3D connectivity
- “Cell-type-specific”: captures TF binding signals that are specific to the training cell type

Prediction of gene expression

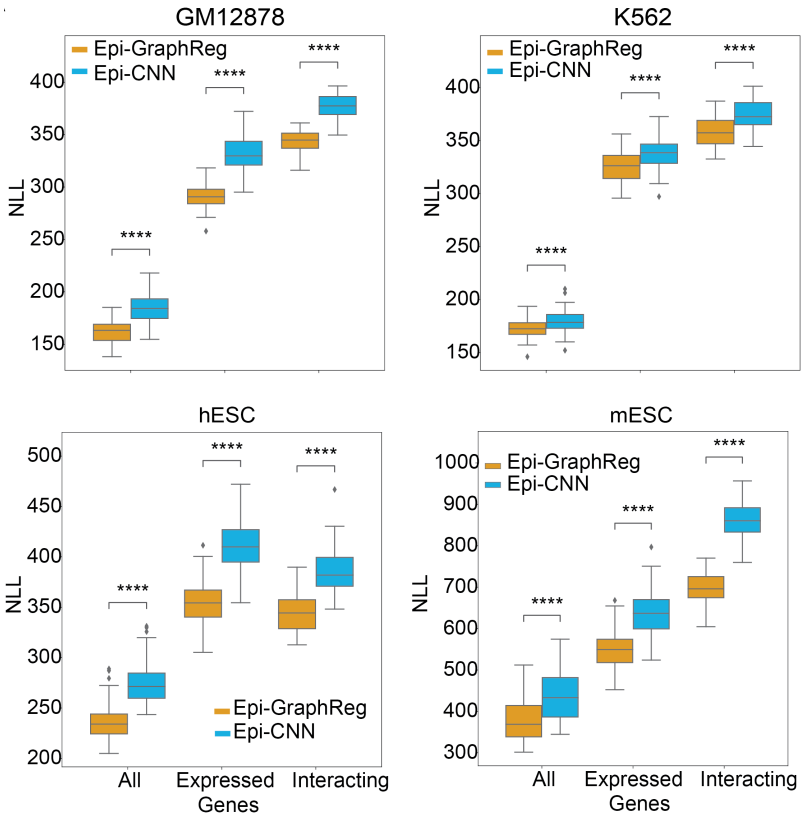
- Train on cell line data, assess performance on held-out chromosomes



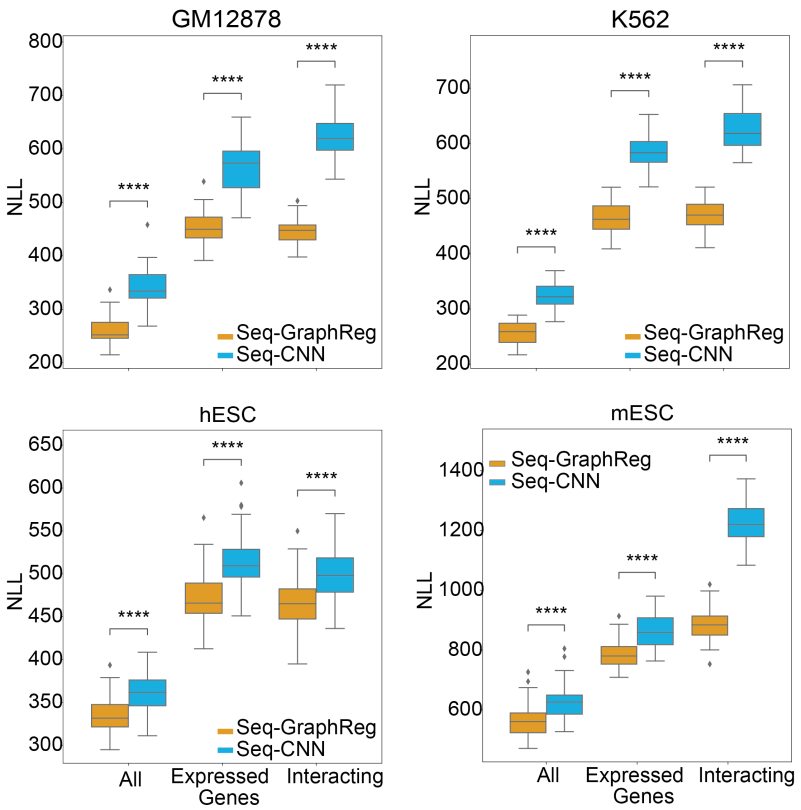


Prediction performance

Epigenome-based models



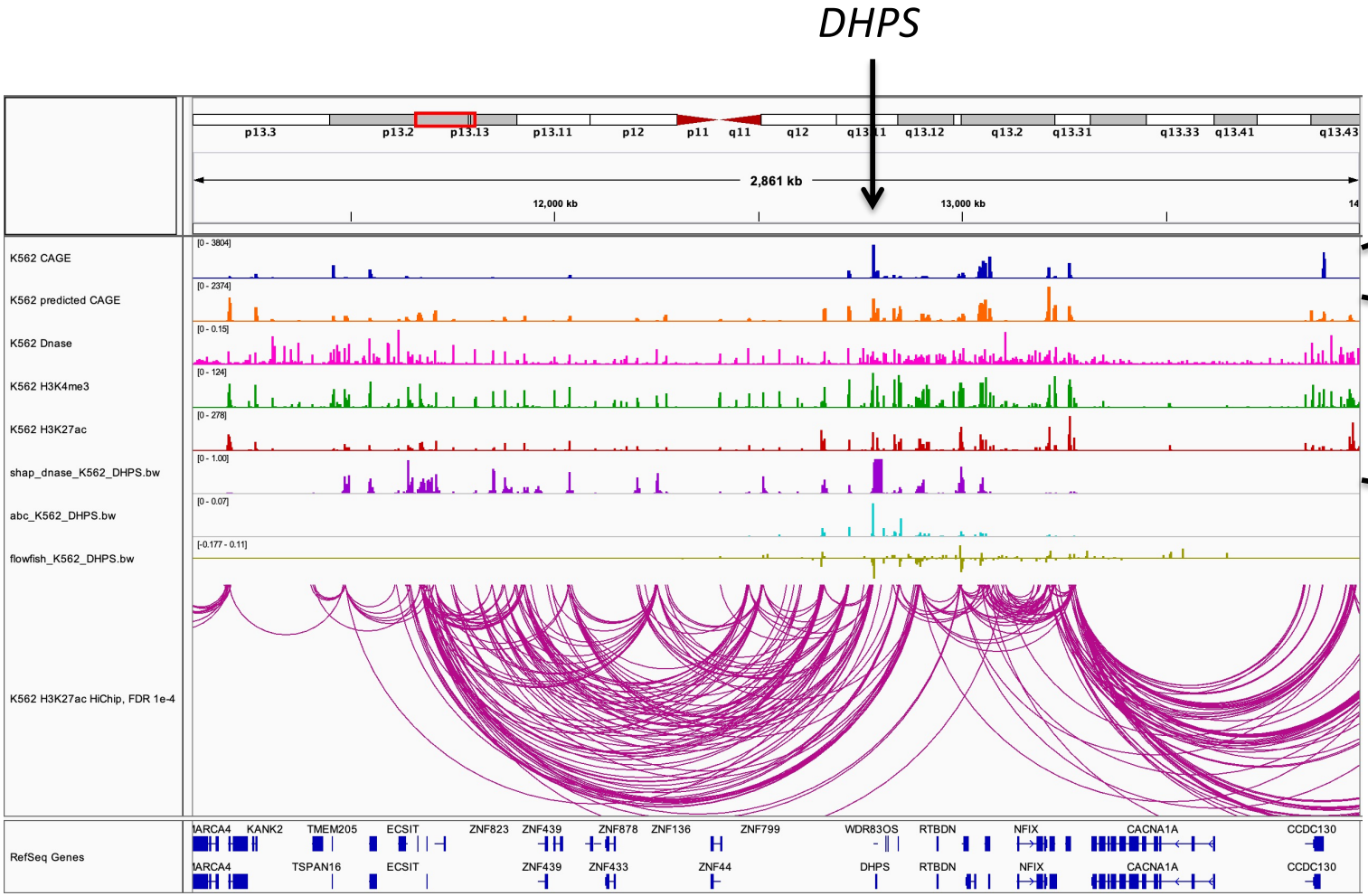
Sequence-based models



- GraphReg models outperform baseline 1D dilated CNNs
- Sequence-based prediction is more difficult
- Prediction of expression *per se* is not the point: want to interpret the model

Feature attribution to predict functional enhancers

- DeepSHAP identifies features/genomic bins that contribute most to specific gene predictions



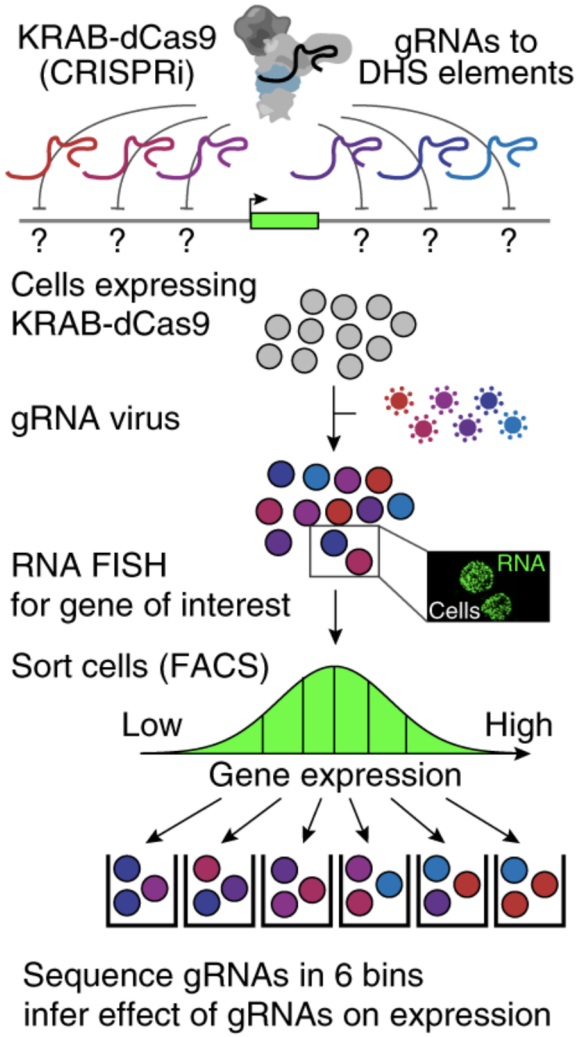
true signal

predicted signal

feature attribution (DeepSHAP) for the gene *DHPS*

Epi-GraphReg

Evaluation of enhancer prediction with FlowFISH

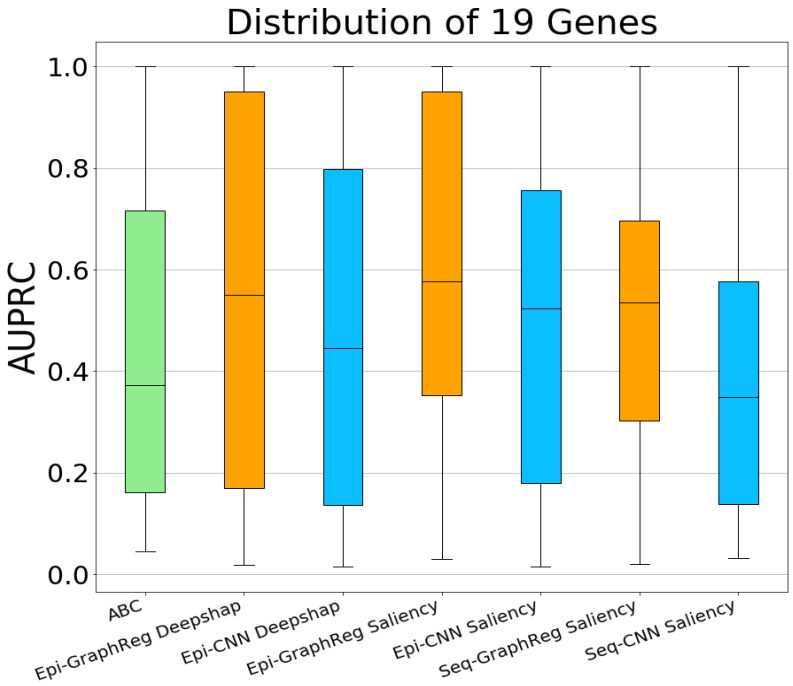


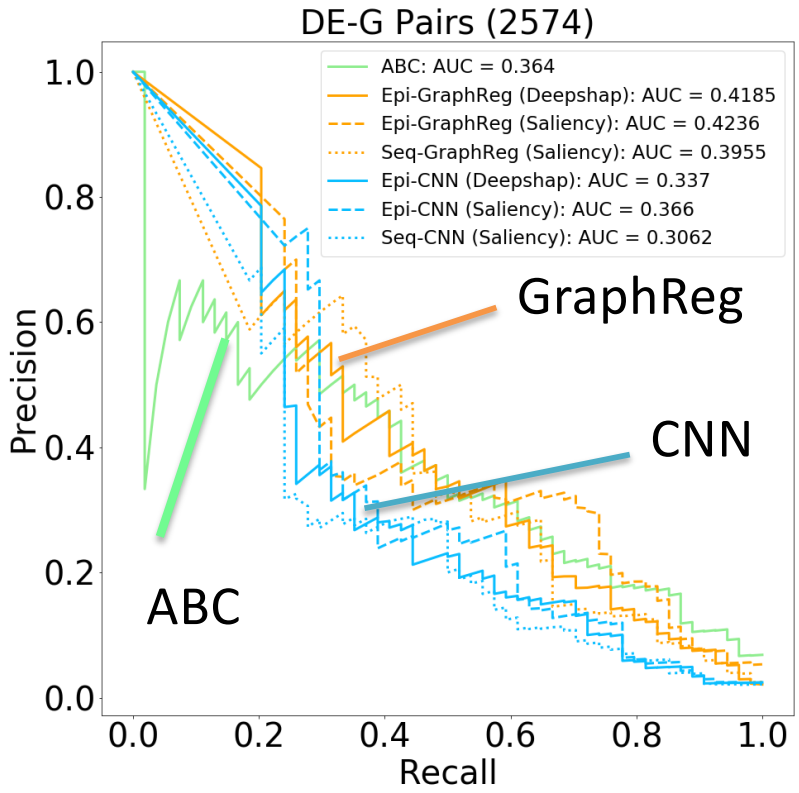
Fulco et al., Nat Genet 2019

- CRISPRi-FlowFISH: CRISPR inactivation screen against candidate enhancers, reads out expression change of target gene
- Activity-by-contact (ABC): score for predicting functional enhancers based on activity (DNase, H3K27ac) and Hi-C contacts

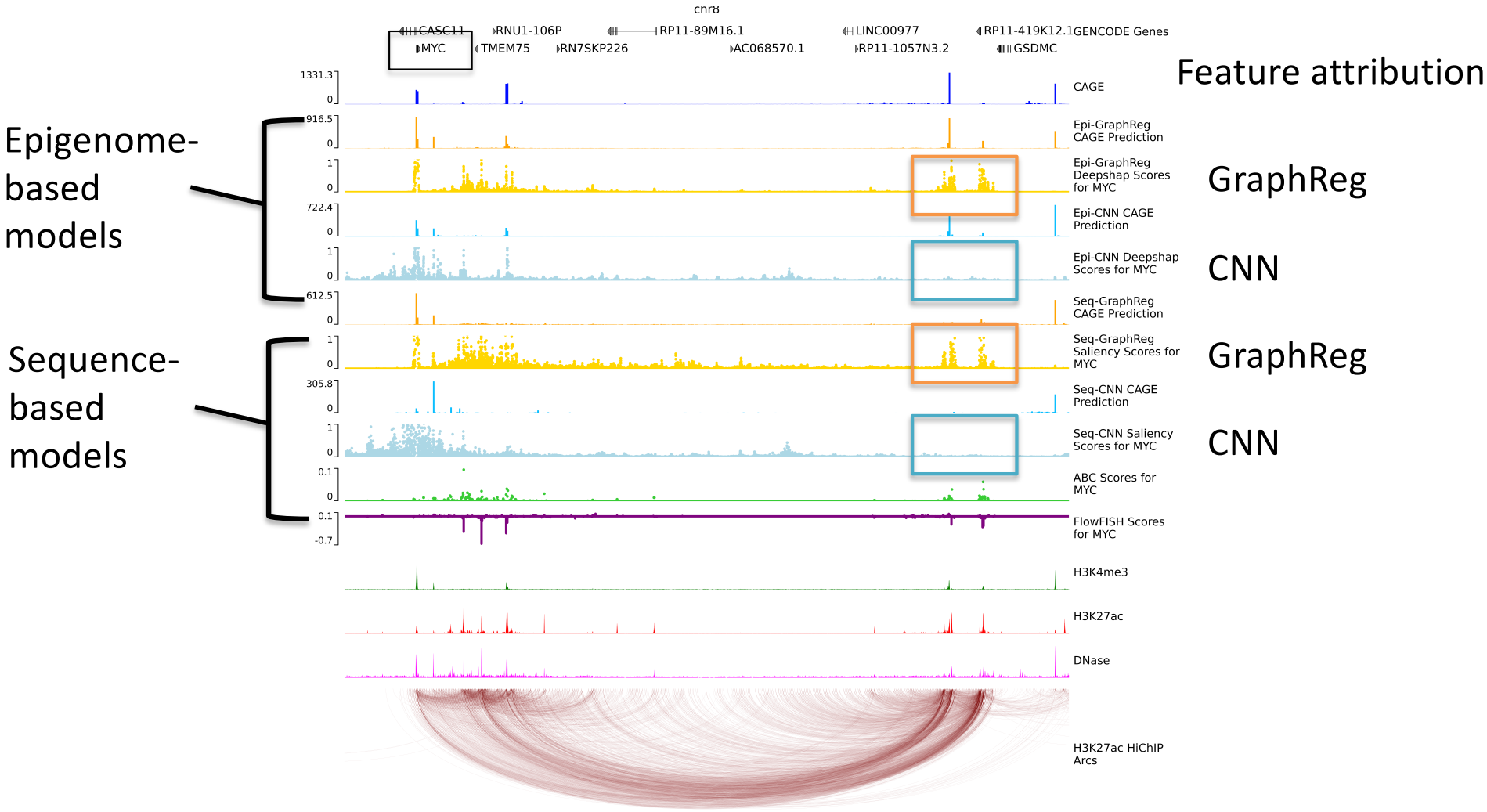
GraphReg improves functional enhancer prediction

- Use FlowFISH experiments sufficient data on distal elements (2574 candidate elements for 19 genes)
- GraphReg models with DeepSHAP or saliency outperform CNN models, ABC





GraphReg models access distal information unavailable to CNNs



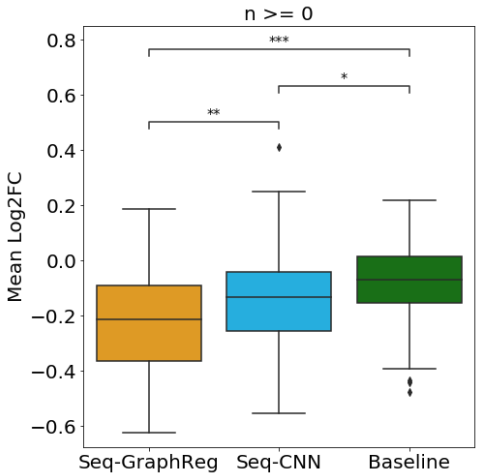
- Dilated CNNs can accept large input region, but feature attribution shows they rely on promoter-proximal signals

GraphReg predicts gene expression changes under TF knockout

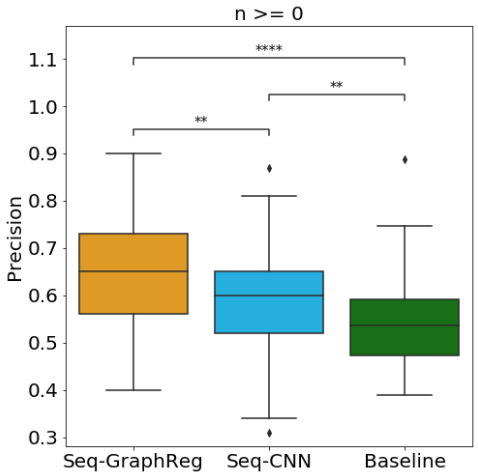
- Can we test that Seq-GraphReg is learning meaningful sequence information?
- *In silico* TF KO via motif ablation:

  - Predict expression of g from original sequence and from sequence with TF hits ablated (set to 0), get predicted logFC

- Compare results to true TF knockdown in K562 cells from ENCODE (29 CRISPRi experiments)



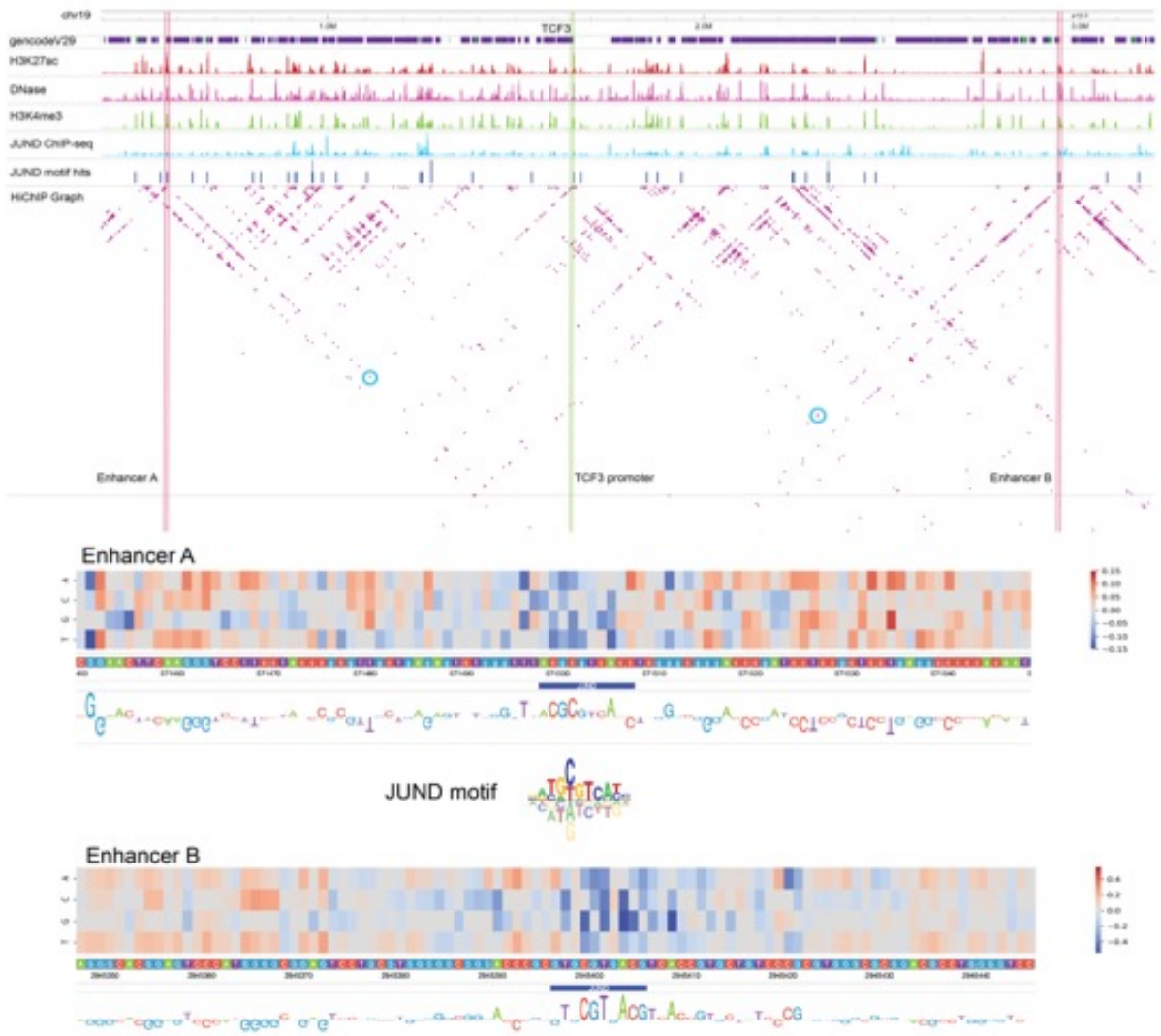
Mean logFC



Precision

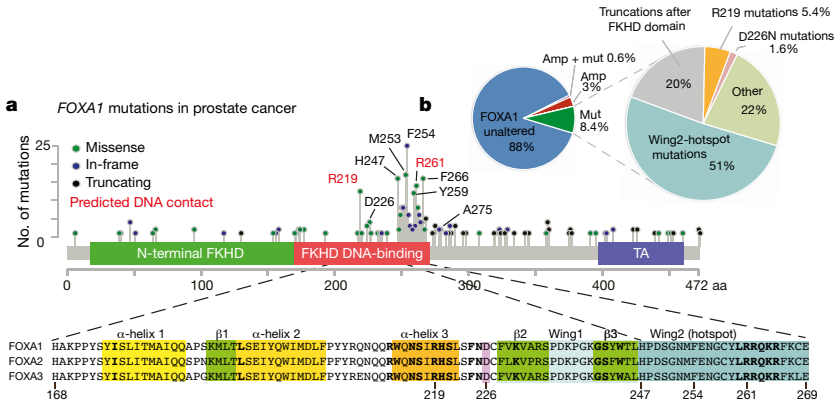
GraphReg identifies TF binding events that contribute to gene regulation

- Gene *TCF3*, downregulated upon JUND KO
- Enhancers A and B have direct HiChIP interactions with promoter
- In silico mutagenesis identifies JUND motifs in both distal enhancers



Coming next: adapting regulatory models to sc-multiome

- High-quality scATAC + scRNA co-assay data enables new algorithmic possibilities
- E.g. Mutant *FOXA1* alleles in prostate organoids (with Charles Sawyers lab)

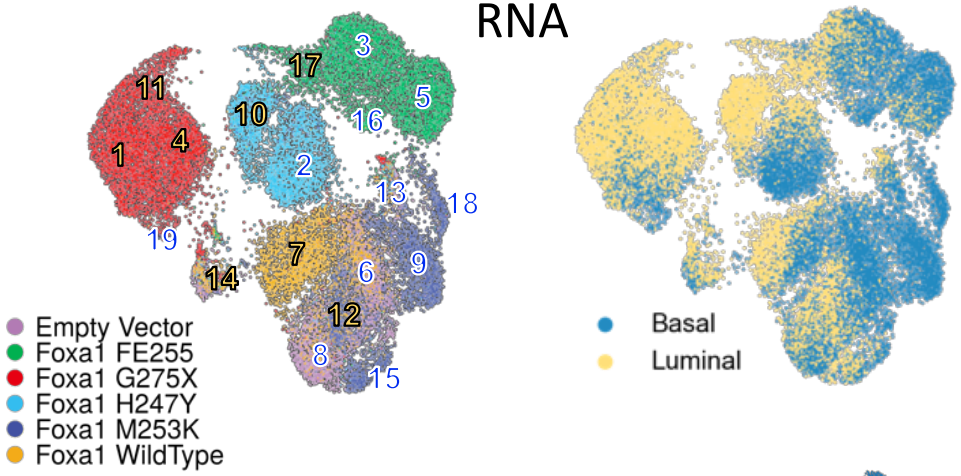


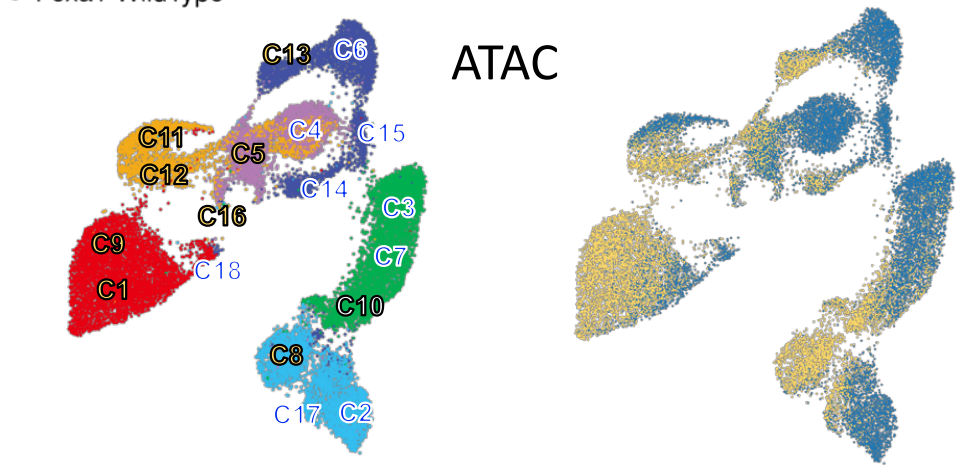
c

MSK-IMPACT 504 mutant frequencies by stage	R219	Wing2-hotspot alterations	Others	Total
Locoregional (*n* = 30)	6.7% (2)	66.7% (20)	26.7% (8)	100%
Metastatic non-castration-resistant (*n* = 13)	0.0% (0)	38.5% (5)	61.5% (8)	100%
Castration-resistant (*n* = 18)	11.1% (2)	44.4% (8)	44.4% (8)	100%

d

Point mutation frequencies in adenocarcinoma vs NEPC	*FOXA1* mutant cases			Non-*FOXA1* mutant cases
	R219 point mutants	Other point mutants	Total	
Adenocarcinoma (*n* = 1,742)	9.5% (8)	90.5% (76)	100% (84)	1,658
NEPC (*n* = 80)	75% (3)	25% (1)	100% (4)	76





Conclusions (GraphReg)

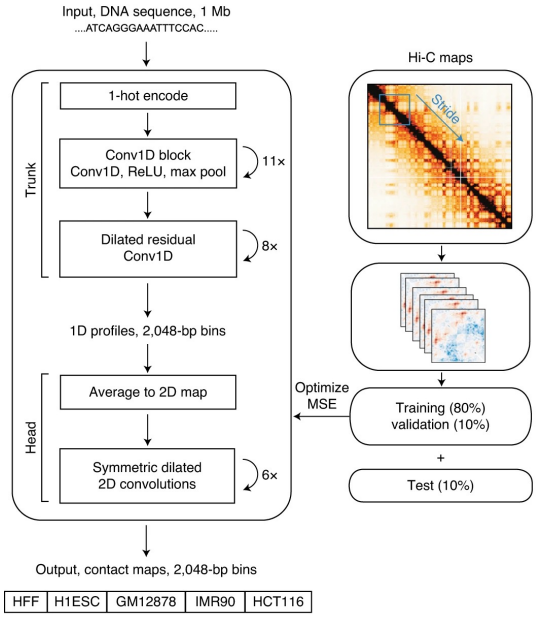
- Graph neural network model can predict gene expression (TSS output) across large genomic regions from 3D and 1D data, or from DNA sequence using 1D epigenomic prediction as auxiliary task
- Epi-GraphReg and Seq-GraphReg outperform baseline dilated 1D CNN models for gene expression prediction
- More importantly, can use feature attribution/in silico mutagenesis to predict functional enhancers for genes
- Epi-GraphReg and Seq-GraphReg outperforms ABC score for identifying enhancer elements, as validated by CRISPRi-FlowFISH
- Next step is to deploy in biologically meaningful contexts, move to single cell multiome data

Predicting the Hi-C contact map

Predicting the 3D contact map from 1D data

- Recent deep learning models like Akita (Fudenberg et al., 2020) and DeepC (Schwessinger et al., 2020) predict the Hi-C contact matrix from DNA sequence
  - Does not generalize to a new cell type
  - Can be expensive to train (~1Mb input sequences)

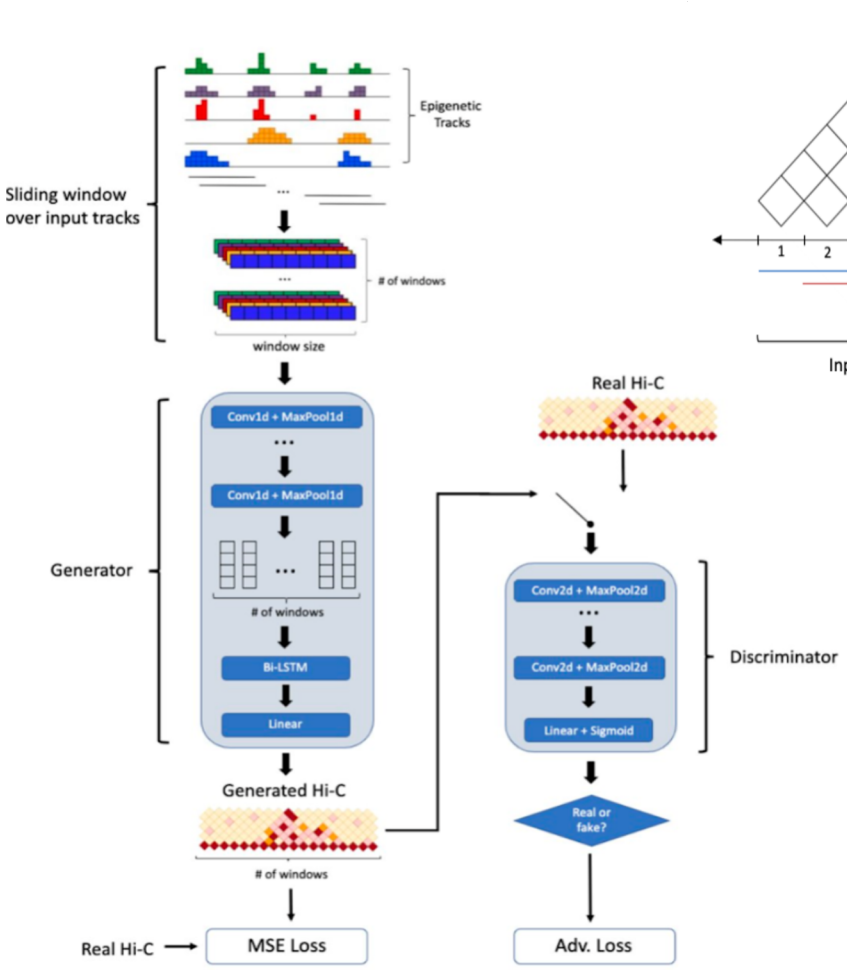
- Can we train a relatively lightweight model on 1D epigenomic data (histone marks, CTCF, DNase) instead?
  - Want to be able to use predicted Hi-C maps quantitatively (e.g. to call interactions or TADs)

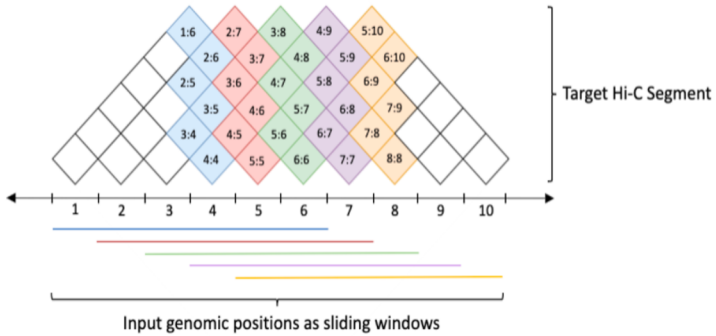


Akita model (2020)

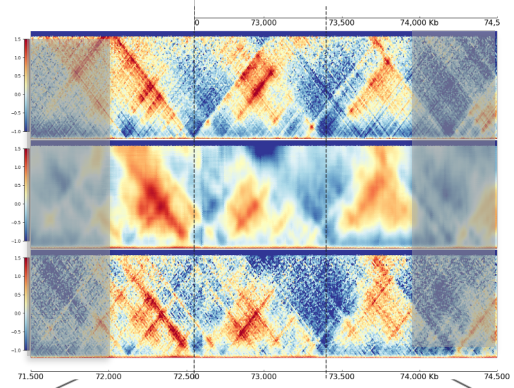
Epiphany predicts the Hi-C contact map from 1D epigenomic tracks

- HiC-DC+ or other preprocessing, MSE + adversarial loss





Yang, Das et al. bioRxiv 2021

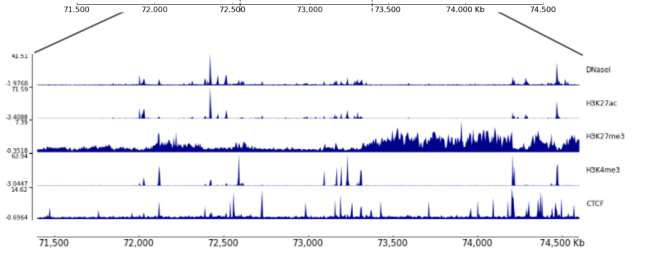


Ground truth

MSE only prediction

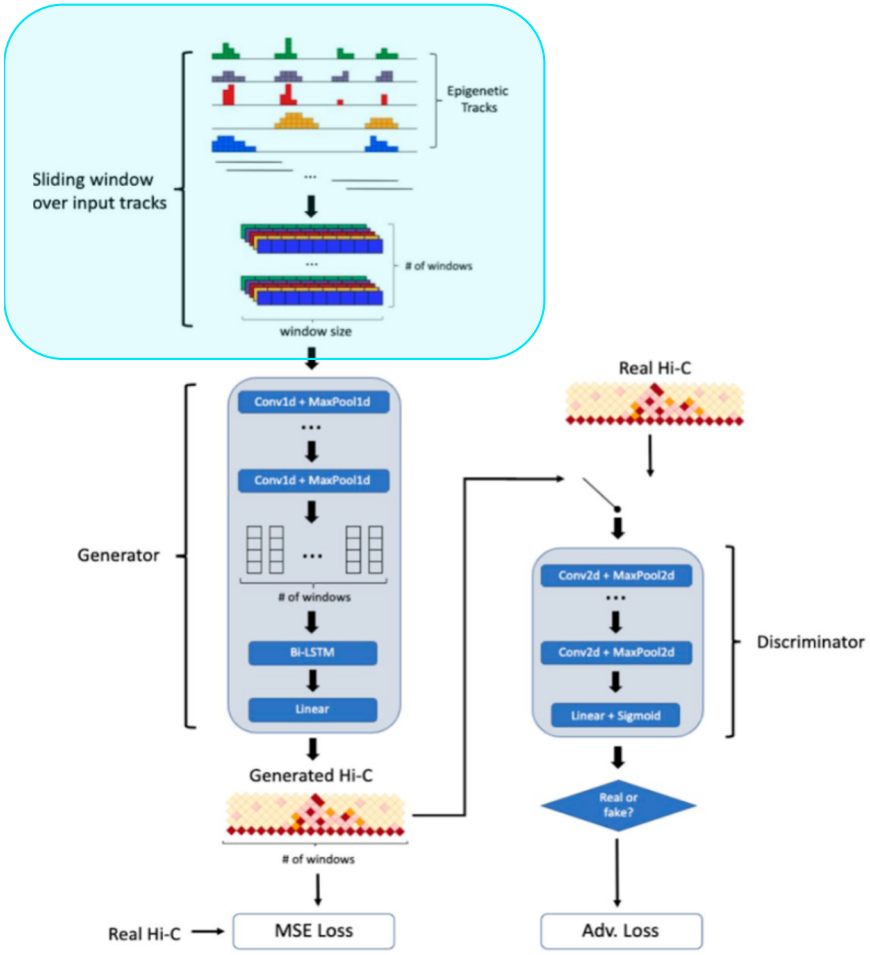
MSE + GAN prediction

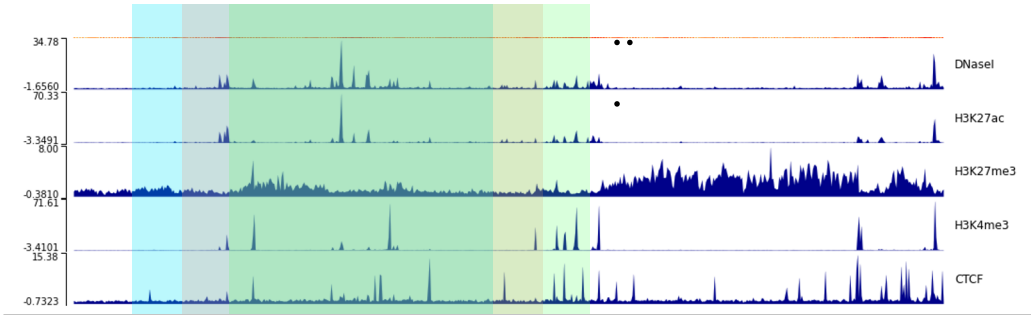
DNase, CTCF, H3K27ac, H3K4me3, H3K27me3

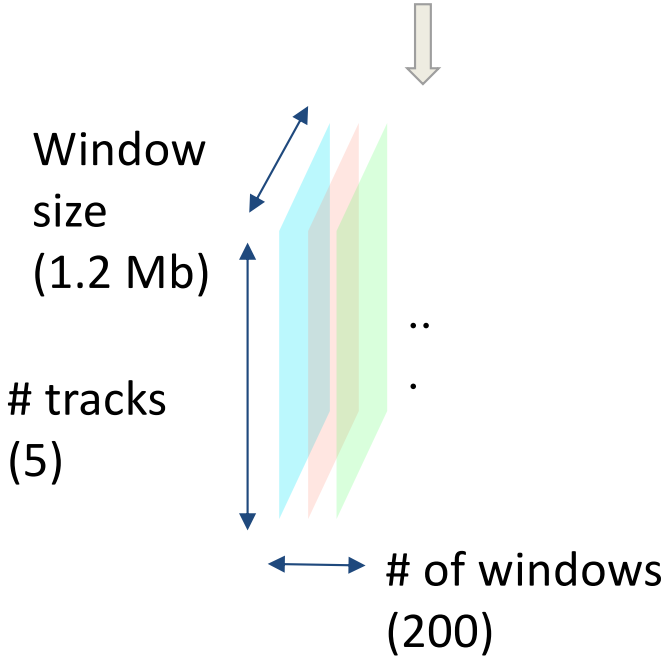


Epiphany predicts the Hi-C contact map from 1D epigenomic tracks

- Sliding window to extract epigenomic inputs

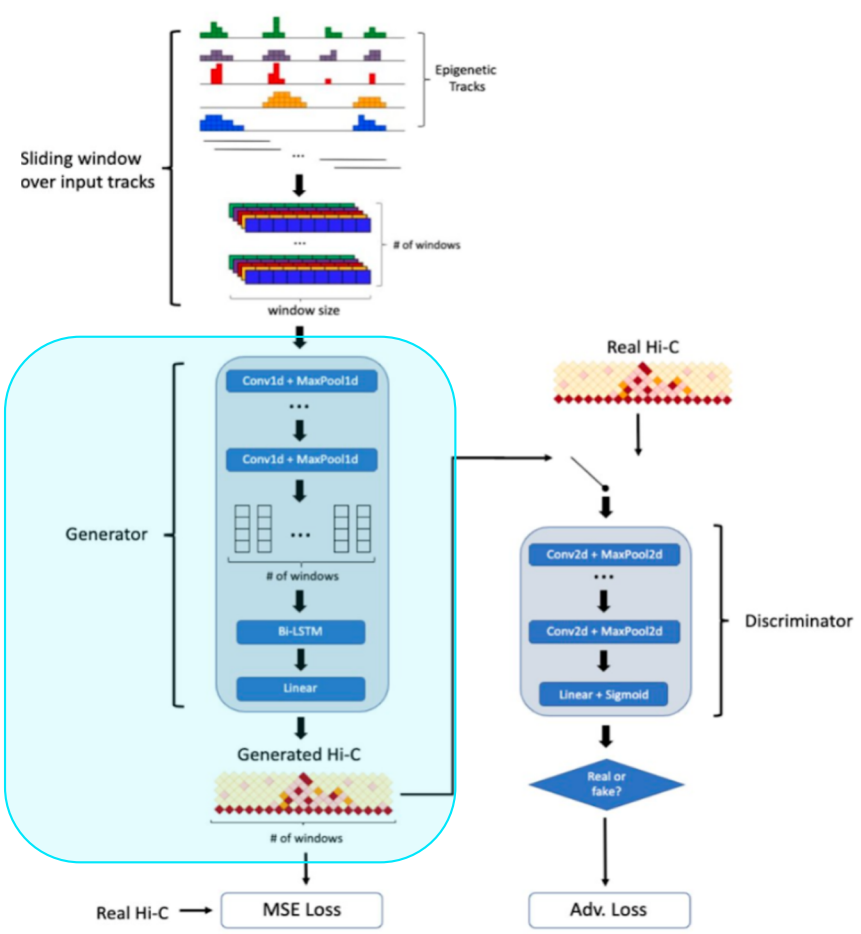


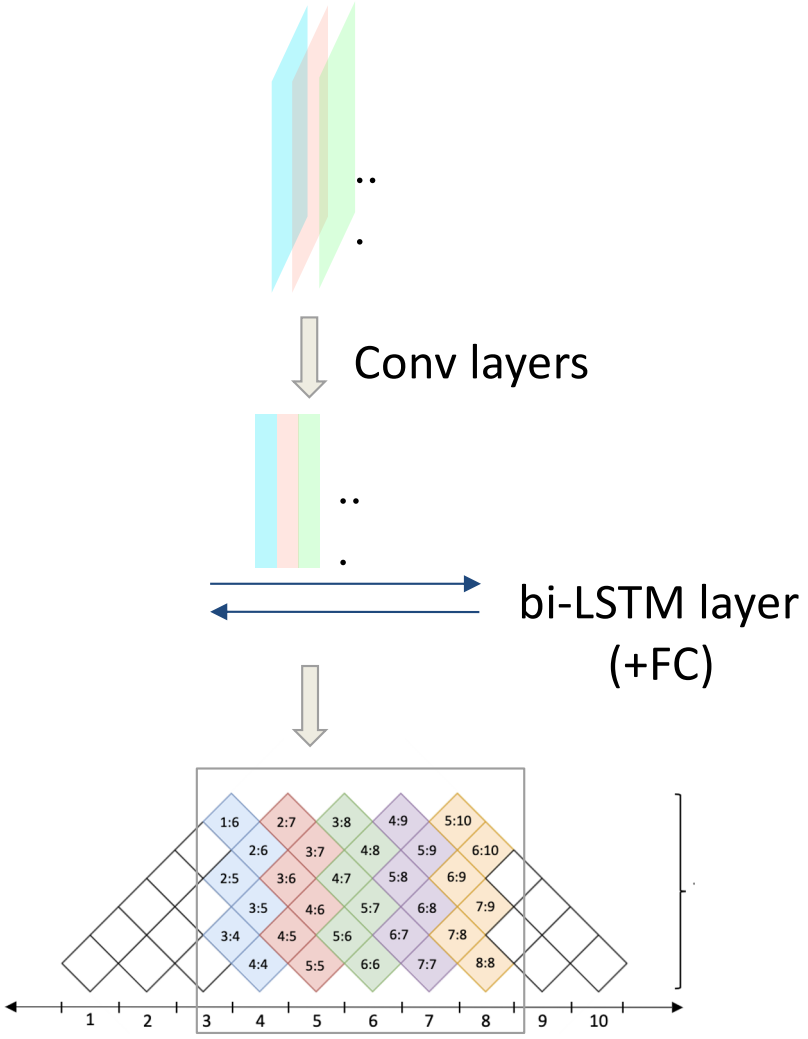




Epiphany predicts the Hi-C contact map from 1D epigenomic tracks

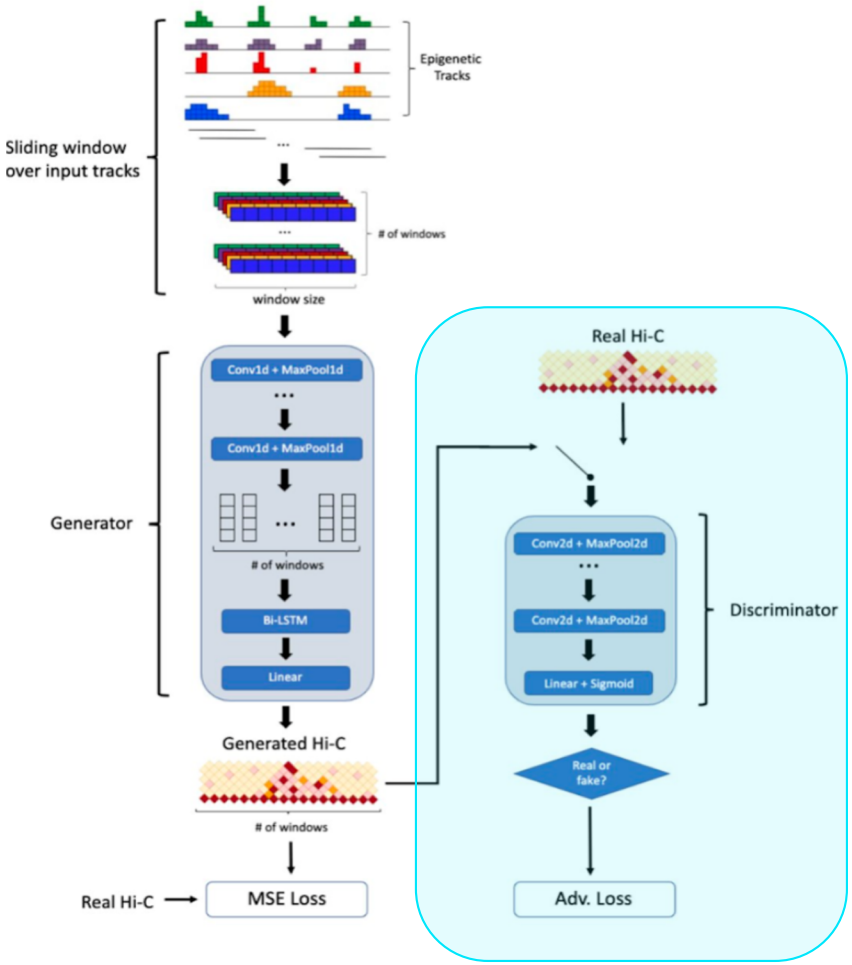
- Bi-LSTM layers to predict “stripes” of Hi-C contact map

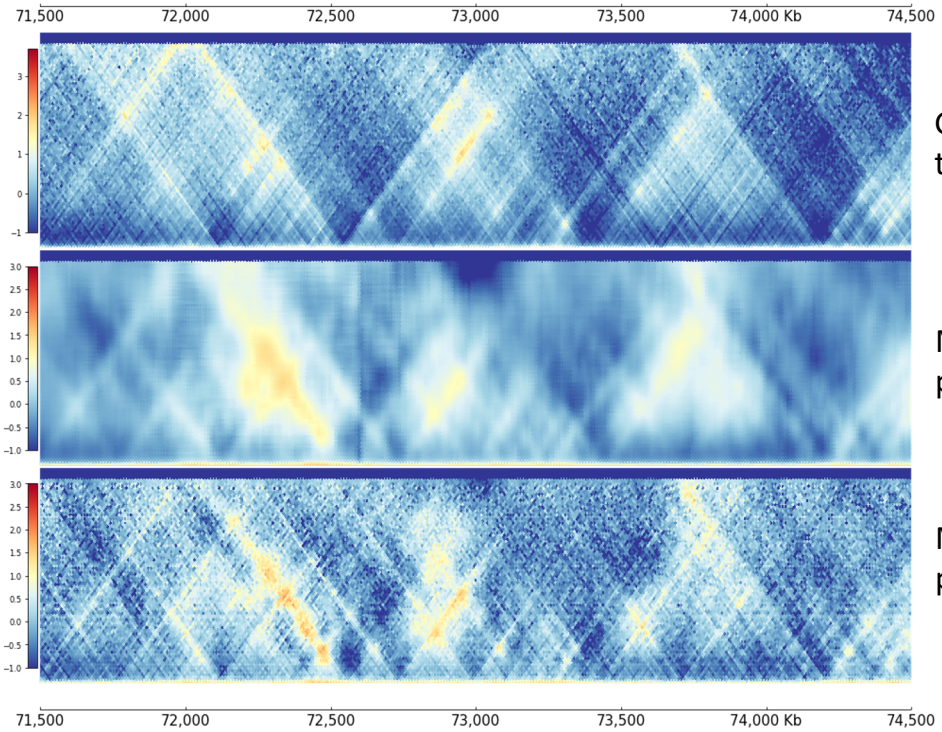




Epiphany predicts the Hi-C contact map from 1D epigenomic tracks

- Generative adversarial network to yield realistic maps





Ground truth

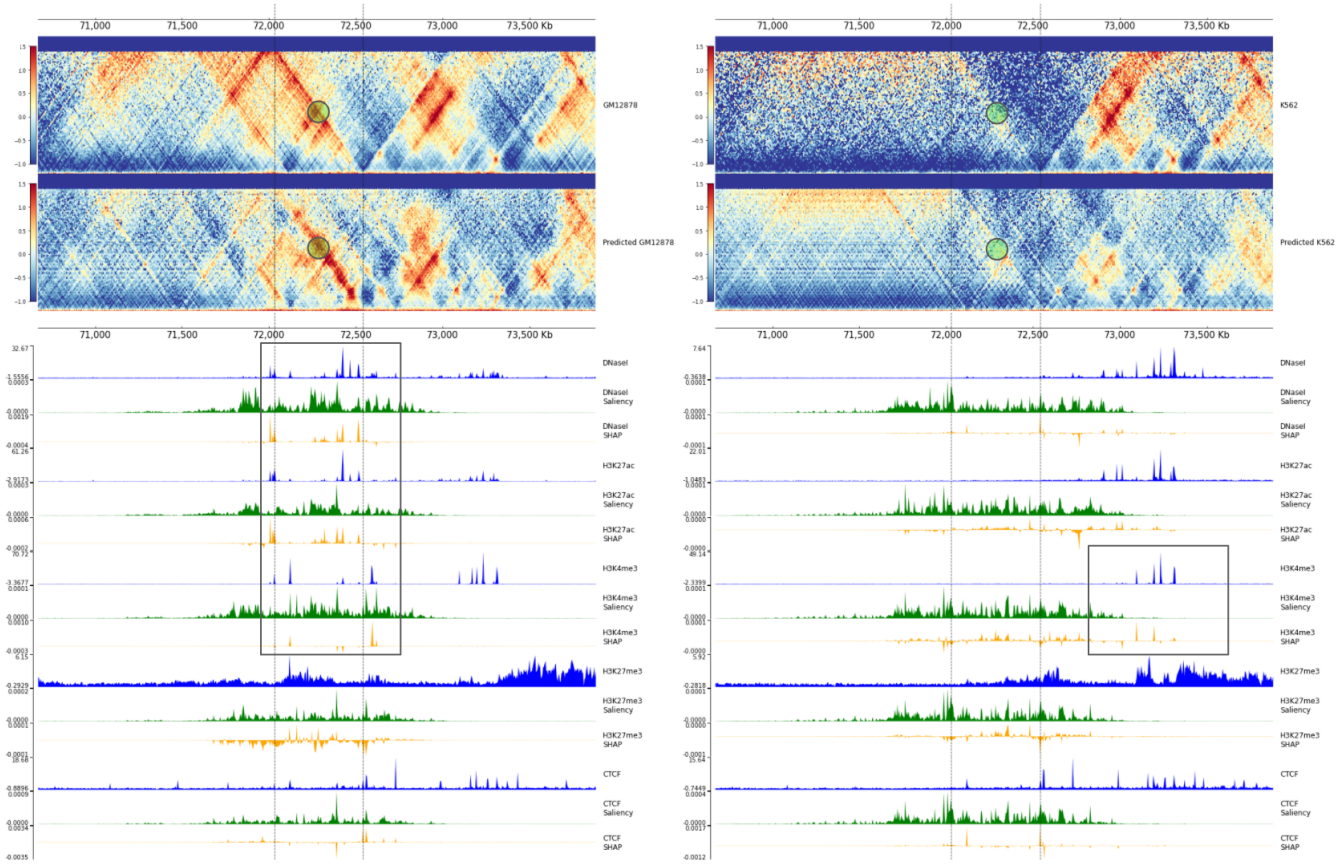
MSE only prediction

MSE + GAN prediction

Epiphany generalizes across cell types, learns cell-type specific contacts

- Example: same locus in GM12878 vs K562 (model trained on GM12878, tested on held-out chromosomes within and across cell types)

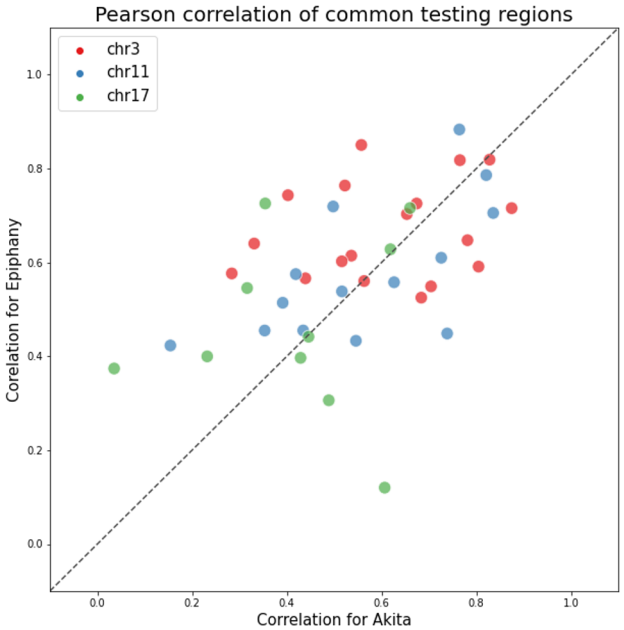
GM12878

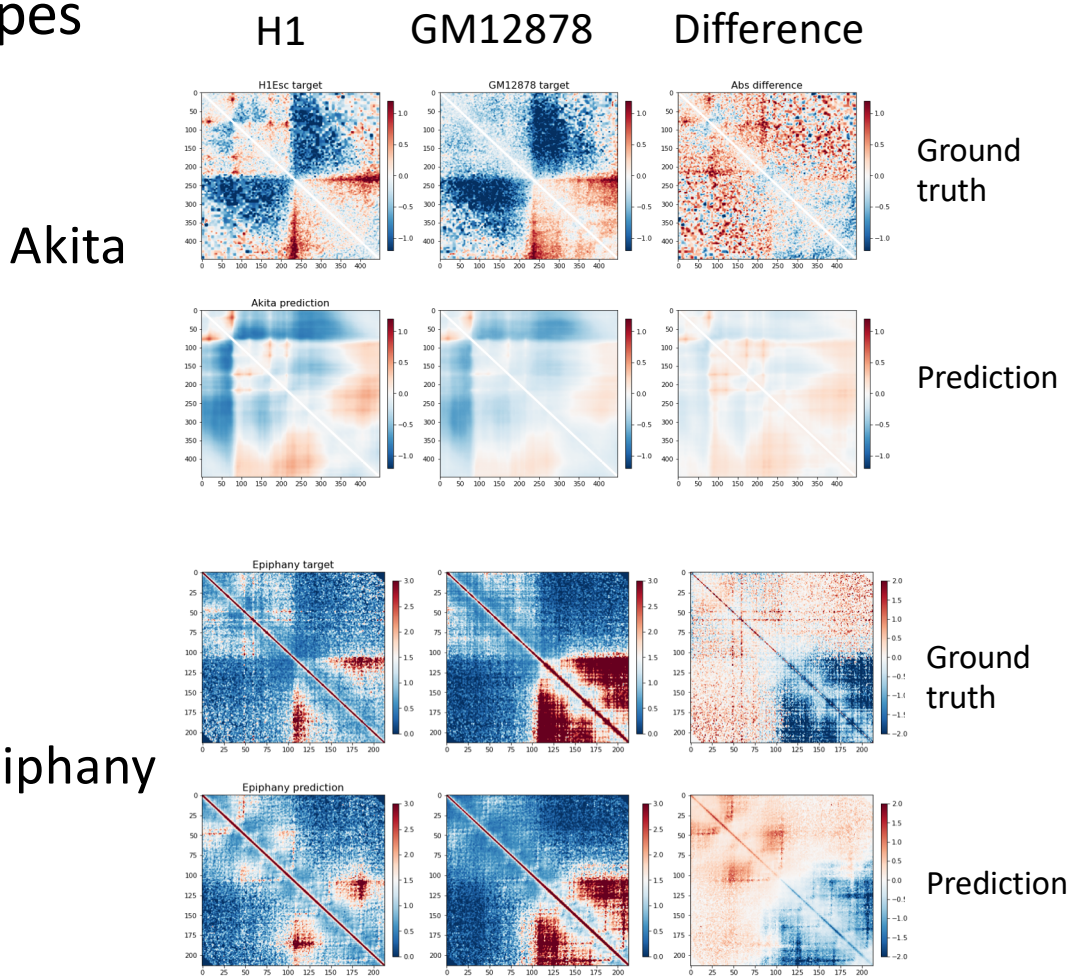


K562

Epiphany can better learn cell-type specific structures

- Similar performance to Akita on common held-out examples
- Epiphany improves cell-type specificify: Akita makes similar predictions across all cell types

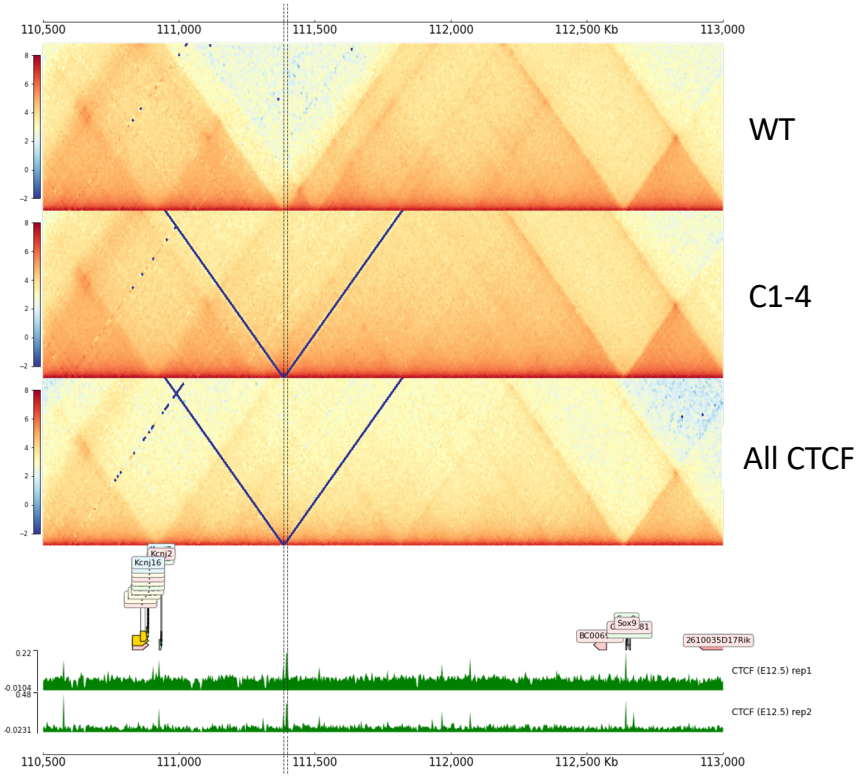




Predicting the impact of CTCF loss at TAD boundary

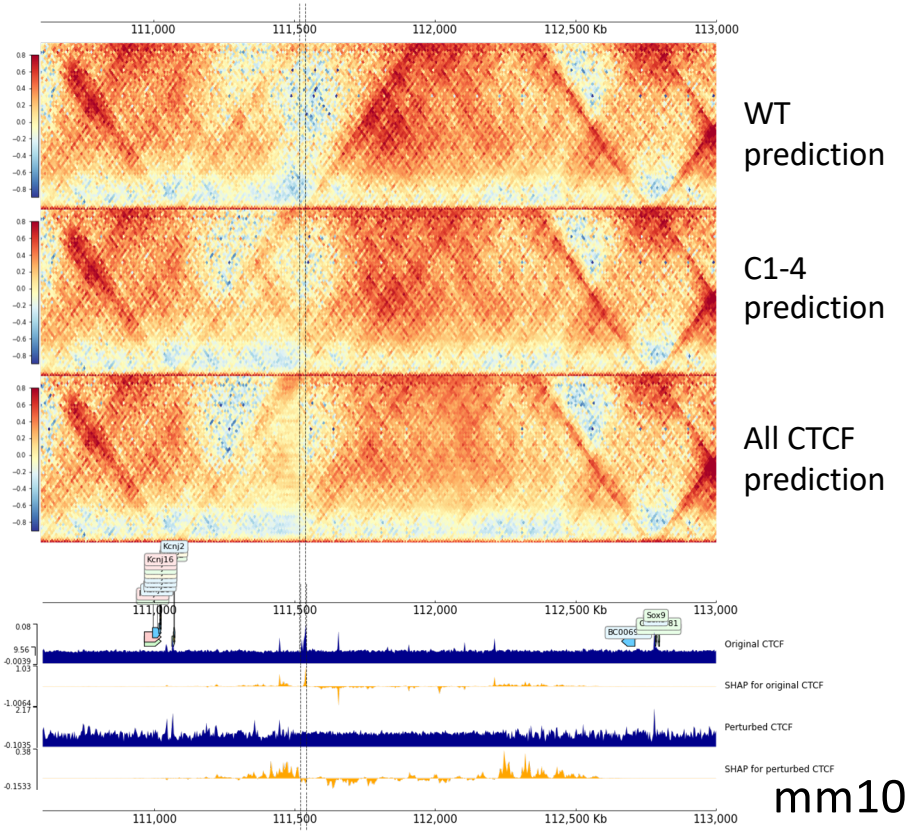
- TAD fusion event due to loss of CTCF binding sites between the *Kcnj2* and *Sox9* genes
- Use model trained in human GM12878 cells, test in mouse limb bud tissue (E11.5, E12.5 data)

Ground truth



mm9

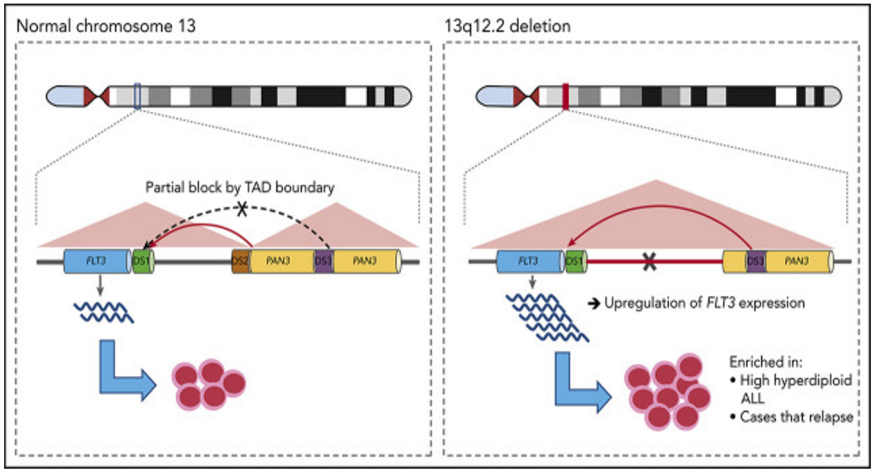
Epiphany prediction

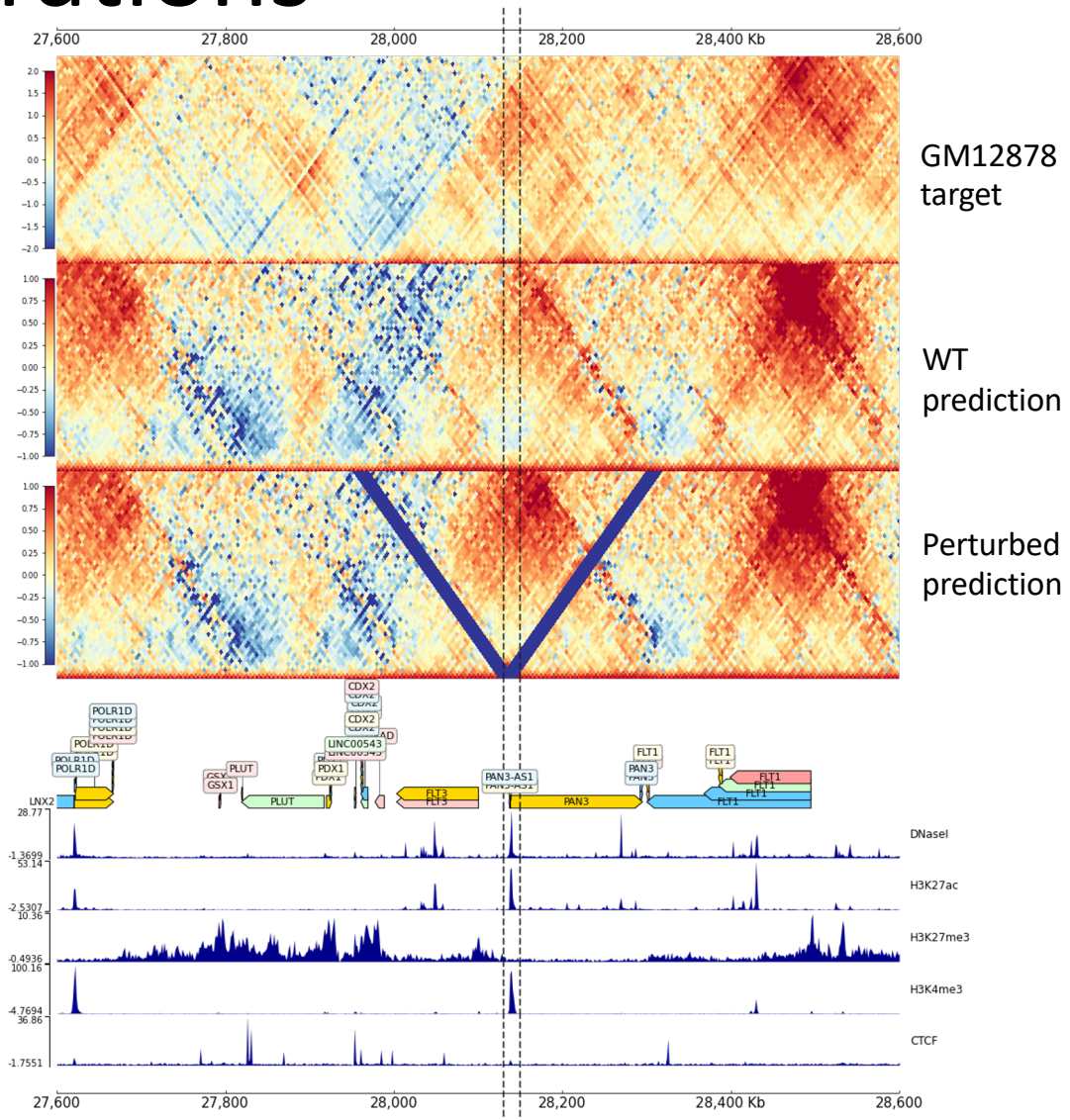


mm10

Predicting 3D changes due to somatic alterations

- Somatic 13q14 deletion results in loss of TAD boundary element in ALL, TAD fusion, and oncogene access to an enhancer
- Predict impact of deletion with Epiphany





Conclusions (Epiphany)

- Epiphany accurately predicts cell-type-specific Hi-C contact map from 1D epigenomic signals
- Bi-LSTM better captures long-range effects of epigenomic inputs on 3D interactions, while generative adversarial network produces more realistic contact maps
- Epiphany can generalize across cell types and species, outperforming sequence-based models
- Can use Epiphany to predict the 3D impact of epigenomic perturbations, like loss of boundary CTCF binding events
- Coming next: predicting of the Hi-C map from single-cell epigenomic data with scOrigami

Acknowledgements

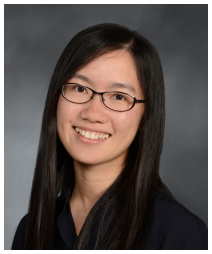
Leslie lab

Alireza Karbalayghareh
Merve Sahin
Erik Ladewig
Rose DiLoreto
Alli Pine
Zakieh Tayyebi
Vianne Gao
Rui Yang
Wilfred Wong
Viraj Rapolu
Aditya Sinha
Vijay Yarlaggada
Preethi Periyakoil
Changlin Wan, Sneha Mitra (PhD interns)

Collaborators

Effie Apostolou
Danwei Huangfu
Jeff Bilmes, Arnav Das
Bill Noble
Charles Sawyers, Abbas Nazir
Ari Melnick
Steve Josefowitz
Alexander Rudensky






We are looking for postdocs!