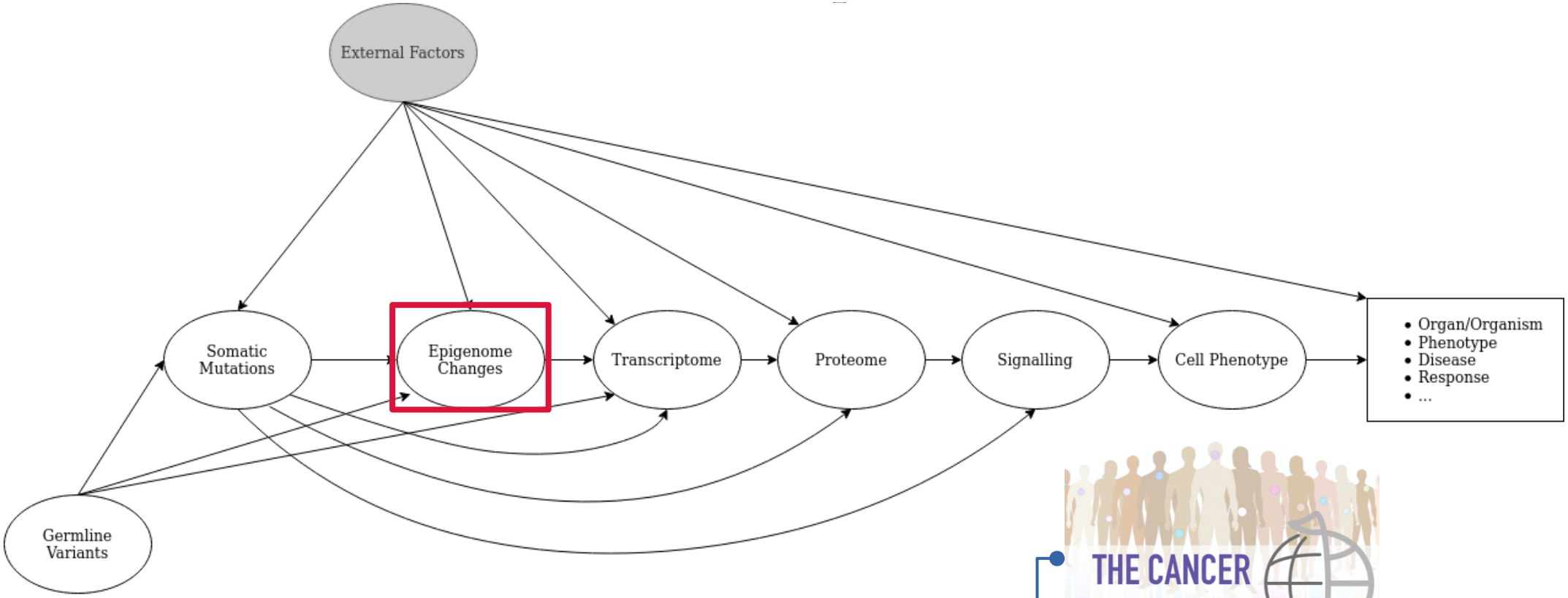
EPI-LOGOS

Getting personal with epigenetics: Towards machine-learning-assisted precision epigenomics

Giovanni Visonà

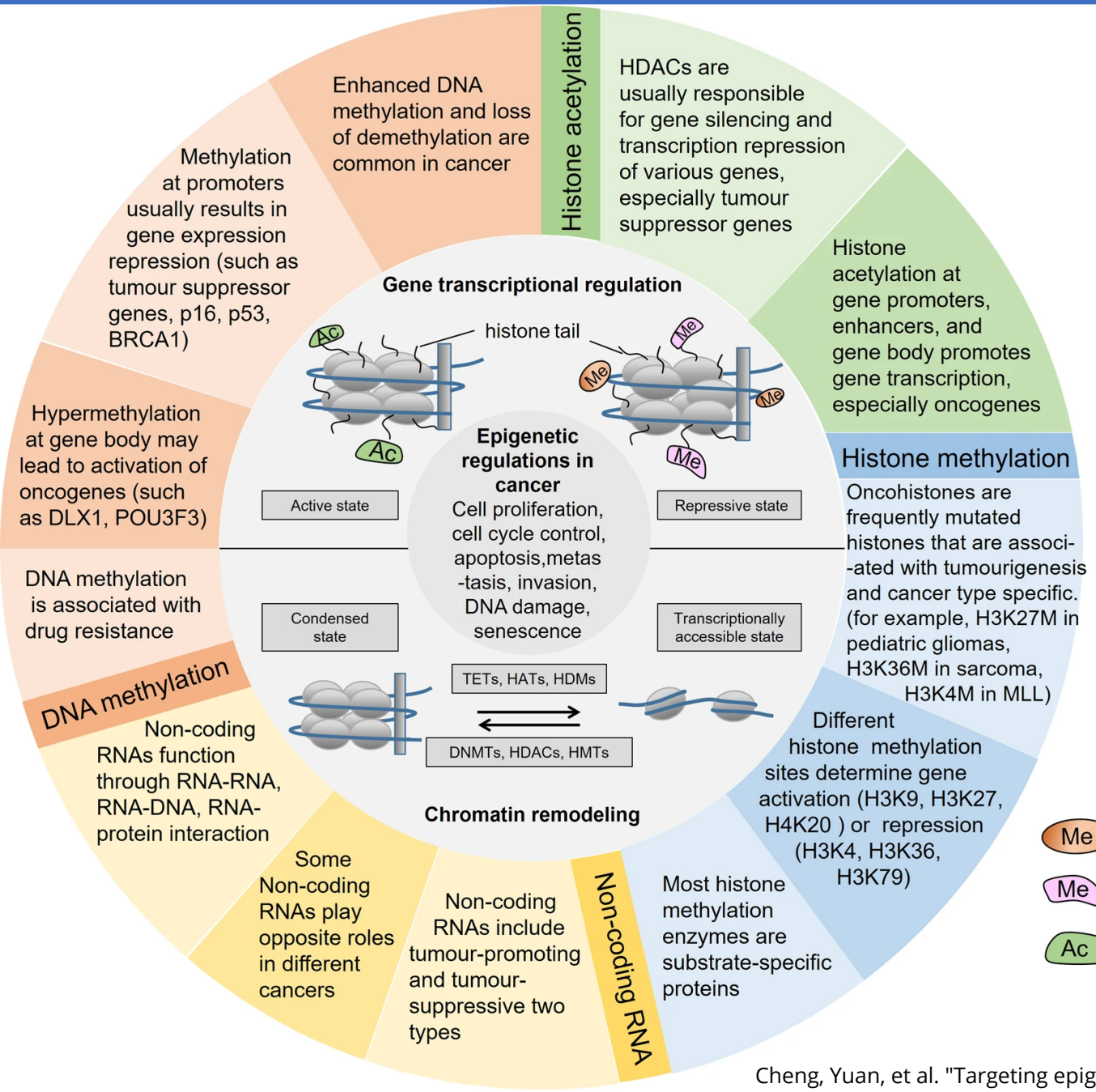
Max Planck Institute for Intelligent Systems
Empirical Inference
Tübingen Campus

MLFPM
Machine Learning Frontiers in Precision Medicine



Large scale efforts conducted by several consortia have been key to the study of cancer and cell biology in general

Epigenetic modifications are **involved in many aspects of carcinogenesis and cancer progression**, which offer compelling **therapeutic targets**



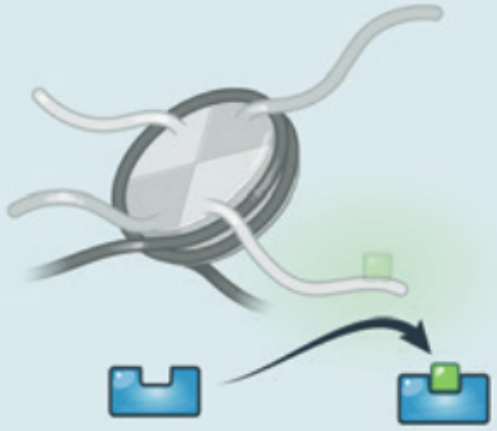
Me DNA methylation

Me Histone methylation

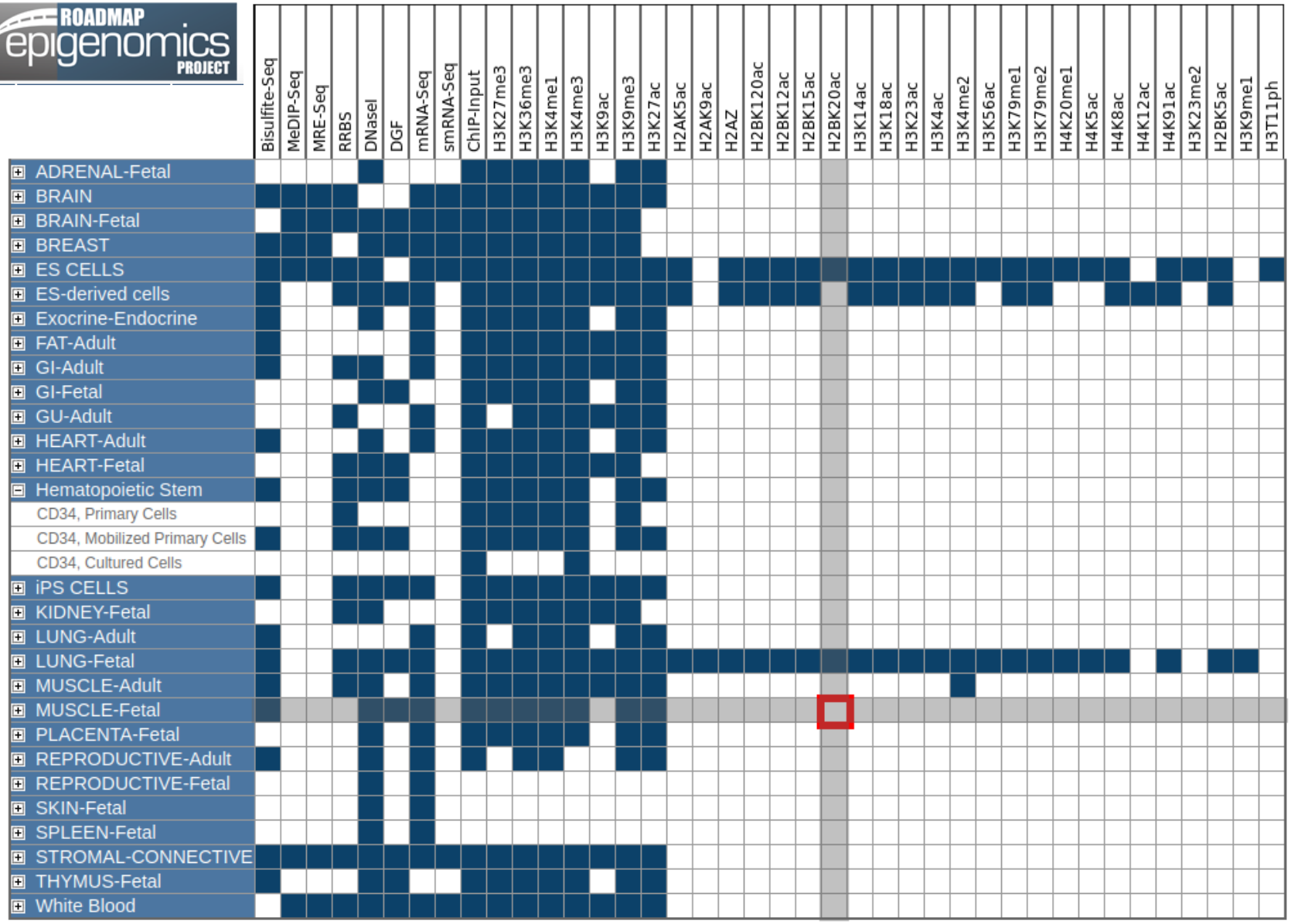
Ac Histone acetylation

Cheng, Yuan, et al. "Targeting epigenetic regulators for cancer therapy: Mechanisms and advances in clinical trials." Signal transduction and targeted therapy 4.1 (2019): 1-39.

EPI-LOGOS

Category	Epigenetic Regulators	Function	FDA-Approved Drug
Writers 	**DNMT1, 3A, and 3B**	Methylates cytosines on DNA, and mutation can lead to aberrant methylation	Azacitidine, decitabine
	EZH2	Methylates histone H3K27	Tazemetostat
	DOT1L	Methylates histone H3K79	
	KMT2A–D, SETD2, NSD1	Methylate histone lysines	
	EP300, CREBBP	Acetylate histone lysines	
Erasers 	**TET2**	Is the first step in cytosine demethylation; is inhibited by 2-hydroxyglutarate (2-HG)	Azacitidine, decitabine
	IDH1, IDH2	Mutated protein produces 2-HG from isocitrate that inhibits TET2 and lysine demethylases	Ivosidenib, enasidenib
	HDAC1–3, 8 HDAC6	Deacetylase removes acetyl groups from histone lysines	Vorinostat, belinostat, panobinostat, romidepsin
	KDM1A, **KDM6A (UTX)**	Demethylates histone lysines	

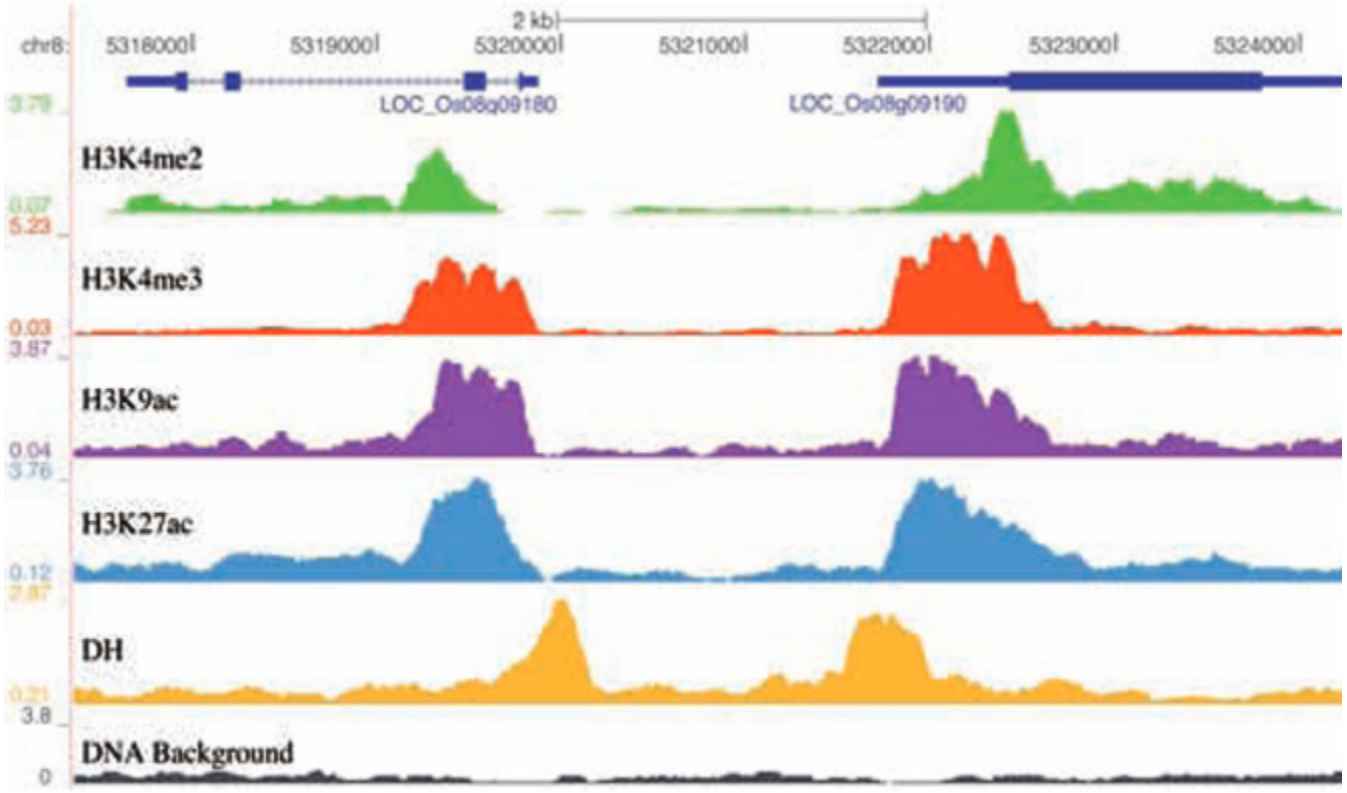
Bates, Susan E. "Epigenetic therapies for cancer." New England Journal of Medicine 383.7 (2020): 650-663.

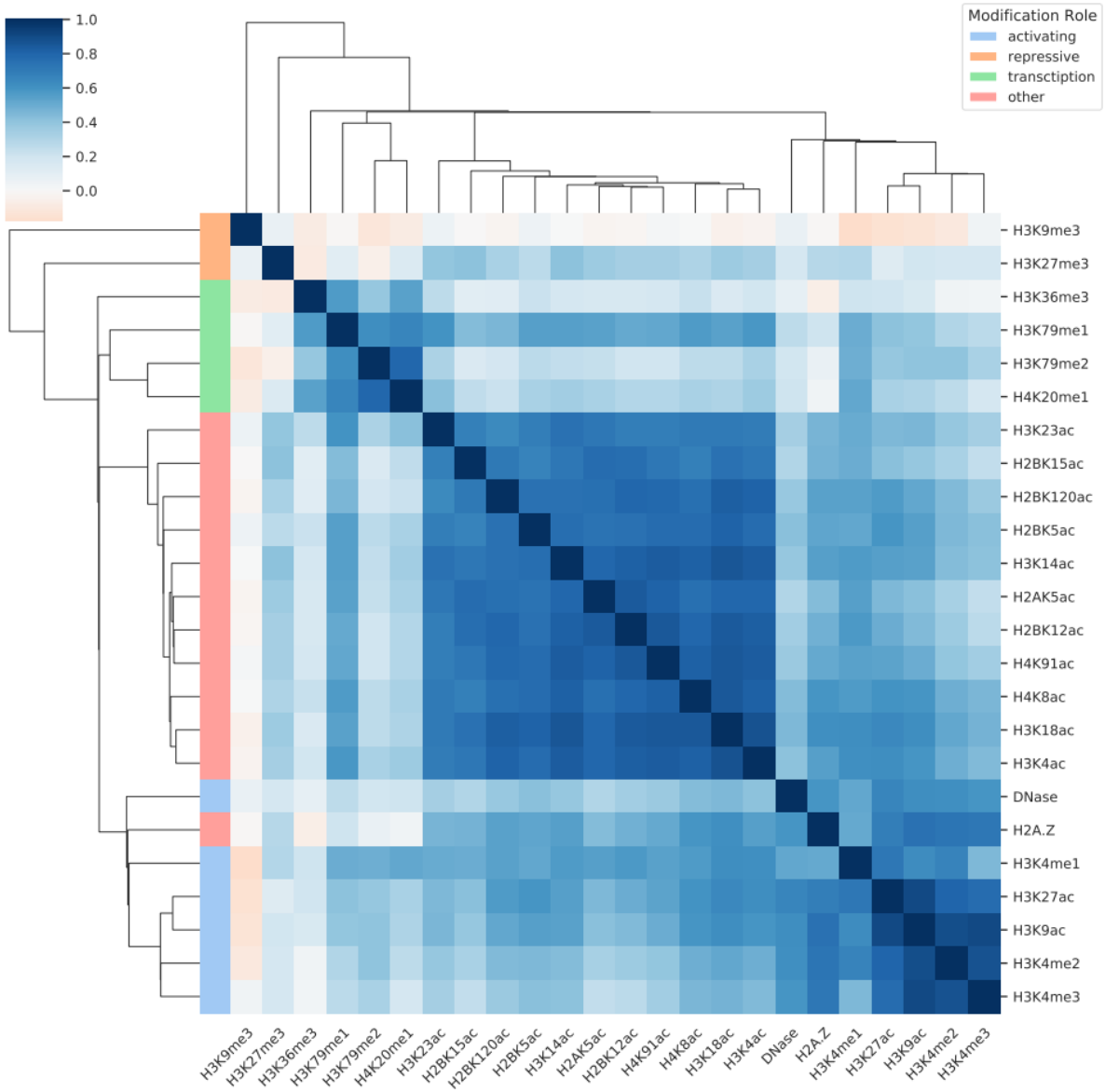


Can we leverage machine learning to **fill the gaps**?

Epigenetic modifications are **cell-type specific**

Epigenetic marks are highly **correlated**

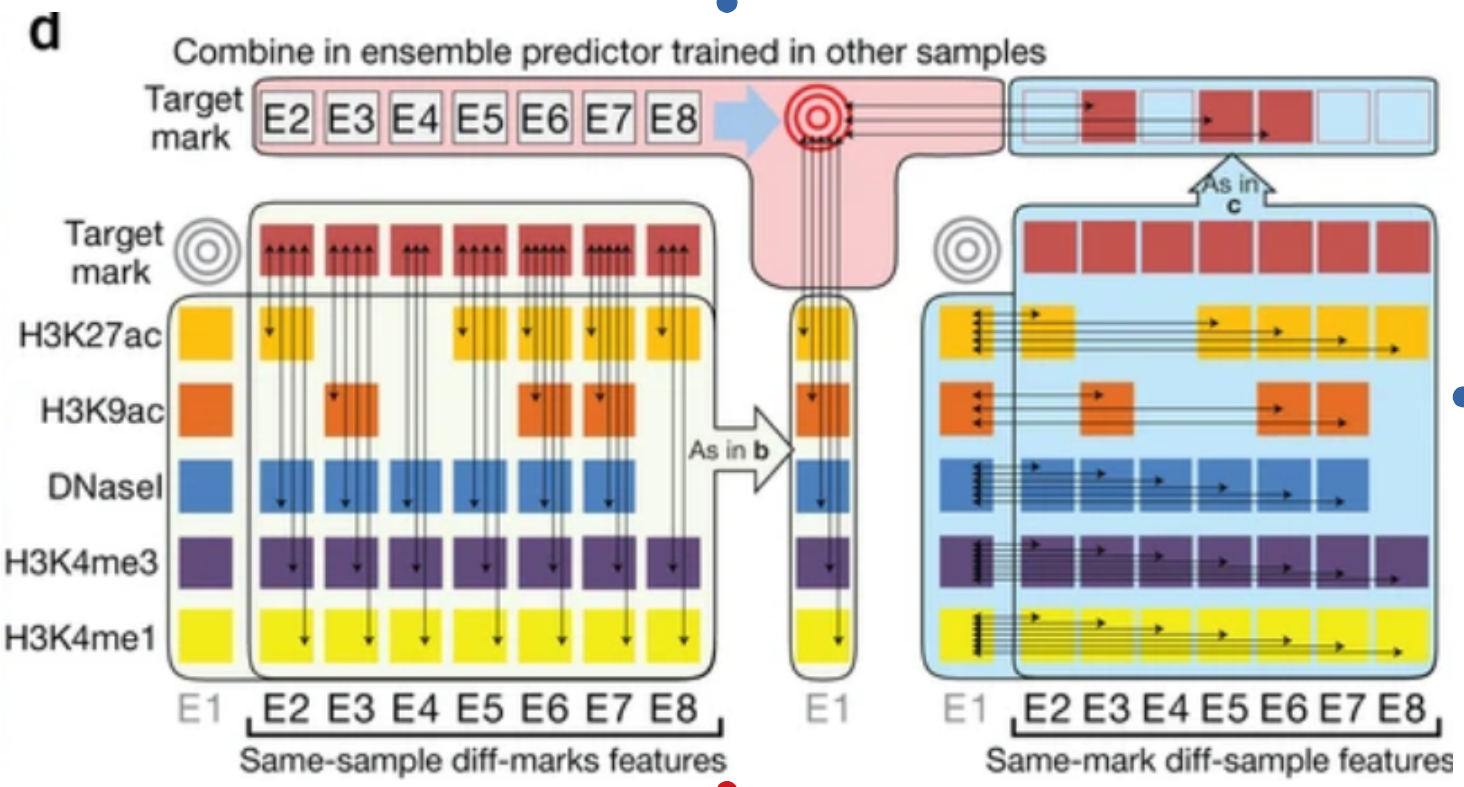




Previous work: ChromImpute

Trains ensembles of **regression trees**

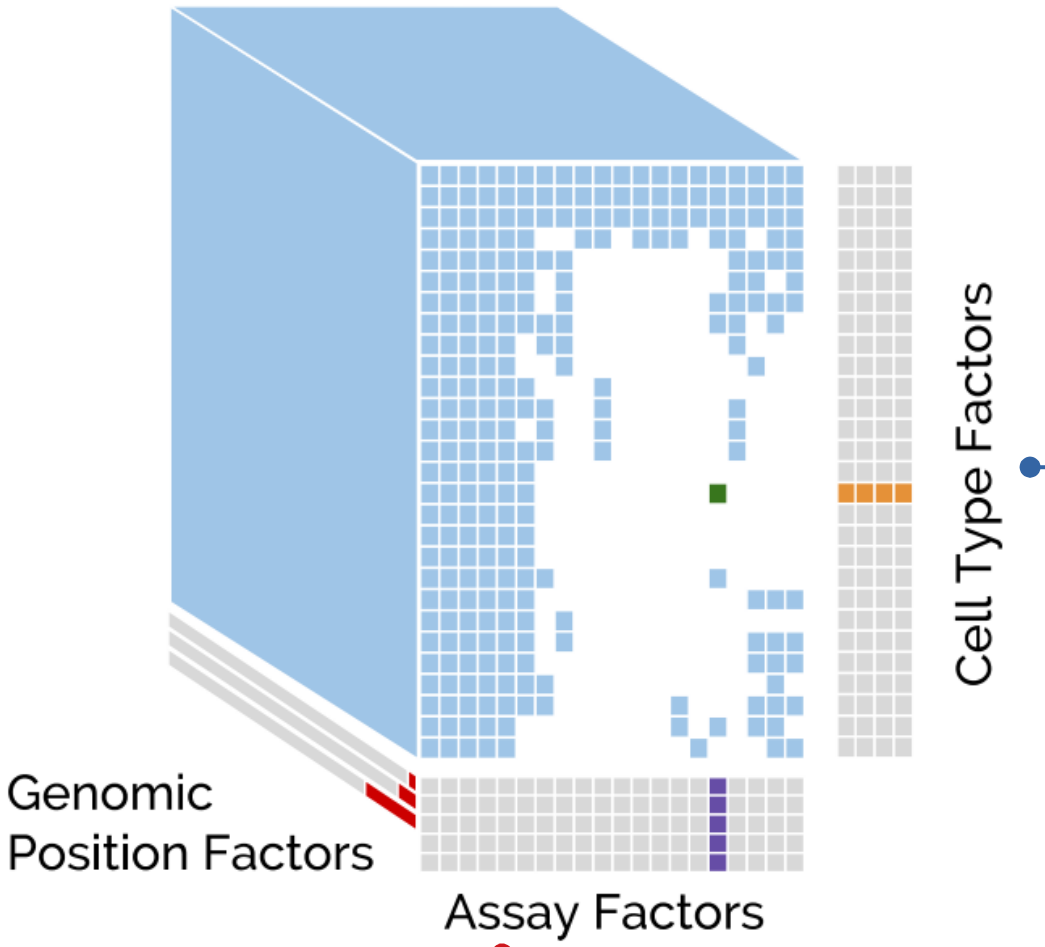
Significant performance improvement compared to all previous methods



For each cell type-assay query CI requires the training of a **new ensemble** of regression trees

Ernst, Jason, and Manolis Kellis. "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues." Nature biotechnology 33.4 (2015)

Previous work: PREDICTD and Avocado



Both models rely on **generalized tensor factorization** methods

Learn embeddings for each assay, cell type, and genomic bin

Extremely **memory expensive**. Using 25bp bins, the model needs to learn ~120.000.000 embeddings for the genome

Durham, Timothy J., et al. "PREDICTD parallel epigenomics data imputation with cloud-based tensor decomposition." Nature communications 9.1 (2018)

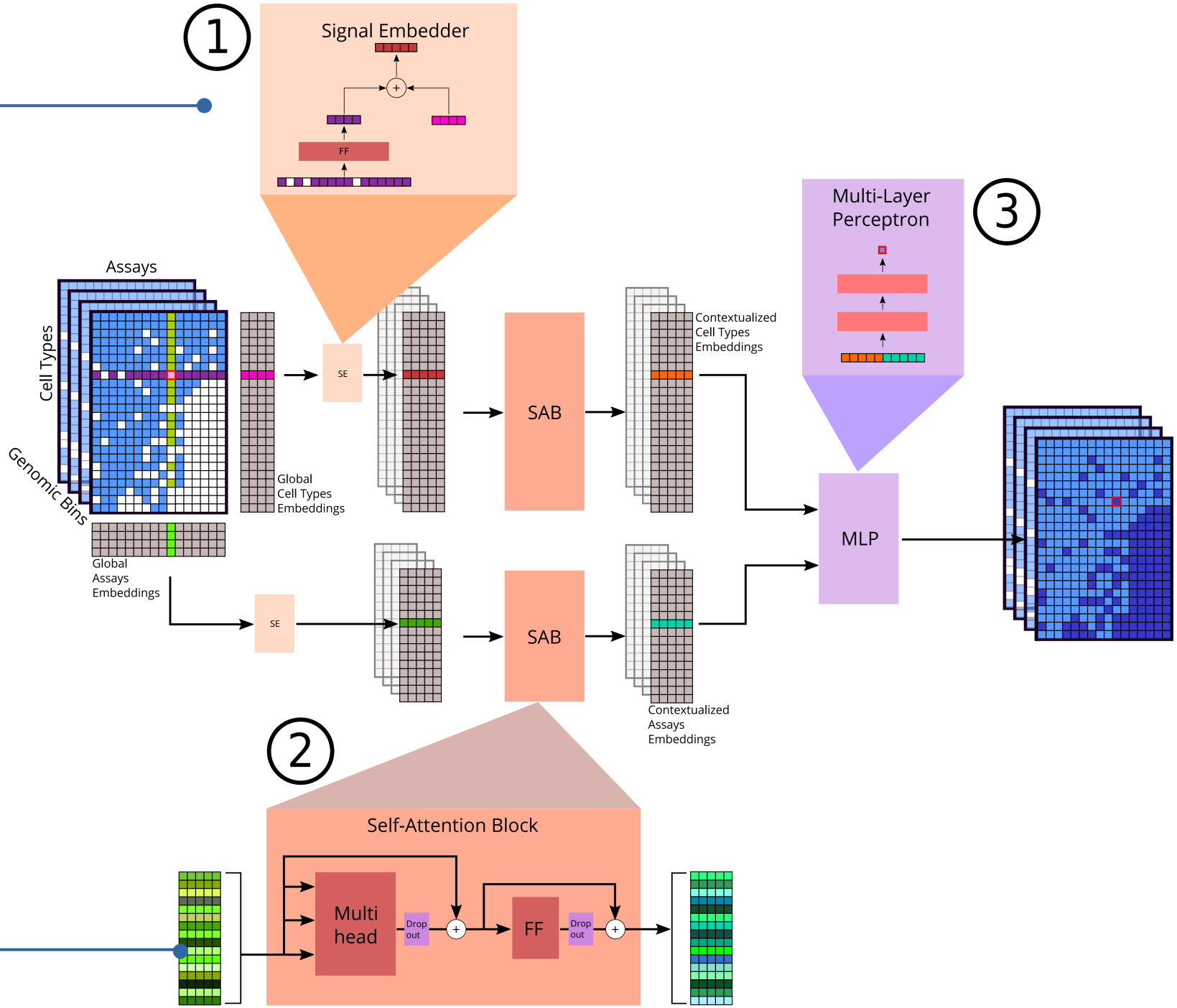
Schreiber, Jacob, et al. "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome." Genome biology 21 (2020)

eDICE

epigenomic **D**ata **I**mputation through **C**ontextualized **E**mbeddings

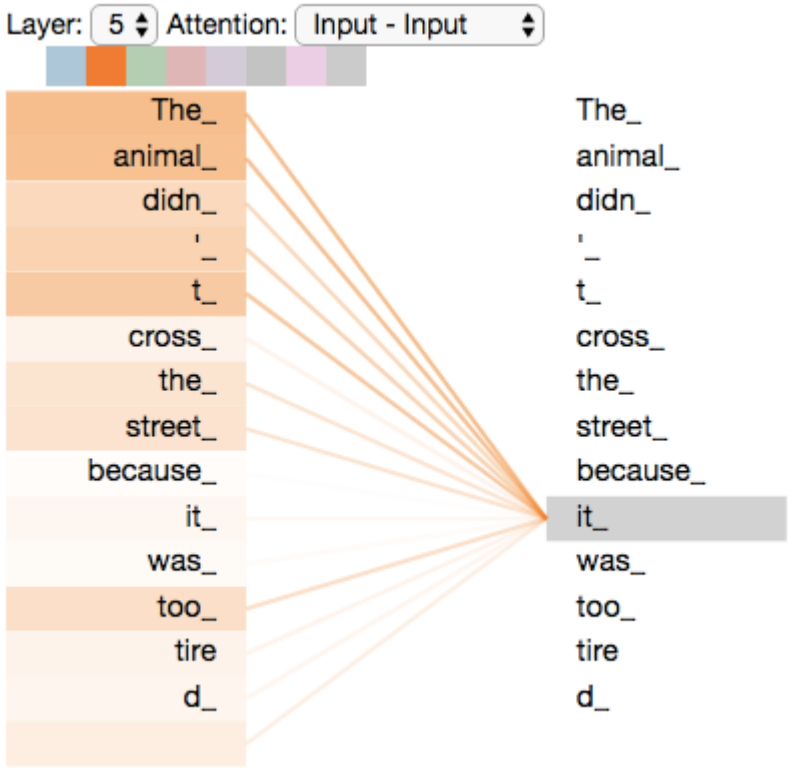
Local signal captures the information of each genomic bin

Self-Attention includes the interaction between assays and between cell types in the model



Hawkins-Hooker, Alex, et al. "Getting Personal with Epigenetics: Towards Machine-Learning-Assisted Precision Epigenomics." bioRxiv (2022).

What is **attention**?



https://jalammar.github.io/illustrated-transformer/

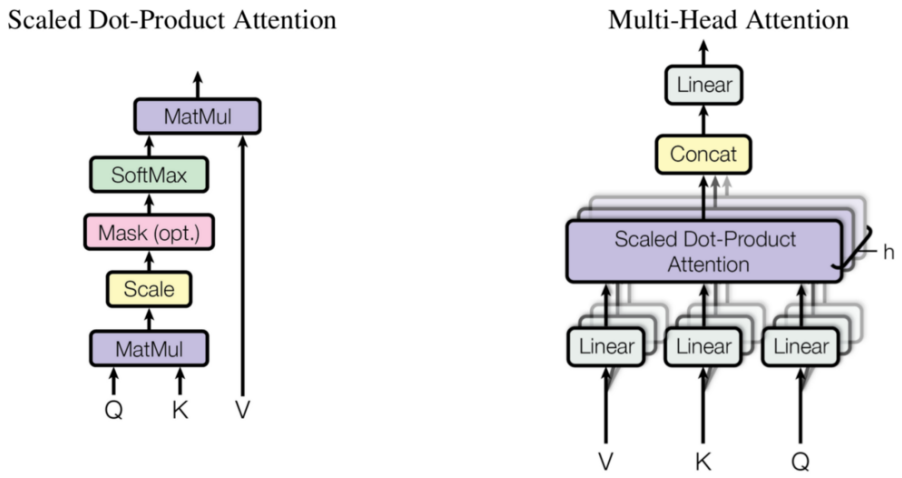
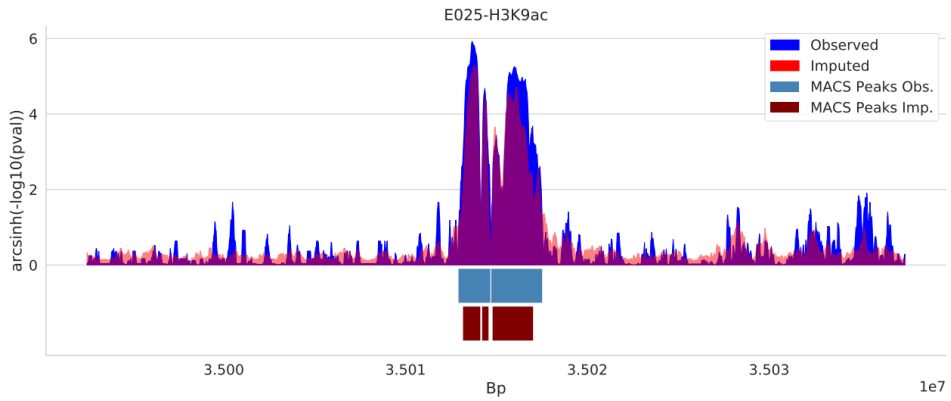


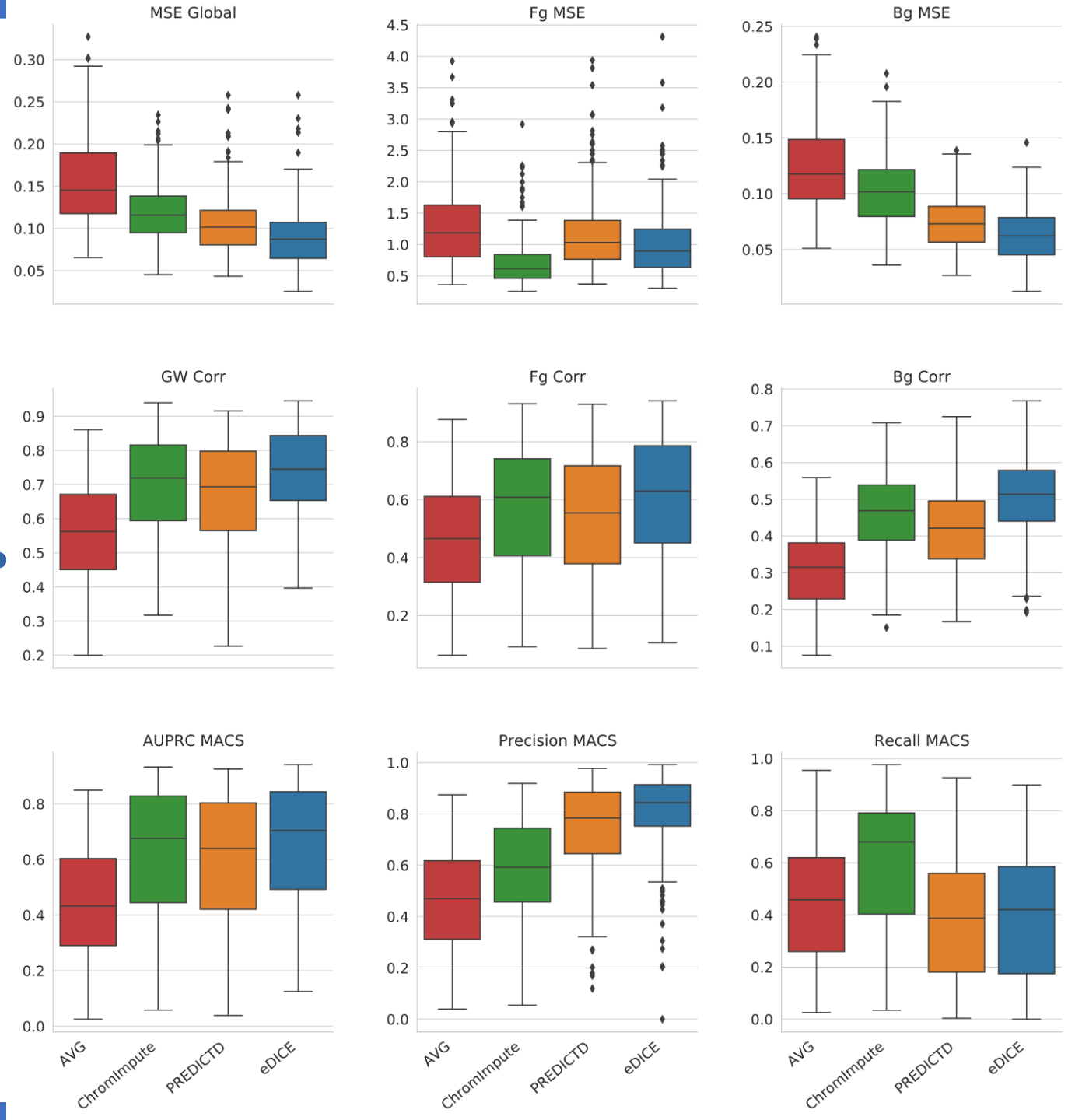
Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

The **Transformer** is one of the most successful attention models

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

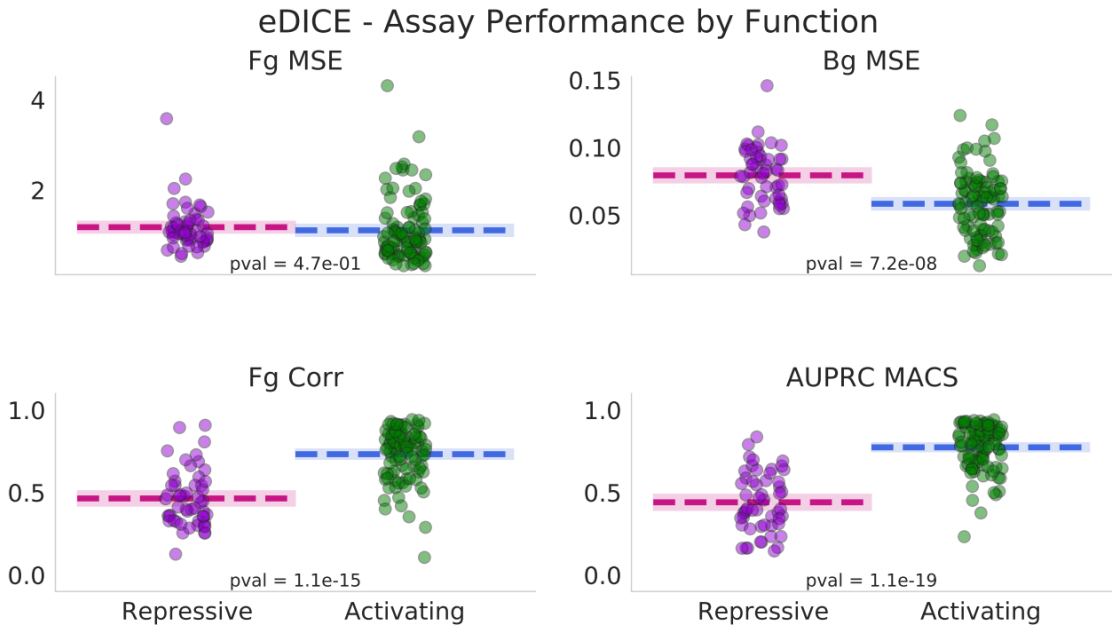
Tested on the Roadmap dataset (chromosome 21),
eDICE outperforms ChromImpute and PREDICTD on almost all metrics





EPI-LOGOS

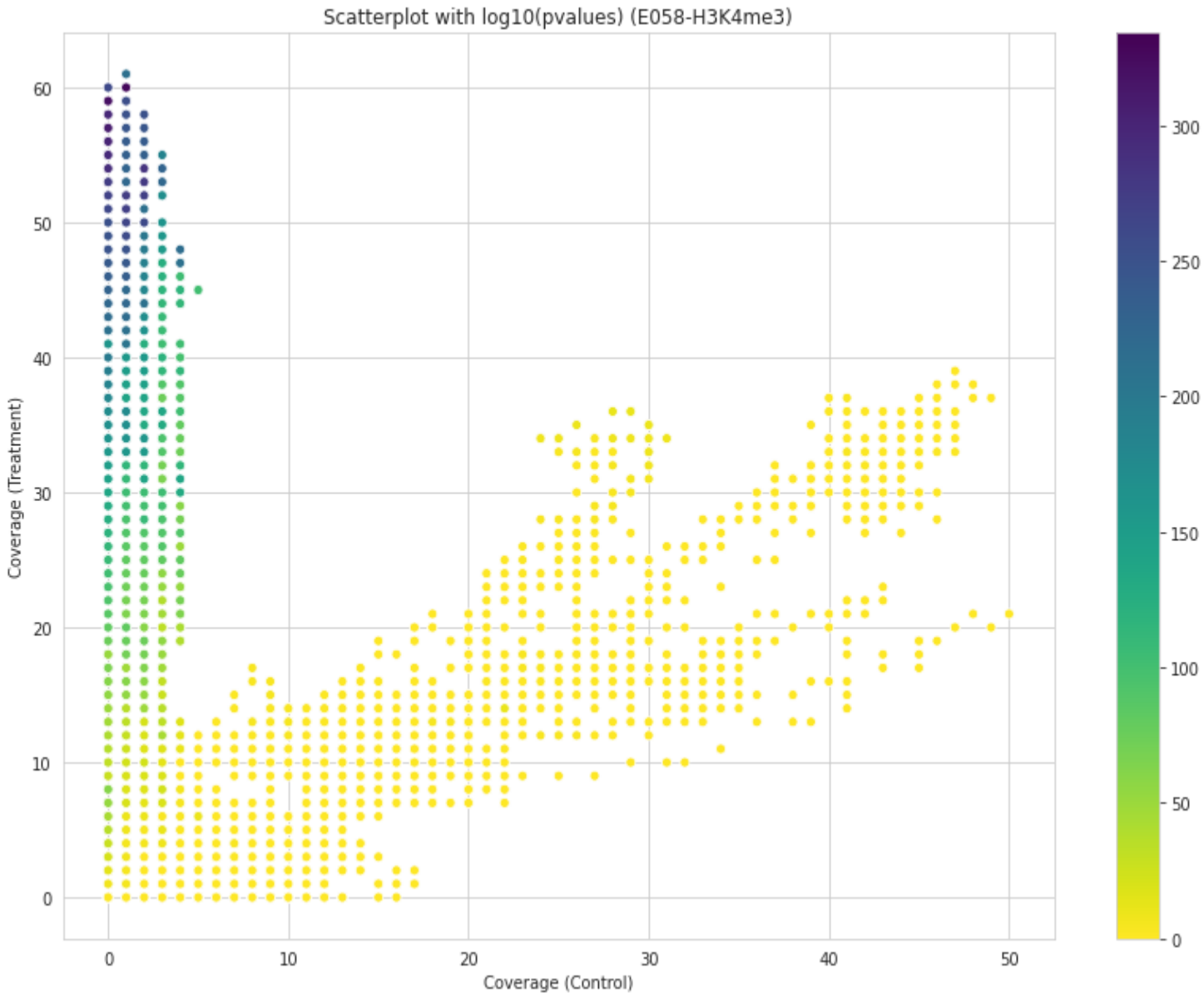
Imputations work better on some assays.
This problem is not unique to eDICE.



The discrepancies are possibly due to biases in the sequencing of heterochromatin-associated marks.

The processing pipeline is affected by **multiple sources of bias**: small number of replicates, low quality control samples, different sequencing platforms.

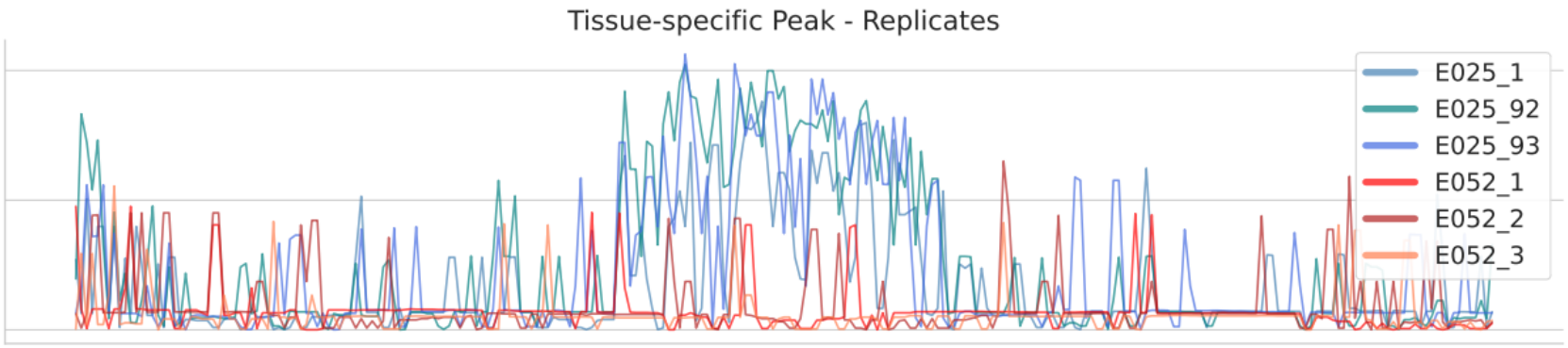
Control Counts	Treatment Counts	-log10(pval)
2	10	5
3	10	3
4	10	2
8	100	72
9	100	67
10	100	63



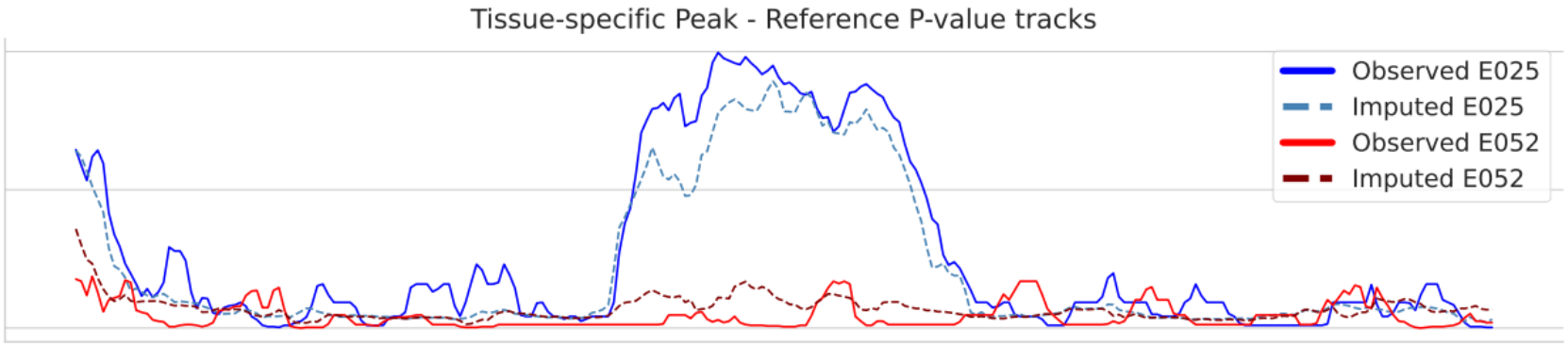
EPI-LOGOS

Identifying meaningful differences between biological samples is crucial to progress our understanding of the regulatory mechanisms of the genome

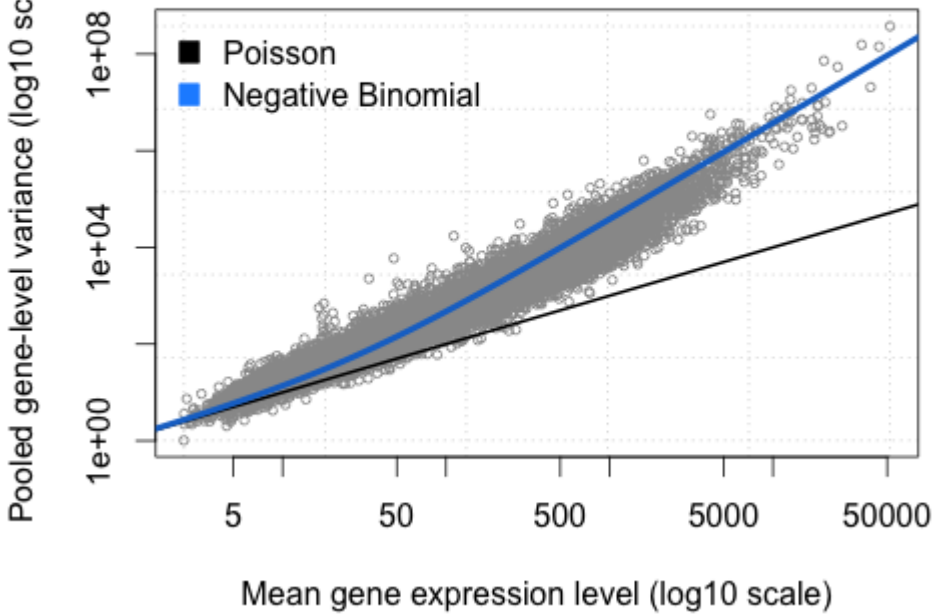
H3K9ac



The use of multiple replicates is fundamental for robust analysis



We simulate pseudo-replicates for two tissues using parameters estimated from the imputations. (Next iteration will explicitly predict the variance of the signal)



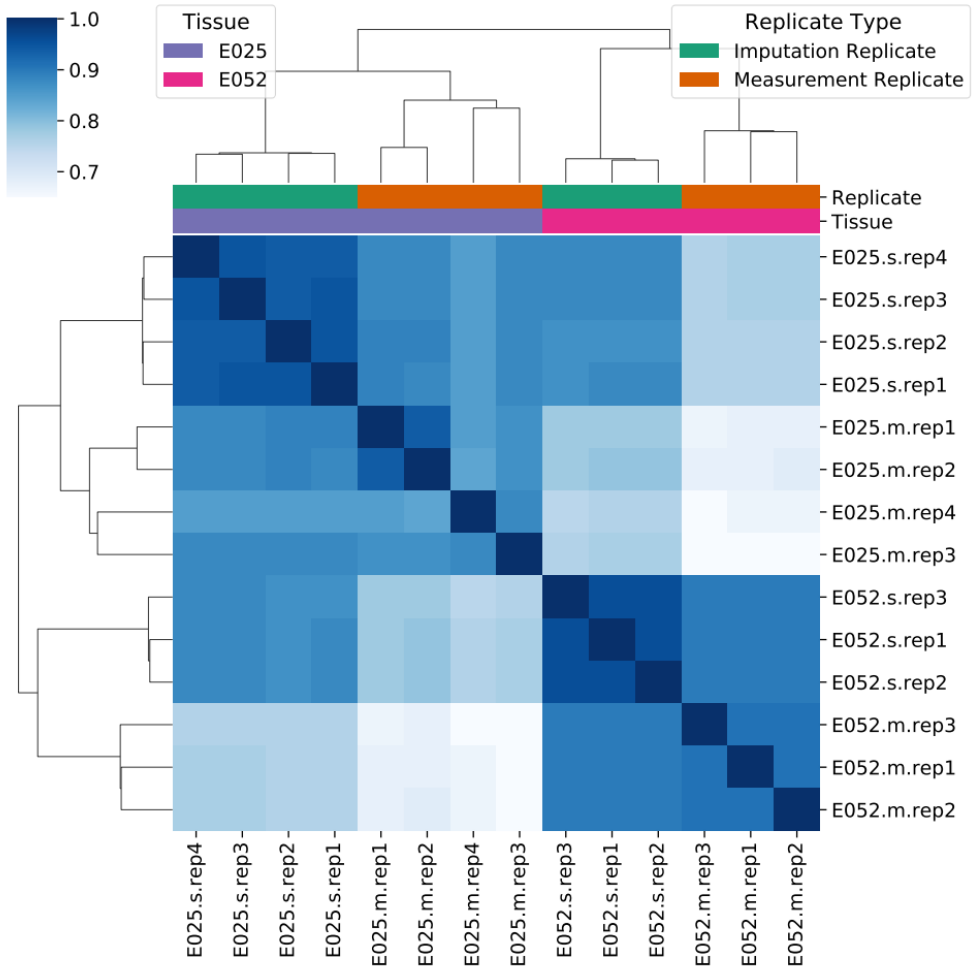
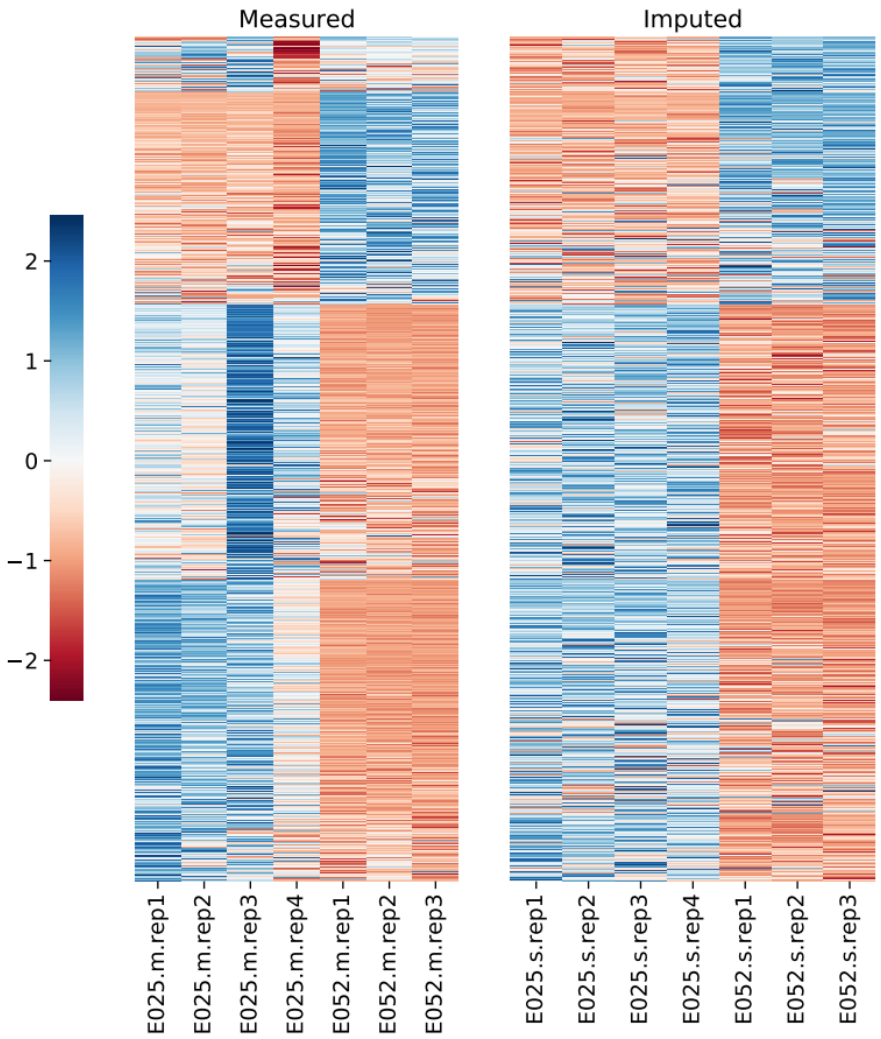


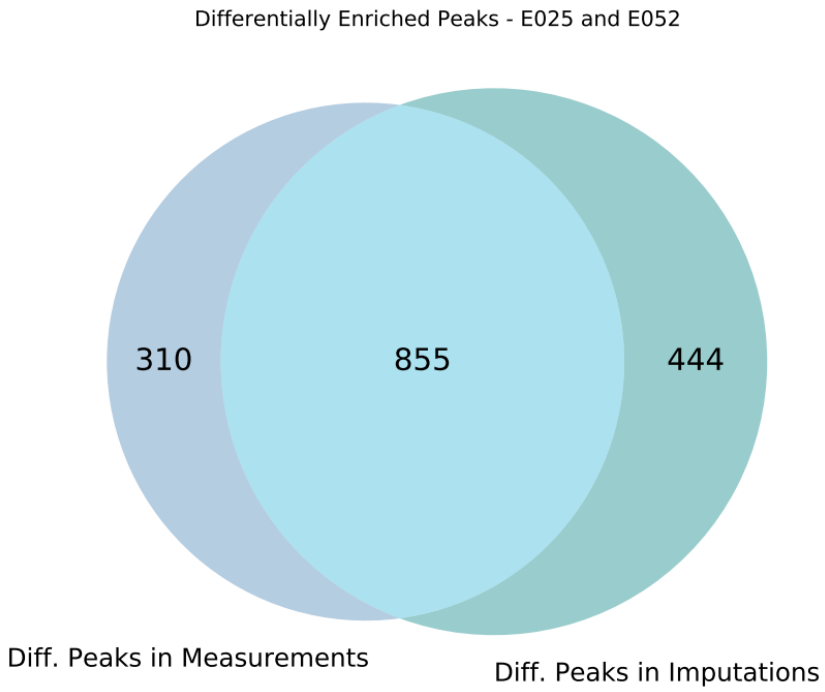
Image from https://bioramble.wordpress.com/2016/01/30/why-sequencing-data-is-modeled-as-negative-binomial/

EPI-LOGOS

Binding affinity scores for the imputed replicates reflect the pattern found in the measurements.

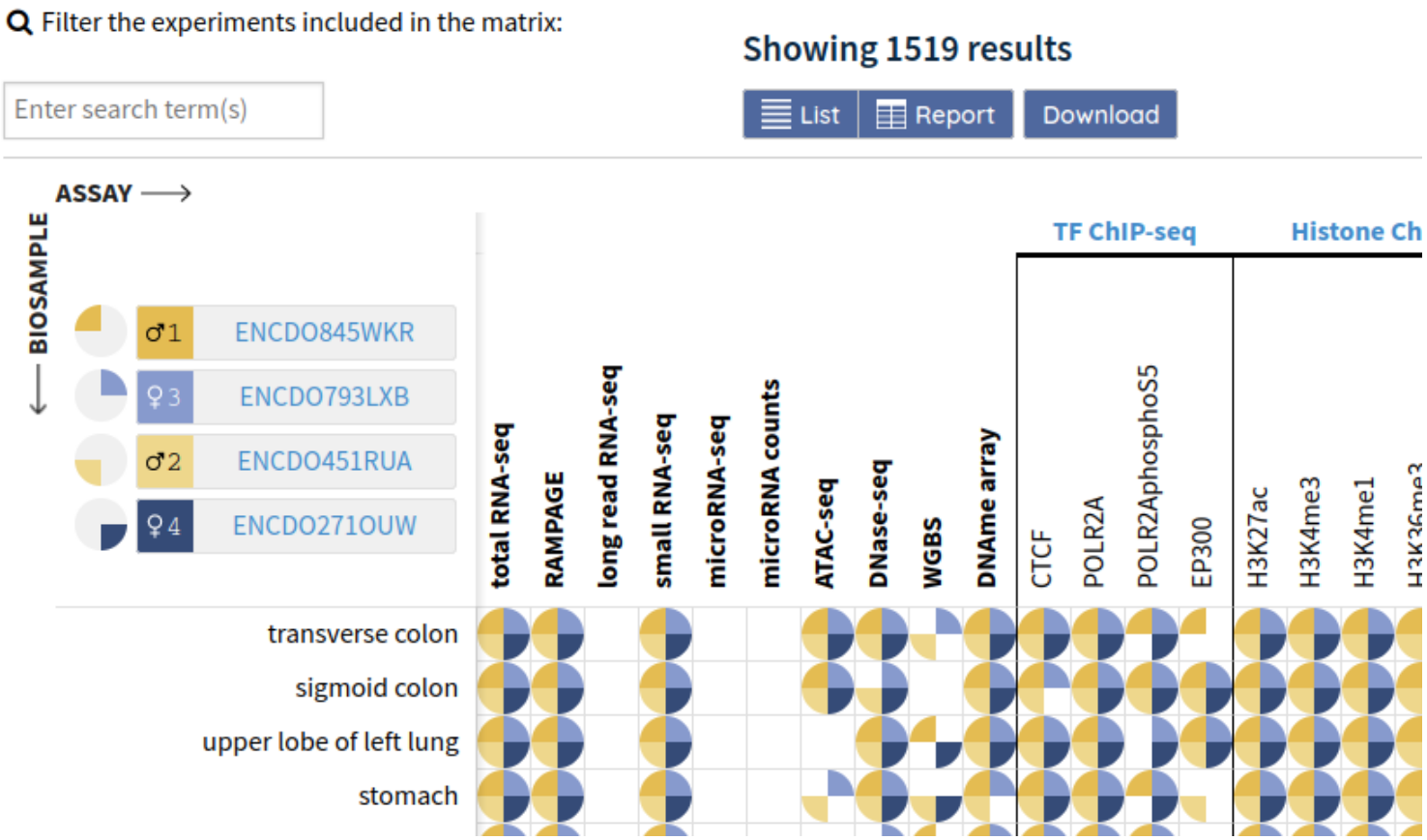


This differential analysis retrieves most of the meaningful differences between sets of replicates.
(PPV ~70%)

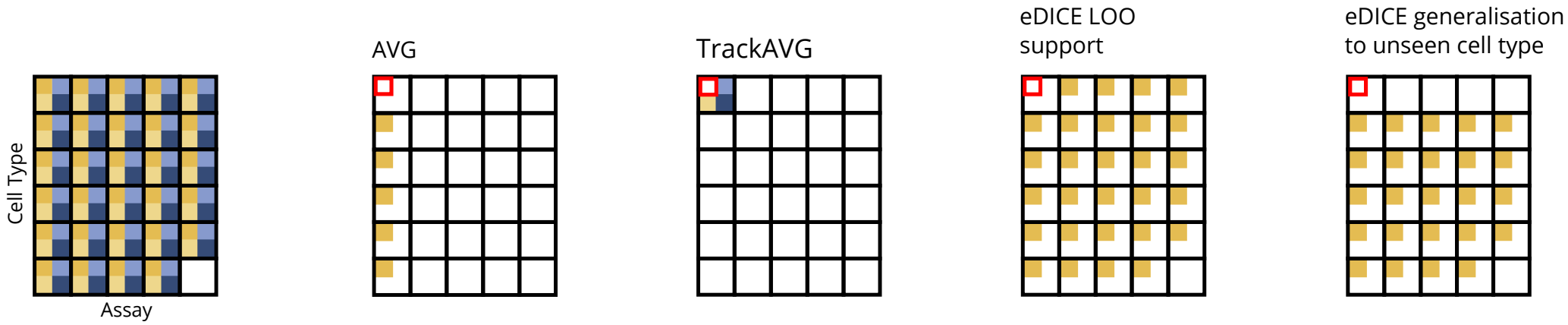


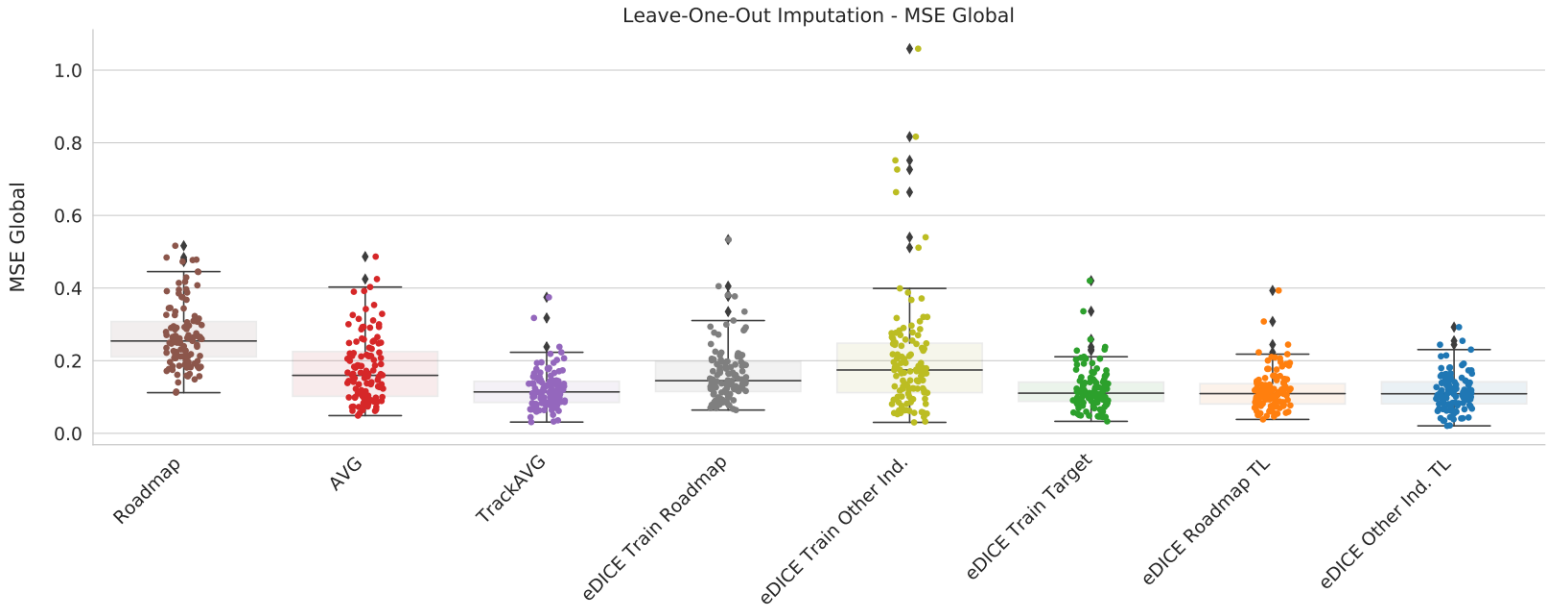
Instead of predicting data from reference epigenomes, can we make personalised predictions?

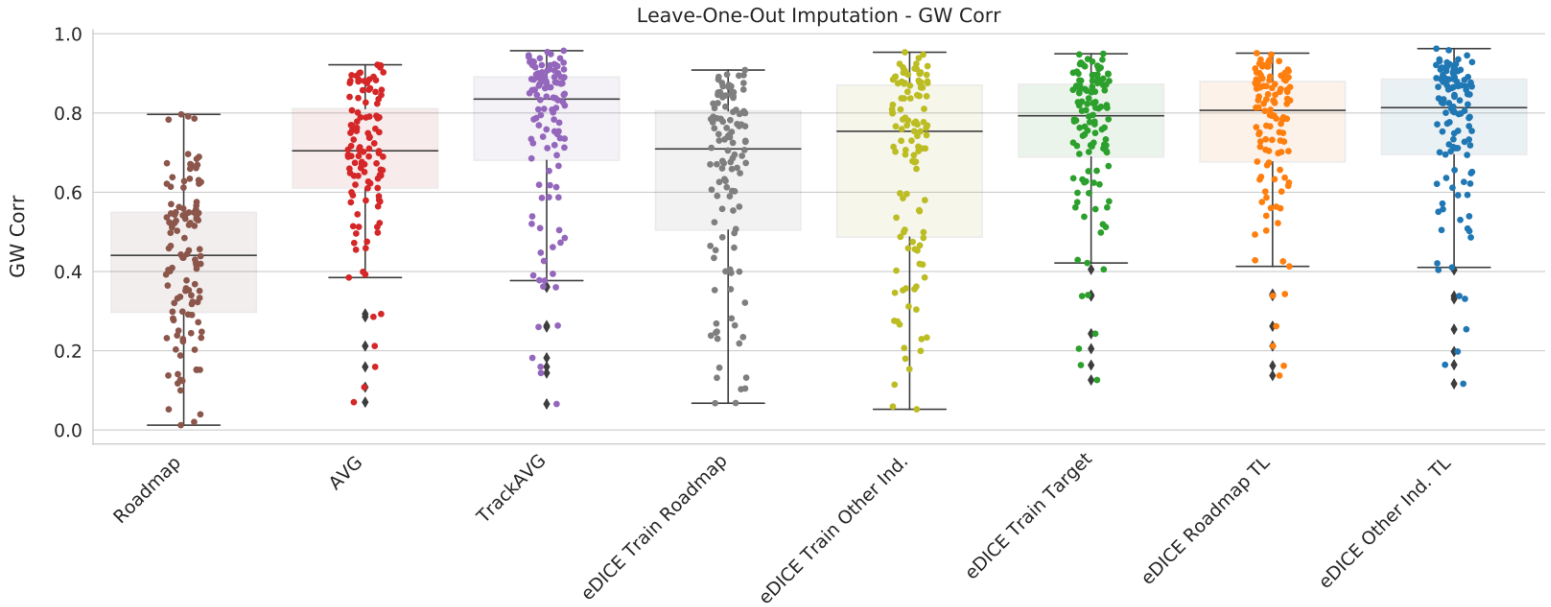
Epigenomes from four individuals (ENTEx)



We compare to two baselines: averaging over tissues (AVG), and averaging over individuals (TrackAVG)



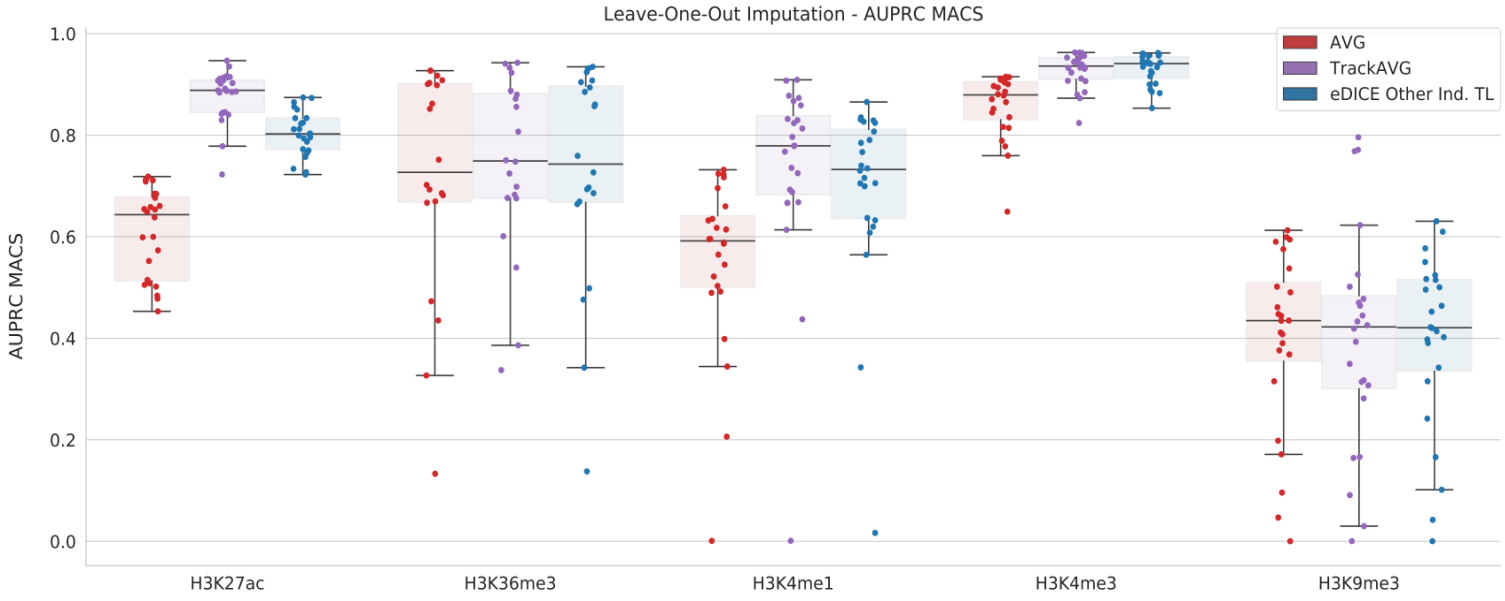




EPI-LOGOS

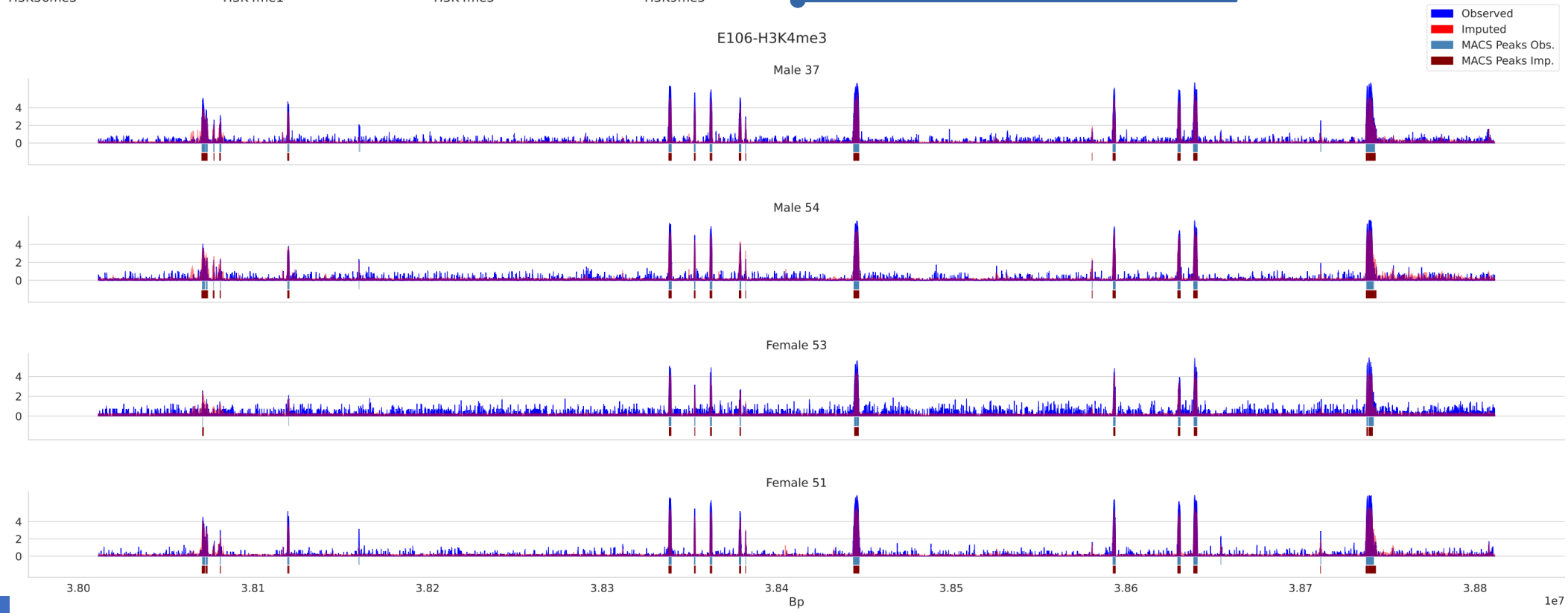
Among eDICE models, transfer learning from other individuals works best

It also offers the lowest computational cost

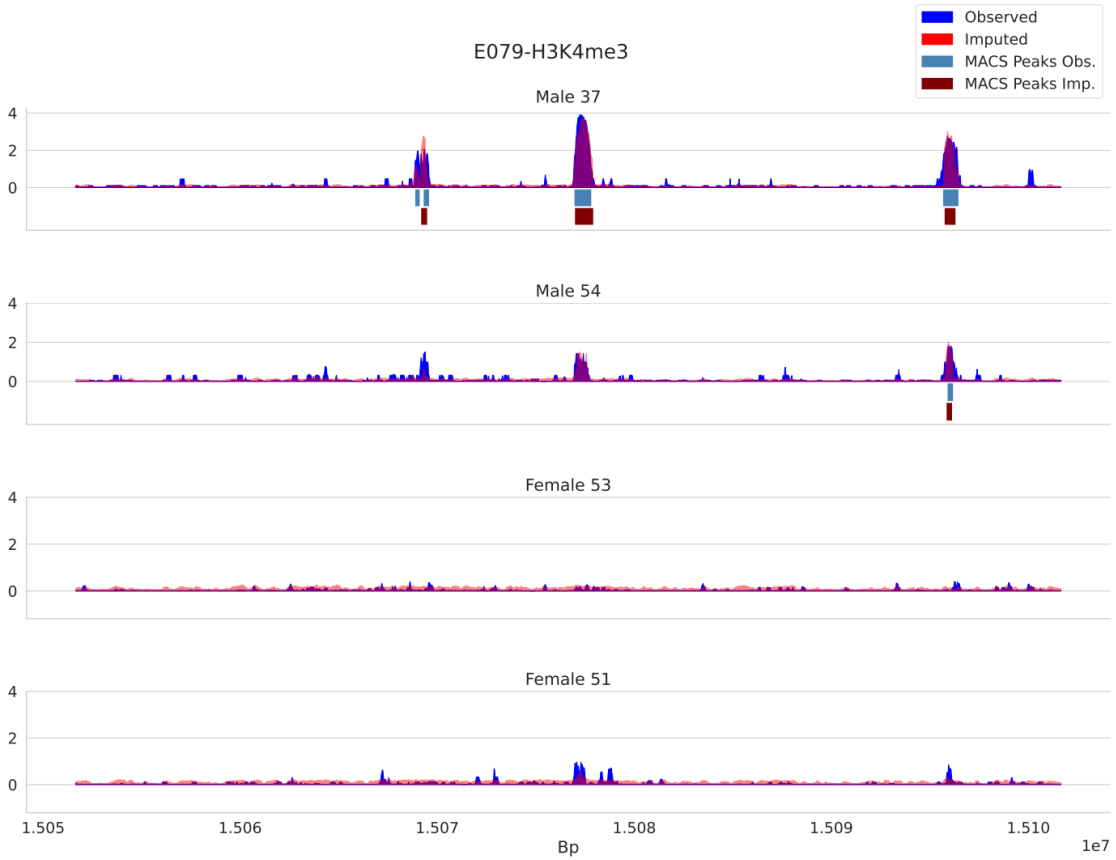


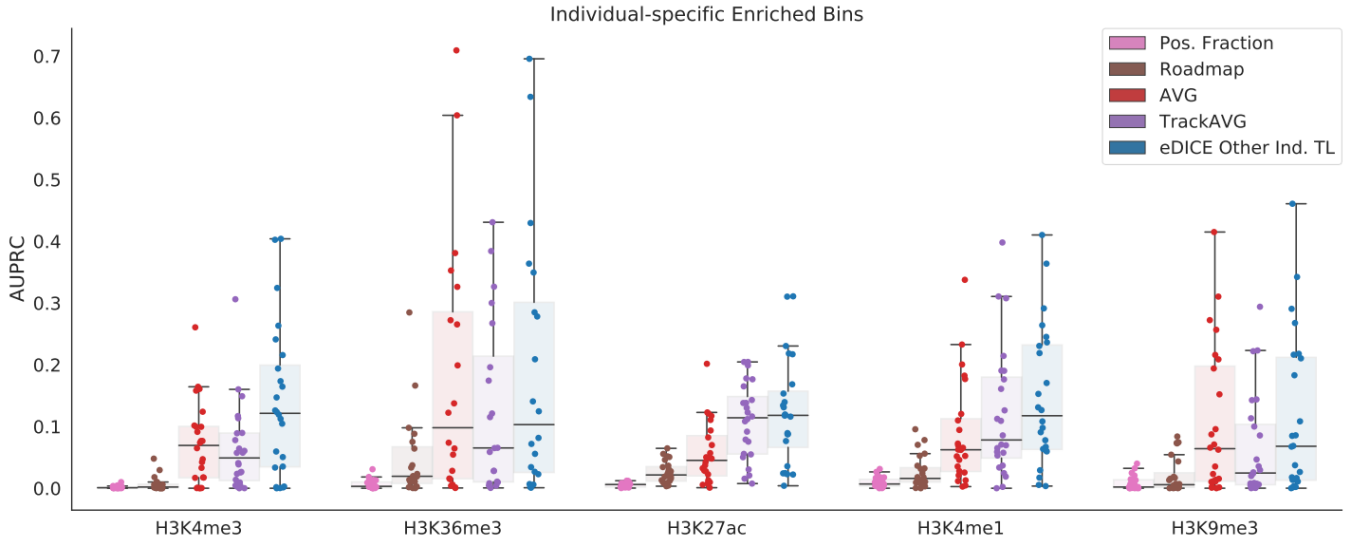
EPI-LOGOS

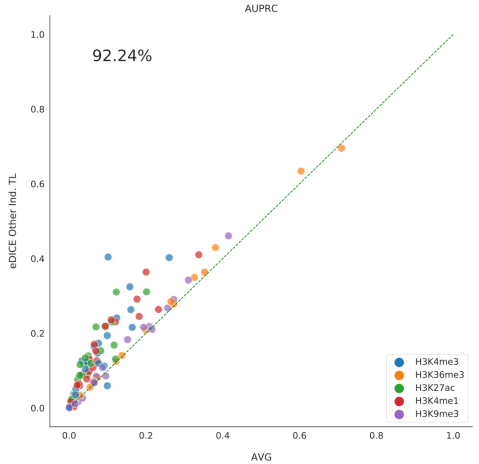
Averaging over individuals finds most of the enriched regions, which are conserved between patients

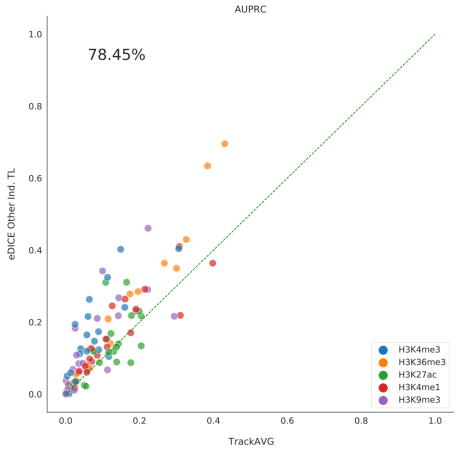


By construction, TrackAVG does not contain personalised information. We examined the performance of the imputations after masking out the shared regions.



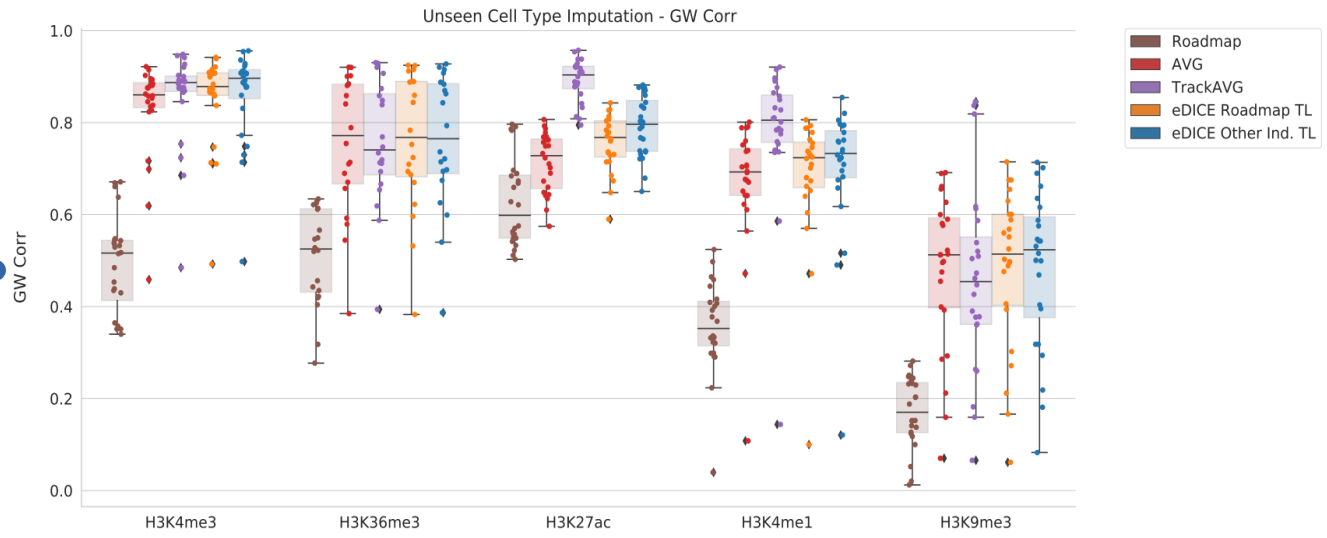


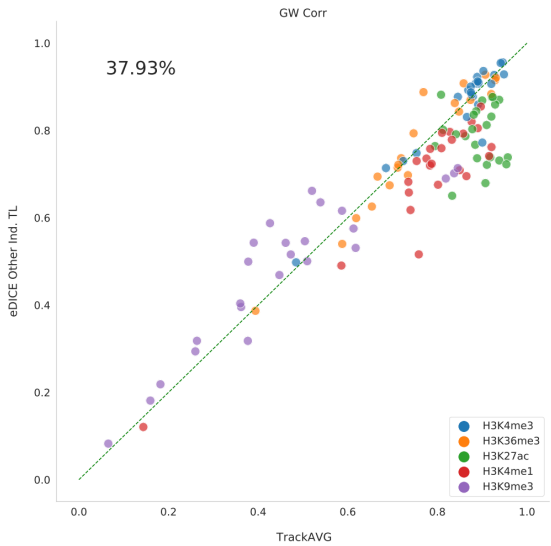


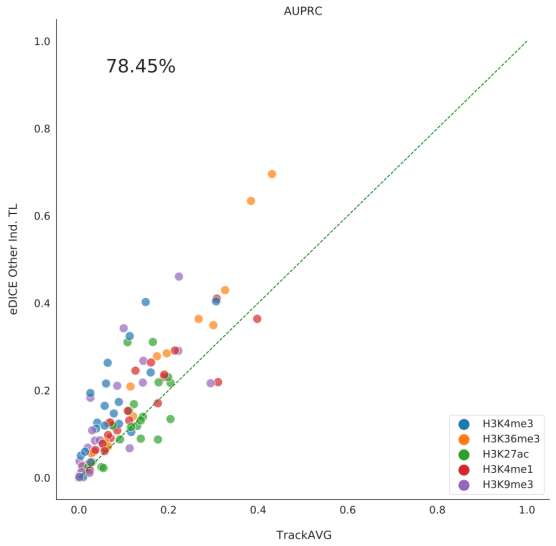


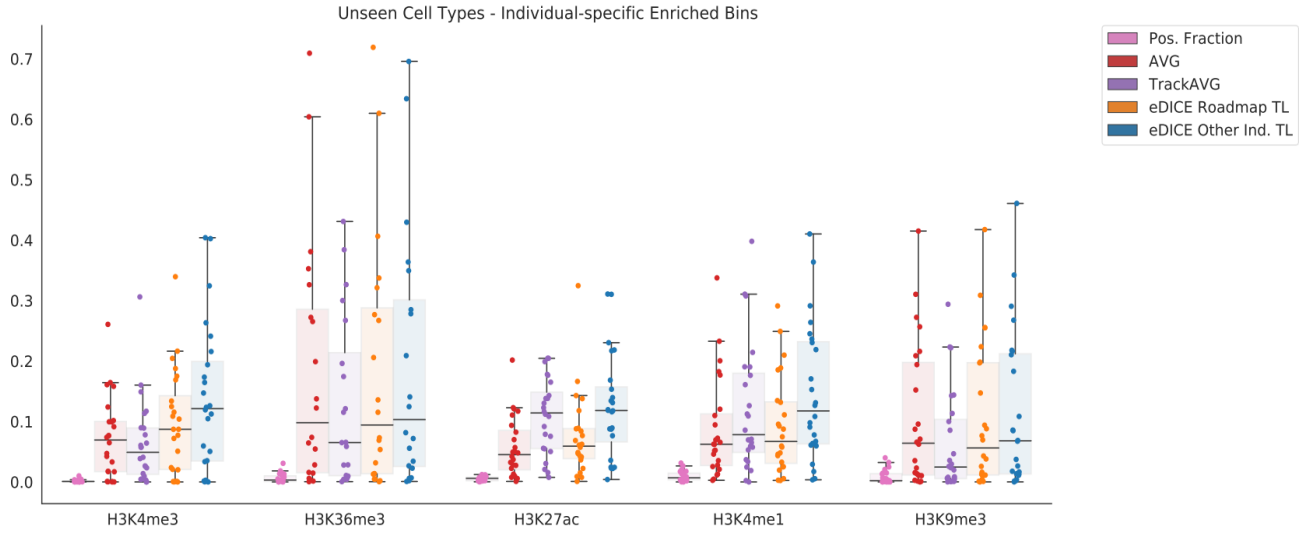
The transfer learning process allows generalisation to unseen tissues

EPI-LOGOS









The OOD imputations capture individual-specific enrichment better than the baseline

Open Questions and Future Work

Predicting **variance** of imputed epigenomes

Improving **robustness of transfer learning**

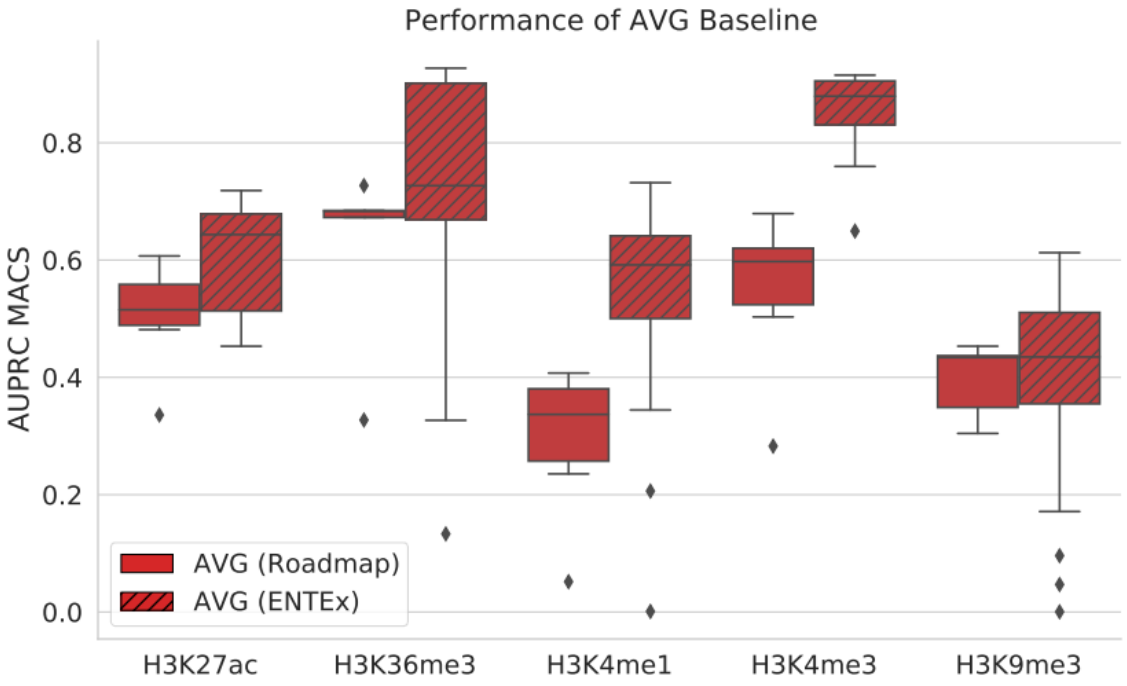
Application to other datasets. How to deal with **covariate shift**?

Acknowledgements

- Alex Hawkins-Hooker, UCL
- Tanmayee Narendra, University of Tübingen
- Mateo Rojas-Carulla, Lakera AI
- Bernhard Schölkopf, Max-Planck Institute for Intelligent Systems
- Gabriele Schweikert, University of Dundee

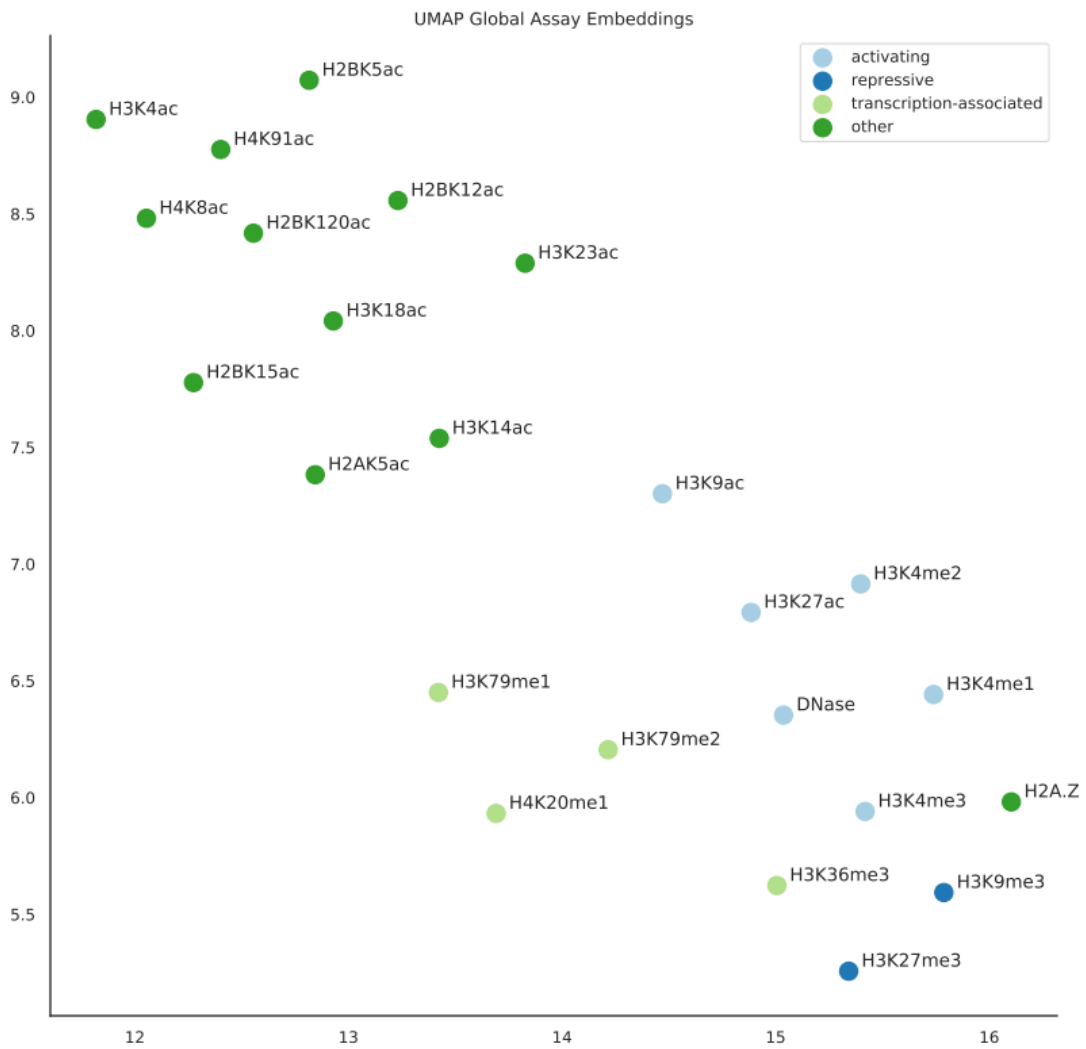
Thank you for your attention

Systematic dataset shifts, differences between assays, and hidden confounders are difficult challenges for personalised imputation



Global embeddings capture tissue similarity and general epigenetic mark function

EPI-LOGOS



EPI-LOGOS

Self-Attention opens up possibilities for **interpretation** of the model

Percentage of attention

$$p_h(\alpha) = \sum_{g \in G} \sum_{a \in A} w_{a\alpha}^{(g,h)} \Big/ \sum_{g \in G} \sum_{a \in A} \sum_{a' \in A} w_{aa'}^{(g,h)}$$

Global Percentage of Attention - Chromosome 21

	DNase	H2A.Z	H2AK5ac	H2BK120ac	H2BK12ac	H2BK15ac	H2BK5ac	H3K14ac	H3K18ac	H3K23ac	H3K27ac	H3K27me3	H3K36me3	H3K4ac	H3K4me1	H3K4me2	H3K4me3	H3K79me1	H3K79me2	H3K9ac	H3K9me3	H4K20me1	H4K8ac	H4K91ac
Attention Head 1	2.1%	5.0%	3.4%	1.8%	2.6%	2.1%	2.3%	2.7%	1.6%	5.9%	2.4%	7.5%	11.3%	0.9%	7.2%	6.5%	7.1%	2.3%	3.1%	3.3%	12.9%	2.9%	2.0%	1.2%
Attention Head 2	0.9%	6.1%	1.5%	2.4%	1.5%	1.4%	0.9%	2.4%	1.8%	1.6%	0.8%	20.9%	15.5%	2.9%	5.9%	0.6%	4.4%	1.4%	1.1%	0.6%	16.9%	1.2%	5.8%	1.5%
Attention Head 3	4.0%	4.6%	3.7%	3.6%	3.5%	3.5%	3.5%	4.5%	4.1%	4.5%	3.8%	4.4%	4.4%	3.9%	4.4%	4.5%	5.3%	3.9%	4.6%	5.0%	4.2%	3.8%	4.6%	3.5%
Attention Head 4	5.8%	4.1%	3.5%	3.3%	2.5%	2.5%	3.6%	3.1%	5.1%	4.2%	6.5%	3.4%	3.1%	4.0%	5.3%	6.2%	6.3%	3.1%	4.6%	8.0%	3.0%	2.8%	2.9%	3.0%

EPI-LOGOS

How does this portion of attention **shift within functional regions** of the genome?

$$d_h^{(R)}(\alpha) = \frac{p_h(\alpha)|_{G \equiv R} - p_h(\alpha)}{p_h(\alpha)}$$

Attention Shifts - Gene

	DNase	H2A.Z	H2AK5ac	H2BK120ac	H2BK12ac	H2BK15ac	H2BK5ac	H3K14ac	H3K18ac	H3K23ac	H3K27ac	H3K27me3	H3K36me3	H3K4ac	H3K4me1	H3K4me2	H3K4me3	H3K79me1	H3K79me2	H3K9ac	H3K9me3	H4K20me1	H4K8ac	H4K91ac
Attention Head 1	-0.4%	2.1%	1.0%	1.0%	1.5%	2.1%	0.3%	0.3%	0.6%	-0.1%	0.1%	1.3%	-4.0%	0.6%	-0.5%	-1.7%	0.7%	-4.0%	-8.3%	0.6%	5.4%	-6.4%	-0.2%	1.2%
Attention Head 2	-2.1%	3.0%	2.1%	2.4%	2.6%	3.1%	2.0%	3.4%	2.5%	1.9%	-6.0%	0.8%	-13.9%	3.8%	-7.3%	-3.1%	-1.2%	4.8%	-10.0%	-6.4%	11.6%	-9.4%	2.0%	2.6%
Attention Head 3	-0.2%	1.1%	-1.1%	-1.0%	-1.7%	-1.2%	-1.2%	0.3%	-0.8%	1.0%	-0.6%	0.5%	0.8%	-0.6%	0.3%	0.4%	1.8%	-0.7%	0.3%	1.8%	-0.3%	-0.9%	0.8%	-1.3%
Attention Head 4	-2.0%	-1.8%	0.2%	0.3%	0.3%	-0.0%	0.8%	1.4%	0.9%	0.1%	-1.0%	-1.2%	0.5%	1.3%	-0.3%	-1.4%	-1.5%	6.4%	1.7%	0.4%	-1.9%	-0.1%	0.4%	0.7%

Attention Shifts - Promoter

	DNase	H2A.Z	H2AK5ac	H2BK120ac	H2BK12ac	H2BK15ac	H2BK5ac	H3K14ac	H3K18ac	H3K23ac	H3K27ac	H3K27me3	H3K36me3	H3K4ac	H3K4me1	H3K4me2	H3K4me3	H3K79me1	H3K79me2	H3K9ac	H3K9me3	H4K20me1	H4K8ac	H4K91ac
Attention Head 1	-22.4%	-33.8%	9.8%	2.6%	12.4%	12.8%	0.5%	0.6%	-4.2%	5.3%	-5.5%	12.4%	24.3%	-2.7%	-0.2%	-57.1%	-26.9%	6.0%	-13.3%	-22.1%	29.4%	11.1%	-4.7%	0.8%
Attention Head 2	-44.2%	-47.4%	7.6%	12.3%	6.3%	13.7%	3.8%	18.5%	8.9%	3.9%	-53.4%	-10.6%	21.6%	14.1%	-39.9%	-70.0%	-74.8%	12.2%	-29.5%	-68.4%	39.4%	-8.7%	11.6%	8.8%
Attention Head 3	-5.6%	11.1%	-8.3%	-9.2%	-12.6%	-10.0%	-10.0%	1.6%	-9.7%	12.8%	-7.2%	6.1%	10.3%	-6.6%	-0.3%	-2.9%	10.2%	-0.1%	4.8%	16.0%	0.5%	-7.7%	6.7%	-11.5%
Attention Head 4	-14.3%	-10.3%	0.4%	7.6%	-1.8%	0.6%	4.5%	11.3%	16.7%	-5.5%	-6.6%	-8.6%	-11.8%	9.8%	-14.4%	-6.7%	5.4%	6.0%	0.1%	17.9%	-12.0%	-8.0%	6.6%	7.3%

EPI-LOGOS

Attention shifts are **consistent with the literature**