# Latent Modeling of the Human Epigenome

Jacob Schreiber

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

William Stafford Noble, Chair

Cole Trapnell

Ali Shojaie

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

**Abstract**

Latent Modeling of the Human Epigenome

Jacob Schreiber

Chair of the Supervisory Committee:
Professor William Stafford Noble
Department of Genome Science

The human epigenome has been experimentally characterized by assays of methylation, histone modification, chromatin accessibility, and protein binding in hundreds of cell lines and tissues ("biosamples"). The result is a huge compendium of data, consisting of thousands of measurements for every basepair in the human genome. These measurements are immensely valuable, in large part because they measure forms of biological activity that differ across biosamples and can help explain many biosample-specific cellular mechanisms that cannot be explained by nucleotide sequence alone, particularly those driving development and disease.

However, these data present two major challenges. The first challenge is that, due primarily to cost, the total number of assays that can be performed is limited. The second challenge is that, despite being incomplete, these compendia are already so large that they can be difficult for either humans or computational methods to make sense of.

In this thesis, we address both of these challenges with a deep tensor factorization method, Avocado, that is trained to impute genome-wide epigenomics experiments. Avocado solves the first challenge by completing the compendium via imputation of all epigenomic experiments that have not yet been performed. Avocado solves the second challenge by learning a compression of the entire compendium into a dense, information-rich, latent representation.

We first applied Avocado to a compendium of data produced by the Roadmap Epigenomics Consortium that contained measurements of chromatin accessibility and histone modification. Our results confirmed the strength of the Avocado model: first, we found that Avocado can impute epigenomic data more accurately than previous methods, and second, we showed that machine learning models that exploit Avocado's learned representation outperform those trained directly on epigenomic data on a variety of genomics tasks.

Next, we applied Avocado to the ENCODE Compendium, which is several times larger than the Roadmap Compendium and additionally includes measurements of protein binding and transcription. We demonstrate that, even in this more difficult setting, Avocado's imputations are of high quality and that the predictions of protein binding outperform the top models in a recent ENCODE-DREAM challenge.

Although the ENCODE compendium currently contains only a small fraction of potential experiments, the human epigenome remains the most characterized epigenome of any species. Accordingly, we extended Avocado to leverage the large number of human epigenomic data sets when making imputations in other species. We found that not only does this extension yields improved imputations of mouse epigenomics, but that the extended model is able to make accurate and biosample-specific imputations for assays that have been performed in humans but not in mice. Further, we found that our extension allows for an epigenomic similarity measure to be defined over pairs of regions across species based on Avocado's learned representations and that this score can be used to identify regions with high sequence similarity whose functions have diverged.

Finally, we sought to demonstrate the utility of these imputations for the challenging task faced by a scientific consortium such as the ENCODE Consortium, "Which experiments should ENCODE perform next?" We demonstrate how to represent this task as an optimization problem carried out using Avocado's imputations. Compared with previous work that

has addressed a similar problem, our approach has the advantage that it can use imputed data to tailor the selected list of experiments based on data collected previously by the consortium. We demonstrate the utility of our proposed method in simulations, and we provide a general software framework, named Kiwano, for prioritizing the order that genomic and epigenomic experiments should be performed.

Taken together, the results presented in this thesis provide strong empirical evidence for the utility and robustness of Avocado. In multiple settings, the imputations generated by Avocado are of high quality, including the novel cross-species settings. The learned latent representations are able to encode epigenomic state in a compact manner, and even result in a way to identify orthologous regions that have diverged across species. Finally, we have shown that the imputations are informative and biosample-specific enough to help guide future experimental efforts.

All of the results from this thesis are publicly available. The imputations can be found on the ENCODE portal (https://www.encodeproject.org). The model files for the first chapter can be found at https://noble.gs.washington.edu/proj/avocado/model/ and the model files for the second chapter can be found at https://noble.gs.washington.edu/proj/mango/models/. The code for Avocado can be found at https://github.com/jmschrei/avocado and has been made available under an Apache v2 license, and the code for Kiwano can be found at https://github.com/jmschrei/kiwano under the MIT license.

# TABLE OF CONTENTS

# LIST OF FIGURES

Figure Number Page

# ACKNOWLEDGMENTS

There are too many people that helped me during my time in graduate school to completely list here; instead, I will partially enumerate them and hope that the others do not read this dissertation.

Foremost, I would like to acknowledge my advisor, Bill Noble, whose patience acted as a foil for my indignance.

I would also like to acknowledge:

- *my committee*, most of whom were kind enough to supervise me on short notice.
- *the many people who provided guidance, thoughtful discussions, and positive reviewer feedback* throughout my studies, including Larry Ruzzo, Scott Lundberg, Eric Mendenhall, Rick Myers, Bryan Moyers, Anshul Kundaje, Carles Boix, Christina Leslie, Gurkan Yardimci and Fabio Navarro.
- *my co-authors*, including Timothy Durham, Maxwell Libbrecht, Ritambhara Singh, Molly Gasperini, Jay Shendure, Will Chen, Deepthi Hegde, and Jeffrey Bilmes.
- *my support network*, including Antoine Bossolut, Mengsha Li, Jiechen Chen, David Wadden, Katie Doroschak, Naozumi Hiranuma, Maaz Ahmad, Niel Lebeck, Ruby Williams, Athena Lin, and most importantly, my father.
- *my fiancée*, Terra Blevins, without whom my writing would have fewer references to natural language processing.

# DEDICATION

To my mother, who wanted to be here.

# Chapter 1

# AVOCADO CAN LEARN A LATENT REPRESENTATION OF HUMAN EPIGENOMICS

## *1.1 Background*

The Human Genome Project, at its completion in 2003, yielded an accurate description of the nucleotide sequence of the human genome but an incomplete picture of how that sequence operates within each cell. Characterizing each basepair of the genome with just two bits of information—its nucleotide identity—yielded many critical insights into genome biology but also left open a host of questions about how this static view of the genome gives rise to a diversity of cell types. Clearly, answering these questions required gathering more data.

In the ensuing 15 years, driven by advances in next-generation sequencing, the research community has developed many assays for characterizing the human epigenome. These include bisulfite sequencing for measuring methylation status, DNase-seq and ATAC-seq for measuring local chromatin accessibility, ChIP-seq for measuring protein binding and histone modifications, RNA-seq for measuring RNA expression, and Hi-C for measuring the 3D structure of the genome. These assays can quantify variation in important biological phenomena across cell types. Accordingly, large consortia such as ENCODE, Roadmap Epigenomics, IHEC, and GTEx, have run many types of assays in many human cell types and cell lines, yielding thousands of epigenomic measurements for each basepair in the human genome. For example, as of May 1, 2018, the ENCODE project (http://www.encodeproject.org) hosts $> 10,000$ assays of the human genome.

Although these data have deepened our understanding of genome biology, we are still far from fully understanding it. Gene annotation compendia such as GENCODE are now quite mature, but cell-type-specific annotations of chromatin state remain only partially interpretable [1, 2, 3]. Other areas of active research include predictive models of gene expression, promoter-enhancer interactions, polymorphism impact, replication timing and 3D conformation (reviewed in [4]).

One part of the analytical challenge arises from the complexity of the genome and its interactions with other physical entities in the cell, but another part stems from biases and noise in the epigenomic data itself. For example, many such data sets exhibit position-specific biases, reflecting variation in local chromatin architecture or GC bias in the sequencer. Furthermore, no high-throughput assay is perfectly reproducible, and run-to-run differences in the same experiment may reflect either biological variation in the cells being assayed or experimental variance arising from sample preparation or downstream steps in the protocol. Finally, many epigenomic assays are highly redundant with one another, and many cell types are closely related to each other, leading to highly redundant measurements.

These measurements are critical for a comprehensive understanding of the human genome, and are central to many efforts to explain complicated phenomena in the cell. For example, researchers have utilized histone modification measurements in promoter regions to explain variation in expression levels of the corresponding gene [5, 6, 7, 8]. The measurements additionally allow researchers to ask more sophisticated questions about regulatory elements, such as explaining how gene expression is regulated through interactions between promoter regions and potentially very distant enhancers [9, 10, 11, 12].

However, this data can be difficult to deal with directly for three reasons. The first is that it is massive and growing. In the context of building machine learning models, each newly performed assay typically corresponds to one or more additional feature for the model, requiring growing computational resources for sophisticated models. The second is that it is

incomplete. While some assays are performed comprehensively over hundreds of cell types, more frequently they are performed in a small number of cell types of interest to the researcher collecting the data. This can make comprehensively analyzing many cell types difficult, as the set of assays performed in them may vary. The third is that biases and noise exist in the experimental data itself. For eample, many data sets exhibit poition-specific biases, reflecting variation in local chromatin architecture or GC bias in the sequencer. Furthermore, no high-throughput assay is perfectly reproducible, and run-to-run differences in the same experiment may reflect either biological variation in the cells being assayed or experimental variance arising from sample preparation or downstream steps in the protocol. Finally, many epigenomic assays are highly redundant with one another, and many cell types are closely related to each other, leading to highly redundant measurements.

To address these challenges, we[1] aim to produce a representation of the human epigenome that is dense and information-rich. Ideally, this representation will reduce redundancy, noise, and bias, so that variance in the representation corresponds to meaningful biological differences rather than technical artifacts. Computationally, this goal can be framed as an embedding task, in which we project the observed collection of thousands of epigenomic measurements per genomic position down to a low-dimensional "latent space." Our aim is to induce a latent representation of the genome that can be used in place of epignomic measurements as input to machine learning models trained to perform a variety of genomic predictive modeling tasks.

To solve this embedding task we combine two mathematical techniques—tensor factorization and deep neural networks. Epigenomic data sets can be represented as a tensor with three orthogonal axes: genomic position, cell type, and assay type. Tensor factorization is

[1]The work in this chapter is based off a paper entitled *Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome* to appear in *Genome Biology* that was written by myself, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble (in the order that authors appear on the paper). In this work, WSN and myself conceived of experiments, I did the coding and analysis, and all authors contributed to writing the text.

thus a natural framework for distilling this data into an informative latent representation [13]. The deep neural network augments this process with the ability to encode nonlinear relationships among the factors and to capture dependencies along the genomic axis at various scales.

In order to learn a latent representation of human epigenomics, we train our model, which we call "Avocado," to perform epigenomic imputation. This task involves computationally "filling in" gaps in a tensor of epigenomic data, where gaps correspond to experiments that have not yet been run. Using data from the Roadmap Epigenomics Consortium, we demonstrate that Avocado yields imputed values that are more accurate than those produced by either ChromImpute [14] or PREDICTD [13], as measured by multiple performance measures. Avocado's imputed data also captures relationships between pairs of histone marks more accurately than these previous approaches. For example, Avocado accurately predicts that activating marks in a promoter region are typically mutually exclusive with repressive marks and are coupled with a higher transcription rate of the associated gene.

Our primary hypothesis is that Avocado's learned representation will be generally useful in the context of a variety of predictive modeling tasks. The idea that representations can be learned in one setting and then used as input for other settings is similar to that of word2vec [15] and is an example of transfer learning. To test this hypothesis, we consider the tasks of predicting gene expression, promoter-enhancer interaction [16], replication timing, and elements known as "frequently interacting regions" (FIREs), defined on the basis of Hi-C data [17]. For each task, we train a supervised machine learning model on each of seven alternate sets of features—experimentally collected epigenomic measurements for the cell type of interest, the three sets of imputed epigenomic assays for the cell type of interest, the latent representation learned by PREDICTD, the latent representation learned by Avocado, or the experimentally collected epigenomic measurements from all cell types and assays contained within the Roadmap compendium. We include the entirety of the Roadmap compendium

as a baseline because, while computationally expensive to train machine learning models on, it contains the full set of information used to learn the Avocado latent representation. In almost every case, we observe that models trained using Avocado's learned latent representation outperform models trained directly on either the primary or imputed data for the cell type of interest. In those remaining cases, the performance of models trained using Avocado's learned latent representation is similar to models trained using either the primary or imputed data for the cell type of interest. Notably, the models that utilize the Avocado latent representation outperform those that utilize the PREDICTD latent representation in every cell type for predicting gene expression and FIREs. However, we notice that the use of the full Roadmap compendium proves to be a surprisingly difficult baseline to beat, and that it also consistently outperforms using either the primary or imputed data from a cell type of interest. When models trained using the Avocado latent factors are compared to those trained using the full Roadmap compendium, there are some contexts in which models trained on the Avocado latent factors perform best and some cases where models trained on the Roadmap compendium perform best. These results suggest the broad utility of Avocado's approach to learning a latent representation of the genome, and that this utility is derived in part from compressing epigenomic assay measurements from all cell types at each genomic position, instead of only a single cell type. Additionally, our results suggest that the process used to learn a latent representation can affect their utility, and that our approach yields a more informative representation than the simpler approach adopted by PREDICTD.

Lastly, we use feature attribution methods to understand the Avocado model. We find that the genomic latent factors encode most of the "peak-like" structure of epigenomic data, while the cell type and assay latent factors serve mostly to sharpen or silence these peaks for a specific track. This observation suggests that the latent representation encodes a rich representation of the functional landscape of the human epigenome.

## 1.2 Results

### 1.2.1 Avocado employs multi-scale deep tensor factorization

To produce a latent representation of the genome, we began with the tensor factorization model employed by PREDICTD. In this model, the 3D data tensor is modeled by three 2D matrices of latent factors, corresponding to cell types, assay types, and genomic positions (Fig. 1.1a). PREDICTD combines these latent factors in a straightforward way by extracting, for each imputed value, the corresponding rows from each of the three latent factors matrices and linearly combining them via generalized dot product operations. Avocado improves upon this approach in two significant ways.

First, Avocado generalizes PREDICTD so that the relationship between the data and the latent factors is nonlinear, by inserting a deep neural network (DNN) into the architecture in place of the generalized dot product operation (Fig. 1.1b). Note that similar "deep factorization" methods have been proposed previously [18, 19]; however, Avocado differs from these methods in an important way: rather than point-multiplying the three pairs of latent factors and putting the resulting vectors through a DNN, Avocado instead concatenates the three latent factor vectors for direct input to the DNN. This more general approach enables Avocado to embed information about cell types, assay types and genomic positions into latent spaces with different dimensionalities.

The concatenation also enables Avocado's second improvement relative to PREDICTD, namely, that the model adopts a multi-scale view of the genome. Avocado employs three sets of latent factors to represent the genome at different scales: one set of factors are learned for each of the 115,241,319 genomic coordinates at 25 bp resolution, another set are learned at 250 bp resolution, and a final set are learned at 5 kbp resolution. These three length scales represent prior knowledge that important epigenomic phenomena operate at fine scale (e.g., transcription factor binding), at the scale of individual nucleosomes, and at

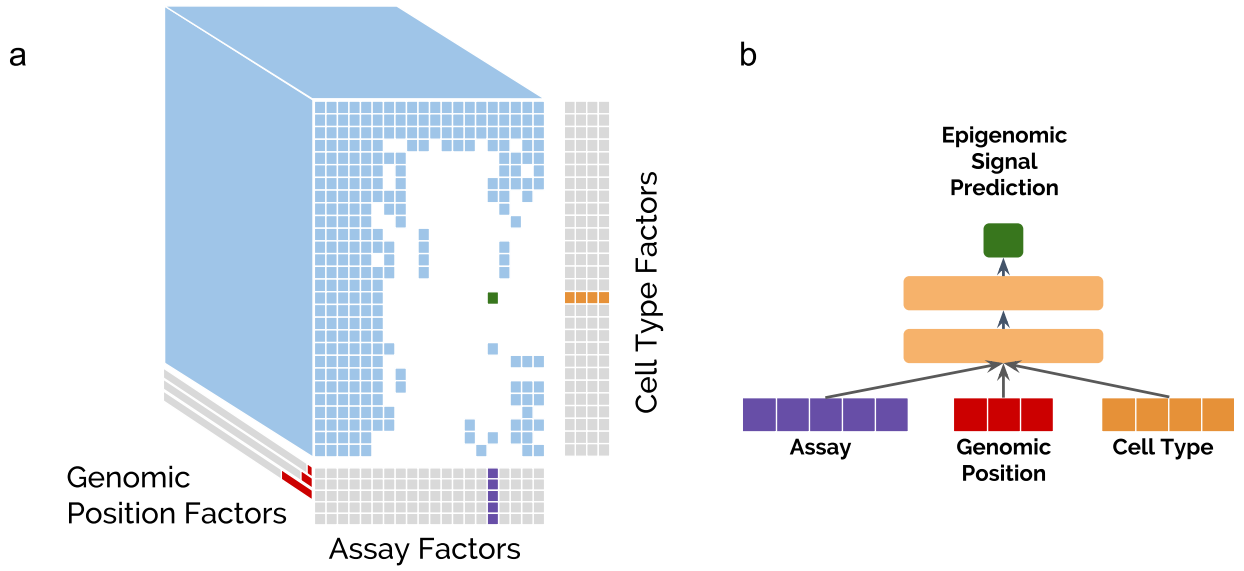



Figure 1.1: **The Avocado deep tensor factorization approach.** (a) A collection of epigenomic data can be visualized as a 3D tensor (blue), in which some experiments (white cells) have not yet been performed. Avocado models the tensor along three orthogonal axes, learning latent factors (gray) that represent the cell types (in orange, with 32 factors each), the assay types (in purple, with 256 factors each), and the genomic axis (in red, with 25, 40, and 45 factors at each of the three resolutions). (b) During the training process, the respective slices from these three axes corresponding to the location of the training sample in the tensor are concatenated together and fed into a neural network comprised of two hidden dense layers each with 2,048 neurons to produce the final prediction (in green).

a broader "domain" scale. Furthermore, by learning the genomic representations at multiple scales, Avocado's genomic latent space can employ far fewer parameters than PREDICTD, requiring only $\sim 3.4$ billion parameters instead of $\sim 11.5$ billion to model each position along the genome. In total, Avocado requires only $\sim 3.7$ percent of the $\sim 92.2$ billion parameters employed by PREDICTD's full ensemble of eight tensor factorization models.

A critical step in developing a model like Avocado involves selecting an appropriate model topology. Avocado's model (see Methods) has seven structural hyperparameters: the number of latent factors representing cell types, assay types, and the three scales of genomic positions, as well as two parameters (number of layers and number of nodes per layer) for the deep neural network. To select these values, we used random search over a grid of hyperparameters, selecting the set that performs best according to the MSE on a validation set when considering the ENCODE Pilot Regions, a selected 1% of the positions in the human genome (Appendix A). The results of this analysis suggest that, among the seven hyperparameters, the two that control the size of the deep neural network are most important, with Avocado performing best with 2 layers and 2,048 neurons per layer (Fig. A.2). We also found that using "drop-out," a form of regularization that involves randomly skipping over some model parameters at each training iteration, significantly boosts Avocado's performance (Fig. B.1).

A full description of the model can be found in Section 2.5.1. This approach requires that many hyperparameters needed to be optimized, such as the number of factors in each dimension and the structure of the neural network, and so hyperparameter selection was performed as described in Appendix A to select these values. Briefly, we used random search and selected the set that performed the best on a validation set when considering only the ENCODE Pilot Regions. Once these hyperparameters were selected, our scheme for training Avocado across the whole genome involves first training the model on the Pilot regions, freezing the aspects of the model that are not genome factors, and then learning only the

| MSE- | global | 1obs | 1imp | Prom | Gene | Enh |
|---|---|---|---|---|---|---|
| **ChromImpute** | 0.113 | **0.941** | 1.09 | 0.325 | 0.149 | 0.316 |
| **PREDICTD** | **0.100** | 1.76 | 0.897 | 0.258 | **0.129** | 0.267 |
| **Avocado** | **0.100** | 1.66 | **0.845** | **0.249** | **0.130** | **0.260** |

Table 1.1: **Evaluation of ChromImpute, PREDICTD, and Avocado.** Six performance measures are reported, reflecting MSE of different subsets of genomic positions. The best result for each metric is in boldface and corresponds to an adjusted two-sided paired t-test p-value $< 0.01$ when compared to both other approaches. For MSE-global and MSE-Gene, both PREDICTD and Avocado are bolded because the difference between the two is not statistically significant, i.e., has a p-value $> 0.01$.

genome factors for each chromosome individually (Section 1.5.4). This approach is meant to overcome the limitation that the full set of genome factors cannot fit in memory by leveraging the fact that the remainder of the model is identical across chromosomes.

### *1.2.2 Avocado imputes epigenomic tracks more accurately than prior methods*

We began our analysis of the Avocado latent representation by measuring its ability to impute epigenomic assays, comparing the overall accuracy of Avocado, as measured by mean-squared error (MSE), to that of ChromImpute and PREDICTD. To this end, we evaluated all three methods on 1,014 tracks of epigenomic data from the Roadmap Epigenomics project. Imputations from Avocado and PREDICTD were made using a five-fold cross validation approach where the folds used for Avocado were the same as those used for PREDICTD. ChromImpute used leave-one-out validation. In each case, signal values across the entire genome were used either for training or testing. When considering the full genome, we first evaluated the three approaches using three performance measures originally defined by Durham

*et al.* [13]: MSEglobal measures the MSE on the full set of positions; MSE1obs measures the MSE on the top 1% of positions according to ChIP-seq signal; and MSE1imp measures the MSE on the top 1% of positions according to the imputed signal. While Avocado and PREDICTD do equally well according to MSEglobal (unadjusted two-sided paired t-test p-value = 0.451), Avocado outperforms PREDICTD on both MSE1obs (p-value = 9.13e-6) and MSE1imp (p-value = 2.60e-10) (Table 1.1). This observation is consistent with the observation by Durham et al. that PREDICTD may systematically underpredict signal values, allowing it to score well on regions of low signal but achieving lower accuracy on peaks. Conversely, ChromImpute performs the best on MSE1obs (Avocado/ChromImpute p-value = 2.37e-22, PREDICTD/ChromImpute p-value = 2.85e-12) but the worst on MSE1imp, suggesting that it may over-call peaks. Additionally, Ernst and Kellis proposed six other evaluation performance measures, which show similar trends as the MSE1imp metric (Fig. B.2). We then focused our evaluation on regions of particular biological interest by implementing three more performance measures: MSEProm, MSEGene, and MSEEnh, that measure the MSE of the imputed tracks across all promoter regions, gene bodies, and enhancers, respectively (Table 1.1). We found that Avocado outperforms the other two methods at MSEProm (Avocado/ChromImpute p-value = 3.98e-32, Avocado/PREDICTD p-value = 8.73e-05) and MSEEnh (Avocado/ChromImpute p-value = 1.72e-30, Avocado/PREDICTD p-value = 1.50e-04), while yielding similar performance to PREDICTD on MSEGene (p-value = 0.875). Taken together, these performance measures suggest that Avocado is able to impute signal well both across the full genome and also at biologically relevant areas (Table SB.1).

All six of the performance measures listed in Table 1.1 consider each epigenomic assay independently at each genomic position. Empirically, however, many of these assays exhibit predictable pairwise relationships. For example, the activating mark H3K4me3 and the repressive mark H3K27me3 tend not to co-localize within a single promoter region. To

measure how well the imputation methods capture such pairwise relationships, we quantitatively evaluated three specific pairwise relationships: negative correlation between H3K4me3 and H3K27me3 in promoter regions, positive correlation between H3K36me3 and RNA-seq in gene bodies, and lack of correlation between H3K4me1 and H3K27me3 in promoter regions. In addition, we considered two pairwise relationships between assays that occur in a promoter and the corresponding gene body: positive correlation between H3K4me3 in promoters and H3K36me3 in the corresponding gene bodies, and the opposite for H3K27me3 and H3K36me3. For each pair of assays, we evaluated how consistent the imputed tracks are with the empirical relationship between the assays (Appendix C). Across all these evaluations we found that Avocado performed the best at reconstructing the pairwise relationship by between 2.73% and 39.6% when compared to ChromImpute, and between 2.89% and 6.64% when compared to PREDICTD, with PREDICTD typically coming in second and ChromImpute coming in last.

We hypothesized that a primary source of error for all three imputation methods comes from the difficulty in predicting peaks that occur in some cell types but not others. Accordingly, for each assay we segregated genomic positions into those for which a peak never occurs, those in which a peak always occurs ("constitutive peaks"), and those for which a peak occurs in some but not all cell types ("facultative peaks"). Intuitively, we expect an algorithm to be able to predict non-peaks or constitutive peaks more easily than predicting facultative peaks. We test this hypothesis by evaluating the performance of each of the three imputation techniques at genomic positions in chromosome 20 that vary in the number of cell types for which a peak is observed (see Methods). This evaluation consists of calculating the MSE, the recall (proportion of true peaks that are imputed), and the precision (proportion of imputed peaks that are true peaks). The recall and precision are calculated by thresholding either the primary or imputed signal at a value of 1.44, corresponding to a signal p-value of 0.01, and evaluating the recovery of MACS2 peak calls. We can determine if a method over-

or under-calls peaks based on the balance between precision and recall.

We find that evaluating the three imputation approaches in this manner explains the discrepancy we observed between the MSE1obs and MSE1imp performance measures. Specifically, we find that ChromImpute routinely achieves the highest recall (measured indirectly by MSE1obs) and that Avocado typically achieves the highest precision (measured indirectly by MSE1imp) in regions that are the most variable (Fig. 1.2, Fig. B.3). Interestingly, ChromImpute shows a higher recall but lower precision than thresholding the ChIP-seq signal directly in positions that exhibit a peak in many cell types. This observation suggests that ChromImpute may impute wider peaks that, when thresholded, encompass the entirety of the called peak by MACS2. ChromImpute's high recall and low precision confirm the hypothesis that ChromImpute is over-calling peaks, and specifically that it is likely to predict a peak at a position that is a peak in another cell type. These results also indicate that Avocado and PREDICTD capture different trends in the model, as Avocado typically has higher recall in facultative peaks and PREDICTD has higher recall in constitutive peaks. This finding suggests that one could consider ensembling the imputations from these approaches to yield even more accurate measurements. Overall, Avocado obtains a balance between precision and recall that frequently allows it to achieve the best MSE.

### *1.2.3 Avocado's latent representation encodes orthogonal views of the data*

Having demonstrated that Avocado's imputed tracks are of high quality, we next investigated Avocado's learned latent representation. This representation consists of separate embeddings for the cell types, the assays, and the genomic coordinates. Because these embeddings are orthogonal to each other, e.g., the cell type embedding does not depend on a particular assay or set of genomic positions, we anticipated that they would each capture a different aspect of the data.

First, we visualized Avocado's representation of promoters, using annotations from GEN-

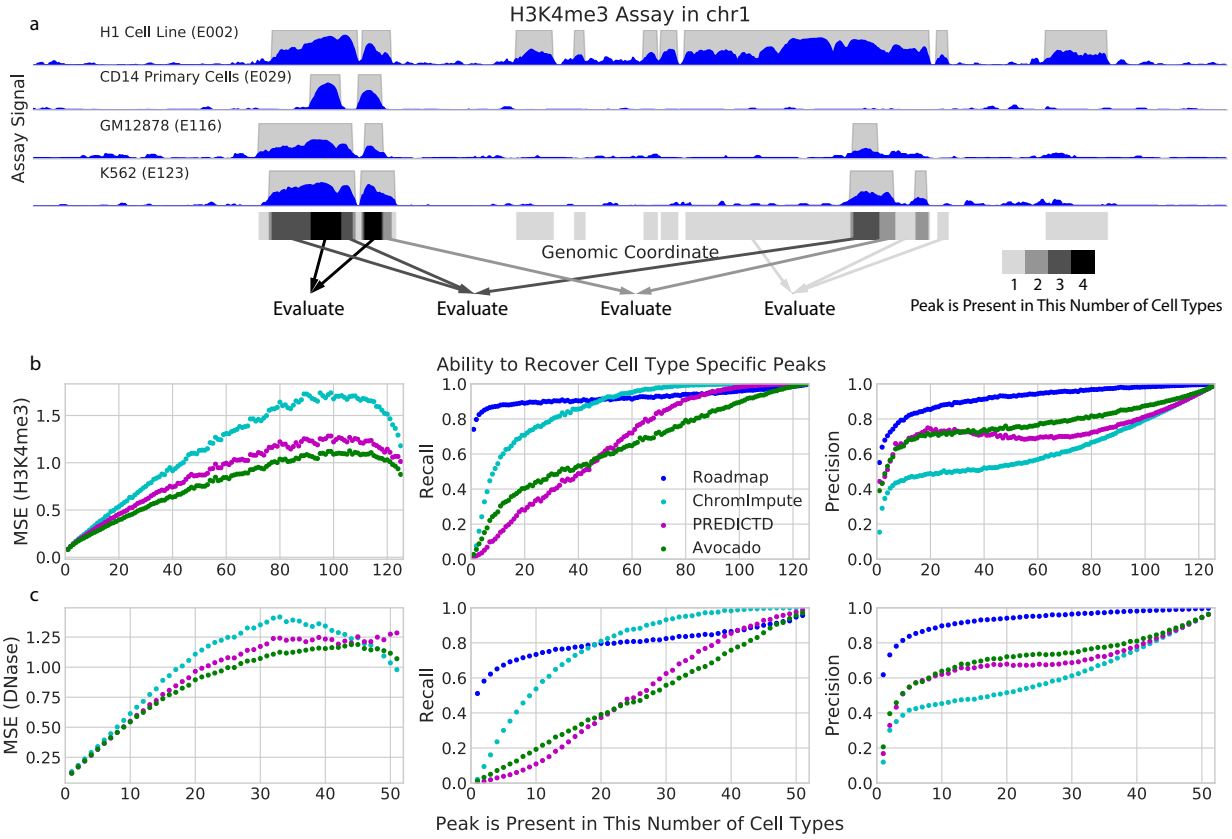



Figure 1.2: **Evaluation of the three imputation approaches at genomic positions that show variation in signal across cell types.** (a) A schematic describing how genomic loci are segregated on an example of four cell types. MACS2 peak calls (in gray) are summed over each of the cell type. Genomic loci are then evaluated separately based on the number of cell types in which a peak occurs. (b) Each panel plots a specified performance measure (y-axis) across varying sets of genomic positions (x-axis) for the H3K4me3 assay. For each point, genomic positions are selected based on the number of cell types in which a peak is called at that position, up to a maximum of 127. MSE is calculated between H3K4me3 ChIP-seq signal and the corresponding imputed signal. Precision and recall are computed by thresholding the imputations at 1.44 and comparing to MACS2 narrow peak calls on the corresponding experimental signal. In the plots, the series labeled "Roadmap" use the experimental Roadmap data likewise thresholded at 1.44 and compared to MACS2 narrow peak calls. (c) Similar to (b), but using DNAse-seq instead of H3K4me3. All analyses are restricted to chromosome 20.

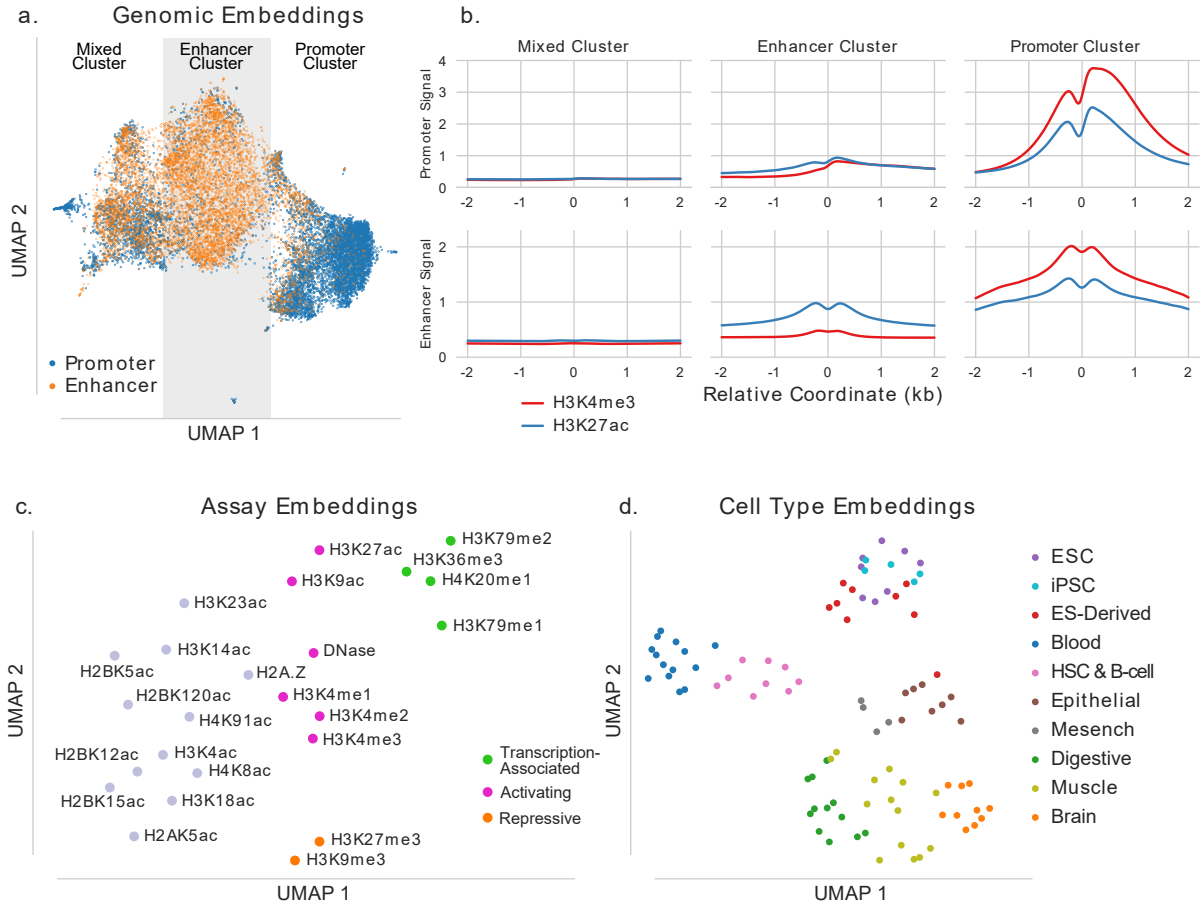



Figure 1.3: **A visualization of Avocado's learned latent representations.** (a) A UMAP projection of the genome embeddings found at promoter (blue) and enhancer (orange) regions. Half of all promoter regions are shown along with an equal number of enhancers, which made up roughly one-fourth of total enhancers. Three manual partitions are shown, one with mainly promoters (85.5%), one with mainly enhancers (85.9%), and one that is mixed (44.9% promoters and 55.1% enhancers). (b) Average epigenomic profiles of H3K4me3 (red) and H3K27ac (blue) in promoters and enhancers in each of the three partitions, with the profiles extended out ±2kb to show additional context. (c) A UMAP projection of the assay embeddings, annotated with their name. Marks are colored to indicate enrichment in transcribed regions (green), association with active expression (pink), or association with repressing expression (orange). Marks that are not well characterized are colored in grey. (d) A UMAP projection of the cell type embeddings, where each cell type has been colored according to its anatomy type.

CODE v19, and enhancers, using the FANTOM5 "robust enhancers" set, by running UMAP [20] on their respective genomic embeddings (Fig. 1.3a). Because each functional element spans several loci, we average the factor values ±250 bp from either the TSS of the gene or the middle of the enhancer. In the figure, we observe three main clusters—one of mostly promoters, one of mostly enhancers, and one that is mixed between the two. Next, we characterized these clusters by their epigenomic signatures. We calculated the average activity of H3K4me3, a mark associated with active promoters, and H3K27ac, a mark associated with active enhancers, in a window ±2 kbp around each locus across all cell types for which experimental data were available. We then averaged these profiles across all enhancers in each cluster and then across all promoters in each cluster (Fig. 1.3b). This ±2 kbp window is larger than the ±250 bp window used for the genomic embedding projection in order to give additional epigenomic context, but we found that projecting genomic embeddings using a ±2 kbp window gave similar results (Fig. B.4). We observe that the promoter cluster consists of loci with high levels of both H3K4me3 and H3K27ac, that the enhancer cluster loci exhibit high levels of H3K27ac but low levels of H3K4me3, and that the mixed cluster has low average levels of both marks. To investigate the loci that compose the mixed cluster more thoroughly, we then clustered the epigenomic signal of these loci across all cell types into "high" signal and "low" signal examples and examined the number of cell types that were deemed high signal (Appendix D). We found that the mixed cluster was made up of some loci that exhibited high signal in a very cell type specific manner and other loci that exhibited low signal across all cell types. These results confirm that the Avocado genomic embeddings are capturing biologically relevant trends across cell types and assays.

Next, we investigated the structure of the assay embeddings. Although histone modifications play diverse roles in regulating gene transcription [21, 22, 23, 24], we found that a UMAP projection of the assay embeddings was able to recapitulate several high-level trends (Fig. 1.3c). For example, the transcription-associated marks H3K36me3, H3K79me2,

H3K79me1 and H4K20me1 are all near one another. Similarly, many marks associated with active gene expression, such as mono-, di-, and tri-methylations of H3K4, are close together. Further, H3K27me3 and H3K9me3, which are both repressive marks, cluster together away from the activating marks. These trends, though admittedly based on projection of a relatively small number of points, suggest that the Avocado latent factors successfully encode some important aspects of histone modification biology.

Lastly, we ran UMAP on the cell type embedding and annotated each cell type with its "anatomy type" as defined in the Roadmap compendium (Fig. 1.3d). We observe a distinct clustering of cell types by anatomy. Furthermore, related cell types such as iPSCs and ESCs lie nearby in the embedded space. Interestingly, despite both residing in bone marrow, hematopoietic stem cells (HSC) lie near blood cells in the projection but mesenchymal stem cells do not. Interestingly, pluipotent stem cells reside on one side of the projection while differentiated cells cluster away from them, suggesting that our embedding may also be capturing some aspects of cellular differentiation. These results are supported by a direct inspection of the Euclidean distances used to make the plot, which show similar clusterings by anatomy type (Fig. B.5).

### *1.2.4 Avocado's latent representation facilitates a variety of prediction tasks*

Having shown that high level trends are captured in Avocado's learned latent representation, we next evaluated its utility as input to machine learning models for tasks for which the representation was not explicitly trained for (Fig. 1.4). This "transfer learning" approach has been used successfully in other domains, such as natural language processing [25] and computer vision [26], and has been more generally described by Pan and Yang [27]. While many techniques can be described as transfer learning, we use the term to refer to training a model for one task and then applying the model (or components thereof) to some other task. Specifically, we hypothesize that Avocado's latent representation can serve as a replacement

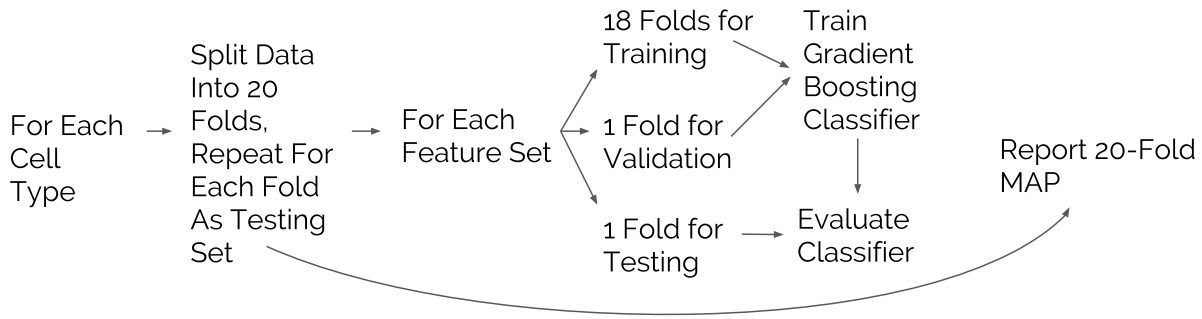



Figure 1.4: **The evaluation procedure for each task.** For each cell type and feature set combination, 20-fold cross validation is performed and the MAP across all 20 folds is returned. At each evaluation, a gradient boosted decision tree classifier is trained on 18 of the folds, convergence is monitored based on performance on a 19th fold, and the performance of the resulting model is evaluated on the 20th fold.

for epigenomic data as the input for machine learning models across a variety of genomic prediction tasks. One reason that transfer learning may be beneficial in this case is that many genomic phenomena are associated with epigenomic signals, and so a representation trained to predict these signals is also likely to be associated with these phenomena.

We then investigated whether Avocado's latent representation has implicitly encoded four different types of important biological activity: gene expression, promoter-enhancer interactions, replication timing, and frequently interacting regions (FIREs). These tasks span a diversity of biological phenomena and data sources to ensure that our findings are robust. For each task we train a supervised machine learning model (see Methods) using one of seven feature sets: (1) all available ChIP-seq assays for the cell types being considered, (2–4) the set of 24 assays imputed by each of the three methods, (5) the genomic position factors from the single model of PREDICTD's ensemble that is highlighted in Figure 3 of Durham *et al.* [13], (6) the genomic position factors in Avocado's latent representation, or (7) the full set

of 1,014 ChIP-seq and DNase-seq assays available in the Roadmap compendium (Fig. 1.4). We include the full set of assays from the Roadmap compendium as a baseline feature set because the Avocado latent representation is learned from this full set, allowing us to test our hypothesis that the learned representation preserves cellular variation while removing redundancy and technical noise. Additionally, we include PREDICTD's learned latent representation to investigate its utility relative to the Avocado latent representation. Lastly, we compare these models to a majority baseline where our prediction for each sample is simply the most prevalent label. We hypothesize that, should the latent representation encode these phenomena well, that the models trained using the latent representation as input will outperform those trained using the other feature sets. Note that the Avocado latent representation is extracted from a model that is trained on the full Roadmap data set. For the prediction of gene expression, replication timing, and FIREs, we use a gradient boosting classifier due to this technique's widespread success in machine learning competitions [28, 29], with a partial list of top performance on Kaggle competitions available at https://bit.ly/2k7W3Jh

Because our goal is not to produce a state-of-the-art classifier for each task but simply to demonstrate the broad utility of Avocado's latent representation, we do not fine-tune the hyperparameters for the gradient boosting classifier, and we extract feature sets from assays or latent factors by taking the average epigenomic signal value or latent factor value in the region/s of interest, rather than considering more complicated representations.

*Gene expression*

The composition of histone modifications present in the promoter region of a gene can be predictive of whether that gene is expressed as measured by RNA-seq or CAGE assays. Accordingly, several prior studies have shown that machine learning models can learn associations between these histone marks and gene expression. Because RNA-seq experiments are cheap enough to be performed in any cell type of interest, the typical goal of building a

machine learning model is not to replace RNA-seq but to better understand the mechanism behind gene expression. While it may be difficult to explain this mechanism through the interaction of complex latent factors, performing well at this task indicates that complex regulatory information comprised of multiple epigenomic marks is being encoded in the latent factors. Furthermore, a gene expression predictor may be useful in hypothesis generation settings, to assist in prioritizing potential RNA-seq experiments or in investigation of the expression behavior of a small number of genes across many cell types for which epigenomic data has been generated. These studies have approached the problem either as a classification task, where the goal is to predict a thresholded RNA-seq or CAGE-seq signal [30, 31], or a regression task, where the goal is to predict RNA-seq or CAGE-seq signal directly [8].

We approach the prediction of gene expression as a classification task and evaluate the ability of the different feature sets derived from the promoter region of a gene to predict whether or not that gene is expressed. This evaluation is carried out in a 20-fold cross validation setting in each cell type individually, and we report the mean average precision (MAP), which is one technique for calculating the area under a precision-recall curve, across all 20 folds. Genes are considered to be active in a cell if the average normalized read-count value from an RNA-seq experiment across the gene body is greater than 0.5 (see Methods).

We find that the Avocado latent factors yield the best models in 34 of 47 cell types (Fig. 1.5a, Fig. B.6, Table B.2). In 11 of the 13 remaining cell types (out of the 47 in total), models trained using the Avocado latent factors are only beaten by those trained using the full Roadmap compendium, and in two cell types (E053 and E054; Cortex derived and ganglionic eminence derived neurosphere cultured cells) Avocado is also beaten by models trained using ChromImpute's imputed epigenomic marks. In no cell type do models trained using the primary data, the typical input for this prediction task, outperform those trained using the Avocado latent representation (unadjusted two-sided paired t-test p-value of 4.62e-153), performing worse by between 0.005 and 0.148 MAP. Additionally, models built

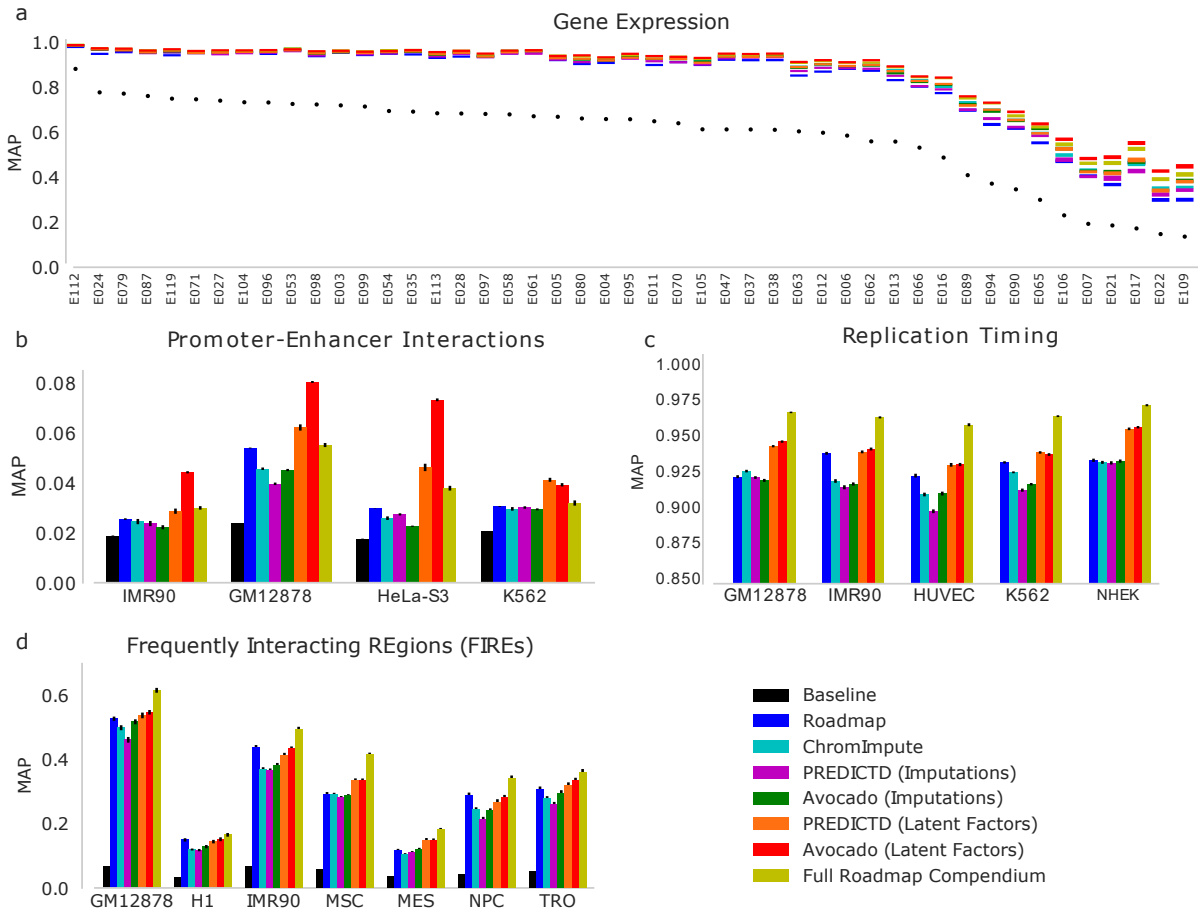



Figure 1.5: **The performance of each feature set when used to for genomic prediction tasks.** In each task, a supervised machine learning model is evaluated separately for each cell type using a 20-fold cross-validation strategy, with the mean average precision reported and standard error of the mean shown in the error bars. Each task considers only genomic loci in chromosomes 1 through 22. The tasks are predicting (a) expressed genes, (b) promoter-enhancer interactions, (c) replication timing, and (d) FIREs. In panel (a) the coloring corresponds to the standard error with the mean average precision lying in the middle, whereas in the other panels the mean average precision is shown as the colored bar with standard error shown in black error bars. The statistical significances of differences observed in this figure are assessed in Tables B.2-5.

using Avocado's latent representation outperform those built using PREDICTD's latent representation in every cell type, ranging from an improvement of 0.002 to an improvement of 0.087 (p-value of 3.86e-101). Overall, the models built using the Avocado latent factors perform 0.006 MAP better than those built using the full Roadmap compendium (p-value of 9.75e-21) and only perform 0.001 MAP worse, on average, in those 13 cell types where they perform worse. While this improvement initially appears to be minor, we note that all feature sets yield models that perform extremely well in most cell types, suggesting that there are cell types where gene expression prediction is simple and those in which it is difficult. Accordingly, when focusing on cell types where prediction is more difficult we notice that the difference in performance between the feature sets is more pronounced. Indeed, when we consider the seven cell types where the majority baseline is lowest, we find that those models trained using the Avocado latent factors outperform those trained using the full Roadmap compendium on average by 0.026 MAP and those built using only Roadmap measurements for a specific cell type by 0.107 MAP. These results show that models built using the Avocado latent representation outperform or are comparable to any other feature set considered.

To confirm these results, we reformulated the problem to be a regression task by removing the threshold on the RNA-seq values used to generate binary labels. We observed similar trends as the classification task, with the Avocado latent factors yielding the best model in 27 of the 47 cell types, the full Roadmap compendium yielding the best model in 19 of the 47 cell types, and ChromImpute yielding the best model in one cell type (Fig. B.7). In each cell type, the Avocado latent factors outperformed using the cell type-specific epigenomic data alone.

*Promoter-enhancer interactions*

One of the many ways that gene expression is regulated in human cell lines is through the potentially long-range interactions of promoters with enhancer elements. Physical promoter-enhancer interactions (PEIs) can be experimentally identified by 3C-based methods such as Hi-C or ChIA-PET. However, the resolution of genome-wide 3C methods can be problematic because high resolution contact maps are expensive to acquire. Consequently, predicting PEIs from more widely available and less expensive data types would be immensely valuable. Accordingly, a wide variety of methods for predicting PEIs have been proposed (reviewed by Mora et al. [32]), including those that pair enhancers with promoters using distance along the genome [33], that use correlations between epigenetic signals in the promoter and enhancer regions [34, 35, 36], and that use machine learning approaches based on epigenetic features extracted from both the promoter and enhancer regions [16].

We consider the task of predicting physical PEIs as a supervised machine learning problem using features derived from both the promoter and enhancer regions. We employ a set of PEIs that were originally created for training TargetFinder [16], a machine learning model that predicted whether given promoter-enhancer pairs interact with each other using epigenomic measurements derived from both regions. These PEIs correspond to ChIA-PET interactions from each of four cell types (HeLa-S3, IMR90, K562, and GM12878) in chromosomes 1 through 22. We further process this data set to remove a source of bias that has been found since the publication of the original data set [37] (Appendix E). TargetFinder was not developed to predict interactions in cell types for which contact maps have not been collected, but rather to better understand the connections within existing contact maps. Likewise, we train our classifier to predict PEIs within each cell type, evaluating a regularized logistic regression model in a cross validation setting. For comparison, we use the same collection of real and imputed data types that we used for the gene expression prediction task.

We find that models trained to predict PEIs using the Avocado latent factors perform

better than any other feature set that we considered (Fig. 1.5b) in IMR-90, GM12878, and HeLa-S3. In K562 using the Avocado latent factors is second only to using the PREDICTD latent factors. These improvements in average precision over the full Roadmap compendium range from 0.007 in K562 to 0.035 in HeLa-S3 (p-values ranging from $6.97 \times 10^{-18}$ to $9.45 \times 10^{-32}$, Table B.3). Interestingly, the PREDICTD latent representation also outperforms the full Roadmap compendium in every cell type (p-value of $2.43 \times 10^{-22}$).

*Replication timing*

The human genome replicates in an orderly replication timing program, in a process that is associated with gene expression and closely linked to the three dimensional structure of the genome [38, 39]. Patterns of replication timing along the genome can be quantified using experimental assays such as Repli-Seq [40], which can be used to segregate loci into early- and late-replicating regions. Because of the slowly varying nature of replication timing along the genome, we choose to make predictions of early- and late-stage replication at 40 kbp resolution.

Consistent with previous tasks, the Avocado latent representation outperforms both primary and imputed epigenomic data from the cell type of interest (Fig. 1.5c). However, in contrast to the previous tasks, the Avocado and PREDICTD latent representations perform similarly to each other. While the Avocado latent representation yields models whose improvement over the PREDICTD latent representation is statistically significant (p-value of 0.004, Table B.4), the effect is small (average precision of 0.9453 vs 0.9442). Further, models that use the full Roadmap compendum yield the best performing models. Taken together, these results suggest that using epigenomic measurements across several cell types can be informative for making predictions even for a single cell type. Additionally, it appears that aggregating these latent spaces to a much coarser resolution (from 25 bp to 40 kbp) may sacrifice valuable information. Potentially, this information loss happens because the latent

space itself is not linearly interpolatable, and so for large spans taking the average factor value is not the optimal way to aggregate the factors.

*Frequently interacting regions*

The three-dimensional structure of the genome can be characterized by experimental techniques that identify contacts between pairs of loci in the genome in a high-throughput manner. In particular, the Hi-C assay [41] produces a contact map that encodes the strength of interactions between all pairs of loci in the genome. Within a typical contact map, blocks of increased pairwise contacts called "topologically associating domains" (TADs) segment the genome into large functional units, where the boundaries are enriched for house-keeping genes and certain epigenetic marks such as the CTCF transcription factor [42]. Recently, a related phenomenon, called "frequently interacting regions" (FIREs), has been identified [17]. These regions are enriched for contacts with nearby loci after computationally accounting for many known forms of bias in experimental contact maps. FIREs are typically found within TADs and are hypothesized to be enriched in super-enhancers [17].

Accordingly, we investigate the utility of the Avocado latent representation in predicting FIREs. Our gold standard is derived from Hi-C measurements in seven human cell types at 40 kbp resolution [17]. We frame each task as a binary prediction task, classifying each genomic locus as a FIRE or not. Note that any state-of-the-art predictive model for elements of chromatin architecture would likely include CTCF data, because this mark is highly enriched at structural elements. However, we do not include this factor in our feature set because transcription factors were not included in the Roadmap compendium and thus not used to train the Avocado model. Further, our goal is not to train a state-of-the-art model for predicting FIREs, but to evaluate the relative usefulness of these feature sets.

The results for predicting FIREs are similar to the results from the replication timing task, with models trained using the Avocado latent factors outperforming both those trained

using cell type specific epigenomic data (p-value of $6.13 \times 10^{-8}$) and the PREDICTD latent factors (p-value of $2.4 \times 10^{-4}$) (Fig. 1.5d and Table B.5). The models trained using the full Roadmap compendium outperform those that use the Avocado latent factors in every cell type except H1 (p-value of $1.85 \times 10^{-33}$). This observation suggests that the inclusion of epigenomic measurements across cell types is important when predicting elements of chromatin architecture, as it was for replication timing, but further suggests that aggregations of these factor values across large genomic loci is not as informative as it was for predicting gene expression or promoter-enhancer interactions.

### *1.2.5 Avocado's genomic representation encodes most peaks*

We next aim to understand why the Avocado latent representation is such an informative feature set across a diversity of tasks. A well-known downside of neural networks is that they are not as easily interpretable as simpler models due to the larger number of parameters and non-linearities involved in the model. In order to understand these models better, feature attribution methods have recently emerged as a means to understand predictions from complex predictive models. These methods, such as LIME [43], DeepLIFT [44], SHAP [45], and integrated gradients [46], attempt to quantify how important each feature is to a specific prediction by attributing to it a portion of the prediction. A useful property of these attributions is that they sum to the resulting prediction, or the difference between the prediction and some reference value.

We chose to inspect the Avocado model using the integrated gradients method, due to its simplicity, in order to understand the role that the various factors play in making predictions. When we run integrated gradients, the input is the set of concatenated latent factors that would be used to make a prediction at a specific position, and the output is the attribution to each factor for that prediction, specifically, the imputed signal at a genomic position for an assay in a cell type. However, the individual factors are unlikely to correspond directly

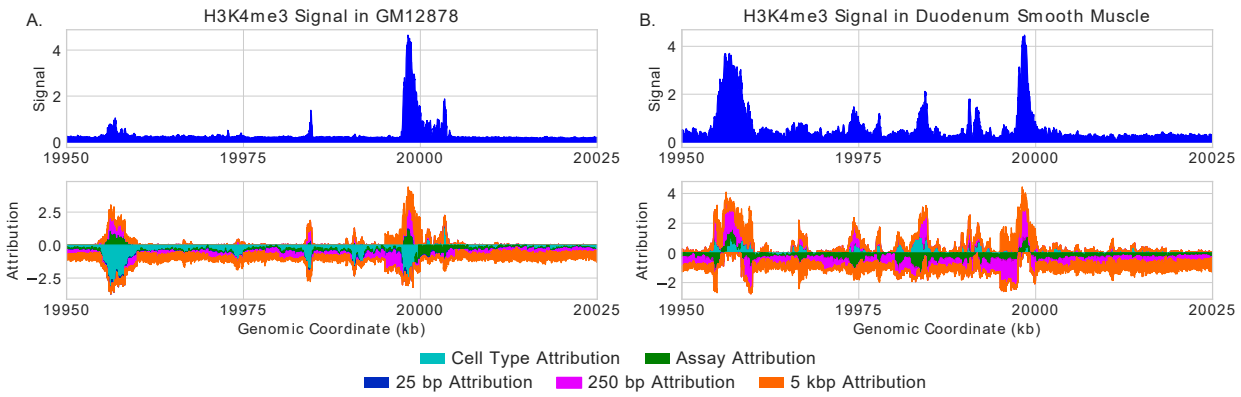



Figure 1.6: **The predicted H3K4me3 signal and corresponding attributions for two cell types in the same region of chromosome 20.** (a) The prediction and attributions for GM12878, where a tall peak on the right is paired with two much smaller peaks to the left. Many short regions have a positive genomic attribution but a negative cell type attribution that masks them. (b) The prediction and attributions for duodenum smooth muscle. A prominent peak is now predicted on the left, corresponding with a swap from a negative cell type attribution to a positive one. The same short regions that previously were masked by the cell type attributions now have positive cell type attributions and exhibit peaks in the imputed signal.

to a distinct biological phenomena. Conveniently, since the attributions sum to the final prediction (minus a reference value), we can sum these attributions over all factors belonging to each component of the model. This aggregation allows us to divide the imputed signals into the cell type, assay, and the three scales of genome attributions.

Upon inspection of many genomic loci, most peaks are encoded in the genomic latent factors, while the cell type and assay factors serve primarily to sharpen or silence peaks. An illustrative example of the role each component plays is to consider a pair of nearby regions in chromosome 20 where a H3K4me3 peak with high signal is imputed near a much weaker peak

for GM12878 with a very narrow spike between them (Fig. 1.6a) Within the imputed peaks the genome factors predominantly increase the signal, whereas the assay factors appear to increase the signal at the cores of both peaks but dampen the signal on the flanks, effectively sharpening the peaks. Interestingly, the weaker peak appears to have a more prominent signal from the genomic latent factors that is mitigated by a large negative signal from the cell type axis. This indicates to us that this region exhibits a peak in some cell types but is being silenced in GM12878. To confirm that this region engages in a peak in some cell types we looked at the same region in duodenum smooth muscle cells (E078, Fig. 1.6b) and observed a strong peak (maximum value 3.70 compared to 1.05 in GM12878) that is bolstered by the cell type factors. In addition, there are many smaller peaks that exist in the duodenum signal that are masked by a negative cell type attribution. This suggests that, while the cell type and assay factors can have positive attributions, they do not fully encode peaks themselves.

We next systematically evaluate the attributions of each component of the model to better understand how Avocado works. Our approach for this analysis is similar to that of analyzing the accuracy of the imputation methods at regions of cellular variability. Specifically, we first segregate positions into bins based on the number of cell types that exhibit a peak at that location, then for each bin we calculate the average attribution in those cell types for which a peak does or does not occur (Fig. B.8). In this manner we can analyze each of the five components of the model in each assay. What we find is that, when peaks are not present in the signal, the average cell type attributions are uniformly negative across assays and the variability of signal at a position. Additionally, these average attributions typically have a larger magnitude at those variable loci in cell types for which a peak is not present, suggesting that the cell type factors are involved in silencing these peaks in the resulting imputations. The only context in which average cell type attributions are positive are when peaks are present at loci that infrequently exhibit peaks suggesting that the cell type factors

may encode infrequent peaks. In contrast, the genomic factors typically have positive values when peaks are present, with negative values correspondingly occuring in infrequent peaks and when peaks are not present. If these rare peaks are a result of technical noise rather than real biology, then this suggests one reason that the genomic factors frequently yield better machine learning models than experimental data. However, this also suggests that the genomic factors may not be useful at identifying biological phenomena that are indicated by these rare peaks. Interestingly, while the assay attribution values can either be positive or negative, these attributions are higher when peaks are not exhibited rather than when they are. It is unclear why this phenomenon occurs, but it further indicates that the genomic components of the model are a critical driver of Avocado predicting a peak.

### *1.3 Discussion*

Avocado is a multi-scale deep tensor factorization model that learns a latent representation of the human epigenome. We find that, when used as input to machine learning models, Avocado's latent representation improves performance across a variety of genomics tasks relative to models trained using either experimentally collected epigenomic measurements or the full set of imputed measurements. This representation is more informative than the one learned through the linear factorization approach taken by PREDICTD, suggesting that latent representations can vary in utility and that more work will need to be done to understand them fully. Additionally, in the context of replication timing and FIRE prediction, we found that aggregating both the PREDICTD and the Avocado latent spaces to much lower resolutions by averaging factor values appeared to diminish their utility, suggesting that perhaps these latent spaces are not linearly interpolatable. We have made the Avocado latent representation available for download from https://noble.gs.washington.edu/proj/avocado/model/.

We hypothesized that a primary reason that this latent representation is so informative is that it distills epigenomic data from all available cell types, rather than representing

measurements for only a single cell type. Indeed, feature attribution methods suggest that the genomic latent factors encode information about peaks from all cell types and assays. However, while verifying this hypothesis, we also found that, contrary to common usage, models that exploit the full Roadmap compendium consistently outperform those that use only measurements available in a single cell type. One explanation for this observation is that cellular context can serve as an implicit regularizer for machine learning models, in the sense that the model can learn to discount peaks that appear in exactly one cell type due to experimental noise or technical error. On the other hand, when the discounted peaks correspond to real biology that is simply very cell type-specific, this tendency may be a source of error.

Although the Avocado latent representation does not outperform using the Roadmap compendium on all tasks, Avocado is much more practical to use. Avocado's representation consists of only 110 features, whereas the full Roadmap compendium has 1,014 experiments. Accordingly, we observed that models could be trained from Avocado's learned genomic representation five to ten times faster than those trained using the full Roadmap compendium. This speedup becomes especially important when the input to a machine learning model is not a single genomic window, but multiple adjacent windows of measurements, as is frequently the case when modeling gene expression. For example, if one were to describe a promoter as eight adjacent 250-bp windows spanning $\pm 2$ kbp from a transcription start site, then the Avocado representation would have only 565 features due to its multi-scale nature, whereas the Roadmap compendium would comprise 8,112 features. We anticipate that the benefits of a low dimensional representation will become even more important once this strategy is applied to even richer data sets, such as the ENCODE compendium, which is composed of $>$10,000 measurements. This number of measurements would make building machine learning models very difficult.

A natural desire is to inspect the Avocado latent representation in order to better under-

stand the genome. Unfortunately, we found that such inspection was difficult, in part because the latent factors do not individually correspond to meaningful biological phenomena. An avenue for future studies is to better understand these latent factors through methods that aim to connect learned latent spaces to interpretable concepts [47]. Potentially, one might apply a semi-automated genome annotation method like ChromHMM [1] or Segway [48] to the latent representation directly, with the goal of producing a model that can translate the latent representation into a cell type-independent annotation of the genome.

This is not the first time that latent representations have been trained on one task with the goal of being broadly useful for other tasks. For example, word embeddings have been used extensively in the domain of natural language processing. These embeddings can be calculated in a variety of manners, but two popular approaches, GLoVE [49] and word2vec [15], involve learning word representations jointly with a machine learning model that is trained to model natural language. In this respect, these embedding approaches are similar to ours because the Avocado latent representation is learned as a result of a machine learning model being trained to impute epigenomic experiments.

Our approach is not the only approach one could take to reducing the dimensionality of the data. Potentially, one could use a technique like principal component analysis or an autoencoder to project the 1,014 measurements down to 110 dimensions. Alternatively, one might consider using a model similar to DeepSEA [50] or Basset [51] that trains an embedding of the genome jointly with a neural network. However, these types of approaches would not easily allow for transfer learning between cell types, would not allow for the imputation of epigenomic experiments, and would not incorporate information about local genomic context through the use of multiple scales of genomic factors. Furthermore, generalizing an unsupervised embedding approach to make cross-cell type predictions would be difficult, whereas Avocado's genomic and cell type factors can be combined in a straightforward way to address such tasks.

In this work, we have only explored the Avocado hyperparameter space with respect to the single dataset employed here; thus, generalizing to a new dataset will require repeating this search. Furthermore, in cases where computational efficiency is critical, our results (Fig. A.3) suggest that models with fewer latent factors might perform nearly as well as the full Avocado model. In such settings, it may be sensible to design an objective function for the hyperparameter search that trades off the predictive accuracy of the model versus the model complexity.

We have emphasized the utility of Avocado's latent genome representation, but the model also solves the primary task on which it is trained—epigenomic imputation—extremely well. In particular, we found that Avocado produced the best imputations when compared with ChromImpute and PREDICTD as measured by five of six performance measures based on MSE for individual tracks, and that these imputed measurements captured pairwise relationships between histone modifications better than either of the other approaches. While investigating why Avocado performed worse than ChromImpute on one of the performance measures, we found that, for all three imputation approaches, much of the empirical error derives from regions where peaks are exhibited in some, but not all, cell types. In the context of identifying which cell types exhibit peaks at these regions of high variability, ChromImpute had the highest recall but the lowest precision, suggesting that it over-calls peaks at a specific region by predicting peaks in more cell types than they actually occur in. In contrast, both Avocado and PREDICTD had lower recall but higher precision, with Avocado frequently managing to balance the two to produce the lowest MSE. Given that these regions are likely the most important for explaining cell type variability, these results suggest that future evaluations of imputation methods should stratify results, as we have done, according to the cell-type specificity of the observed signals. Such investigations might suggest different Avocado hyperparameter settings, focusing on either improved precision or recall, depending upon the end user's needs.

Finally, we anticipate that researchers may wish to extend the imputation abilities of Avocado to a new cell type or assay using their own experimental data but lack the computational resources to retrain Avocado from scratch. In follow-up work, we describe a simple transfer learning approach for adding in new cell types or assays to a pretrained Avocado model [52]. This approach involves freezing the parameters of a pretrained model and fitting only the new cell type or assay factors. Our analysis suggests that one can achieve good quality imputations with as little as a single track of training data in a given biosample. Further, because very few parameters need to be trained, this process can be done without relying on a GPU.

## *1.4 Conclusion*

Avocado employs a multi-scale deep tensor factorization approach to compress large compendia of epigenomics experiments into a low dimensional latent representation. This latent representation is trained to impute genome-wide epigenomics experiments, and we find that the resulting model outperforms prior methods at that task. We further demonstrate that the resulting latent representation captures important aspects of the three orthogonal axes of the data—the cell types, the assays, and the genomic loci. Accordingly, when we use the genomic latent factors directly as input into machine learning models, we find that they yield models that are much more accurate than the traditional setting of using cell type-specific epigenomic data across a variety of predictive tasks. We anticipate that this model and its associated latent factors will serve as valuable tools for researchers studying human epigenomics.

## 1.5 Methods

### 1.5.1 Datasets

The Roadmap ChIP-seq and DNase-seq epigenomic data was downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/pval/. Only cell types that had at least five experiments done, and assays that had been run in at least five cell types, were used. These criteria resulted in 1,014 histone modification ChIP-seq tracks spanning 127 cell types and 24 assays. The assays included 23 histone modifications and DNase sensitivity. RNA-seq bigwigs containing unstranded normalized read counts across the entire genome for 47 cell types were also downloaded for the purpose of downstream analyses, rather than for inclusion in the imputation task. The full set of 24 assays imputed by ChromImpute were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidatedImputed/, and the full set of 24 assays from PREDICTD were downloaded from the ENCODE portal at https://www.encodeproject.org/.

The specific ChIP-seq measurements downloaded were the $-\log_{10}$ p-values. These measurements correspond to the statistical significance of an enrichment at each genomic position, with a low signal value meaning that there is unlikely to be a meaningful enrichment at that position. Tracks that encode statistical significance, such as the $-\log_{10}$ p-value of the signal compared to a control track, typically have a higher signal-to-noise ratio than using fold enrichment. Furthermore, to reduce the effect of outliers, we use the arcsinh-transformed signal

$$\sinh^{-1} x = \ln\left(x + \sqrt{1 + x^2}\right)$$

for both training of the Avocado model and all evaluations presented here. Other models, such as PREDICTD [13] and Segway [48], also use this transformation, because it sharpens the effect of the shape of the signal while diminishing the effect of large values.

Gene bodies were defined as GENCODE v19 gene elements (https://www.gencodegenes.

org/releases/19.html) from chromosomes 1 through 22 that had one of their transcripts annotated as the primary transcript for that gene. This resulted in 16,724 gene bodies.

Promoter regions were defined at the transcription start site for each of the GENCODE v19 gene elements that gene bodies were identified for, accounting for the strand of the gene. For the purpose of the MSEProm metric and for the gene expression prediction task, the span of the promoter was defined as 2 kbp upstream from the transcription start site. For the purpose of the visualization of promoters and enhancers, promoters were defined as ±250 bp from the transcription start site.

Enhancer elements were defined using two sets of enhancers defined by the FANTOM5 consortium. For the purpose of the MSEEnh metric, the set of "permissive" enhancers was used, in order to get a wider view of potential enhancer activity. For the purpose of visualization of promoters and enhancers, enhancers were defined using ±250 bp from the middle of each enhancer in the "robust" enhancer set. Both enhancer sets are available at http://slidebase.binf.ku.dk/human_enhancers/presets.

Anatomy types for each cell type were downloaded from https://docs.google.com/spreadsheets/d/1yikGx4MsO9Ei36b64yOy9Vb6oPC5IBGlFbYEt-N6gOM/edit#gid=15.

Promoter-enhancer interactions were obtained from the public GitHub repository for [16], available at https://github.com/shwhalen/targetfinder/tree/master/paper/targetfinder/combined/output-epw. This data set includes promoter-enhancer interactions as defined by ChIA-PET interactions for four cell lines—GM12878, HeLa-S3, IMR90, and K562. To correct a recently identified bias in this particular benchmark [37], the data set was further processed as described in Appendix E.

Replication timing data was downloaded from http://www.replicationdomain.org. The resulting tracks encode early- and late-stage timing as continuous values, which are subsequently binarized using a threshold of 0.

FIRE scores were obtained from the supplementary material of [17] for the seven cell lines

TRO, H1, NPC, GM12878, MES, IMR90, and MSC. These measurements are composed of binary indicators at 40 kbp resolution, resulting in 72,036 loci for each cell type.

### *1.5.2 Network topology*

Avocado is a deep tensor factorization model, i.e., a tensor factorization model that uses a neural network instead of a scalar product to combine factors into a prediction. The tensor factorization component is comprised of five matrices of latent factors, also known as embedding matrices, that encode the cell type, assay, 25 bp genome, 250 bp genome, and 5 kbp genome factors. These matrices represent each element as a set of latent factors, with 32 factors per cell type, 256 factors per assay, 25 factors per 25-bp genomic position, 40 factors per 250-bp genomic position, and 45 factors per 5-kbp genomic position. For a specific prediction, the factors corresponding to the respective cell type, assay, and genomic position, are concatenated together and fed into a simple feed-forward neural network. This network has two intermediate dense hidden layers that each have 2,048 neurons before the regression output, for a total of three weight matrices to be learned. The network uses the ReLU activation function, $\mathrm{ReLU}(x) = \max(0, x)$, on the hidden layers and no activation function on the prediction. The training process jointly optimizes the latent factors in the tensor factorization model and the neural network, rather than switching between optimizing each.

The model was implemented using Keras [53] with the Theano backend [54], and experiments were run using Tesla K40c and GTX 1080 GPUs. For further background on neural network models, we recommend the comprehensive review by J. Schmidhuber [55].

### *1.5.3 Inputs and outputs*

Avocado takes as input the indices corresponding to a genomic position, assay, and cell type, and outputs an imputed data value. The indices for each dimension are a set of sequential

values that uniquely represent each of the possibilities for that dimension, e.g., a specific cell type, assay, or genomic position. Any data value in the Roadmap compendium can thus be uniquely represented by a triplet of indices, specifying the cell type, index, and assay.

### *1.5.4 Training*

Avocado is trained using standard neural network optimization techniques. The model was fit using the ADAM optimizer due to its widespread adoption and success across several fields [56]. Avocado's loss function is the global mean-squared error (MSE). Most training hyperparameters are set to their default values in the Keras toolkit. For the ADAM optimizer, this corresponds to an initial learning rate of 0.01, beta1 of 0.9, beta2 of 0.999, epsilon of $10^{-8}$, and a decay factor of $1 - 10^{-8}$. The embedding matrices are initialized with random uniform weights in the range $[-0.5, 0.5]$. Dense layers are initialized using the "glorot uniform" setting [57]. Using these settings, our experiments show that performance, as measured by MSE, was similar across different model initializations.

Avocado does not fit a single model to the full genome because the genome latent factors could not fit in memory. Instead, training is performed in two steps. First, the model is trained on the selected training tracks but with the genomic positions restricted to those in the ENCODE Pilot Regions [58]. Second, the weights of the cell type factors, assay factors, and neural network parameters are frozen, and the genome factors are trained for each chromosome individually. This training strategy allows the model to fit in memory while also ensuring consistent parameters for the non-genomic aspects of the model across chromosomes, and for the latent factors learned on the genomic axis to be comparable across cell types. Both of the stages involve the same set of training experiments. During cross-validation this procedure is repeated separatedly for each fold. We did not find that this procedure was sensitive to using other equally sized regions for the initial training step (Appendix F).

The two steps of training have the same initial hyperparameters for the ADAM optimizer but are run for different numbers of epochs. Each epoch corresponds to a single pass through the genomic axis such that each 25 bp position is seen exactly once, with cell type and assays chosen randomly for each position. This definition of "epoch" ensures that the entire genome is seen the same number of times during training. Training is carried out for 800 epochs on the ENCODE Pilot regions and 200 epochs on each chromosome. No early-stopping criterion is set, because models converge in terms of validation set performance for all chromosomes in fewer than 200 epochs but do not show evidence of over-fitting if given extra time to train.

### *1.5.5 Evaluation of variable genomic loci*

For each assay, we evaluated the performance of Avocado, PREDICTD, and ChromImpute, at genomic positions segregated by the number of cell types in which that genomic locus was called a peak by MACS2. We first calculated the number of cell types that each genomic locus was called a peak by summing together MACS2 narrow peak calls across chromosome 20 and discarded those positions that were never a peak. This resulted in a vector where each genomic locus was represented by the number of cell types in which it was a peak, ranging between 1 and the number of cell types in which that assay was performed. For each value in that range, we calculated the MSE, the recall, and the precision, for each technique. Because precision and recall require binarized inputs, the predictions for each approach were binarized using a threshold on the -log10 p-value of 2, corresponding to the same threshold that Ernst and Kellis used to binarize signals as input for ChromHMM.

### *1.5.6 Supervised machine learning model training*

We performed three tasks that involved training a gradient boosted decision tree model to predict some genomic phenomenon across cell types. In each task, we used a 20-fold cross validation procedure, where the data from a single cell type is split into 20 folds, 19 are

used for training and 1 is used for model evaluation. This procedure was performed for each cell type, feature set, and task. These models were trained using XGBoost [59] with a maximum of 5000 estimators, a maximum depth of 6, and an early stopping criterion that stopped training if performance on a held out validation set, one of the 19 folds used for training, did not improve after 20 epochs. No other regularization was used, and the remaining hyperparameters were kept at their default values.

For the task of predicting promoter-enhancer interactions, we used logistic regression as an additional safeguard against the bias issue described in Appendix E. Rather than perform 20-fold cross-vallidation, we performed 5-fold cross-validation 20 times, shuffling the data set after each cross-validation. We adopted this approach due to the small number of positive samples in each cell type, such that there would be fewer than 10 positive samples in each fold of a 20-fold cross-validation. Additionally, we tuned the regularization strength in the default manner for scikit-learn, which considers 10 regularization strengths evenly spaced logarithmically between $10^{-4}$ and $10^{4}$ and choosing the strength that performs best on an internal 3-fold cross-validation on the training set.

We evaluate each model in each task according to the average precision (AP) on the test set, which summarizes a precision-recall curve in a single score. The score is calculated as

$$AP = \sum_{n} \left(\text{Recall}_n - \text{Recall}_{n-1}\right) \text{Precision}_n$$

where $\text{Recall}_n$ and $\text{Precision}_n$ are the recall and the precision at the n-th calculated threshold, with one threshold for each data point.

# Chapter 2

# AVOCADO CAN COMPLETE THE ENCODE3 COMPENDIUM

## *2.1 Background*

Recently, several scientific consortia have generated large sets of genomic, transcriptomic and epigenomic data. For example, since its inception in 2003, the NIH ENCODE Consortium [60] has generated over 10,000 human transcriptomic and epigenomic experiments. Similar efforts include Roadmap Epigenomics [61] (which we examined in Chapter 1), modENCODE [62], the International Human Epigenome Consortium [63], mouseENCODE [64], PsychENCODE [65], and GTEx [66]. These projects have varied motivations, but all spring from the common belief that the generation of massive and diverse high-throughput sequencing datasets can yield valuable insights into molecular biology and disease.

Unfortunately, the resulting datasets are usually incomplete. In the case of ENCODE, this incompleteness is by design. Faced with a huge range of potential cell lines and primary cell types to study (referred to hereafter using the ENCODE terminology "biosample"), ENCODE investigators made the strategic decision to perform "tiered" analyses. Thus, some "Tier 1" biosamples were analyzed using a large number of different types of sequencing assays, whereas biosamples assigned to lower tiers were analyzed in less depth. This strategy allowed ENCODE to cover many biosamples while also allowing researchers to examine a few biosamples in great detail. In other cases, even for a consortium such as GTEx, which aims to systematically characterize a common set of tissue types across a set of individuals using a fixed set of assays, missing data is unavoidable due to the cost of sequencing and loss of samples during processing. Given the vast space of potential biosamples to study and the fact that new types of assays are always being developed to characterize new phenomena,

the sparsity of these compendia is likely to increase over time.

This incompleteness can be problematic. For example, many large-scale analysis methods have trouble handling missing data. Despite the benefit that additional measurements may offer, many analysis methods discard assays that have not systematically been performed in the biosamples of interest. More critically, many biomedical scientists want to exploit these massive, publicly funded consortium datasets but find that the particular biosample type that they study was relegated to a lower tier and hence is only sparsely characterized.

Imputation methods address this problem by filling in the missing data with computationally predicted values. Imputation is feasible in part due to the structured nature of consortium-style datasets, in which data from high-throughput sequencing experiments can be arranged systematically along axes such as "biosample" and "assay." The first epigenomic imputation method to be applied at a large scale, ChromImpute [14], trains a separate machine learning model for each missing experiment, deriving input features from the same row or column in the data matrix, i.e., training from experiments that involve the same biosample but a different assay or the same assay but a different biosample. A second method, PREDICTD [13], takes a more wholistic approach, first organizing the entire dataset into a 3D tensor (assay $\times$ biosample $\times$ genomic position) and then training an ensemble of machine learning models that each jointly decompose all experiments in the tensor into three matrices, one for each dimension. PREDICTD imputes missing values by linearly combining values from these three matrices. Most recently, a third method, Avocado [67], extends PREDICTD by replacing the linear combination with a non-linear, deep neural network, and by modeling the genomic axis at multiple scales, thereby achieving significantly more accurate imputations without the need to train an ensemble of models.

All three of these existing imputation methods rely upon a common dataset. In creating ChromImpute, Ernst and Kellis utilized what was, at the time, one of the largest collections of uniformly processed epigenomic and transcriptomic data, derived from 1,122 experiments

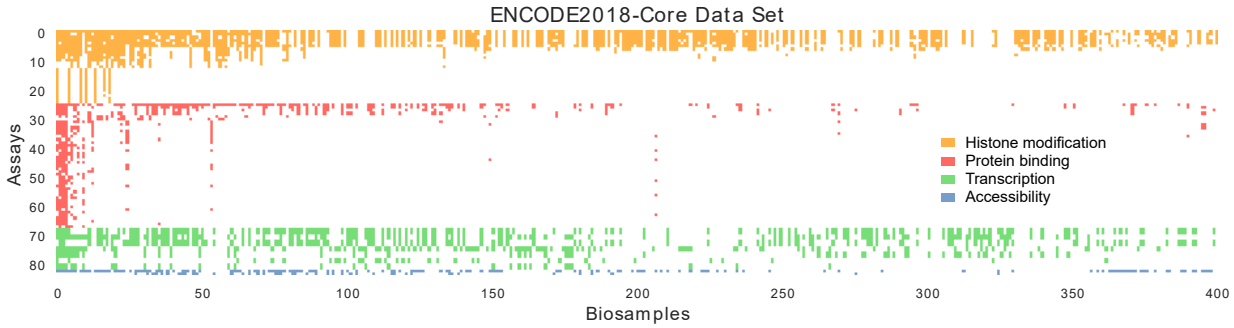



Figure 2.1: **The ENCODE2018-Core data matrix.** In the matrix, columns represent biosamples and rows represent assays. Colors correspond to general types of assays (histone modification ChiP-seq in orange, transcription factor ChIP-seq in red, RNA-seq in green, and chromatin accessibility in blue). Biosamples are sorted by the total number of assays performed in them, and assays are first grouped by their type before being sorted by the number of biosamples that they have been performed in.

from the Roadmap Epigenomics and ENCODE consortia. To allow for direct comparison between methods, both PREDICTD and Avocado relied upon a subset of 1,014 of those experiments. Since 2015, however, the amount of available data has increased tremendously. Here, we[1] report the training of Avocado on a dataset derived from the ENCODE compendium that contains 3,814 tracks from 400 biosamples and 84 assays (Fig. 2.1). This ENCODE2018-Core dataset is 3.4 times larger than the original ChromImpute dataset. We demonstrate that this increase in size leads to a concomitant improvement in predictive accuracy.

Furthermore, whereas the ChromImpute dataset included only chromatin accessibility,

[1]The work in this chapter is based off a paper entitled *Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples* to appear in *Genome Biology* that was written by myself, Jeffrey Bilmes, and William Stafford Noble (in the order that authors appear on the paper). In this work, WSN and myself conceived of experiments, I did the coding and analysis, and all authors contributed to writing the text.

histone modification, and RNA-seq data, the ENCODE2018-Core dataset also includes ChIP-seq measurements of the binding of transcription factors (TF) and other proteins, such as CTCF and POLR2A (referred to hereafter, for simplicity, as "transcription factors," despite the differences in their biological roles). Accurate prediction of TF binding in a cell type-specific fashion is an extremely challenging and well-studied problem (reviewed in [68]). We demonstrate that, by leveraging the large and diverse ENCODE2018-Core dataset, Avocado achieves high accuracy in prediction of TF binding, outperforming several state-of-the-art methods.

Finally, we demonstrate a practically important feature of the Avocado model, namely, that the model can be easily extended to apply to newly or very sparsely characterized biosamples and assays via a simple transfer learning approach. Specifically, we demonstrate how a new biosample or assay can be added to a pre-trained Avocado model by fixing all of the existing model parameters and only training the new assay or biosample factors. We do this using experiments from a second dataset, ENCODE2018-Sparse, that contains 3,056 experiments from biosamples that are sparsely characterized and from assays that have been performed in only few biosamples. We find that the model can yield high quality imputations for transcription factors that are added in this manner, and that these imputations can outperform the ENCODE-DREAM challenge participants even when trained using a single track of data. Finally, we find that when biosamples are added using only DNase-seq experiments, the resulting imputations for other assays can still be of high quality.

As a resource for the community, we have made the AvocadoENCODE imputations publicly available via the ENCODE portal (http://www.encodeproject.org).

## 2.2 Results

### 2.2.1 Avocado's imputations are accurate and biosample specific

We first aimed to evaluate systematically the accuracy of Avocado's imputed values on the ENCODE2018-Core dataset. One challenge associated with this assessment is that no competing imputation method has yet been applied to this particular dataset, making a direct comparison of methods difficult. Further, the size of the dataset makes training competing methods difficult, with ChromImpute requiring the training of thousands of different models. However, we have shown recently that the average activity of a given assay across many biosamples is a good predictor of that activity in a new biosample [69]. Admittedly, this predictor is scientifically uninteresting, in the sense that it makes the same prediction for every new biosample and so, by construction, cannot capture biosample specific variation. However, we reasoned that improvement over this baseline indicates that the model must be capturing biosample specific signal. Furthermore, because the signal from most epigenomic assays is similar across biosamples, the average activity predictor serves as a strong baseline that any cross-cell type predictor must beat. Accordingly, we compare the predictions made by Avocado to the average activity of that assay in the training set that was used for model training.

Overall, we found that Avocado is able to impute signal accurately for a variety of different types of assays. We compared Avocado's imputations to those of the average activity predictor across 37,249,359 genomic loci from chromosomes 12–22 using five-fold cross-validation among epigenomic experiments in the ENCODE2018-Core dataset. Qualitatively, we observed strong visual concordance between observed and imputed values across a variety of assay types (Fig. 2.2A, Fig. G.2). In particular, the imputations capture the shape of peaks in histone modification signal, such as those exhibited in H3K27ac and H3K4me3, the shape of peaks found in assays of transcription factors like ELF1 and CTCF, and exon-specific

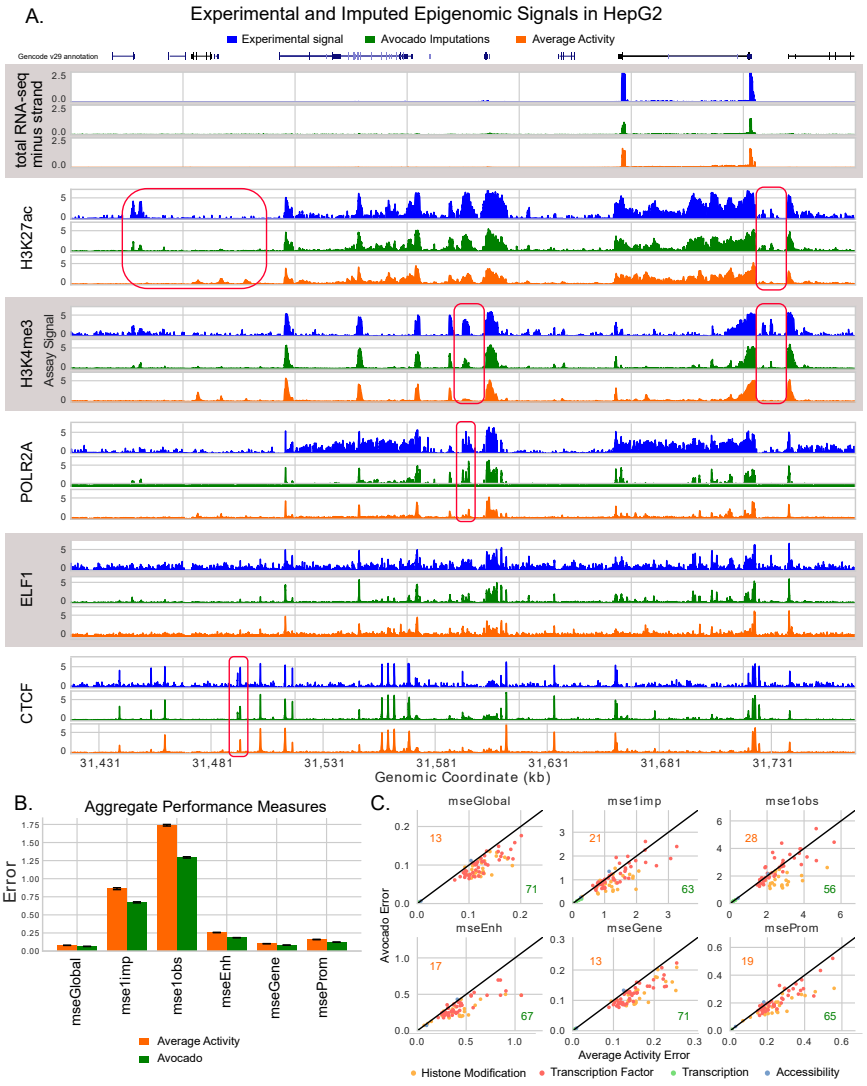



Figure 2.2: **Avocado imputes epigenomic experiments accurately.** (A) Example signal, corresponding imputations, and the average activity of that assay, for six assays performed in HepG2. The figure includes representative tracks for RNA-seq, histone modification, and factor binding. The data covers 350 kbp of chromosome 20. (B) Performance measures evaluated in aggregate over all experiments from all biosamples in chromosomes 12 through 22. Orange bars show the performance of the average activity baseline and green bars show the performance of Avocado's imputations. (C) Performance measures evaluated for each assay, with Avocado's error (y-axis) compared against the error of the average activity (x-axis). The number of assays in which Avocado outperforms the average activity is denoted in green for each metric, and the number of assays in which Avocado underperforms the average activity is denoted in orange.

activity in gene transcription assays. As our primary quantitative measure, we compute the global mean-squared error (MSE) between the observed and imputed values. This value reduces from 0.0807 to 0.0653 (paired t-test p-value of 1e-157), a reduction of 19.1%, between the average activity predictor and Avocado (Fig. 2.2B).

We also compute five complementary quantitative measures. Two measures emphasize the ability of an imputation method to correctly identify peaks in the data. One of these (mse1obs), defined as the MSE in the positions with the top 1% of observed signal, corresponds to a notion of recall. The complementary measure (mse1imp), defined as the MSE in position with the top 1% of imputed signal, corresponds to precision. Three additional measures focus on the MSE in regions of biological activity: the MSE in promoters (mseProm), gene bodies (mseGene), and enhancers (mseEnh). In aggregate, Avocado outperforms the average activity baseline on all six performance measures (p-values between 8e-65 for mse1imp and 1e-157 for mseGlobal) (Fig. 2.2B/C).

When grouped by assay, we find that Avocado outperforms the average activity in 71 of the 84 experiments in our test set according to mseGlobal. Further investigation suggested that these problematic assays were mostly of transcription, indicating a weakness of the Avocado model, or assays that may have been of poor quality (Appendix H).

The primary benefit of the ENCODE2018-Core dataset, in comparison to previous datasets drawn from the Roadmap Compendium, is the inclusion of many more assays and biosamples. We hypothesized that not only will this dataset allow us to make a more diverse set of imputations, but that these additional measurements will improve performance on assays already included in the Roadmap Compendium. We reasoned this may be the case because, for example, previous imputation approaches have imputed H3K36me3, a transcription associated mark, but have not utilized measurements of transcription to do so. A direct comparison to previous work was not simple due to differences in the processing pipelines and reference genomes, and so we re-trained Avocado using the same five-fold cross-validation strategy

after having removed all experiments that did not originate from the Roadmap Epigenomics Consortium. Additionally, we removed all RNA-seq and methylation datasets, as they had not been used as input for previous imputation methods. This resulted in 1,072 tracks of histone modification and chromatin accessibility.

We found that the inclusion of additional assays and biosamples lead to a clear improvement in performance on the tracks from the Roadmap compendium. The MSE of Avocado's imputations dropped from 0.115 when trained exclusively on Roadmap datasets to 0.107 when trained on all tracks in the ENCODE2018-Core dataset, an improvement of 7% (p-value of 8e-45). When we grouped the error by assay, we observed that tracks appeared to range from a significant improvement to only a small decrease in performance (Fig. G.3A). When aggregating these performances across assays, we similarly observe large improvements in the performance of most assays, and small decreases in a few (Fig. G.3B/C). These results indicate that the inclusion of other phenomena do, indeed, aid in the imputation of the original tracks.

### *2.2.2 Comparison to ENCODE-DREAM participants*

Predicting the binding of various transcription factors is particularly important due both to these proteins' critical roles in regulating gene expression and the sparsity with which their binding has been experimentally characterized across different biosamples. For example, of the 43 transcription factors included in the ENCODE2018-Core dataset, only 9 have been performed in more than 10 biosamples. The most performed assay measures CTCF binding and has been performed 136 times, which is almost twice as high as the next most performed assay, measuring POLR2A binding, at 70 assays. In contrast, 13 of the 25 histone modifications in the ENCODE2018-Core dataset have been measured in more than 10 biosamples, and the top six have all been performed in more than 200 biosamples. The sparsity of protein binding assays is exacerbated in the ENCODE2018-Sparse dataset, where an additional 704

assays measuring protein binding have been performed in fewer than five biosamples.

A recent ENCODE-DREAM challenge focused on the prediction of transcription factor binding across biosamples, and phrased the prediction task as one of classification where the aim is to predict whether binding is occuring at a given locus (https://www.synapse.org/#!Synapse:syn6131484). The challenge involved training machine learning models to predict signal peaks using nucleotide sequence, sequence properties, and measurements of gene expression and chromatin accessibility. The participants trained their models on a subset of chromosomes and biosamples, and were evaluated based on how well their models generalized both across chromosomes and in new biosamples. We acquired predicted probabilities of binding from the top four teams, Yuanfang Guan [70], dxquang [71], autosome.ru, and J-TEAM [72], for 13 tracks of epigenomic data. Four of the assays, E2F1, HNF4A, FOXA2 and NANOG, were excluded from the ENCODE2018-Core data set because they had been performed in fewer than five biosamples. Consequently, Avocado could not make predictions for these four assays. Thus, we used only nine tracks for this evaluation.

We compared Avocado's predictions of transcription factor binding to the predictions of the top four models from the ENCODE-DREAM challenge to serve as an independent validation of Avocado's quality. We used both the average precision (AP) and the point on the precision-recall curve where precision and recall are equal (EPR) to evaluate the methods. In order to provide an upper limit for how good Avocado's predictions could be after the conversion process, we included as a baseline the experimental ChIP-seq data that the peaks were called from (called "Same Biosample"). Additionally, we compared against the average activity of that assay in Avocado's training set for that prediction. This baseline serves to show that Avocado is learning to make biosample-specific predictions. Further, when we investigated the training sets for the various experiments, we noted that there were two liver biosamples, male adult (age 32), and female child (age 4), that had similar assays performed in them. To ensure that Avocado was not simply memorizing the signal from one

| Biosample<br>Assay<br>Method | iPSC<br>CTCF | PC-3<br>CTCF | liver<br>EGR1 | liver<br>FOXA1 | liver<br>GABPA | liver<br>JUND | liver<br>MAX | liver<br>REST | liver<br>TAF1 |
|---|---|---|---|---|---|---|---|---|---|
| Yuanfang Guan | 0.729 | 0.600 | 0.397 | 0.282 | 0.353 | 0.533 | 0.441 | 0.319 | 0.281 |
| dxquang | 0.866 | 0.783 | 0.274 | 0.400 | 0.347 | 0.260 | 0.330 | 0.312 | 0.264 |
| autosome.ru | 0.778 | 0.486 | 0.331 | 0.243 | 0.342 | 0.416 | 0.384 | 0.264 | 0.221 |
| J-TEAM | **0.812** | 0.747 | 0.363 | **0.462** | 0.344 | 0.415 | 0.377 | 0.196 | 0.272 |
| Avocado | 0.723 | **0.791** | **0.530** | 0.354 | **0.396** | **0.660** | **0.574** | **0.477** | **0.384** |
| Similar Biosample | — | — | 0.363 | 0.389 | 0.226 | 0.568 | 0.446 | 0.408 | — |
| Same Biosample | 0.741 | 0.878 | 0.648 | 0.716 | 0.573 | 0.731 | 0.622 | 0.622 | 0.556 |
| Average Activity | 0.574 | 0.735 | 0.240 | 0.299 | 0.253 | 0.223 | 0.349 | 0.124 | 0.140 |

Table 2.1: **Comparison of methods on ENCODE-DREAM challenge test set.** The average precision (AP) computed across nine epigenomic experiments in the ENCODE-DREAM challenge test set in chromosome 21. For each track, the score for the best-performing predictive model is in boldface.

of these biosamples and predicting it for the other liver biosample, we compare against the signal from the related biosample as well (denoted "Similar Biosample").

We observed that Avocado's predictions outperform all of the challenge participants in all tracks except for CTCF in iPSC and FOXA1 in liver (Table 2.1, Table G.1). The most significant improvement comes in predicting REST, a transcriptional factor that represses neuronal genes in biosamples that are not neurons, and the highest overall performance is in predicting CTCF binding. This high performance is due in part to the large number of CTCF binding sites, but is likely also because CTCF binding is similar across most biosamples. Importantly, the REST assay for both liver biosamples were in the same fold, and TAF1 was only performed in one of the liver biosamples, so Avocado's good performance on those tracks are strong indicators of its performance. Visually, we observe that some of the participants models appeared to overpredict signal values, suggesting that a source of

error for these models is their lack of precision, corresponding to rapid drop in precision for predicting REST (Fig. G.4). Interestingly, Avocado appears to underperform using the related liver biosample as the predictor for FOXA1, suggesting that perhaps the factors for FOXA1 are poorly trained. However, this result is further evidence that Avocado is not simply memorizing related signal. We also note that, in the case of CTCF in iPSCs, the ChIP-seq signal from iPSC appears to underperform two challenge participants, suggesting that the conversion process may limit Avocado's performance.

We did our best to ensure a fair comparison between Avocado and the challenge participants, but the comparison is necessarily imperfect, for several reasons. Two factors make the comparison easier for Avocado. First, Avocado is exposed to many epigenomic measurements that the challenge participants did not have available, including measurements of the same transcription factor in other cell types. Second, as an imputation approach, Avocado is trained on the same genomic loci that it makes predictions for, whereas the challenge participants had to make predictions for held-out chromosomes. On the other hand, three factors skew the comparison in favor of the challenge participants. First, unlike the challenge participants, Avocado was not directly exposed to any aspect of nucleotide sequence or motif presence. Second, Avocado makes predictions at 25 bp resolution in hg38, whereas the challenge was conducted at 200 bp resolution in hg19. We were able to use liftOver to convert between assemblies, followed by aggregating the signal from 25 bp resolution to 200 bp resolution, but both steps blurred the signal. Third, Avocado is trained to predict signal values directly, whereas the challenge participants are trained on the classification task of identifying whether a position is a peak. Evaluation is done in a classification setting. In particular, Avocado is penalized for accurately predicting high signal values in regions that aren't labeled as peaks, exemplifying the discordance between the regression and classification settings. For all these reasons, Avocado would not have been a valid submission to the challenge. Finally, it is perhaps worth emphasizing that whereas the challenge was truly

blind, our application of Avocado to the challenge data is only blind "by construction." We emphasize that we did not adjust Avocado's model or hyperparameters based on looking at the challenge results: the comparison presented here is based entirely on a pre-trained Avocado model.

We investigated the effect that these differences may have had on predictive performance. First, we evaluated the performance of Avocado and the challenge participants at predicting the test set challenge tracks on chromosome 17, whose loci were used for training the challenge models. This evaluation resulted in similar trends as in Table 2.1 (Appendix I), and suggests that the loci used for evaluation are not a significant factor for Avocado's improved performance over the challenge participants. Next, we removed from Avocado's training set all experiments from biosamples which appeared in the challenge test set, except for those experiments that the challenge participants had—namely, DNase-seq and RNA-seq experiments. This restricted Avocado to only being able to make predictions on the challenge tracks using the same epigenomic information that the participants had. In this setting we observed poor performance of Avocado on the liver test set tracks, but even better performance on the CTCF tracks in iPSC and PC-3 than the original Avocado model. However, as described in Appendix I, it was difficult to ensure a fair comparison on the biosamples noted as being from liver, and these reasons may potentially explain the poor results. Finally, we trained an Avocado model using only DNase-seq and RNA-seq from the biosamples used in the challenge, as well as the transcription factor binding tracks available in the training set. Again, performance on liver biosamples was poor. While performance also degraded on the CTCF tracks, it was still competative with the top four participants. These results indicate that a source of Avocado's power is leveraging the diverse data in the massive ENCODE compendium.

### *2.2.3 Extending Avocado to more biosamples and assays*

*Adding new assays*

Despite including 3,814 epigenomic experiments, the ENCODE2018-Core dataset does not contain all biosamples or assays that are represented in the ENCODE compendium. Specifically, the dataset does not include 667 biosamples where fewer than five assays had been performed, and it does not include 1,281 assays that had been performed in fewer than five biosamples. The missing biosamples primarily include time courses, genetic modifications, and treatments of canonical biosamples, such as HepG2 genetically modified using RNAi. However, several primary cell lines and tissues such as amniotic stem cells, adipocytes, and pulmonary artery, were also not included in the ENCODE2018-Core dataset due to lack of sufficient data. The majority of the missing assays corresponded to transcription measurements after gene knockdowns/knockouts (shRNA and CRISPR assays) or to binding measurements of eGFP fusion proteins. Yet some transcription factors, such as NANOG, FOXA2, and HNF4A, were excluded as well. We collect these experiments into a separate dataset, called ENCODE2018-Sparse (see Methods).

We constructed the ENCODE2018-Sparse dataset to attempt to address some of the problems of missingness in ENCODE2018-Core. This sparse version of the data has 99.7% missing entries, in comparson to 88.6% missing in ENCODE2018-Core. Within ENCODE2018-Sparse, we identified four main groups of biosamples: (1) 417 biosamples that only had DNase-seq performed on them, with 58 additional biosamples that had DNase and one or more other assays performed in it, (2) 112 biosamples that had various measurements of transcription performed in them, (3) 7 biosamples that were well characterized by at least 50 sparsely performed assays of transcription factor binding, and (4) biosamples derived from HepG2 and K562 that were well characterized by various knockouts (Fig. G.1).

In general, handling sparsely characterized assays or biosamples in a model like Avocado is challenging. Hence, we designed a three-step process that we hypothesized would allow

us to make accurate imputations for additions with few corresponding tracks (Fig. G.5). This approach is conceptually similar to our main approach for training Avocado. First, we trained the Avocado model on all 3,814 experiments in ENCODE2018-Core. Second, we froze all of the weights in the model, including both the neural network weights and all five of the latent factor matrices. Third, we fit the new biosample or assay factors to the model using only the experimental signal derived from the ENCODE Pilot Regions. This resulted in a model whose only difference was the inclusion of a set of trained assay or biosample factors that were not present in original model. This training strategy has the benefit of allowing for quick addition of biosamples or assays to the pre-trained model, without requiring retraining of any of the existing model parameters.

In order to test the effectiveness of this approach, we extended Avocado to include assays that were in the ENCODE-DREAM challenge but not in the ENCODE2018-Core dataset. For the four assays that we did not compare against (HNF4A and FOXA2 in liver, NANOG in iPSC, and E2F1 in K562), all but E2F1 had been performed in a biosample other than the one included in the challenge. Accordingly, we fit these three new assay factors using the procedure above. This fitting was done using HNF4A and FOXA2 from HepG2, and NANOG from h1-hESC. We then used the new assay factors, coupled with the pre-trained network, genome factors, and relevant biosample factors, to impute three remaining tracks in the challenge.

We observed that Avocado's imputed tracks for HNF4A and FOXA2 in liver were of high quality and outperformed several baselines (Fig. 2.3). Most notably, both of these tracks outperformed all four challenge participants in their respective settings according to both EPR and AP. Second, both Avocado tracks outperformed simply using the track that they were trained on as the predictor, indicating that the model is leveraging the pre-trained biosample latent factors to predict biosample-specific signal.

However, we also observed that Avocado's imputations for NANOG in iPSCs are of

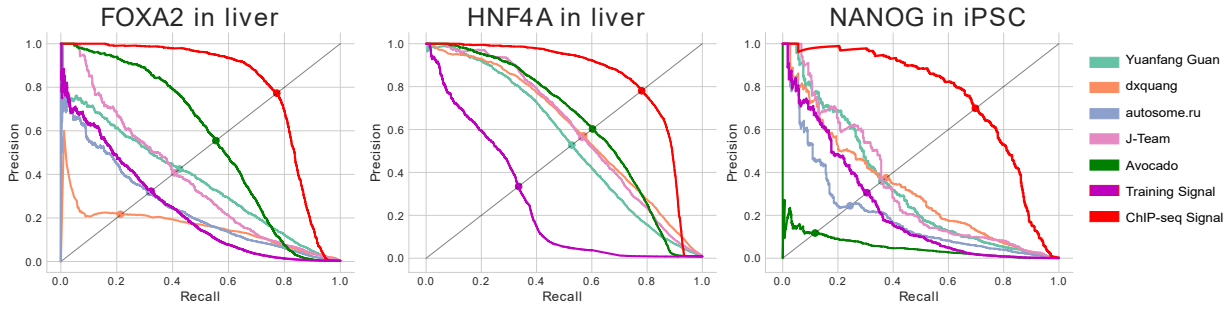



Figure 2.3: **Avocado's performance when adding new transcription factors to a pre-trained model.** Precision-recall curves for three transcription factors that were added to a pre-trained model using a single track of data each from the ENCODE2018-Sparse dataset. Similar to the previous comparisons against the ENCODE-DREAM participants, the evaluation was performed in chromosome 21.

particularly poor quality. Avocado's predictions underperform all four challenge participants. Notably, Avocado also underperforms using the signal from h1-hESC that it was trained on as the predictor. One potential reason for this poor performance is that relevant features of the NANOG binding sites are not encoded in the genomic latent factors. Alternatively, given that Avocado also underperformed the challenge participants at predicting CTCF in iPSC, it may be that the iPSC latent factors are not well trained, leading to poor performance in predictions of any track.

*Adding new biosamples*

We then tested the ability of the three-step process in Fig. G.5 to make accurate predictions for biosamples that the model was not originally trained on. To do so, we began by training biosample factors for 475 biosamples not in the ENCODE2018-Core dataset that had DNase-seq performed in them. We then evaluated Avocado's ability to predict other assays that

were performed in these biosamples. A large number of these biosamples had only DNase-seq performed in them, so we also evaluated Avocado's ability to predict DNase-seq as well. We reasoned that because the biosample factors were trained using the ENCODE Pilot regions, but the predictions were evaluated in chromosome 20 without re-training the corresponding genomic latent factors, this would be a fair evaluation.

We observed good performance of the imputations for these biosamples. Visually, we noticed the same concordance between the imputed and the experimental signal, and we observed that biosample-specific elements are being captured (Fig. 2.4a). We then evaluated the performance of Avocado on the mseGlobal metric compared to the average activity baseline for each assay. We observed that Avocado appears to produce high quality predictions for several assays, including CTCF, H3K27ac, and POLR2A (Fig. 2.4b). However, for other assays, such as H3K9me3 an H3K36me3, the average activity dominates. It is possible that this phenomenon speaks to the ability of DNase to recover these other approaches. Overall, we observe a decrease in error from 0.027 when using the average activity to 0.024 when using the imputations from Avocado.

While these evaluations have thus far used only DNase-seq to fit new biosamples to the model, it is not necessarily the case that performing a single assay is sufficient to optimally fit new biosamples to a model. Unfortunately, it would be computationally expensive to identify the combination of assays that yielded optimal performance. To investigate whether there was a general trend that biosamples fit with more assays performed better than those fit with fewer assays, we partitioned the 3,814 experiments from the five-fold cross-validation on ENCODE2018-Core by the number of assays performed in the biosample of the experiment (Fig. G.6). When we plotted the average error of each type of activity, we did not observe a noticable trend between the number of assays used to fit biosample factors and performance at imputing experiments. This evaluation is limited by not considering the composition of experiments used to fit each biosample or by considering experiments that had fewer than

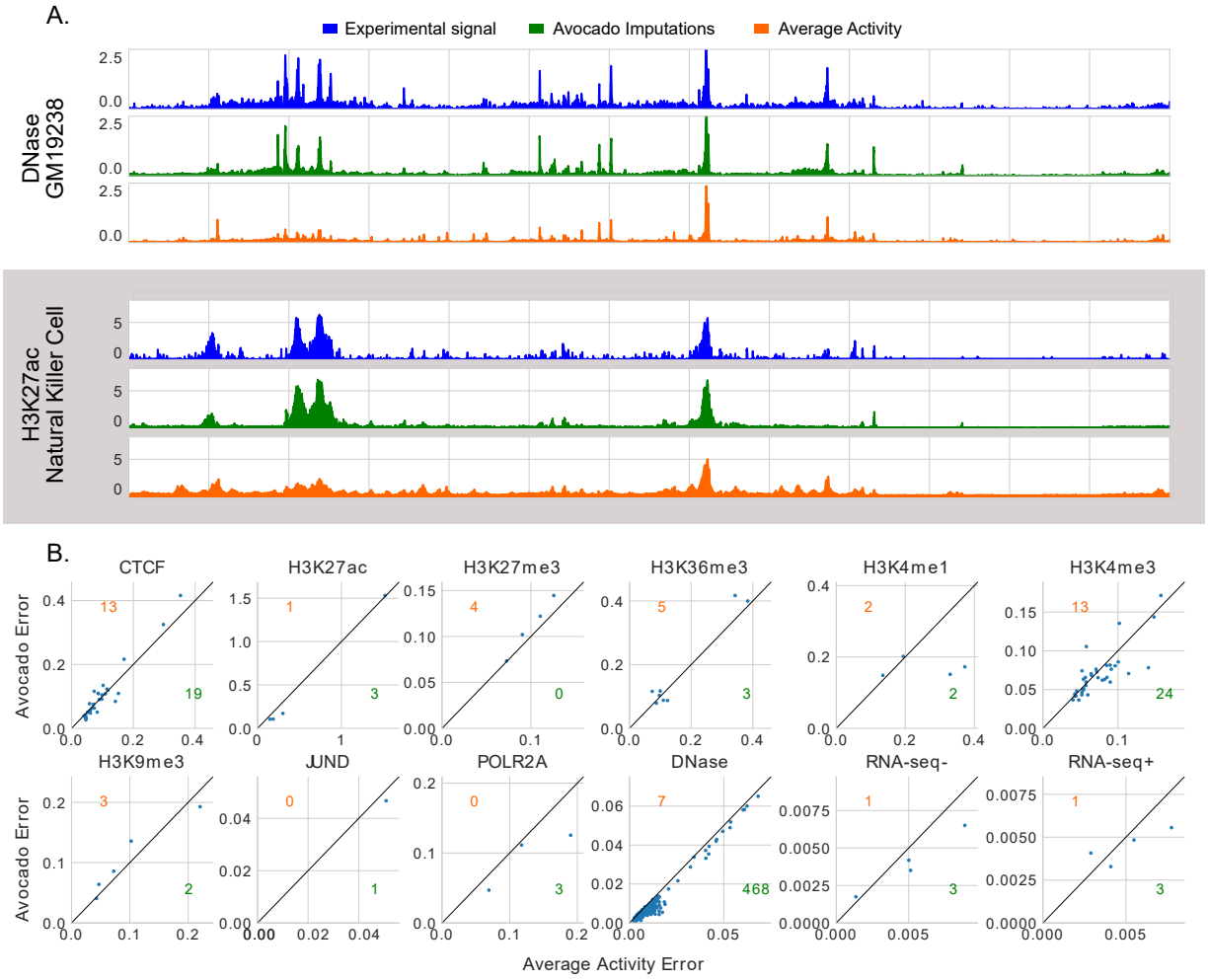



Figure 2.4: **Imputations and performance when adding biosamples to a pre-trained model** (A) Imputations for two tracks of data in the ENCODE2018-Sparse data set on chromosome 20 after fitting the biosample factors using only DNase-seq signal from the ENCODE Pilot Regions. (B) Performance of Avocado at imputing tracks on chromosome 20 after fitting the biosample factors using only DNase-seq signal from the ENCODE Pilot regions.

four assays performed in the respective biosample. However, these results do not suggest that simply performing more experiments will yield better performance overall, or that biosamples with many experiments performed in them will necessarily have better imputations than those that are more sparsely assayed.

*Evaluating alternate training methods*

Our strategy for incorporating new biosamples and assays into a pre-trained Avocado model involves first freezing almost all of the parameters of the model. In practice, large consortia and other providers of imputations are likely to be interested in this approach because it would allow for continuous incorporation of new biosamples and assays without affecting the imputations that have already been released. However, it is unlikely that keeping these parameters frozen during training would yield performance as high as updating them using the new data, because the new experiments may point to interesting loci, novel forms of activity, or important cell-type specific signatures that are not captured in the frozen parameters. To test this hypothesis, we compared the performance of our strategy for incorporating new biosamples and assays to two alternate approaches: retraining Avocado from scratch, and fine-tuning a pre-trained Avocado model (see Methods, Section 2.5.2).

We first simulated the setting where one has trained a model on a set of "original" experiments and would now like to extend the model to include assays and biosamples contained in a set of "additional" experiments. We used four of the five folds used in Section 2.2.1 from the ENCODE2018-Core data set as the original experiments, and half of the experiments from ENCODE2018-Sparse, after filtering, as the set of additional experiments. This filtering step consisted of removing all experiments where the assay or biosample had only been performed once. We created two separate test sets: the first was the second half of the experiments in the ENCODE2018-Sparse data set, and the second was the fifth fold from the ENCODE2018-Core data set to use as validation that the models were still performing

well on the original data. The experiments from ENCODE2018-Sparse were split such that, for each assay, the biosamples were evenly partitioned into the training and test sets.

Overall, we found that our strategy of freezing parameters underperformed both retraining Avocado and fine-tuning a pre-trained model (Table 2.2). In particular, we observed a large difference in performance between the freezing strategy and the other two strategies on the ENCODE2018-Sparse test set. However, upon inspecting the errors more closely, we observed that the majority of errors on the second half of the Sparse data sets come from short-hairpin RNA-seq (shRNA) experiments (Fig. G.7). These experiments involve knocking out a target gene using RNA interference and are not present at all in the ENCODE2018-Core data set. Thus, it makes sense that a model that had been trained on ENCODE2018-Core and then had most of its parameters frozen would perform poorly at imputing shRNA experiments, because neither the genome factors nor the neural network were trained using this type of activity. When we remove these experiments from the ENCODE2018-Sparse test set, we find that the gap in performance between the different methods diminishes significantly. Further, we observe that the error of the frozen model decreases, whereas the errors of the other models increases, confirming that models that were able to train on this type of activity could capture it well. We then validated the resulting models by checking their performance on the fifth fold of the ENCODE2018-Core data set that had been held out. We observed that the models all performed similarly both to each other and to the split in the original five-fold cross-validation when the fold used here as the test set was held out.

## *2.3 Discussion*

To our knowledge, we report here the largest imputation of epigenomic data that has been performed to date. We applied the Avocado deep tensor factorization model to 3,814 epigenomic experiments in the ENCODE2018-Core dataset. The resulting imputations cover a diverse set of biological activity and cellular contexts and are publicly available

| Test set | Retrain from Scratch | Fine-tune | Freeze | Five-fold |
|---|---|---|---|---|
| ENCODE2018-Sparse | 0.058 | 0.056 | 0.091 | — |
| ENCODE2018-Sparse (w/o shRNA) | 0.069 | 0.068 | 0.080 | — |
| ENCODE2018-Core | 0.049 | 0.050 | 0.051 | 0.050 |

Table 2.2: **Comparison of approaches for extending Avocado to new cell types and assays.** The MSE of three approaches for adding new biosamples or assay on three test sets. The test sets are half of the experiments in the ENCODE2018-Sparse data set, that same data set with shRNA experiments removed, and the fifth fold from the ENCODE2018-Core data set. The MSE from the five-fold cross-validation is shown for the ENCODE2018-Core test fold as a reference.

at http://www.encodeproject.org. Due to the cost of experimentation and the increasing sparsity of epigenomic compendia we anticipate that imputations of this scale will serve as a valuable community resource for characterizing the human epigenome.

We used multiple independent lines of reasoning to confirm that Avocado's imputations are both accurate and biosample specific. First, we compared each imputed data track to the average activity of that assay and found that, for almost all assays, that Avocado's imputations were more accurate. A current weakness in Avocado's imputations is imputing transcription, likely due to the sparse, exon-level activity of these assays along the genome. Second, we compared imputations of transcription factor binding tracks to the predictions made by the top four models in the recent ENCODE-DREAM challenge. In almost all cases, the Avocado imputations were significantly more accurate than the imputations produced by the challenge participants. Notably, Avocado is not exposed to nucleotide sequence at all during the training process, and so its ability to correctly impute transcription factor binding is based entirely on local epigenomic context, rather than binding motifs.

Ongoing characterization efforts regularly identify new biosamples of interest and develop

assays to measure previously uncharacterized phenomena. These efforts aid in understanding the complexities of the human genome but pose a problem for imputation efforts that must be trained in a batch fashion. Given that it took almost a day to fit the Avocado genomic latent factors for even the smallest chromosome, re-training the model for each inclusion is not feasible. We demonstrated that, by leveraging parameters that had been pre-trained on the ENCODE2018-Core dataset, new assays and biosamples could be quickly added to the existing Avocado model. In contrast, extending imputations to cover a single new assay using ChromImpute would require training a new model for each of the 400 biosamples in the ENCODE2018-Core data set, or each of the $> 1,000$ biosamples in the combined ENCODE2018-Core/ENCODE2018-Sparse data set. Our observations suggest that not only is the Avocado approach computationally efficient, with three new assays taking only a few minutes to add to the model, but that the resulting imputations are highly accurate.

One potential reason that this pre-training strategy works well is that the genomic latent factors efficiently encode information about regions of biological activity. For example, rather than memorizing the specific assays that exhibit activity at each locus, the latent factors may be organizing general features of the biochemical activity at that locus. We have previously demonstrated the utility of Avocado's latent genomic representation for several predictive tasks [67]. Investigating the utility and meaning of the latent factors from this improved Avocado model is ongoing work.

Notably, however, the encoding of relevant information in the latent factors may lead to a potential weakness in Avocado's ability to generalize to novel biosamples or assays. Specifically, if the signal in a novel biosample or assay is not predictable from the tracks that were used to train the initial genomic latent factors, then it is unlikely that Avocado will make good imputations for the new data. For example, if a transcription factor is dissimilar to any factors in the training set, then the genomic latent factors may not have captured features relevant to the novel factor. This may explain why Avocado fails to generalize well

to NANOG.

A strength of large consortia, such as ENCODE, is that they are able to collect massive amounts of experimental data. This amount of data is only possible because many labs collect it over the course of several years. Inevitably, this results in some data that is of poor quality. While quality control measures can usually identify data that is of very poor quality, they are not perfect, and the decision of what to do with such data can be challenging. Unfortunately, data of poor quality poses a dual challenge for any large scale imputation approach. When an imputation approach is trained on low quaity data, then the resulting imputations may be distorted by the noise. Furthermore, when the approach is evaluated against data that is of poor quality, imputations that are of good quality may be incorrectly scored poorly. Thus, when dealing with large and historic data sources, it is important to ensure the quality of the data being used.

### *2.4 Conclusion*

In this work, we describe the training of an imputation approach that can predict a variety of epigenomic phenomena, including histone modification, protein binding, transcription, and chromatin accessibility, across hundreds of human biosamples. The resulting model is capable of imputing 33,600 genome-wide epigenomic experiments, representing the largest imputation effort performed to date both in terms of the number of tracks imputed and in terms of biological phenomena that are jointly modeled. We found that these tracks were of high quality, with a 19.5% decrease in overall error when compared to the strong average activity baseline. Empirically, the imputations of transcription factor binding significantly outperformed the top participants in a recent ENCODE-DREAM transcription factor binding challenge, further indicating their quality.

We anticipate that this work will be impactful in several ways. The simplest application of these imputations is to enable analyses or prediction in biosamples where the required

epigenomic experiments have not yet been performed. Another approach is to look for inconsistencies between the imputed and primary data for experiments that have been performed, with the anticipation that these regions may prove biologically interesting. Further, one could use imputed tracks where there is no corresponding experimental data to determine what experiments should be performed next, priotizing imputed tracks that appear to encode interesting phenomena.

The imputation approach offered by Avocado has great potential to be extended both to precision medicine and to single cell datasets. In the precision medicine setting, a biosample is sparsely assayed in a variety of individuals, and the goal is to correctly impute the inter-individual variation, particularly in regions associated with disease. We anticipate that Avocado could either be applied directly in this setting, with biosamples including the annotation of the individual they came from, or extended to accommodate a 4D data tensor, where the fourth dimension corresponds to distinct individuals. In the single cell setting, the biosample axis would be replaced with a cell axis where each entry would correspond to a single cell. This approach could potentially be used as a computational co-assay, leveraging a shared genomic axis to impute multiple types of experiments in each individual cell.

## *2.5 Methods*

### *2.5.1 Avocado*

#### *Avocado topology*

Avocado is a multi-scale deep tensor factorization model. The tensor factization component is comprised of five matrices of latent factors that encode the biosample, assay, and three resolutons of genomic factors at 25 bp, 250 bp, and 5 kbp resolution. Having multiple resolutions of genomic factors means that adjacent positions along the genome may share the same 250 bp and 5 kbp resolution factors. We used the same model architecture as in the original Avocado model [67], with 32 factors per biosample, 256 factors per assay, 25

factors per 25-bp genomic position, 40 factors per 250-bp genomic position, and 45 factors per 5-kbp genomic position. The neural network model has two hidden dense layers that each have 2,048 neurons, before the regression output, for a total of three weight matrices to be learned jointly with the matrices of latent factors. The network uses ReLU activation functions, $\text{ReLU}(x) = \max(0, x)$, on the hidden layers, but no activation function on the prediction.

*Avocado training*

Avocado is trained in a similar fashion to our previous work [67]. This procedure involves two steps, because the genome is large and the full set of genomic latent factors cannot fit in memory. The first is to jointly train all parameters of the model on the ENCODE Pilot regions, which comprise roughly 1% of the genome. After training is complete, the neural network weights, the assay factors, and the biosamples are all frozen. The second step is to train only the three matrices of latent factors that make up the genomic factors on each chromosome individually. In this manner, we can train comparable latent factors across each chromosome without the need to keep then all in memory at the same time.

Avocado was trained in a standard fashion for neural network optimization. All initial model parameters and optimizer hyperparameters were set to the defaults in Keras. In this work, Avocado was trained using the Adam optimizer [56] for 8,000 epochs with a batch size of 40,000. This is longer than our original work, where the model was trained for 800 epochs initially and 200 epochs on the subsequent transfer learning step. Empirical results suggest that this longer training process is required to reach convergence, potentially because of the large diversity of signals in the ENCODE2018-Core dataset. When adding in additional biosample or assay factors, due to the small number of trainable parameters, the model was trained for only 10 epochs with a batch size of 512. Due to the large dataset size, one epoch is defined as one pass over the genomic axis, randomly selecting experiments at each position,

rather than one full pass over every experiment.

The model was implemented using Keras (https://keras.io) with the Theano backend [54], and experiments were run using GTX 1080 and GTX 2080 GPUs. For further background on neural network models, we recommend the comprehensive review by J. Schmidhuber [55].

### *2.5.2 Data and evaluation*

#### *ENCODE dataset*

We downloaded 6,870 genome-wide tracks of epigenomic data from the ENCODE project (https://www.encodeproject.org). These experiments were all processed using the ENCODE processing pipeline and mapped to human genome assembly hg38, except for the ATAC-seq tracks, which were processed using an approach that would later be added to the ENCODE processing pipeline. The values are signal p-value for ChIP-seq data and ATAC-seq, read-depth normalized signal for DNase-seq, and plus/minus strand signal for RNA-seq. When multiple replicates were present, we preferentially chose the pooled replicate; otherwise, we chose the second replicate. The experimental signal tracks were then further processed before being used for model training and evaluation. First, the signal was downsampled to 25 bp resolution by taking the average signal in each 25 bp bin. Second, an inverse hyperbolic sin transformation was applied to the data. This transformation has been used previously to reduce the effect of outliers in epigenomic signal [48, 13].

We divided these experiments into two datasets, the ENCODE2018-Core dataset and the ENCODE2018-Sparse dataset. The ENCODE2018-Core dataset contains 3,814 experiments from all 84 assays that have been performed in at least five biosamples, and all 400 biosamples that have been characterized by at least five assays. Hence, ∼88.6% of the data in the ENCODE2018-Core data matrix is missing. The ENCODE2018-Sparse dataset contains 3,056 experiments, including 1,281 assays that have been performed in fewer than

five biosamples and 667 biosamples that have been characterized by fewer than five assays, yielding a matrix that is ∼99.7% missing.

We adopted a similar strategy to Durham *et al.* for partitioning these experiments into folds for cross-validation. Specifically, we partitioned entire genome-wide experiments into five folds such that a model would be trained on all genomic loci and then evaluated on its ability to predict entirely held-out experiments, because this is the most realistic evaluation setting. However, randomly assigning tracks to each fold may inadvertantly leave some folds without seeing some assays or some biosamples, meaning that the model would not learn anything for those embeddings and thus perform poorly on imputation. Unfortunately, even after only keeping experiments from assays and biosamples where five experiments had been performed, it is not always possible to partition a set of experiments into folds such that each assay and biosample are seen. Thus, Durham *et al.* adopted a simple optimization approach that randomly assigned experiments to partitions and evaluated each partition by the total number of biosamples and assays covered by each partition. We empirically found that this approach underperformed a simple greedy approach that uses a counter to sequentially assign folds to random experiments within biosamples, one biosample at a time, preserving the location in the cycle from one biosample to the next.

*ENCODE-DREAM challenge datasets*

For our comparisons with the ENCODE-DREAM challenge participants, we acquired from the challenge organizers both genome-wide model predictions from the top four participants and the binary labels (https://www.synapse.org/#!Synapse:syn17805945). The predictions and labels were defined at 200 bp resolution, with a stride of 50 bp, meaning that each 50 bp bin was included in four adjacent bins. The labels corresponded to conservative thresholded irreproducible discover rate (IDR) peaks called from multiple replicates of ChIP-seq signal.

*Comparison to ENCODE-DREAM predictions*

Avocado's predictions had to be processed in several ways to make them comparable with the data format for the challenge. First, because Avocado's predictions are in hg38 and the challenge was performed in hg19, the UCSC liftOver command (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates across reference genomes. Unfortunately, many of the 25 bp bins in hg38 mapped to the middle of bins in hg19, blurring the signal. Further, ∼27% of positions on chromosome 21 of hg38 could not be mapped to positions in hg19, so those positions were discarded from the analysis. Lastly, because the challenge was performed at 200 bp resolution, the average prediction in the 200 bp region was used as Avocado's predictions for that bin. We then filtered out all regions that were marked as "ambiguous"' by the challenge organizers. These regions included both the flanks of true peaks as well as regions that were considered peaks in some, but not all, replicates.

The evaluation of each model was performed using both the average precision, which roughly corresponds to the area under a precision-recall curve, and the point along the precision-recall curve of equal precision and recall (EPR). The EPR corresponds to setting the decision threshold so that the number of positive predictions made by the model is equal to the number of positive labels in the dataset. This is also called the "break-even point". A strength of EPR, in comparison to taking the recall at a fixed precision, is that it accounts for the true sparsity in the label set.For example, if it is known beforehand that an experimental track generally has between 100 and 200 peaks across the entire genome, then a reasonable user may use the top 150 predictions from a model. However, if an experimental track had between 10,000 and 20,000 peaks, then a user may use the top 15,000 predicted peaks.

*Calculation of average activity*

In several of our experiments we compared model performance against the average activity of an assay. In all instances involving the ENCODE2018-Core dataset, "average activity"

refers to the average signal value at each locus across all biosamples in the training set for that particular experiment. Because the predictions across the entire ENCODE2018-Core dataset are made using five-fold cross-validation, the training set differ for tracks from different folds. This approach ensures that the track being predicted is not included in the calculation of average activity which would make the baseline unfair. In instances involving the ENCODE2018-Sparse dataset, "average activity" refers to the average activity across all tracks of that assay that were present in the entire ENCODE2018-Core dataset.

*Incorporating new experiments*

We evaluated the performance of three approaches for handling the incorporation of additional biosamples or assays into a model: retraining the model from scratch, fine-tuning the parameters of a pre-trained model, and freezing most parameters of a pre-trained model and training the remaining subset. These approaches were evaluated using four of the five folds from the ENCODE2018-Core data set as the set of "original" experiments and half of the experiments in the ENCODE2018-Sparse data set as the "additional" experiments. When training Avocado from scratch the model was trained on both the original and additional experiments for a total of 8,000 epochs, just like our normal training approach. When fine-tuning a pre-trained model we first created a pre-trained model by training Avocado for 6,000 epochs on just the original experiments and then training on botht he original and the additional experiments for an additional 2,000 epochs. This ensured that differences in performance between the retrained model and the fine-tuned model did not arise simply due to a different number of epochs of training. Lastly, we trained Avocado for 8,000 epochs on just the original experiments, froze the neural network and genomic position parameters, and proceeded with training the assay and biosample factors using only the additional experiments for 100 epochs.

# Chapter 3

# AVOCADO CAN BE EXTENDED TO MODEL MULTIPLE SPECIES

## *3.1 Background*

A common way for researchers to investigate questions in genomics is by designing and performing high-throughput assays that quantify various forms of biochemical activity along a genome. Such assays include ChIP-seq, which has been used to measure measure histone modification and protein binding, RNA-seq, which measures transcription, and ATAC-seq, which measures chromatin accessibility. The resulting experimental data quantify cell type-specific activity and so can be used to help explain the basis for cellular mechanisms such as differentiation and, when things go wrong, disease. Consequently, individual investigators perform these assays to answer specific research questions, and large consortia—such as the Roadmap Epigenomics Consortium, the ENCODE Project, and the International Human Epigenomics Consortium—perform and collect thousands of them into compendia that broadly characterize human epigenomics across a variety of primary cells and tissues.

Despite the value of this data, the cost of these assays means that such compendia are rarely complete. For example, the Roadmap Compendium [61] contains 1,122 experiments spanning 34 assays and 127 different human cell types and tissues (which we refer to as "biosamples"), making it only 26% filled in. This incompleteness can be an obstacle for computational methods and for investigators studying biosamples that are incompletely assayed. Recently, there have been efforts to build machine learning models that can impute these missing experiments by leveraging learned associations among the experiments that have been performed [14, 13, 67].

In practice, this problem of incompleteness is even worse in compendia collected for species other than humans. As of January 24th, 2020, the ENCODE Project portal (https://www.encodeproject.com) hosts only only 1,814 epigenomic experiments mapped to the mouse reference genome mm10, in contrast to the 9,111 experiments mapped to the human reference genome hg38. The experiments performed in mice span fewer assays and biosamples than the human experiments, and each mouse biosample is generally less well assayed than a typical human biosample. Perhaps most importantly, the overall characterization of protein binding is far sparser in mouse than in human, despite proteins like transcription factors playing crucial roles in the cell. To illustrate this difference in sparsity, the best characterized human biosample, K562, has 504 protein binding experiments mapped to hg38, whereas the best characterized biosample in mouse, MEL, has only 49 assays mapped to mm10. Further, only 36 mouse biosamples have been assayed for protein binding at all, while hundreds of human biosamples have been assayed for the binding of at least one protein. Because imputation approaches are restricted to the set of assays and biosamples in a given compendium, the lack of protein binding experiments poses a significant challenge that current imputation approaches cannot yet overcome.

Fortunately, many types of biochemical activity play similar roles in the cell across evolutionarily related species. For instance, the histone modification H3K4me3 is enriched in active promoters in both humans and mice [73], and the transcription factor MYC is associated with cell growth in both species [74]. These similarities suggests that it is possible for an imputation model to transfer knowledge of these types of activity across species. The concept of transfer learning has been used in other domains, such as natural language processing, where machine learning models have transferred knowledge from a high-resource setting, such as a language with plentiful annotated training examples, to a low-resource setting, where there are fewer annotated examples [75]. In our setting, the epigenomic compendia available for humans can be viewed as a high-resource setting and the sparser compendia

available for other species as the low resource setting.

In this work, we[1] propose an extension to the imputation approach Avocado [67] to enable the joint modeling of human and mouse epigenomics. In its original formulation, Avocado first organizes a compendium of data into a 3D tensor with axes corresponding to the biosamples, assays, and genomic positions along the genome. Then Avocado uses a deep tensor factorization approach to learn latent representations of each of these axes. This process is similar to a standard matrix factorization, except that the dot product operation on the latent factors is replaced with a neural network. Our extension involves merging the human and the mice compendia into a single tensor by taking the union of assays as one axis, many of which have been performed in both species, the union of biosamples as the second axis, which are disjoint across species, and the concatenation of genomic positions as the third axis. This factorization procedure is similar to independently factorizing the human and mouse compendia, except that the assay representations and the neural network parameters are tied across species.

We demonstrate that this joint optimization procedure allows Avocado to make higher quality and more comprehensive imputations for mice than one could by modeling mouse epigenomics alone. First, we show that incorporating human data into the standard imputation task improves model performance, particularly for experiments that measure transcription. Then we show that this procedure allows for the imputation of assays that have been performed in humans but not in mice. In machine learning terminology, this is an example of a "zero-shot" problem, where a model is asked to make predictions despite not having any labeled training data for a particular class. Specifically, only 64 assays have been performed in both mice and humans, but we aim to make zero-shot predictions for a further 735 assays

[1]The work in this chapter is based off a paper entitled *Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics* that has been submitted to *ISMB 2020* that was written by myself, Deepthi Hedge, and William Stafford Noble (in the order that authors appear on the paper). In this work, WSN and myself conceived of experiments, DH and I did the coding, I did the analysis, and myself and WN contributed to writing the text.

have been performed only in humans (excluding assays that involve performing RNA-seq after CRISPR editing or short-hairpin RNA interference).

Lastly, we show that jointly modeling human and mouse epigenomic data can encode epigenomic state into a latent representation in a manner that is shared across species. Conveniently, these representations can be learned regardless of the number of experiments in each species and without the need to either explicitly match biosamples between species or be restricted to orthologous regions. We exploit these representations to define a similarity score between pairs of regions that is based on their epigenomic similarity. To demonstrate that this similarity score differs from simple sequence similarity, we highlight an epigenomically dissimilar pair of regions with high sequence similarity that corresponds to an exon in the gene EEF1D that is alternatively spliced and potentially an alternate promoter in humans but not in mice.

## *3.2 Methods*

**Data sets** In total, we downloaded and processed 7,986 epigenomic experiments hosted on the the ENCODE project portal (https://www.encodeproject.org). These experiments were partitioned into three data sets: (1) the ENCODE2018-Full data set comprising the 6,870 epigenomic experiments that measure activity in humans, (2) a subset of 3,814 of those experiments, called the ENCODE2018-Core data set, that only include experiments from biosamples or assays that had at least five experiments performed in them, and (3) the 1,116 epigenomic experiments that measure activity in mice.

The data processing pipeline is similar to previous work involving Avocado [67, 52]. All experiments were processed using the ENCODE Processing Pipeline and mapped to either human genome assembly hg38 or mouse genome assembly mm10. The signal values were $-\log 10$ p-values for the ChIP-seq and ATAC-seq data, read-depth normalized signal for DNase-seq, and normalized read coverage for the RNA-seq experiments. When multiple

replicates were present for an experiment, we preferentially chose the pooled replicate; otherwise, we chose the second replicate if two had been performed, and the first (and only) replicate otherwise. The data were then further processed before model training. First, the signal was downsampled to 25 bp resolution by taking the average signal in each non-overlapping 25 bp window. Second, an inverse hyperbolic sine transformation was applied to the downsampled data. This transformation has been used previously to reduce the effect of ourliers in epigenomic signal [48, 13] and is similar to a log transformation except that it is defined at 0 and is almost linear in the range between 0 and 1. The transformed tracks are used both for training and evaluating the models.

**Calculation of average activity** "Average activity" here refers to the average signal value across a set of training tracks at each position in the genome. Specifically, in the context of three-fold cross-validation, for each fold the average activity is calculated for each assay using the experiments in the other two folds (the training set). This baseline is generally much stronger than the simple average signal value across all loci, which is a more traditional baseline. Formally, the average activity $AA$ for an assay $a$ from the set of all training set experiments of that assay $E$ at position $i$ in the genome is calculated as $AA_{a,i} = \frac{1}{|E|} \sum_{e \in E} e_i$

**Avocado model** We kept the general topology of the Avocado model the same as previous work, but in some experiments we made two modifications to enable the joint modeling of two species. The first modification is that we treated the genomic axis as the concateation of positions from both the mouse and human genome (Figure 3.1). In our experiments, this meant that either mouse chromosome 11 or 19 was concatenated to the ENCODE Pilot Regions from the human genome. The second modification is that the total set of experiments modeled was the union of the mouse experiments and the human experiments. This meant that the biosample axis contained the union of mouse and human biosamples and that the assay axis contained the union of assays performed in mice and humans.

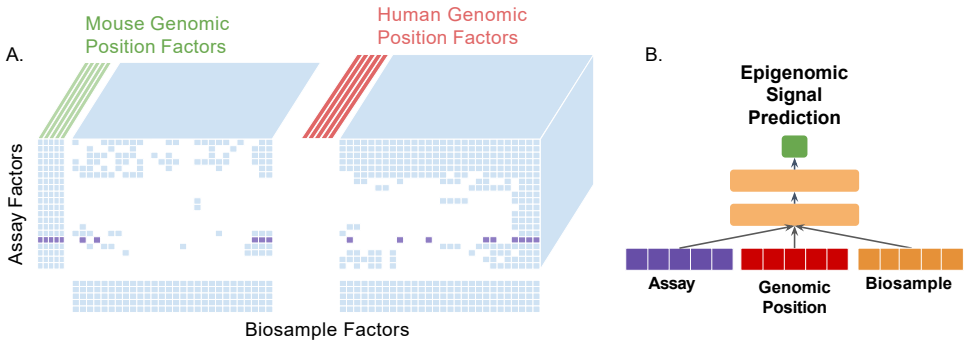



Figure 3.1: **The cross-species Avocado model.** (A) A schematic of the mouse and human compendia, each represented as a 3D tensor of data, aligned on the assay axis. The Avocado model learns latent representations of each dimension of these tensors but maintains a single shared assay representation across both compendia. (B) The neural network component of the model takes in factor values for a single biosample, assay, and genomic position at a time and predicts the signal for that assay in that biosample at that position. Both the biosample and the genomic position factors come either from human or mouse compendia.

These modifications required changes in the training strategy. In its original formulation, Avocado would sequentially sample positions along the genomic axis and randomly select an experiment at each position to train on. However, this strategy would not work with two disjoint sets of experiments that were performed on disjoint sets of loci. Thus, for each genomic position, an experiment was selected at each locus from the set of experiments performed on that locus, i.e., mouse experiments were selected for positions on the mouse chromosome and human experiments were selected for positions in the ENCODE Pilot Regions. This procedure is similar to simply performing a separate factorization for each species except that the assay embeddings and the neural network parameters are tied across species. We also observed empirically that permuting the order of the genomic positions that were sampled, rather than passing over them sequentially, led to better convergence of the model.

**Model evaluation** We evaluate the presented models on only mouse chr19 and chr11 for computation efficiency, where chr19 was the smallest chromosome and chr11 was randomly selected. These evaluations were primarily done using the mean-squared-error (MSE) criterion. For our initial cross-validation experiment we randomly split experiments into three folds. For the protein binding sections, proteins were included or excluded during cross-validation.

**Identification of mm10–hg38 aligned regions** We extracted pairs of regions with high sequence similarity from the mm10ToHg38 chain file (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/liftOver/). This file contains a series of "chains," which are the boundaries of adjacent local alignments. Each chain begins with a line that indicates the starting position for each species of the local alignments. Subsequently lines indicate the locations of each of the aligned regions within the chain. Documentation can be found at https://genome.ucsc.edu/goldenPath/help/chain.html. We took all aligned regions with high sequence similarity that were longer than 500 bp and mapped to the positive

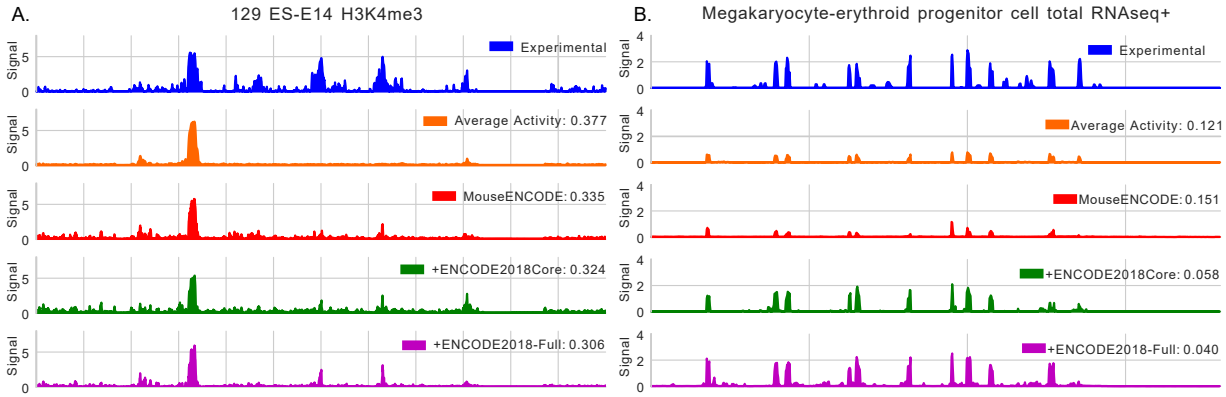



Figure 3.2: **Examples of real and imputed signal.** (A) An example of experimental signal for H3K4me3 in ES-E14 cells, the average activity baseline (orange), and the imputed signal from three different Avocado-based imputation models. The Avocado-based models used only mouse epigenomic data (red), mouse and the human data in the ENCODE2018-Core data set (green), or mouse and the human data in the ENCODE2018-Full data set (magenta). The MSE of each approach for the visualized region is shown in the legend. (B) The same as A except for predicting total RNA-seq in megakaryocyte-erythroid progenitor cells.

strand on both genomes. The requirement for the positive strand is not necessary for our analysis but made extracting the aligned regions easier. This process resulted in 3,188 pairs of regions.

## *3.3 Results*

### *3.3.1 Joint optimization improves imputations in mice*

Our primary hypothesis is that an imputation approach that jointly models epigenomic state in humans and mice will perform better in the less characterized species than an

approach that considers each species independently. Accordingly, our first evaluation involves imputing epigenomic experiments in mice. We used three-fold cross-validation to measure model performance, where the three folds came from partitioning the 1,116 experiments in MouseENCODE2019. While some models were trained using both human and mouse epigenomic experiments, the models were only evaluated on their ability to impute mouse epigenomic experiments.

We evaluated three different Avocado models that were trained using different numbers of epigenomic tracks. The first model was trained using only epigenomic experiments from mice and so represented model performance on the standard imputation task. The second model was trained using the joint optimization procedure (see Methods for details) on the same amount of mouse data as the first model but also the smaller of the two human data sets. The third model is similar to the second model except that it was trained using the larger of the two human data sets. We train models using two human data sets to assess the effect that including additional human experiments has on performance. Importantly, when performing cross-validation, models that include human epigenomic data are given access to the entirety of the human data sets for each fold.

As a baseline for the imputation approaches described above, we calculated the average activity of each assay (see Methods for details). The average activity for an assay is the average signal at each position exhibited by the training set experiments that are of that assay. The average activity baseline represents a simple rule that regions of consistently high or low signal in the training set will exhibit similar behavior in the test set [69]. Accordingly, improvement over the average activity baseline generally indicates the prediction of cell-type specific activity.

A visual inspection of the imputations made during cross-validation (Figure 4.1) showed that, consistent with prior work on epigenomic signal imputation, a source of error was locations that are peaks in some, but not all, biosamples [67]. These locations can be

| MSE | Overall | Histone Modifications | Protein Binding | Transcription | Accessibility |
|---|---|---|---|---|---|
| Average Activity | 0.10030 | 0.13021 | 0.10957 | 0.00296 | 0.05512 |
| Mouse Only | 0.07125 | 0.09199 | **0.07628** | 0.00260 | 0.04618 |
| Mouse + ENCODE2018-Core | 0.06992 | 0.09004 | 0.07654 | 0.00217 | **0.04542** |
| Mouse + ENCODE2018-Full | **0.06986** | **0.08984** | 0.07683 | **0.00209** | 0.04617 |

Table 3.1: **Imputation performance with and without including human epigenomic data.** The mean squared error (MSE) computed both overall across all experiments and for each of the four main forms of biological activity in our data set. For each measure, the score for the best-performing model is in boldface.

identified as those where the average activity is lower than experimental signal, i.e. those where the signal in this biosample is above average. When we visualized the imputations of H3K4me3 in HS-E14 cells (Figure 4.1A) and the imputation of total RNA-seq in progenitor cells (Figure 4.1B), we observed that the model trained on mouse data alone generally makes imputations that are similar, but not identical, to the average activity. In contrast, the models trained using human data appeared to make more accurate predictions at these biosample-specific regions, suggesting that our joint training procedure reduces the risk of fitting too closely to the average activity in the less well characterized species.

We then comprehensively calculated the overall performance of each of the models during cross-validation using MSE. The two models that leveraged human epigenomic data sets produced a small but consistent improvement in performance in comparison to the model that used only mouse epigenomic data (Table 3.1, Wilcoxon signed-rank test p-values of 5.9e-16 and 7.4e-17 respectively). Overall, using more human epigenomic data, in the form of ENCODE2018-Full, gave a larger decrease in MSE relative to the model trained using only mouse epigenomic data than using only the ENCODE2018-Core data set. Proportionally, the largest improvement is observed in the experiments that measure transcription, likely

because the human epigenomics data sets contain a large number of transcription-measuring experiments. A similar trend was observed when building models that predict epigenomic state across species using nucleotide sequence alone [76]. Interestingly, we observe a small decrease in performance at predicting protein binding when using human epigenomics data. Potentially, this decrease in performance could arise from proteins whose binding affinities differ in the human and mouse genomes. We found that the greatest improvement came when imputing the histone modification H3K79me2, which had been assayed 6 times in mice and 37 times in humans, with a decrease from 0.128 MSE when only using mouse epigenomic data to 0.108 MSE when also using the ENCODE2018-Full data set. Conversely, we found that the greatest decrease in performance came from imputing the binding of ZC3H11A, which has been assayed only 3 times in mice and 2 times in human, with an increase from 0.041 to 0.049 MSE. These results support our hypothesis that joint modeling can yield improved performance when an assay is sparsely characterized in mice but well characterized in humans.

### *3.3.2 Joint optimization enables zero-shot imputations*

Encouraged that our joint training procedure led to an improvement in overall performance, we hypothesized that the same procedure could allow a model to make predictions in one species for assays that have only been performed in the other species. We refer to this as the "zero-shot" setting because the model has no training data in mice for the assays that it is making imputations of. This setting is particularly relevant because there are 735 assays, most of which measure protein binding, that have been performed in humans that have not yet been performed in mice.

We investigated the zero-shot performance of models trained in two related settings, both of which involved imputing protein binding. The first setting involved holding out the experiments from some, but not all, protein binding assays performed in mice, while keeping

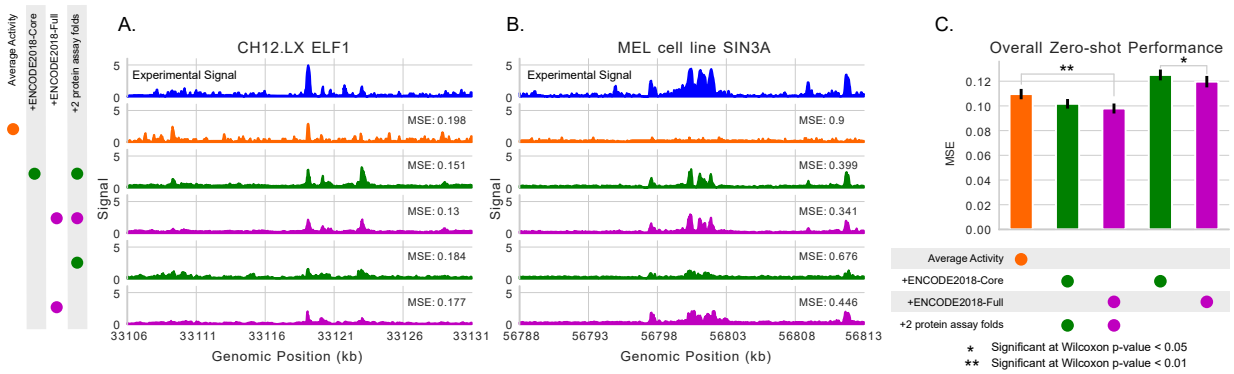



Figure 3.3: **Examples and evaluation of zero-shot imputations.** (A) The experimental signal, average activity, and imputed signal of four models for binding of the protein ELF1 in CH12.LX. The MSE for each approach compared to the experimental signal in the displayed window is also shown. The legend to the left shows the set of experiments used to train each model, with the top two using human experimental data as well as two of the three folds of proteins in mice, whereas the bottom two only use human experimental data. (B) The same as (A) but measuring the binding of the protein SIN3A in the MEL cell line. (3) The performance of each approach overall on 140 tracks of protein binding data on chr19 and chr11. Two statistically significant relationships are shown.

the experiments that were performed in humans. In this setting, protein binding assays, rather than experiments, were divided into three folds for cross-validation. The second setting, which was more challenging, involved holding out all protein binding experiments in mice. While the first setting is the more realistic one, because some protein binding experiments have already been performed in mice, the second setting allows us to investigate the extent to which the tied assay and neural network parameters can be utilized. We focused our evaluations on imputing protein binding experiments due both to the importance that protein binding plays in regulating gene expression and because there are many proteins whose binding has been characterized in humans but not in mice.

We found that these models were capable of outperforming the average activity baseline even in the zero-shot setting. Notably, models in both settings were capable of imputing signal peaks both at regions with high average activity (Figure 3.3A) and low average activity, indicating biosample-specific peaks (Figure 3.3B). Models trained using the ENCODE2018-Core data set yielded imputations that beat the average activity baseline from the previous cross-validation, and models trained using the ENCODE2018-Full data set yielded even better performance. These results indicate that the model managed to learn not only the general locations where protein binding occurs, which is measured by the average activity, but also biosample-specific signal. This result is particularly striking because the model has not been exposed to nucleotide sequence, motif presence, or examples of that particular protein binding in other mouse cell types. Thus, the model is able to discern where particular proteins bind both from their epigenomic context and from the binding profiles of other proteins included in training.

### *3.3.3 Unified genomic embeddings identify epigenomic divergence*

A key property of Avocado is that the learned latent representations encode a compression of epigenomic information that is useful for more than just imputing epigenomic signal. In

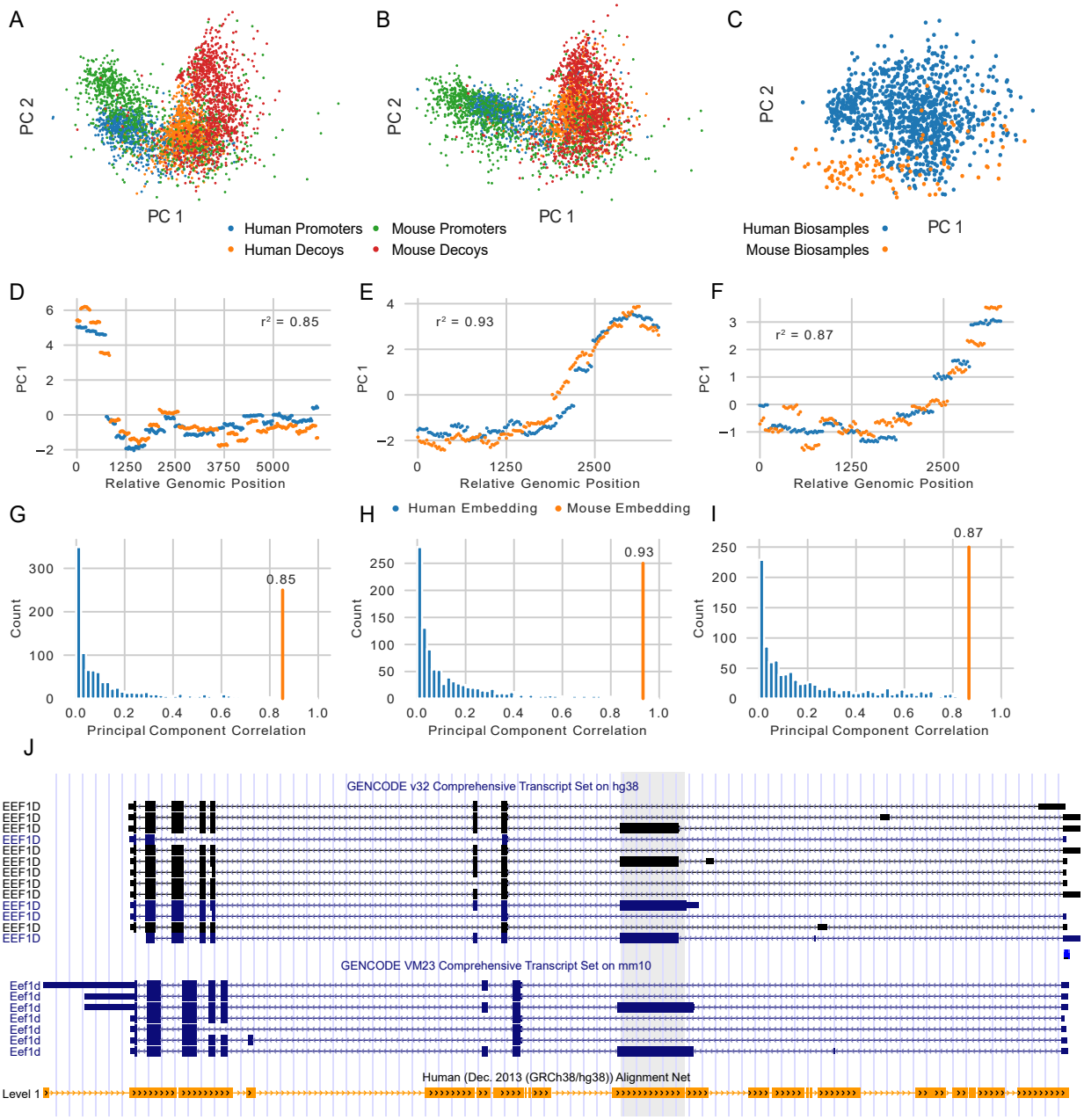



Figure 3.4: **Genome embeddings across species.** (A) The first and second principal components of a PCA projection of the Avocado embeddings (excluding 5 kbp factors) for human promoters and decoy regions are shown superimposed on those components for mouse promoters and decoy regions. (B) The same as (A) except as a single PCA projection of both human and mouse regions instead of separate PCA projections per species. (C) A projection of the biosample embeddings learned from the original human model and from the mouse model. (D-F) The first principal component of the Avocado embeddings (excluding 5 kbp factors) for three pairs of regions whose sequences align across mice and humans. (G-I) The

initial work where we trained Avocado using the Roadmap compendium we showed a clustering of biosamples by anatomy type in the learned biosample representation and distinct clusters of promoters and enhancers in the learned genome embedding [67]. We hypothesized that training Avocado using epigenomic data from multiple species would result in representations of epigenomic context that were comparable regardless of the number or composition of epigenomic experiments available for each species. Learning comparable representations in such a manner would greatly simplify the epigenome alignment problem [77, 78].

We tested this hypothesis by extended the strategy proposed in our initial work for training Avocado to the setting of modeling multiple species. As before, we began by fitting a model to the ENCODE2018-Full data set using only the ENCODE Pilot Regions. Then we froze the biosample and assay embeddings as well as the neural network parameters and fit only the genome embeddings for each human chromosome. At this point, we add in the additional step of taking the frozen assay embedding and neural network parameters and training the biosample and genome embeddings for each mouse chromosome individually.

To assess whether the learned representations were comparable across species, we first inspected promoter regions in both the mouse and human genomes. We took 1,000 random protein coding genes from the GENCODE v33 annotation file and an equal number of random protein coding genes from the GENCODE M23 annotation file for mm10 and extracted the corresponding $\pm$1kb windows surrounding the respective transcription start sites. As a reference for distance in the projected space, for each promoter we also extracted a "decoy" region that was on the same chromosome and the same size but at a randomly selected position. When we projected all 4,000 of these regions using principal component analysis (PCA), we observed that while the promoters were separated from decoy regions, neither the promoters nor the decoy regions aligned across speices (Figure 3.4A). We suspected that this may be because the mouse chromosomes had two free axes being fit simultaneously (genomic position and biosample) and so a linear shift occured across both axes. To factor

this effect out, we ran PCA separately for the examples from each species, and observed a much better degree of alignment between both the promoter and the decoy regions across species (Figure 3.4B). A projection of the biosample embeddings shows that those from the human and mouse models do not overlap, indicating that a shift had occured (Figure 3.4C). Overall, these results suggested that, despite this shift, our approach led to similar structure being encoded in the learned genome representations.

Having shown that similar structure is being encoded across species in the genomic representations, we used the representations to develop a measure of epigenomic similarity between a pair of regions. Our procedure for calculating similarity involves first defining a pair of regions, one in human and one in mice, that are of equal size. Then, similar to our approach for aligning promoter and decoy regions, we extract the first principal component separately for each of the two regions. The principal component can serve as a coarse-grained summary of epigenomic activity at each genomic position in the region. Finally, we calculate the squared Pearson correlation between the two principal components—squared because the sign can flipped for principal components—and use this correlation as our measure of similarity.

This similarity measure is complementary to a sequence-based similarity measure, e.g. alignment score, because the genome representations were not learned using nucleotide sequence. Thus, a natural comparison is to compare the epigenomic and sequence-based similarity scores for pairs of regions. Because it is likely that regions that have dissimilar sequence will also have dissimilar epigenomic state, we focus our search on regions with high sequence similarity. We extracted 3,188 aligned regions from the mm10ToHg38 chain file (see Methods for details) and began by visualizing the first principal component for three of the longest aligned regions (Figures 3.4B-D). We note that the principal components appear to overlap well for all three of the displayed regions and that the squared Pearson correlation values are high. However, to make sure that these values are significant, we perform a permutation test by calculating the epigenomic similarity score between the same human region and 1000

randomly selected and equally sized regions on the same mouse chromosome. For each of the three regions, we see that the epigenomic similarity score on these random regions is generally much smaller (Figures 3.4E-G), indicating that the similarity scores at these aligned regions are statistically significant (p-value $< 0.01$).

Next, we consider the more interesting setting where sequence-based similarity and epigenomic similarity differ. This form of analysis may be particularly useful for identifying the large number of orthologous regions that have functionally diverged [79]. To automatically find these regions, we calculate the epigenomic similarity score between all 3,188 pairs from the same set. We identified a representative pair of regions with low score (0.063) that corresponded to a region on chain 9 that mapped chr8:143,589,453-143,589,975 on hg38 to chr15:75,903,147-75,903,669 on mm10. This pair of regions corresponds to the largest exon on the EEF1D gene (Eef1d on mice), which encodes a protein that is a subunit of elongation factor 1 involved in transferring aminoacyl tRNAs to the ribosome [80]. Interestingly, while this exon is alternatively spliced in both humans and mice, it only serves as the initial promoter for a splice variant in humans, suggesting a difference in regulation that is not conserved between species.

## *3.4 Discussion*

In this work, we introduce and analyze an optimization approach for training imputation models like Avocado by jointly modeling human and mouse epigenomics. Our experiments show that the impact of this extension can be significant. Not only does leveraging the large number of human epigenomic experiments result in improved imputations of mouse epigenomics, but the same procedure allows for zero-shot imputations to be made for assays that have been performed in human but not in mouse.

A straightforward application of these imputations will be to help prioritize the order of future epigenomic experiments in mice. We have previously described an approach for

experiment prioritization based on minimizing redundancy within imputed versions of the experiments [81]. Applying that approach to the imputations generated here would likely be even more valuable because they include proteins that have never been assayed in mice. Thus, the ordering would reflect not only what further biosamples to perform assays in, but what assays should even be performed.

Our approach assumes that the assays measure phenomena that are conserved across species. While this is generally true for evolutionarily related species, such as humans and mice, it is not always true. A source of error in the imputations may come from cases where these forms of activity differed between mice and humans. Potentially, instead of using the same assay representations across species, one could use a soft-tying approach where the model learns species-specific assay representations that are regularized to be similar to each other. On the other hand, careful analysis of errors made by the current approach may yield a data-driven way of identifying what forms of activity differ across species and where they differ along the genome. Regardless, it will be important to keep in mind that this procedure will likely need to be modified to work well on evolutionarily distant pairs of species.

Some proteins that are co-factors or from the same family exhibit similar binding profiles along the genome. We intentionally did not take these relationships into account when constructing folds for our zero-shot imputation setting. Our reasoning was that it is not the case that these sets of proteins (co-binders or families) are assayed together and so it is not a realistic assumpton that experimental data will or will not be available for all members. Rather, it is generally the case that the binding of some proteins in these sets have been assayed and a researcher is interested in imputing the binding behavior of the other family members. Indeed, for the general task of imputing the binding of proteins, it would be expected that an algorithm would make use of close relationships. One would be disappointed to, for instance, give a model an experiment profiling MYC binding and see poor performance at predicting MAX binding.

An element of this work that deserves further inspection is the process for learning representations when jointly modeling mouse and human epigenomics. Our preliminary work resulted in an epigenomic similarity score that can measure the degree to which a pair of regions are epigenomically similar. Because the epigenomic similarity score is not directly based on nucleotide sequence, it serves as a complementary measure to sequence similarity. Indeed, we found that some pairs of regions with aligned sequence have a very low epigenomic similarity score, suggesting a divergence in function in these regions. A comprehensive analysis of these regions will likely help researchers better identify and understand regions whose function have diverged. We anticipate that designing optimization strategies that result in improved comparability of the representations across species will be an important piece of future work.

Imputations, models, and latent factors produced by this project will be made freely available at https://github.com/jmschrei/avocado.

# Chapter 4

# AVOCADO CAN HELP PRIORITIZE EXPERIMENTAL EFFORTS

## *4.1 Background*

Experimental characterization of the genomic and epigenomic landscape of a human cell line or tissue ("biosample") is expensive but can potentially yield valuable insights into the molecular basis for development and disease. Fully measuring the epigenome involves, in principle, assaying chromatin accessibility, transcription, dozens of histone modifications, and the binding of over a thousand DNA-binding proteins. Even after accounting for the decreasing cost of high-throughput sequencing, such an exhaustive analysis is expensive, and systematically applying such techniques to diverse cell types and cell states is simply infeasible. Essentially, we cannot afford to fill in an experimental data matrix in which rows correspond to types of assays and columns correspond to biosamples.

Several approaches have been proposed to address this challenge. Some scientific consortia, such as GTEx and ENTEX, aim to completely fill in a submatrix of selected assays and biosamples. In contrast, other consortia, such as the Roadmap Epigenomics Mapping Consortium [61] and ENCODE [8], adopted a roughly "L"-shaped strategy, in which consortium members focused on carrying out many assays in a small set of high-priority biosamples, and some assays were carried out over a much larger set of biosamples. Recently, computational approaches have been proposed that rely on using machine learning models to impute the experiments that have not yet been performed [14, 13, 67, 52]. While the imputation strategy can relatively easily complete the entire matrix, a drawback is that the imputed data is potentially less trustworthy than actual experimental data.

In this work, we[1] address a variant of the matrix completion problem. Specifically, we consider the scenario that we, as a field, find ourselves in currently, having performed many assays in many biosamples and trying to figure out which of the remaining assay/biosample combinations ("experiments") we should perform next. In many cases, the choice of which experiments to perform is driven by intuitition and guesswork. We hypothesize that a data-driven approach to this problem can increase the rate of scientific discovery.

Previous work by Wei et al. [82] has addressed a closely related problem. Wei et al. studied the problem of filling in a new row (or column) of the matrix. Say that we have decided to begin performing experiments on a new biosample, but we can only afford to carry out a fixed number $k$ of assays. Then the question is, "Which set of $k$ assays is likely to yield the most information?" Wei et al. answer this question by framing the task as an optimization problem, where we attempt to maximize a function $f(\cdot)$ that quantifies the joint quality of a given subset of size $k$ relative to the full collection of possible assays.

An important feature of the method proposed by Wei et al. is that the approach is "cell-type agnostic," in the sense that it yields a single set of suggested assays, irrespective of what biosample will be analyzed. This property arises because the set quality function $f(\cdot)$ measures the similarity of a given pair of assays by averaging across all cell types in which both assays have already been performed. Wei et al. explicitly consider the scenario in which a specified set of assays has already been performed in a given biosample, and the task is to select the next $k$ assays to perform. However, even in this setting, the proposed approach is cell-type agnostic: the method yields the same answer for any biosample in which the specified set of assays has been performed.

In this work, we propose a method that can select experiments that span a diverse set of

[1]The work in this chapter is based off a paper entitled *Prioritizing transcriptomic and epigenomic experiments by using an optimization strategy that leverages imputed data* that has been submitted to *Bioinformatics* that was written by myself, Jeffrey Bilmes, and William Stafford Noble (in the order that authors appear on the paper). In this work, WSN and myself conceived of experiments, I did the coding, I did the analysis, and myself and WN contributed to writing the text.

biosamples and assays jointly. The key idea is to apply the quality function $f(\cdot)$ to similarities calculated using imputed, rather than real, data. The resulting method, implemented in a software package called "Kiwano," is far more flexible and powerful than the original method. Most importantly, rather than restricting our selection to a single row or column of the data matrix, using imputed data allows us to address the global question, "Among all possible experiments, which one should I do next?" Furthermore, even in the case where we want to select assays to perform in a given biosample, our imputation-based approach selects a set that is tailored to this particular biosample, by computing similarities between the imputed values for potential experiments and experiments that have already been performed in that biosample.

We validate Kiwano in several ways, using ENCODE data. First, we illustrate via visualization that the imputation-based similarity matrix encodes meaningful biological relationships among assay types and biosamples. We then apply the optimization procedure to this similarity matrix and show that the resulting subset of experiments is representative of the full set, both qualitatively and through simulation experiments. We also illustrate how to apply the objective function used in our optimization to ascertain which biosamples are currently undercharacterized or which assays are underutilized. We have made a tool available at https://www.github.com/jmschrei/kiwano/ that can order experiments based on the pre-calculated similarity matrix we use here.

## *4.2 Methods*

### *4.2.1 Submodular optimization and facility location*

Submodular optimization is the discrete analog of convex optimization and operates on submodular set functions. A function is submodular if and only if it has the property of diminishing returns; i.e., the incremental gain in function value associated with adding an element s to a set A becomes smaller as the size of the set A becomes larger. More formally,

given a finite set $S = \{s_1, s_2, ..., s_n\}$, a discrete set function $f : 2^S \rightarrow R$ is submodular if and only if

$$f(A \cup s) - f(A) \geq f(B \cup s) - f(B), \forall A \subseteq B \subset S, s \notin B.$$

In this work, we employ a submodular function whose value is inversely related to the redundancy within a given set. Thus, optimizing such a function, subject to a cardinality constraint, involves identifying the subset whose elements are minimally redundant with each other. For further reading on submodular optimization, we suggest [83, 84, 85].

Kiwano relies on optimizing a particular submodular function called facility location. Facility location takes the form

$$f(X) = \sum_{y \in Y} \max_{x \in X} \phi(x, y) \tag{4.1}$$

such that $Y$ is the full set of experiments, $X$ is the selected subset of experiments such that $X \subseteq Y$, $x$ and $y$ are individual experiments in $X$ and $Y$ respectively, and $\phi(x, y)$ is the squared correlation between $x$ and $y$. The facility location function is optimized using the accelerated greedy algorithm [86]. We use apricot v0.3.0 to perform this selection [87].

### *4.2.2 Model training*

The models used for evaluation are implemented using keras (v2.2.4) [53] with a Theano (v1.0.4) [54] backend. Each model is a multi-task linear regression where a single weight matrix is learned that transforms inputs into outputs. Training is performed using the Adam optimizer [56] with a mean-squared-error loss. The optimizer hyperparameters and the weight initializations are set to the keras defaults with no explicit regularization.

### *4.2.3 Datasets*

We generated our imputations using an Avocado model that had previously been trained on the ENCODE2018-Core data set [52]. The model is available at https://noble.gs.washington.edu/ jm-

schr/mango/models/. This model was trained on 3,814 experiments across 400 biosamples and 84 assays where the signal was $-\log_{10}$ p-values that had subsequently been arcsinh transformed to reduce the effect of outliers. The resulting imputations are in the same space. Due to the large size of the genome, we only imputed the ENCODE Pilot Regions, comprising $\sim$ 1% of the genome [58], for each of the 33,600 potential experiments. This 1% is comprised of a handful of manually selected regions that were deemed of particular biological interest, combined with a randomly selected set of 30 1-Mb regions that systematically vary in terms of gene density level of non-exonic conservation.

An important detail is that, at the time of accession, experiments meausuring transcription had been divided into plus-strand signal and minus-strand signal on the ENCODE portal. Consequently, each strand was counted as a separate assay when training the Avocado model. While the strand that transcription occurs on is important for an imputation approach to capture, this distinction is not helpful for prioritizing experiments because one generally cannot perform an experiment measuring transcription on only one of the strands. Thus, we combine the plus- and minus-strand experiments for both the imputed and the primary epigenomic data by simply adding the tracks together. This process reduced the total number of assays from 84 to 77, the total number of performed experiments from 3,814 to 3,510, and the total number of potential experiments from 33,600 to 30,800.

## *4.3 Results*

### *4.3.1 Imputations cluster according to known biological patterns*

Our approach for prioritizing experimental characterization relies on a similarity matrix that is calculated on imputed experiments. To produce this matrix, we first generated imputations of epigenomic and transcriptomic experiments using a recently developed imputation approach based on deep tensor factorization, named Avocado. These imputations span 400 human biosamples and 77 assays of biological activity for a total of 30,800 imputed tracks.

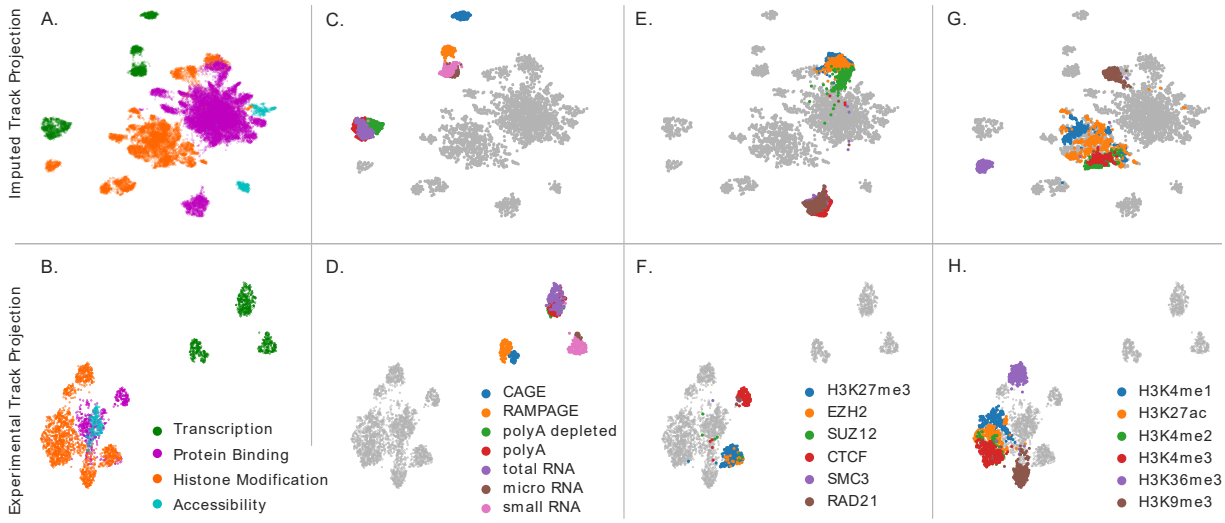



Figure 4.1: **A projection of imputed and experimental epigenomic tracks.** Each panel shows a UMAP projection of 30,800 imputed experiments (top row) or of 3,150 tracks of primary data (bottom row). In each column, a different set of experiments is highlighted based on their biological activity. (A/B) Experiments are highlighted based on broad categorization of the assayed activity. (C/D) Transcription measuring experiments are colored according to different types of assays. (E/F) Experiments are highlighted that measure H3K27me3 and two polycomb subunits, as well as CTCF and two cohesin subunits. (G/H) Experiments are highlighted showing several histone modifications that are enhancer-associated, such as H3K4me1 (blue) and H3K27ac (orange), promoter-associated such as H3K4me2 (green) and H3K4me3 (red), transcription-associated such as H3K36me3 (purple), or broadly repressive such as H3K9me3 (brown).

After acquiring these imputations, we calculated the squared Pearson correlation between all pairs of imputed experiments for use as a similarity measure, resulting in a 30,800 by 30,800 matrix. For efficiency, these correlations were computed with respect to the ENCODE Pilot Regions [58], comprising 1% of the genome.

After calculating the similarity matrix, we investigated whether the similarity matrix was able to capture high level biological trends that would be crucial for prioritization. We began by visually inspecting a two-dimensional UMAP projection [20] of the similarity matrix down to two dimensions (Figure 4.1A). The clearest trend in this projection is a separation of experiments based on a broad categorization of the type of activity measured by the assay. We observed that one cluster contained mostly protein binding experiments, one contained mostly histone modification experiments, and several neighboring clusters were composed exclusively of transcription-measuring experiments. Initially, one might expect that experiments in the same biosample where the assays measure the same underlying phenomena might cluster together. However, we observed that in some cases a pair of experiments may exhibit low correlation when the shape of their signals along the genome differ, even when the assays used in the experiments both measure the same underlying biological activity. For example, the histone modification H3K36me3 is known to be associated with transcription but generally forms broad peaks across the entire gene body, whereas assays such as CAGE or RAMPAGE form punctate peaks.

In order to confirm that the separation according to assay categorization was not an artifact of the imputation process, we used the same process to calculate a similarity matrix and subsequent UMAP projection for the 3,150 tracks of the experimental (or "primary") data (Figure 4.1B). The major trends present in the projection of imputed data are consistent with those in the primary data. In particular, transcription experiments form distinct clusters, protein binding experiments are mostly distinct from histone modification ones, and chromatin accessibility experiments localize closer to protein binding experiments than to

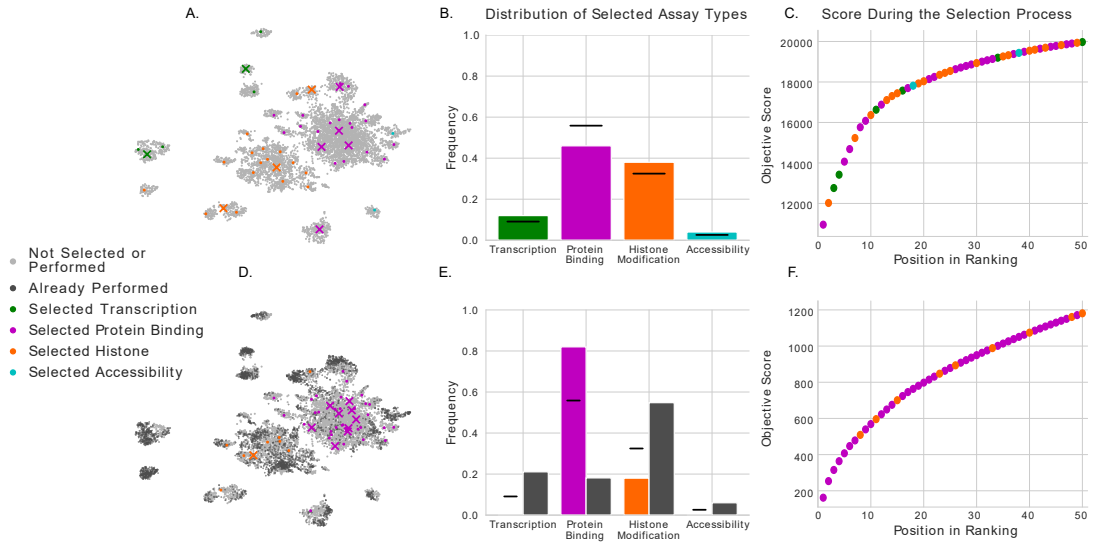



Figure 4.2: **A selection of experiments before and after accounting for those that have already been performed.** (A) The same projection of imputed experiments as shown in Figure 4.1A, where the first 50 experiments selected using Kiwano are colored by the type of activity that they measure. The first 10 experiments selected are marked using an X, and the remaining 40 are marked with a dot. (B) A bar chart showing the frequency that experiments of each type of activity are selected in the first 50 experiments. (C) The facility location objective score as the first 50 experiments are selected, with each point colored by the type of activity measured by that experiment. (D) The same as (A), but with the selection procedure initialized with the experiments that have already been performed, and with those experiments displayed in dark grey. (E) The same as (B), but with dark grey bars showing the frequency of experiments of each type that have already been performed. (F) The same as (C), but with the selection procedure initialized with the experiments that have already been performed.

histone modification experiments. Note that although the figure may appear to show that accessibility experiments overlap with protein binding experiments, a closer examination reveals that the protein binding experiments mostly surround the accessibility experiments.

Next, we more closely examined four sets of assays that, a priori, we expected to show distinctive patterns. The first set of experiments were those that measured transcription. When we highlighted experiments by assay type, we observed CAGE and RAMPAGE experiments forming distinct cluters, micro- and small-RNA-seq experiments forming a third cluster, and polyA-, polyA-depleted-, and total-RNA-seq experiments forming a fourth (Figure 4.1C–D). The second and third sets of experiments involved triplets of assays whose activity are usually associated, specifically, with CTCF and the cohesin subunits, SMC3 and RAD21, as well as H3K27me3 and two polycomb subunits, EZH2 and SUZ12 (Figure 4.1E–F). In both cases we observe distinct clusters of experiments, which is particularly interesting for H3K27me3 and the polycomb subunits because one assay measures a histone modification and the other two measure protein binding. The fourth set of experiments focused on six well-studied histone modifications (Figure 4.1G–H). The clustering of these six marks coincide with the genomic element in which they are typically enriched. In particular, experiments measuring H3K36me3 and H3K9me3 form their own clusters, with the two assays respectively measuring activity enriched in gene bodies and constitutive heterochromatin. Further, the primary cluster of histone modification experiments exhibited a separation between the promoter-associated marks, H3K4me2 and H3K4me3, and the enhancer-associated marks, H3K4me1 and H3K27ac. We observed similar patterns across both the imputed and primary data for each of these four sets of assays. Taken together, these observations suggest that a similarity matrix derived from imputed experiments is successfully capturing important aspects of real biological activity.

### *4.3.2 Submodular selection of imputations flexibly prioritizes assays across cellular contexts*

Having shown that the similarity matrix captures several high level trends in the data, we turn to the task of experimental prioritization. Our strategy for prioritizing experiments relies on submodular selection, which is a technique for reducing a set of elements to a minimally redundant subset through the optimization of a submodular function that captures the quality, or "representativeness," of a given subset relative to the full set (see Methods for details). Submodular selection has been used previously to select genomics assays [82], to select representative sets of protein sequences [88], and to choose genomic loci for characterization by CRISPR-based screens [89]. Specifically, we optimize a "facility location" objective function, which operates on pairwise similarities between elements and so is well suited to leverage our similarity matrix (see Methods). A critical property of submodular functions is that greedy optimization will yield a subset whose objective value is within $1 - e^{-1}$ of the optimal subset, and that this is the best approximation one can make unless P=NP [90]. This greedy optimization algorithm iteratively selects the single element whose inclusion in the representative set leads to the largest gain in the objective function. Thus, when applied to our similarity matrix, the submodular selection procedure will yield an ordering over all experiments that attempts to minimize redundancy among those experiments that are selected early in the process.

In order to demonstrate that submodular selection results in a representative subset of assays, we applied it to our calculated similarity matrix. Visually, we observe that the first 50 selected experiments appear to cover the space well and include selections from many of the small clusters of experiments (Figure 4.2A, Supplementary Table 1). When we count the number of assays selected for each type of biological activity, we find that protein binding assays are the most commonly selected with 23 experiments, followed by histone modification assays with 19 experiments, transcription assays with 6 experiments, and, finally, accessibility assays with 2 experiments (Figure 4.2B). However, when we compare the number of selected

experiments of each type to the number that one would expect by randomly selecting with replacement, we observe that protein binding experiments are underrepresented, whereas histone modification experiments are overrepresented. We note that the first 10 experiments are at the centers of large clusters of experiments and that the subsequent 40 experiments are selected from smaller clusters. This finding corresponds with the gain in the facility location objective score from each successive experiment significantly diminishing by the tenth experiment (Figure 4.2C).

A weakness in simply applying submodular selection to the full set of imputed experiments is that the procedure does not account for the thousands of epigenomic and transcriptomic experiments that have already been performed. Fortunately, there are two ways that one can account for these experiments. The first is to remove those experiments that have already been performed from the similarity matrix and perform selection on the remaining experiments. While this approach is simple, it does not account for the content of the experiments that have already been performed. For example, if transcription has already been measured in hundreds of biosamples, then it may be beneficial to focus experimental efforts on characterizing other types of biological activity. A second approach takes advantage of the fact that the selection process is greedy by initializing the set of selected experiments with those that have already been performed. This ensures that the selected experiments cover types of activity that are not already well characterized.

Accordingly, we proceeded with the second approach. We initialized a facility location function with the 3,150 experiments that had already been performed and ranked the remaining 27,650 experiments. We observed that the selected experiments lie primarily in areas of the UMAP projection that do not already have many experiments performed (Figure 4.2D, Supplementary Table 2). When we counted the number of selected experiments of each type, we found that the number of protein binding experiments increased from 23, when not accounting for the experiments that had already been performed, to 41, when

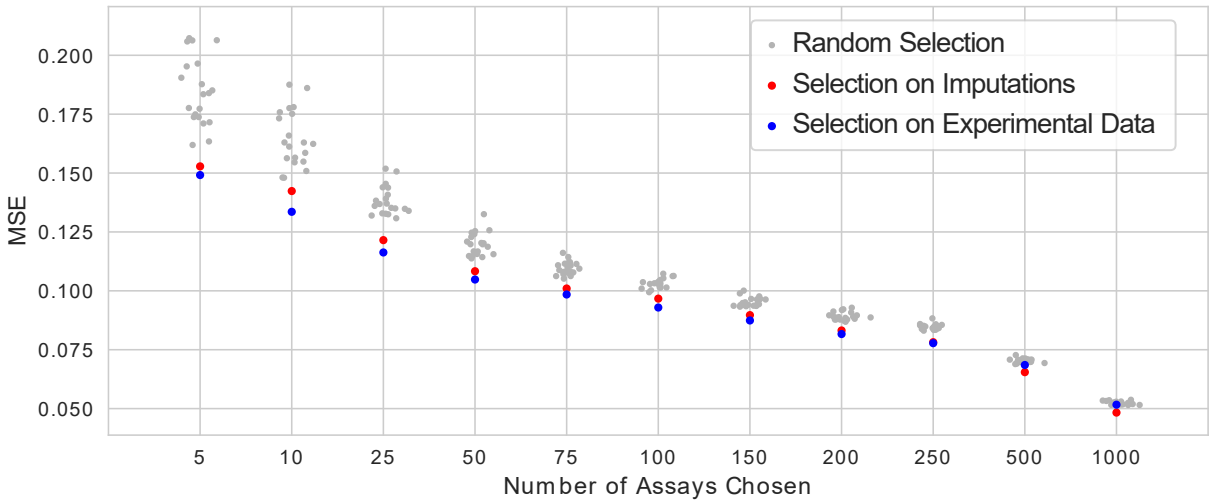



Figure 4.3: **Imputation performance using different panels of assays** The performance of regression models (in terms of mean-squared-error, MSE) as a function of the number of assays chosen as the input. These panels range in size from 5 assays to 1000 assays, and are selected either randomly (grey), through a facility location function applied to imputed experiments (red), or through a facility location function applied to experimental data (blue).

accounting for them (Figure 4.2E). Correspondingly, the number of histone modification experiments decreased from 19 to 9. This change in coverage is likely because 1,726 experiments measuring histone modification have already been performed, whereas only 571 experiments measuring protein binding have been performed. Further, none of the first 50 selected experiments measure transcription or accessibility, likely because those forms of activity are already much better measured than protein binding. In this setting, the gain in the facility location objective function of each successive experiment is much lower, due in large part to the experiments that have already been performed (Figure 4.2F).

### *4.3.3 Selection on imputed experiments identify diversity in primary data*

Our next step was to evaluate the quality of the selected experiments in a quantitative way. Following Wei et al; we reasoned that the signal contained in a representative subset of experiments would be well suited for reconstructing the signal in all experiments. We formulated the problem of quantitatively measuring how representative a subset is as a multi-task regression problem, with the input features being the signal from the selected subset of experiments and the outputs being the signal from the full set of experiments (see Methods). Importantly, to ensure that this validation measured how representative a subset is of the primary data, despite subset selection having been performed on the imputations, we used the primary data as both the input and target for this task.

We selected a subset of experiments in three ways. The first was through the submodular selection procedure described in Section 4.3.2, applied to the 3,150 imputed experiments for which primary data had already been collected. The second was by applying the submodular selection procedure to the 3,150 tracks of primary data themselves. Naturally, selecting subsets based on the primary data cannot be extended to experiments that have not yet been performed, and so the purpose of evaluating models trained using this subset is to measure the effect that the imputation process itself has on selecting a representative subset

of experiments. The third was selecting subsets of the 3,150 performed experiments at random. This random process was repeated 20 times to obtain a distribution of scores.

We observe that the subsets of experiments selected using submodular selection consistently outperform those selected at random (Figure 4.3). Each comparison is statistically significant at a p-value threshold of 0.01 according to a one sample t-test. Further, for smaller subsets, applying submodular selection to the imputed tracks performs nearly as well as the panels selected on the primary data itself, showing that the distortion introduced by the imputation process is small. Interestingly, when the subsets become much larger, those selected using imputed tracks appear to outperform those selected using the primary data. This trend may arise because imputed tracks can serve as denoised versions of the primary data [14]. At the beginning of the selection process, this denoising is not necessary to select experiments that are very different from each other. However, once many experiments have been selected, the denoised experiments may be better at identifying real differences between experiments.

### *4.3.4 Prioritization can be performed for individual biosamples or assays*

A potential weakness in selecting experiments across both biosamples and assays is that, because the selection process is driven to select experiments with very different signal profiles, differences in the shape of the signal from each assay may dominate over differences in meaningful biology. This phenomenon is reflected in the observation that assay type is the predominant determinant of location in the UMAP embedding (Figure 4.1A). For example, although a difference in expression of a small number of important genes across two biosamples may be critical for certain cellular processes, the resulting assay signals are likely still more similar to each other than to an assay that measures histone modification. Thus, there may be cases where it is useful to focus prioritization efforts on a particular assay or biosample in order to factor out these effects. In our proposed approach, both can be

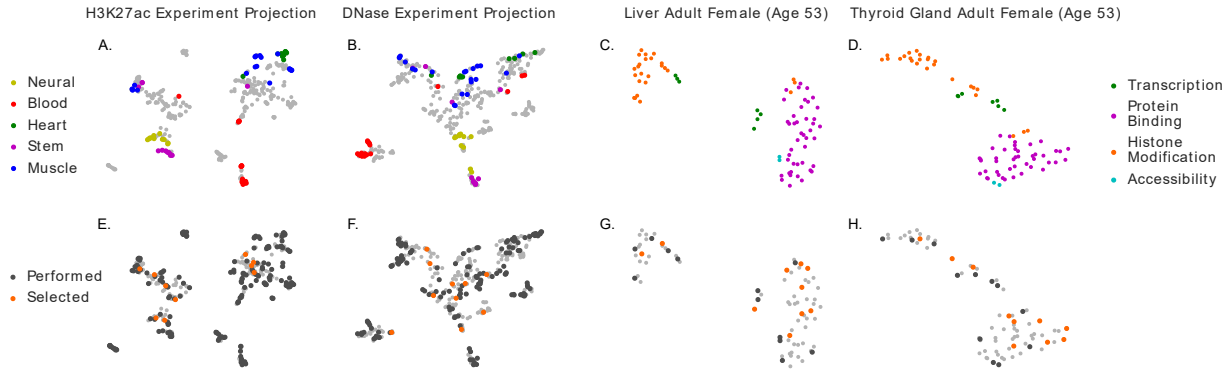



Figure 4.4: **Projections and selections of all experiments containing a specific biosample or assay.** UMAP projections for sets of experiments that each contain a particular biosample or assay. (A) A projection of H3K27ac experiments in all 400 biosamples, with some experiments colored according to anatomy type. Not all experiments are colored because ENCODE biosamples do not have anatomies assigned to them, and only some could be unambiguously determined. (B) Same as (A), but with DNase experiments. (C) A projection of all assays performed in liver biosample, with assays colored by activity type. (D) Same as (C), but in a thyroid gland biosample from the same individual. (E-H) The same projections as (A-D), but performed experiments are colored in dark grey, the next 10 selected experiments are colored in orange, and experiments that are not selected and have not yet been performed are colored in light grey.

accomplished by simply prioritizing different subsets of experiments.

First, we considered the task of prioritizing the order of biosamples in which to run a given assay. We focused on H3K27ac, a histone modification that is enriched in active enhancer elements, and chromatin accessibility as measured by DNase-seq. For both of these assays, from the full correlation matrix we extracted the submatrix of all experiments that include the assay. Reassuringly, UMAP projections of the extracted submatrices show that the experiments appear to cluster by the anatomy type of the biosample that they were performed in (Figure 4.4A–B). Much of this structure was not apparent in the joint projection of all experiments (Figure 4.1A), likely because a single visualization cannot easily capture the full complexity of such a data set. Interestingly, we note that projections share similarities across assays, such as neural and stem cell biosamples forming nearby clusters, but that there are also differences, such as heart biosamples forming a more compact cluster in the H3K27ac experiments than in the DNase experiments. When we initialize our selection procedure with the biosamples that the assays have already been performed in, we confirm that the experiments that are selected next are dissimilar to those that have already been assayed (Figure 4.4E–F).

Next, we can also prioritize the order that assays should be performed in a given biosample by using the submatrix of experiments that include the relevant biosample. In order to highlight that Kiwano accounts for the content of the performed experiments rather than just those that have been performed, we selected two biosamples from the same individual which had the same set of experiments performed. A projection of their respective experiments resembles our joint visualization of all experiments in some aspects (Figures 4.4C–D). Specifically, accessibility and protein binding assays appear to form one large cluster, histone modification experiments form another cluster, and transcription assays are separate from the others but do not form their own cluster.

Similar to the joint selection procedure, we can account for experiments that have al-

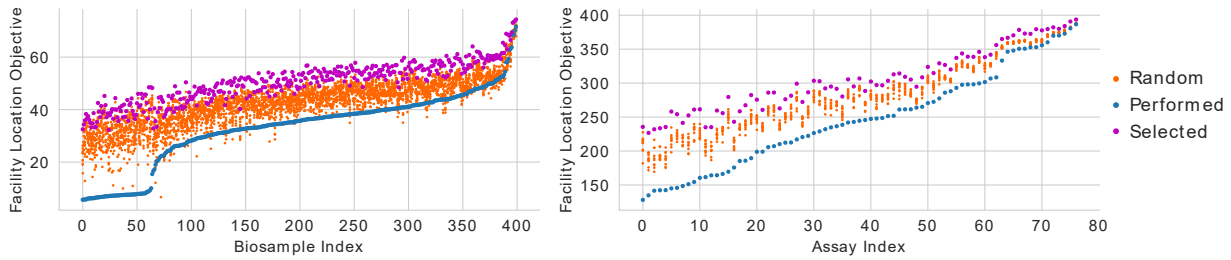



Figure 4.5: **Scoring biosamples and assays according to their captured diversity.** (A) The facility location objective score for each biosample when applied to the set of experiments that investigators have performed in that biosample (blue), the set of experiments identified by optimizing the objective function (magenta), and the sets of randomly selected experiments (orange), ordered by the score of the performed experiments. (B) The same as (A), but for each assay instead of each biosample.

ready been performed. When we prioritize the order that biosamples should be assayed, we observe that the experiments that are selected appear to be dissimilar than those that have already been assayed, and that well characterized areas are not selected from (Figure 4.4E–F). Likewise, despite the same set of assays having been performed in the two biosamples we considered, we observe that our procedure selects different assays (Figure 4.4G–H).

### *4.3.5 Calculating the coverage of each biosample and assay*

Thus far we have focused our efforts on prioritizing individual experiments but have provided little guidance for how to prioritize entire biosamples or assays. We next considered a scenario where an investigator is looking to either assay undercharacterized biosamples or to run underperformed assays, but is unsure which biosamples or assays to focus on. A simple approach would be to count the number of experiments that each biosample or assay is involved in and choose the ones with the fewest experiments. However, this approach does not

account for the content of the performed experiments, which can be extremely similar in some cases. For example, in the ENCODE data several biosamples have been assayed extensively for transcription but not assayed at all for histone modifications or protein binding.

A final component of our methodology is the ability to quantify the extent to which each biosample has been characterized and each assay has been performed using the facility location objective function. Because the objective function takes in a set of experiments and returns a score corresponding to the diversity of the set, this function can be used to assess the diversity obtained by an existing set of experiments, corresponding to a single biosample or a single type of assay. In our setting, where similarity is measured via squared correlation, this score ranges from zero up to the total number of experiments that have been performed. Thus, for each biosample, the maximum value is 77 due to the 77 assays in the data set, and for each assay, the maximum value is 400 due to the 400 biosamples in the data set.

We applied this approach to score each of the biosamples and assays in the ENCODE2018-Core data set. Not surprisingly, we find that the three ENCODE Tier 1 cell lines—H1-hESC, K562, and GM12878—are the three best scoring biosamples, with scores of 71.2, 70.2, and 68.6, respectively. These biosamples are followed by several ENCODE Tier 2 cell lines, such as HepG2, IMR90, and HeLa-S3. We found a rank correlation of 0.82 between the number of assays performed in a biosample and the objective score, confirming that while in general there is increase in coverage as more assays are performed, the composition of those assays is also captured by the objective function. Next, we scored the assays and found that the highest scoring ones were H3K4me3, H3K36me3, and CTCF, whereas the lowest scoring assays are H2BK15ac and FOXK2. We found a weaker, but still very significant, rank correlation of 0.66 between the number of biosamples that an assay was performed in and the objective score.

We next sought to contextualize the scores we obtained for each biosample and assay by comparing them to scores obtained if one had used alternate methods to select experiments.

For each element, i.e., a particular assay or biosample, we scored 10 randomly selected panels of the same size as the number of experiments involving that element. Additionally, we score the panel of experiments that would have been selected using submodular selection. We observe a striking result, which is that the set of experiments that were actually performed not only underperforms the set selected through submodular selection, but also generally underperform random selection (Figure 4.5). This trend is consistent across both biosamples and assays. We note that the 64 biosamples with the worst scores were assayed almost exclusively for transcription, supporting the notion that biosamples with more assays performed in them are not always better characterized.

### *4.4 Discussion*

In this work, we describe an approach for the prioritization of epigenomic and transcriptomic experiments that has the potential to increase the rate of scientific discovery by focusing characterization efforts on those experiments that are expected to yield the least redundant information. To our knowledge, this is the first approach that enables the global prioritization of experiments across both biosamples and assays. We anticipate that, due to the time it takes to perform experiments and the simplicity of Kiwano, investigators may use our prioritization methods even when they plan eventually to perform all potential experiments in order to begin analyses sooner.

An important consideration is that, due to the reliance on imputed experiments, Kiwano cannot be applied directly to a biosample or assay type when no experiments have yet been performed. However, because a diversity of biosamples have already been experimentally characterized, in many cases it would be simple to identify closely related experiments that imputations have already been generated for. While these imputations may not capture activity specific to an experiment, it is likely that the resulting similarity matrix is still a reasonable approximation. Unfortunately, similar imputations are unlikely to be readily

available in cases where one is performing experiments that are very unlike anything that have been performed before. In this setting, it would likely be necessary to first perform a subset of experiments that include all assays and biosamples for use in training an imputation model, and then use the resulting imputations to prioritize the remaining experiments.

While the primary question that we address is how to prioritize experiments across both biosamples and assays, we recognize that this approach may not always result in a practical set of experiments to perform. Specifically, it is generally more difficult to culture and maintain a variety of biosamples than it is to maintain a large quantity of a single biosample, making sets of experiments that span several biosamples harder to perform than those in the same biosample. This difficulty may cause investigators to prefer performing batches of experiments within a biosample, and so we have provided methods both for choosing biosamples that currently are not well characterized and for prioritizing assays within a given biosample. Alternatively, the objective function can be easily modified to include weights on biosamples that are inversely proportional to their difficulty to culture. This modification would encourage the selection process to focus on biosamples that were easy to culture, but still allow more difficult to culture biosamples to be selected when assays performed in them would yield important information.

More generally, the broad applicability of Kiwano may hinge on its ability to incorporate additional considerations into its objective function. In practice, selection of which experiments to perform next often depends on factors such as the availability of samples, the importance of particular cell lines in research, and the relative costs of different types of assays. In principle, extending Kiwano to account for such considerations is straightforward, by generalizing the facility location objective function in the vast space of possible submodular objectives. For example, in the context of protein representative set selection, different classes of submodular objectives yield qualitatively different results with respect to a gold standard based on protein structure, and a convex combination of multiple objectives

ended up yielding a good trade-off between two different goals. Similar approaches would be interesting to explore in the context of genomic experiment prioritization.

When we scored the biosamples in the ENCODE2018-Core data set using the facility location objective function, we noted that the actual set of assays performed in many biosamples performed worse than randomly selecting an equally sized panel of assays. However, this trend is not entirely surprising. The experiments that are included in our data set were intentionally devised to investigate specific research questions, and generally these questions do not aim to broadly characterize the human epigenome. Thus, these results serve primarily to demonstrate that the current strategy for selecting experiments is not well suited for the goal of characterizing the overall diversity of activity in the human epigenome.

A weakness in Kiwano is that mistakes in the imputation process are propagated to the selection process. These mistakes can be simple errors in predicting certain peaks, but can also involve more systematic trends. For example, REST is a transcription factor that is involved in suppressing neuronal genes in non-neuronal tissues. However, the ENCODE2018-Core data set does not have examples of REST in neuronal tissue, and so an imputation model trained on this data set would likely be unaware of this property of REST. Consequently, the prioritization process is unlikely to capture that REST binding in neuronal tissues is significantly different than in non-neuronal tissues. In general, unexpected patterns in data that has not yet been collected will be difficult for any prioritization method to account for.

The flexibility of Kiwano allows for several extensions that we did not consider here, but may nonetheless prove valuable to those prioritizing experiments. The first is that, in the setting where one is prioritizing experiments within a particular biosample, one could measure the gain that each successive experiment adds to the objective function to determine when to stop performing experiments. This would serve as a data-driven indicator of when further experimental efforts are mostly redundant. A second potential extension is to add regularization to the selection process itself to encourage successive experiments to come from

the same biosample. While there are many ways that one could do this, a simple approach would be to rephrase the objective function as $f(X) + \lambda G(X)$, where $f(X)$ is the facility location function as used here, $\lambda$ is the regularization strength, and $G(X)$ is a submodular function counting the number of biosamples not considered by this set. Because the sum of two submodular functions is itself submodular, a similar greedy optimization approach could be applied here. A third extension is that one could calculate the similarity matrix using only a specific genomic locus or set of loci of interest. For example, if an investigator was aiming to experimentally quantify the activity surrounding an important gene across many biosamples, one could restrict the similarity calculation to a window surrounding that gene. Overall, Kiwano is a simple yet powerful way to prioritize experiments in a wide variety of contexts.

# Chapter 5

# CONCLUSION

It is an exciting time to be studying genomics. This excitement is fueled by a revolution in the amount, variety, and quality of data being generated. Methods that were originally developed to sequence genomes in a high-throughput manner are being adapted to give increasingly detailed genome-wide readouts about protein binding, chromatin accessibility, transcription, and three-dimensional genome structure. At the same time, novel experimental techniques are capable of simultaneously profiling millions of individual cells and providing a readout for each one. These methods are an auspicious sign of discoveries to come.

In parallel, the field of machine learning has seen an explosion of theoretical discovery centered around the optimization of neural network. These ideas, paired with freely available implementations, have led to the rapid adoption of neural networks across fields as diverse as computer vision, diagnostic testing in healthcare, and the processing of text and voice signals. A key strength of neural networks is that their multi-layered nature can replace expert-designed features with those learned directly from raw data—the very thing that fields far and wide are experiencing a deluge of. While an unfortunate amount of hype has been generated about the practical capabilities of neural networks, it is clear that modern ideas have dramatically increased the number of tasks that machine learning models can do well at [1].

However, a takeaway from the hype surrounding machine learning is that modern algo-

[1]The practical consequence of this hype is that students of machine learning are cursed to go on to soul-crushing jobs at tech companies, comforted only by their six-figure salaries, whereas I get to look forward to the intellectually stimulating and stress-free life of post-doctoral studies.

rithmic developments are well suited to solving problems when massive data sets are available. This thesis describes a method, named Avocado, that integrates these computational methods with the massive data sets that have been collected using recently developed (and some not-so-recently developed) genomics assays. The resulting model aggregates hundreds of gigabytes of genomics data into a comprehensive representation of the human genome that can be used by humans or computational methods directly, or used predict the output of the tens of thousands of experiments that have not yet been performed.

There are countless applications for an integrative model like Avocado outside the settings considered in this thesis. For instance, it is undoubtedly important to consider the three-dimensional structure of the genome when modeling cell state and function. Assays developed in the last decade, such as Hi-C and ChIA-PET, provide a way to measure pairwise contacts between regions of the genome in a high-throughout manner. However, these pairwise measurements cannot be easily incorporated into the Avocado model proposed here. Extending Avocado to incorporate pairwise measurements will likely improve the quality of epigenomic imputations resulting from the model and also enable the imputation or cleaning of expensive structural measurements in a manner that leverages thousands of epigenomic experiments. Likewise, single-cell measurements have been invaluable for understanding variability within tissues and cell populations, and a framework like Avocado could be a powerful approach to solving the challenges of working with single-cell measurements by grafting them onto existing bulk measurements. A third project involves extending the zero-shot imputation setting described in Chapter 3 to the setting of imputing protein binding for proteins that have not yet been assayed at all. This could be done by replacing the fixed assay embeddings with a neural network that takes in the sequence of a protein, and perhaps its connections in a protein-protein interaction network, and outputs an embedding. A successful method to make these imputations would mitigate the apparent impossibility of profiling the binding of thousands of proteins to each genome of interest. Further, such a method could be used to

inspect the effect that mutating proteins would have on their ability to bind to the genome.

There are also several important problems that could potentially be addressed by a model like Avocado, but for which the experimental data is currently lacking. Likely the most impactful of these problems involve connecting changes in genomic sequence, such as single nucleotide polymorphisms and structural re-arrangements, to overall epigenomic state. A potential solution to these problems involves augmenting the genome embeddings with nucleotide sequence directly so that mutations can be directly mapped to differences in imputed measurements. Knowing these differences would be useful in the study of disease and could lead to approaches that engineer CRISPR targets that result in a desired set of epigenomic imputations. Unfortunately, there is only limited experimental data measuring the changes epigenomic state that result from sequence mutations. Fortunately, as the cost of performing assays continues to decrease, the challenges associated with collecting such data will likely diminish.

# BIBLIOGRAPHY

[1] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3):215–216, 2012.

[2] M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, and W. S. Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–41, 2013.

[3] M. W. Libbrecht, O. Rodriguez, Z. Weng, M. Hoffman, J. A. Bilmes, and W. S. Noble. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types (preprint in advance of publication). *bioRxiv*, 2016.

[4] M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6):321–332, 2015.

[5] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, 2007.

[6] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008.

[7] R. Karlic, H. R.Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2926–2931, 2010.

[8] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[9] E. M. Smith, B. R. Lajoie, G. Jain, and J. Dekker. Invariant tad boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the cftr locus. *American Journal of Human Genetics*, 98:185–201, 2016.

[10] Y. Zhang, C. H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, F. H. Mulawadi, W. K. Sung, S. Nicolis, N. Ahituv, Y. Ruan, and C. L. Wei. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–10, 2013.

[11] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbanian, M. Kasowski, M. Q. Zhang, and M. P. Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Research*, 12:1905–1917, 2014.

[12] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. Mei Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. O., S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, Jan 2012.

[13] T. J. Durham, M. W. Libbrecht, J. J. Howbert, J. A. Bilmes, and W. S. Noble. PREDICTD: PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. *Nature Communications*, 9, 2018.

[14] J. Ernst and M. Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICML*, 2013.

[16] S. Whalen, R. M. Truty, and K. S. Pollard. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, 48:488–496, 2016.

[17] A. D. Schmitt, M. Hu, I. Jung, Z. Xu, Y. Qiu, C. L. Tan, Y. Li, S. Lin, Y. Lin, C. L. Barr, and B. Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Reports*, 17:2042–2059, 2016.

[18] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller. A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 417–429, 2017.

[19] J. Fan and J. Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34–41, 2018.

[20] L. McInnes and J. Healy. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

[21] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2011.

[22] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[23] T. Suganuma and J. L. Workman. Signals and combinatorial functions of histone modifications. *Annual Review of Biochemistry*, 80(473–499), 2011.

[24] T. Suganama and J. L. Workman. Crosstalk among histone modifications. *Cell*, 135(604–607), 2008.

[25] H. Daumé III. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics*, 2007.

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pages 512–519, Washington, DC, USA, 2014. IEEE Computer Society.

[27] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[28] V. Sandulescu and M. Chiru. Predicting the future relevance of research institutions - the winning solution of the KDD cup 2016. *CoRR*, abs/1609.02728, 2016.

[29] M. Volkovs, G. W. Yu, and T. Poutanen. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17, pages 7:1–7:6, New York, NY, USA, 2017. ACM.

[30] R. Singh, J. Lanchantin, G. Robins, and Y. Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i649, 2016.

[31] R. Singh, J. Lanchantin, A. Sekhon, and Y. Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. *Advances in Neural Information Processing Systems*, pages 6788–6798, 2017.

[32] A. Mora, G. K. Sandve, O. S. Gabrielsen, and R. Eskeland. In the loop: promoter-enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, 17(6):980–995, 2015.

[33] N. D. Heintzmann, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108–112, 2009.

[34] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[35] RE Thurman, E Rynes, R Humbert, J Vierstra, MT Maurano, E Haugen, NC Sheffield, AB Stergachis, H Wang, B Vernot, K Garg, S John, R Sandstrom, D Bates, L Boatman, TK Canfield, M Diegel, D Dunn, AK Ebersol, T Frum, E Giste, AK Johnson, EM Johnson, T Kutyavin, B Lajoie, BK Lee, K Lee, D London, D Lotakis, S Neph, F Neri, ED Nguyen, H Qu, AP Reynolds, V Roach, A Safi, ME Sanchez, A Sanyal, A Shafer, JM Simon, L Song, S Vong, M Weaver, Y Yan, Z Zhang, Z Zhang, B Lenhard, M Tewari, MO Dorschner, RS Hansen, PA Navas, G Stamatoyannopoulos, VR Iyer, JD Lieb, SR Sunyaev, JM Akey, PJ Sabo, R Kaul, TS Furey, J Dekker, GE Crawford, and JA Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[36] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, The FANTOM Consortium, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507:455–461, 2014.

[37] W. Xi and M.A. Beer. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLOS Computational Biology*, 14(12):1–7, 2018.

[38] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 20(6):761–770, 2010.

[39] Vishnu Dileep, Ferhat Ay, Jiao Sima, Daniel L Vera, William S Noble, and David M Gilbert. Topologically-associating domains and their long-range contacts are established during early g1 coincident with the establishment of the replication timing program. *Genome Research*, pages gr–183699, 2015.

[40] Claire Marchal, Takayo Sasaki, Daniel Vera, Korey Wilson, Jiao Sima, Juan Carlos Rivera-Mulia, Claudia Trevilla-García, Coralin Nogues, Ebtesam Nafie, and David M Gilbert. Genome-wide analysis of replication timing by next-generation sequencing with e/l repli-seq. *Nature protocols*, 13(5):819, 2018.

[41] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[42] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[43] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[44] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017.

[45] S. Lundberg and S. Lee. An unexpected unity among methods for interpreting model predictions. In *Neural Information Processing Systems*, 2017.

[46] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[47] S. Dumančić and H. Blockeel. Demystifying relational latent representations. In *Inductive Logic Programming*, pages 63–77. Springer International Publishing, 2018.

[48] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.

[49] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[50] J. Zhou and O. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12:931–934, 2015.

[51] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, 2016.

[52] J. M. Schreiber, J. Bilmes, and W. S. Noble. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biology*, 2020. In press.

[53] François Chollet et al. Keras. https://keras.io, 2015.

[54] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[55] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[56] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[57] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9, pages 249–256, 2010.

[58] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[59] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[60] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306(5696):636–640, 2004.

[61] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, and Michael J Ziller. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[62] The modENCODE Consortium. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, 330:1775–1787, 2010.

[63] D. Bujold, D. A. Morais, C. Gauthier, C. Cote, M. Caron, T. Kwan, K. C. Chen, J. Laperle, A. N. Markovits, T. Pastinen, B. Caron, A. Veilleux, P. E. Jacques, and G. Bourque. The international human epigenome consortium data portal. *Cell Systems*, 3:496–499, 2016.

[64] F. Yue and The Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515:355–364, 2014.

[65] S. Akbarian et al. The PsychENCODE project. *Nature Neuroscience*, 18:1707–1712, 2015.

[66] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550:204–213, 2017.

[67] J. M. Schreiber, T. J. Durham, J. Bilmes, and W. S. Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *Genome Biology*, 2020. In press.

[68] X. Lai, A. Stigliani, G. Vachon, C. Carles, C. Smaczniak, C. Zubieta, K. Kaufmann, and F. Parcy. Building transcription factor binding site models to understand gene regulation in plants. *Molecular Plant*, 2018.

[69] J. M. Schreiber, R. Singh, J. Bilmes, and W. S. Noble. A pitfall for machine learning methods aiming to predict across cell types. *bioRxiv*, 2019. https://www.biorxiv.org/content/10.1101/512434v1.

[70] H. Li, D. Quang, and Y. Guan. Anchor: Trans-cell type prediction of transcription factor binding sites. *Genome Research*, 29(2):281–292, 2019.

[71] D. Quang and X. Xie. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 2019.

[72] J. Keilwagen, S. Posch, and J. Grau. Accurate prediction of cell type-specific transcription factor binding. *Genome Biology*, 20(9), 2019.

[73] C. W. Hanna, H. Demond, and G. Kelsey. Epigenetic regulation in development: is the mouse a good model for the human? *Human Reproduction Update*, 24:556–576, 2018.

[74] J. P. Morton and O. J. Sansom. Myc-y mice: From tumour initiation to therapeutic targeting of endogenous myc. *Molecular Oncology*, 7(2):248–258, 2013.

[75] J. Hana, A. Feldman, and C. Brew. A resource-light approach to Russian morpholoy: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 222–229, 2004.

[76] D. R. Kelley. Cross-species regulatory sequence activity prediction. *bioRxiv*, 2019.

[77] J. Lu, X. Cao, and S. Zhong. EpiAlignment: alignment with both dna sequence and epigenomic data. *Nucleic Acids Research*, 47:W11–W19, 2019.

[78] R. F. Lowdon, H. S. Jang, and T. Wang. Evolution of epigenetic regulation in vertebrate genomes. *Trends Genet*, 32:269–283, 2016.

[79] W. H. Gharib and M. Robinson-Rechavi. When orthologs diverge between human and mouse. *Brief Bioinform*, 12:436–441, 2011.

[80] T. Kaitsuka and M. Matsushita. Regulation of transcription factor eef1d gene function by alternate splicing. *Int J Mol Sci*, 16:3970–3979, 2015.

[81] J. M. Schreiber, J. Bilmes, and W. S. Noble. Prioritizing transcriptomic and epigenomic experiments by using an optimization strategy that leverages imputed data. *bioRxiv*, 2019. https://www.biorxiv.org/content/10.1101/708107v1.

[82] K. Wei, M. W. Libbrecht, J. A. Bilmes, and W. S. Noble. Choosing panels of genomics assays using submodular optimization. *Genome Biology*, 17(1):229, 2016.

[83] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier Science, 2005.

[84] A. Krause and D. Golovin. Submodular function maximization., 2014.

[85] L. Lovász. Submodular functions and convexity. In M. Grotchel A. Bachem and B. Korte, editors, *Mathematical Programming – The State of the Art*, pages 235–257. Springer-Verlag, 1983.

[86] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, pages 234–243, 1978.

[87] J. M. Schreiber, J. Bilmes, and W. S. Noble. apricot: Submodular selection for data summarization in python. *arXiv*, 2019. https://arxiv.org/abs/1906.03543.

[88] M. W. Libbrecht, J. A. Bilmes, and W. S. Noble. Choosing non-redundant representative subsets of protein sequence data sets using submodular optimization. *Proteins*, 86(4):454–466, 2018.

[89] M. Gasperini, A. J. Hill, J. L. McFaline-Figueroa, B. Martin, S. Kim, D. Jackson, A. Leith, J. Schreiber, W. S. Noble, C. Trapnell, N. Ahituv, and J. Shendure. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176:377–390, 2019.

[90] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[91] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, Sep 2012.

# Appendix A

# HYPERPARAMETER SELECTION

Avocado's model has seven structural hyperparameters: the number of latent factors representing cell types, assay types, and the three scales of genomic positions, as well as two parameters (number of layers and number of nodes per layer) for the deep neural network.

We optimized these hyperparameters via random search. The search considered the following grid of values: cell type factors $\in$ (16, 32, 64, 128, 256), assay factors $\in$ (16, 32, 64, 128, 256), 25 bp resolution genome factors $\in$ (5, 10, 15, 20, 25), 250 bp resolution genome factors $\in$ (10, 20, 30, 40, 50), 5 kbp resolution genome factors $\in$ (15, 30, 45, 60, 75), number of layers in the neural network $\in$ (0, 1, 2, 3, 4), and number of neurons in the neural network $\in$ (128, 256, 512, 1024, 2048). Note that setting the number of layers to 0 corresponds to training a linear regression model on top of the learned factors. These ranges were selected based on experimental results from Durham et al. [13], suggesting that 100 latent factors for each of the three axes performed well. In this grid we trained 1,000 models out of a possible $\sim$61,000. Each model was trained on the ENCODE Pilot Regions, which are comprised of 44 regions of 0.5-2 Mb length that jointly make up approximately 1% of the full genome. The data were split into a training set of 764 tracks, a validation set of 100 tracks, and a test set of 150 tracks. We selected the final set of hyperparameters based on performance on the validation set, as measured by mean-squared error (MSE).

The different hyperparameter settings displayed a wide variance in performance, with most performing better than ChromImpute and many performing better than PREDICTD on the validation set (Additional file 1: Figure A.1). Once the hyperparameters were set, the model was then retrained on both the training and validation sets and tested on the held-out

test set. Note that the training, validation, and test sets used here correspond to the same splits used for the PREDICTD approach. The resulting model had a MSE of 0.1130 on the test set, which represents an 18.5% improvement over ChromImpute (MSE 0.1387) and a 4.9% improvement over PREDICTD (MSE 0.1188).

We next investigated the effect that each hyperparameter had on the overall predictive performance of Avocado. To do this, we considered each hyperparameter individually and, for each value that the hyperparameter could take, we plotted the MSE of each model that used that value (Additional file 1: Fig. A.2). The clearest trend was that the performance of the model increased as the size of the neural network increased, both in terms of the number of layers and the number of neurons per layer. In contrast, the number of latent factors did not show a clear trend of improvement over any of the three axes.

To attempt to better understand where the allocation of parameters was most beneficial, we considered performance when compared to the total number of parameters in the neural network and when compared to the total number of parameters in the embedding matrices (Additional file 1: Fig. A.3). We see that the validation set error decreases steadily with an increase in the number of network parameters until leveling off around $10^7$ parameters. In particular, having no hidden layer, i.e., learning a linear regression on top of the tensor factorization, leads to very poor models. However, adding more than two layers does not yield much gain. When considering the number of parameters at each genomic position in the tensor factorization, we see no similar trend of increased complexity leading to increased performance. We focus on the number of parameters per genomic position rather than the total number of parameters in the model because otherwise the genomic axis would dominate. We can see that having one hidden layer improves model performance; however, we do not see a trend where deeper models are able to better utilize more complex tensor factorization models. Overall, these results suggests that the use of a neural network coupled with the tensor factorization can significantly boost the performance of the model, but that the model

is not very sensitive to the complexity of the tensor factorization component.

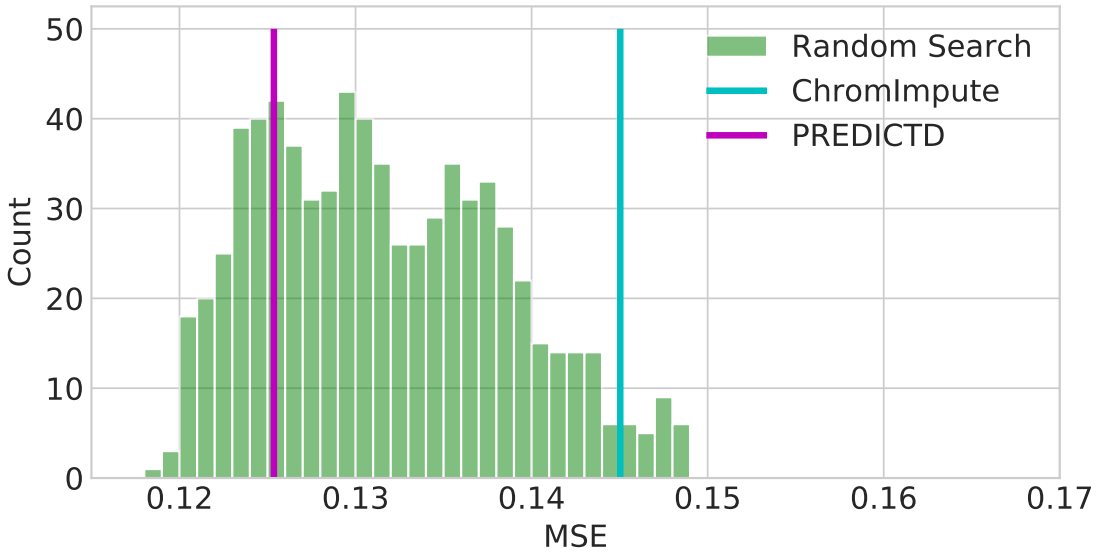



Figure A.1: **Random search results on ENCODE pilot regions.** The figure plots a histogram of Avocado validation set MSE values across each hyperparameter setting. For reference, MSE values on the same data set for ChromImpute and PREDICTD are depicted as vertical lines.

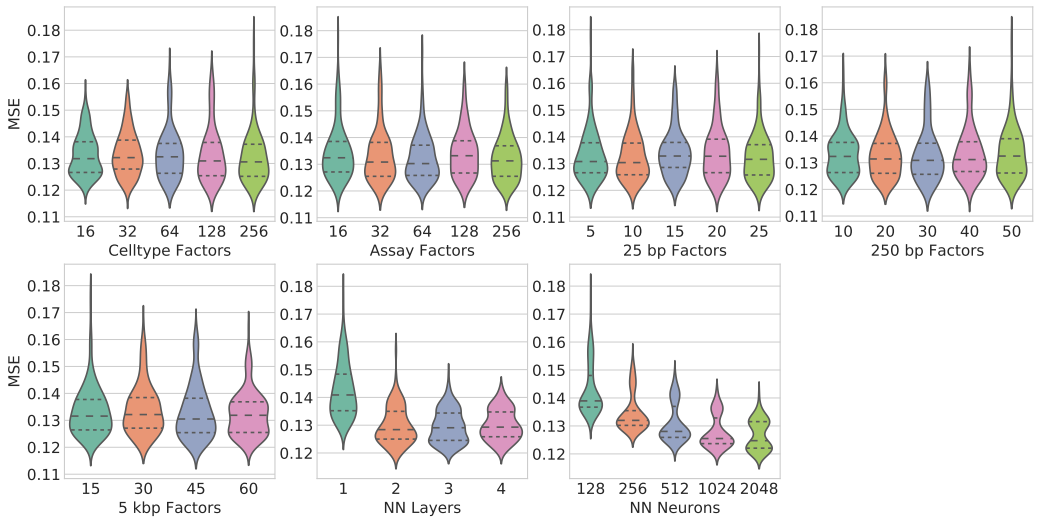



Figure A.2: **The performance of the Avocado models learned during random search when stratified by values for each hyperparameter individually.** Each panel shows results for all models that had at least one hidden layer in the neural network. The median is indicated in each violin plot with the longer dashed lines, with the shorter dashed lines indicating the inter-quartile range. The performance seems to be fairly constant across hyperparameter values, except for those hyperparameters related to the neural network. Increasing the number of neurons per layer seemed to increase performance consistently, whereas past two layers the model did not appear to learn significantly more. Models with no hidden layers are not shown, because their performance was uniformly poor.

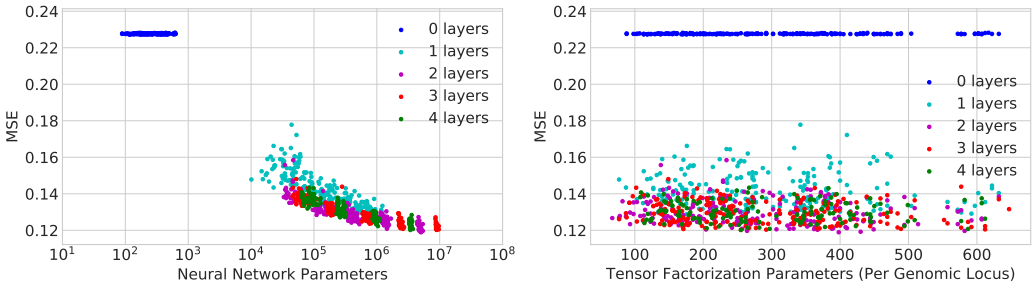



Figure A.3: **The number of parameters in each model considered as a part of the random search procedure compared to validation set performance for both the neural network and the tensor factorization aspects.** Left: The trend appears to be that the greater the number of parameters, the better the performance of the model. Models with no hidden layers still have parameters in the form of a linear regression on top of the tensor factorization. The models are colored by the number of layers that they have. Right: The number of parameters in the tensor factorization component at each genomic position. This corresponds to the number of cell type factors plus the number of assay factors plus the number of genomic factors at each resolution. The models are colored by the number of layer in the neural network.

# Appendix B

# CHAPTER 1 SUPPLEMENT

| | **A/C** | **A/P** | **P/C** |
|---|---|---|---|
| **MSEglobal** | 1.21e-59 | **4.51e-01** | 2.00e-60 |
| **MSE1imp** | 1.97e-152 | 2.60e-10 | 1.95e-151 |
| **MSE1obs** | 2.37e-22 | 9.13e-06 | 2.85e-12 |
| **GWcorr** | 7.96e-119 | 2.12e-05 | 8.53e-110 |
| **match1** | 1.59e-138 | 1.71e-05 | 3.38e-105 |
| **catch1obs** | 1.04e-154 | 8.45e-50 | 3.79e-90 |
| **catch1imp** | 3.44e-68 | 1.63e-09 | 2.16e-51 |
| **aucobs1** | 5.00e-96 | 3.59e-52 | 4.55e-58 |
| **aucimp1** | 2.60e-25 | 9.22e-04 | 2.29e-18 |
| **MSEProm** | 3.98e-32 | 8.73e-05 | 1.04e-25 |
| **MSEGene** | 1.09e-49 | **8.75e-01** | 7.66e-48 |
| **MSEEnh** | 1.72e-30 | 1.50e-04 | 3.25e-23 |

Table B.1: **Statistical significances of imputation performance measures.** Unadjusted p-values from a two-sided paired t-test that compares the average metric value across all 1,014 tracks of data for each pair of imputation methods (Avocado (A), PREDICTD (P), and ChromImpute (C)) and performance metric. The two highlighted values are the only two >0.01, indicating that all other comparisons result in statistically significant differences between the two methods.

| | **B** | **R** | **CI** | **P(I)** | **A(I)** | **P(LF)** | **A(LF)** | **FRC** |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | — | | | | | | | |
| **Roadmap** | 0.0 | — | | | | | | |
| **ChromImpute** | 0.0 | 5.31e-135 | — | | | | | |
| PREDICTD (I) | 0.0 | 2.48e-27 | 1.34e-99 | — | | | | |
| Avocado (I) | 0.0 | 7.99e-116 | 1.25e-01 | 5.08e-75 | — | | | |
| PREDICTD (LF) | 0.0 | 1.19e-107 | 7.57e-02 | 2.66e-71 | 7.59e-01 | — | | |
| Avocado (LF) | 0.0 | 4.62e-153 | 1.91e-76 | 8.33e-134 | 3.13e-104 | 3.86e-101 | — | |
| FRC | 0.0 | 2.52e-168 | 1.84e-69 | 1.59e-153 | 1.25e-96 | 2.13e-93 | 9.75e-21 | — |

Table B.2: **Statistical significances of performance when predicting gene expression.** Unadjusted p-values from a two-sided paired t-test that compares the average precision across all 20 folds from all 47 cell types for a total of 940 measurements. Column names are abbreviated versions of the row names, in the same order. "(I)" stands for "imputations," "(LF)" stands for "latent factors," and "FRC" stands for "full Roadmap compendium." P-values >0.01 are in boldface.

| | **B** | **R** | **CI** | **P(I)** | **A(I)** | **A(LF)** | **(LF)** | **FRC** |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | — | | | | | | | |
| **Roadmap** | 2.46e-22 | — | | | | | | |
| **ChromImpute** | 1.15e-23 | 3.29e-10 | — | | | | | |
| **PREDICTD (I)** | 2.82e-32 | 1.66e-08 | **0.0127** | — | | | | |
| **Avocado (I)** | 9.56e-19 | 7.4e-16 | 0.000176 | **0.502** | — | | | |
| **PREDICTD (LF)** | 9.35e-32 | 1.31e-20 | 1.34e-25 | 8.11e-26 | 9.51e-28 | — | | |
| **Avocado (LF)** | 9.45e-32 | 9.54e-26 | 1.53e-26 | 8.33e-27 | 9.32e-27 | 6.97e-18 | — | |
| **FRC** | 1.57e-30 | 2.35e-09 | 2.02e-19 | 1.17e-18 | 2.43e-22 | 1.25e-12 | 1e-24 | — |

Table B.3: **Statistical significances of performance when predicting promoter-enhancer interactions.** Unadjusted p-values from a two-sided paired t-test that compares the average precision across all 20 runs from all 4 cell types for a total of 80 measurements. Column names are abbreviated versions of the row names, in the same order. "(I)" stands for "imputations," "(LF)" stands for "latent factors," and "FRC" stands for "full Roadmap compendium." P-values >0.01 are in boldface.

| | **B** | **R** | **CI** | **P(I)** | **A(I)** | **P(LF)** | **A(LF)** | **FRC** |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | — | | | | | | | |
| **Roadmap** | 6.91e-143 | — | | | | | | |
| **ChromImpute** | 2.42e-149 | 3.8e-13 | — | | | | | |
| **PREDICTD (I)** | 6.93e-146 | 7.04e-22 | 2.13e-20 | — | | | | |
| **Avocado (I)** | 1.1e-150 | 1.48e-21 | 7.83e-09 | 4.57e-09 | — | | | |
| **PREDICTD (LF)** | 5.37e-154 | 2.35e-22 | 2.73e-62 | 9.98e-75 | 3.77e-83 | — | | |
| **Avocado (LF)** | 5.53e-154 | 2.47e-22 | 4.23e-58 | 6.26e-76 | 1.78e-74 | 0.00406 | — | |
| **FRC** | 6.64e-156 | 5.33e-70 | 3.52e-95 | 5.85e-82 | 1.96e-97 | 1.73e-69 | 2.8e-63 | — |

Table B.4: **Statistical significances of performance when predicting replication timing.** Unadjusted p-values from a two-sided paired t-test that compares the average precision across all 20 runs from all 5 cell types for a total of 100 measurements. Column names are abbreviated versions of the row names, in the same order. "(I)" stands for "imputations," "(LF)" stands for "latent factors," and "FRC" stands for "full Roadmap compendium."

| | B | R | CI | P(I) | A(I) | P(LF) | A(LF) | FRC |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | — | | | | | | | |
| **Roadmap** | 3.76e-50 | — | | | | | | |
| **ChromImpute** | 2.89e-47 | 5.04e-21 | — | | | | | |
| **PREDICTD (I)** | 3.17e-48 | 2.80e-29 | 1.18e-08 | — | | | | |
| **Avocado (I)** | 3.79e-48 | 2.17e-12 | 9.62e-06 | 2.15e-17 | — | | | |
| **PREDICTD (LF)** | 7.67e-53 | 4.69e-02 | 6.92e-27 | 1.28e-37 | 1.75e-18 | — | | |
| **Avocado (LF)** | 6.15e-54 | 6.13e-08 | 4.39e-39 | 1.34e-49 | 1.32e-34 | 2.40e-04 | — | |
| **FRC** | 3.54e-56 | 6.72e-41 | 4.26e-62 | 7.37e-61 | 2.07e-55 | 6.94e-39 | 1.85e-33 | — |

Table B.5: **Statistical significances of performance when predicting FIREs** Unadjusted p-values from a two-sided paired t-test that compares the average precision across 20 folds from all 7 cell types, for a total of 140 measurements. Column names are abbreviated versions of the row names, in the same order. "(I)" stands for "imputations," "(LF)" stands for "latent factors," and "FRC" stands for "full Roadmap compendium." P-values >0.01 are in boldface.

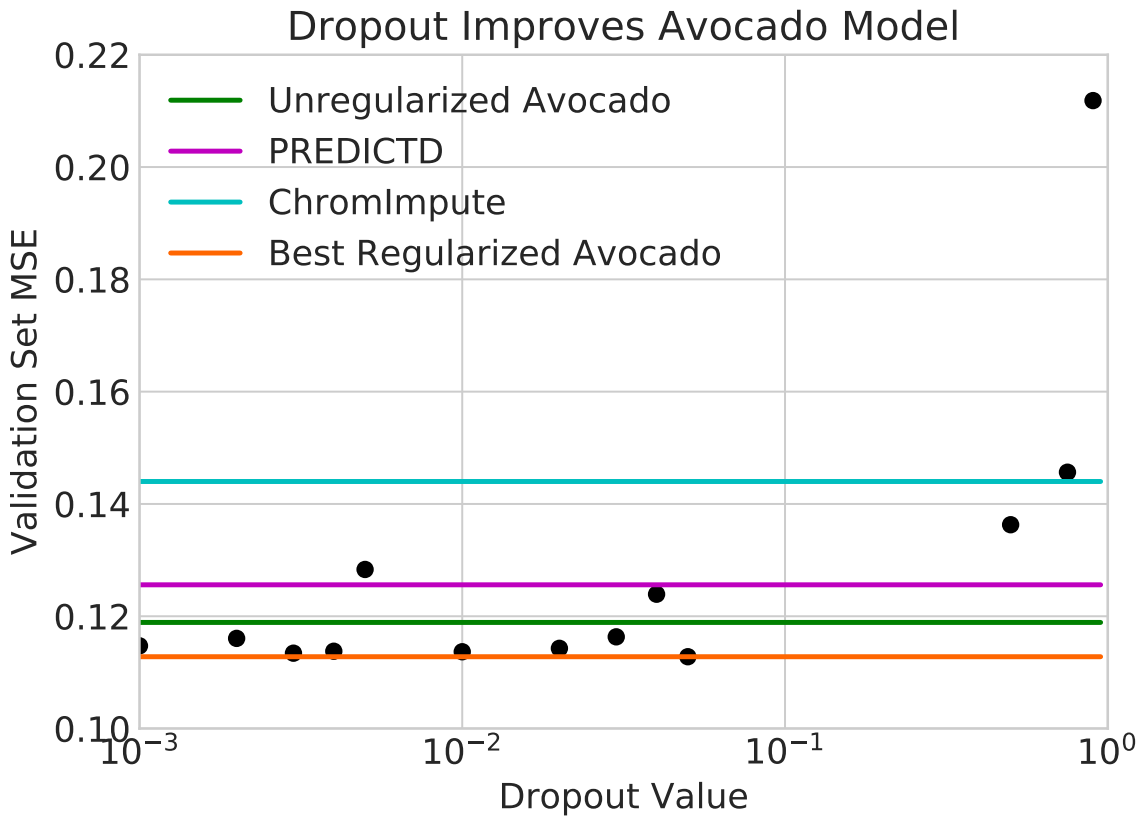



Figure B.1: **Dropout improves the validation set performance of Avocado.** Each point corresponds to the performance of an Avocado model trained with a given dropout probability in the two hidden layers. The best performing model (in orange) outperforms not only the unregularized model (in green) but further improves over PREDICTD (in magenta) and ChromImpute (in cyan).

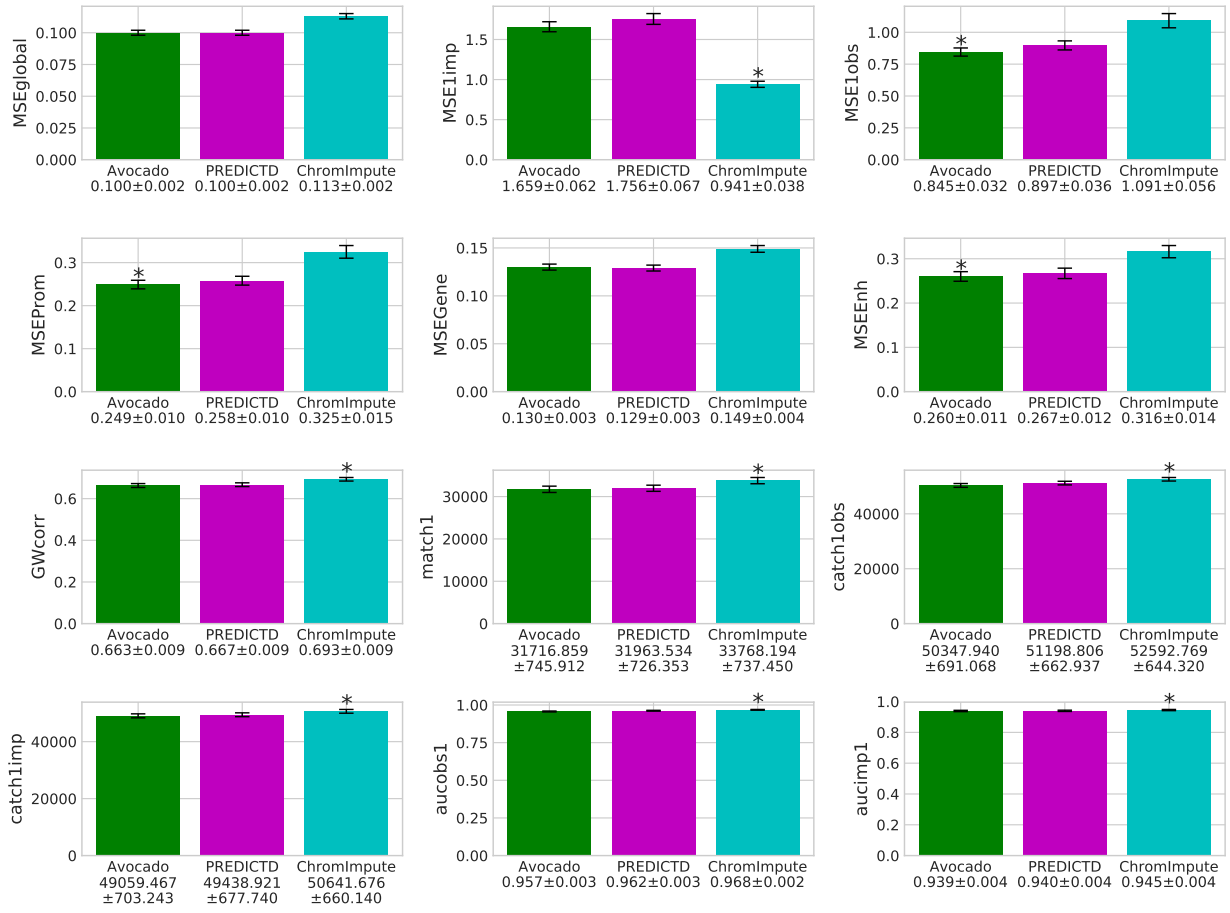



Figure B.2: **Twelve performance measures evaluated across the full genome for each imputation approach**. Each panel plots the value of a specified performance measure (y-axis), averaged across all 1,014 tracks. Nine of the performance measures correspond to those proposed by either Durham et al. or Ernst and Kellis. Error bars display the 95% confidence interval. The best performing approach for each performance measure is denoted with an asterisk above the bar if that result is statistically significant when compared to the next highest performing approach, i.e., p-value $< 0.01$ on a two sided paired t-test, adjusted for the three comparisons.

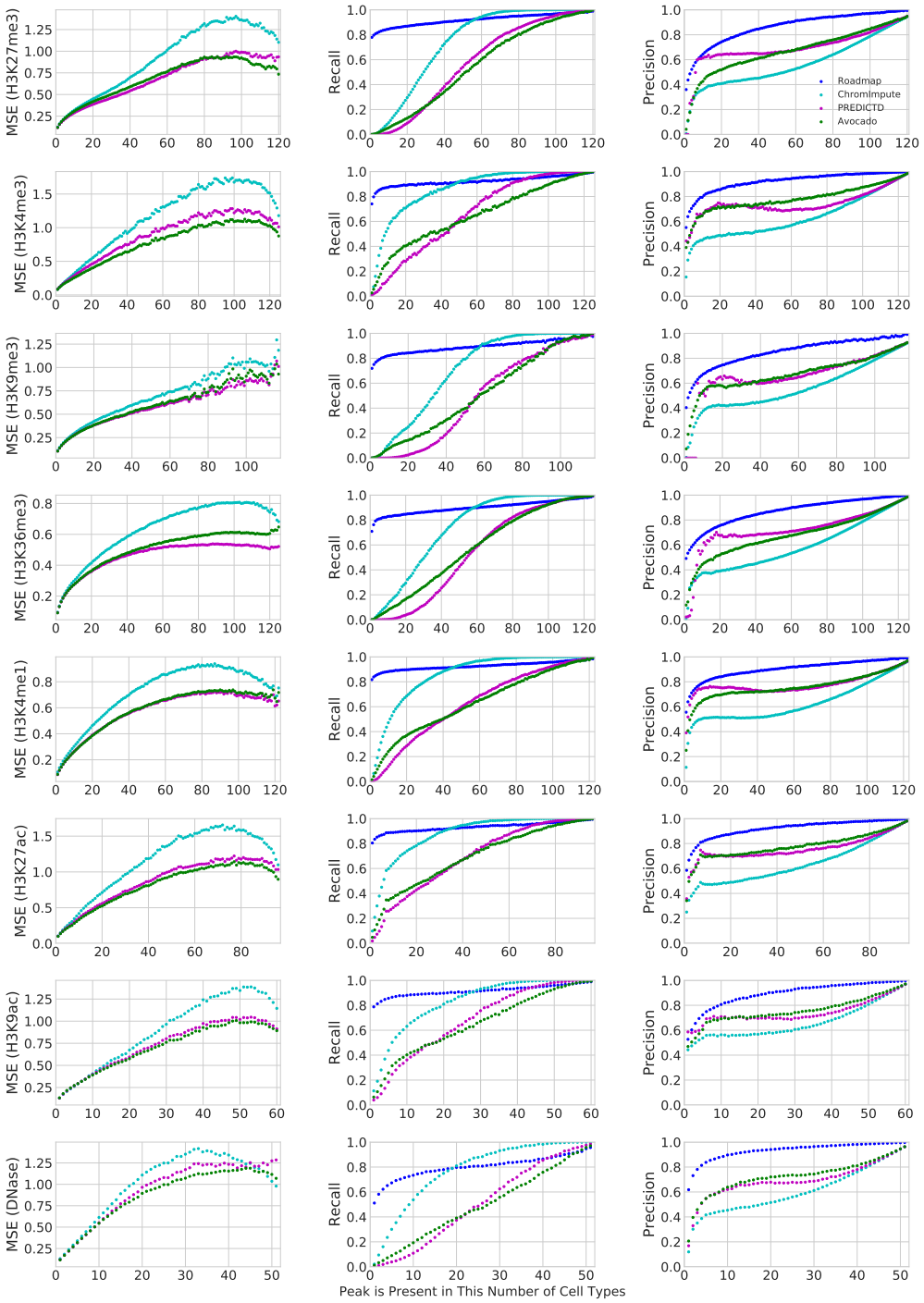



Figure B.3: **Ability to recover cell type-specific peaks.** Each panel plots, for a given assay type, the MSE (left column), recall (middle column) or precision (right column) as a function of the number of cell types in which a given peak occurs. Only the 12 assays that have been performed in more than 10 cell types are shown.

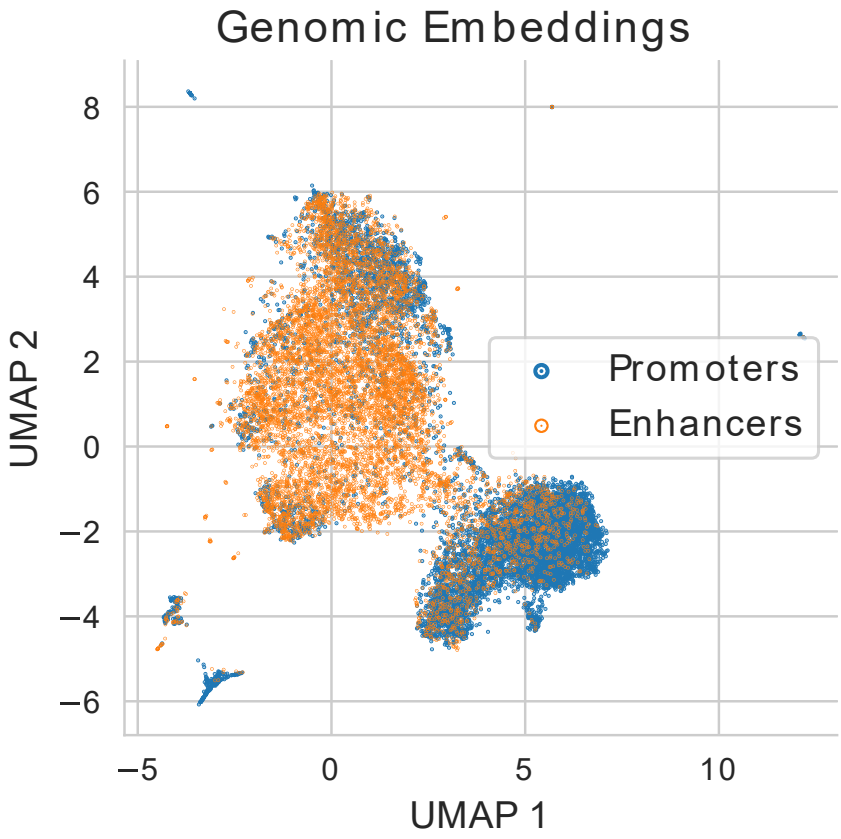



Figure B.4: **A projection of Avocado's genome embeddings with a ±2kbp window.** This plot shows the same procedure as Fig. 3a, except that the window used here is ±2kbp rather than ±250bp.

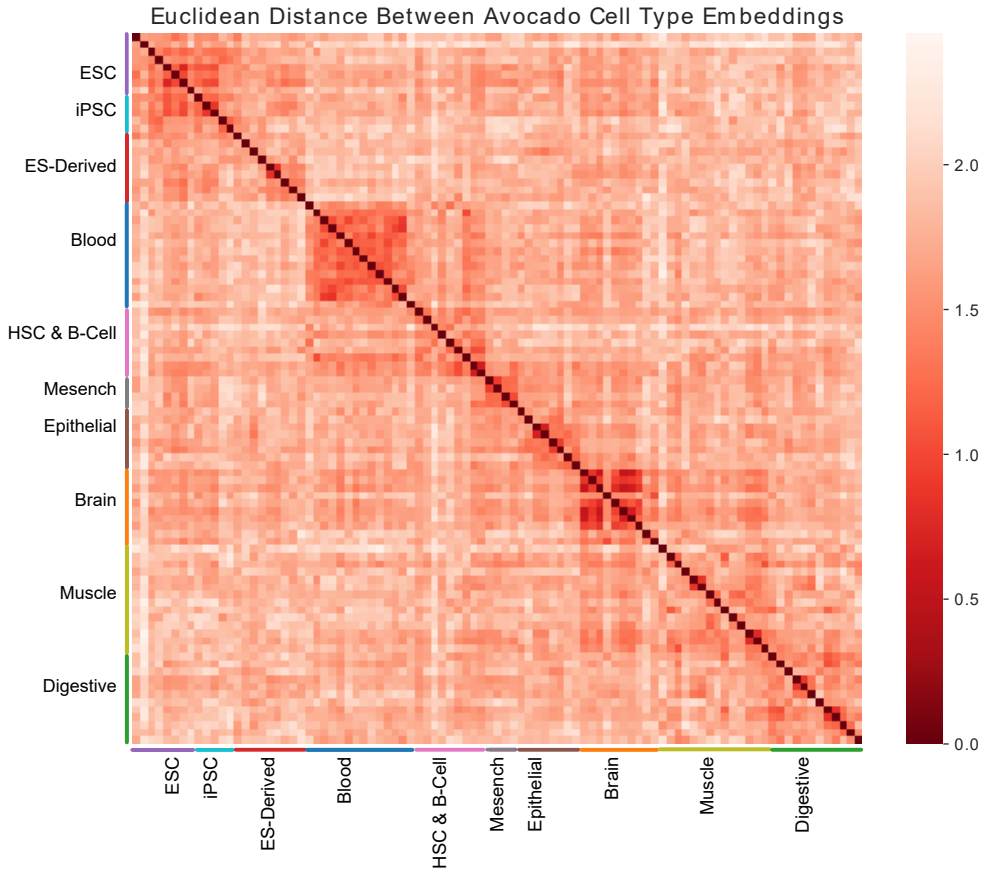



Figure B.5: **Euclidean distance matrix between the cell type embeddings learned by Avocado.** The euclidean distances between 93 cell type embeddings learned by Avocado and inspected in Fig. 3d. Cell types are grouped by anatomy type, as denoted on the axes, with anatomy type colored the same as Fig. 3d.

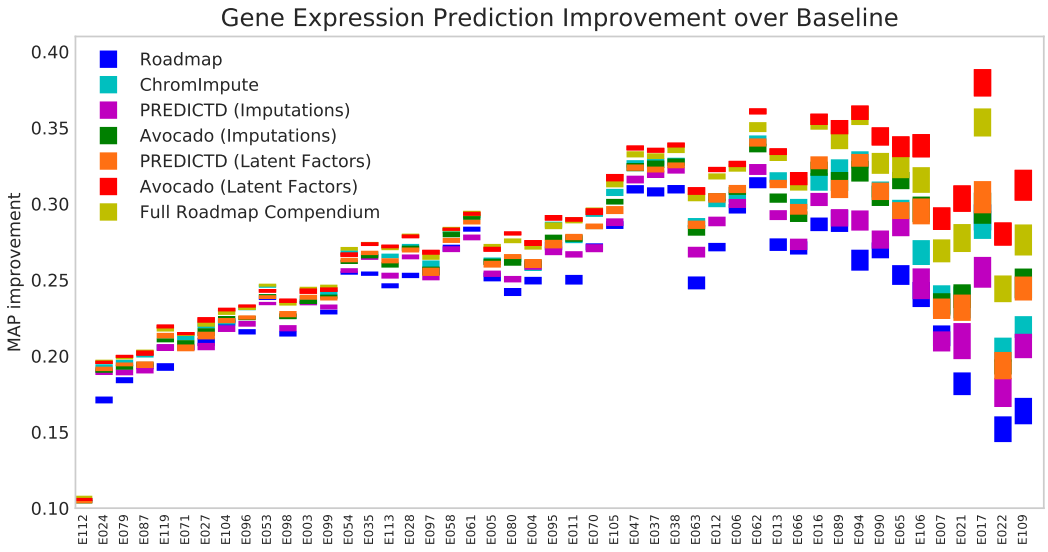



Figure B.6: **Relative improvement over a random baseline for each feature set at predicting gene expression.** This plot shows the same values as Fig. 5a except that the values for each cell type have the majority baseline subtracted out. This view provides a more detailed look at the relative performance of each of the feature sets, even when the performance of all metrics is high.

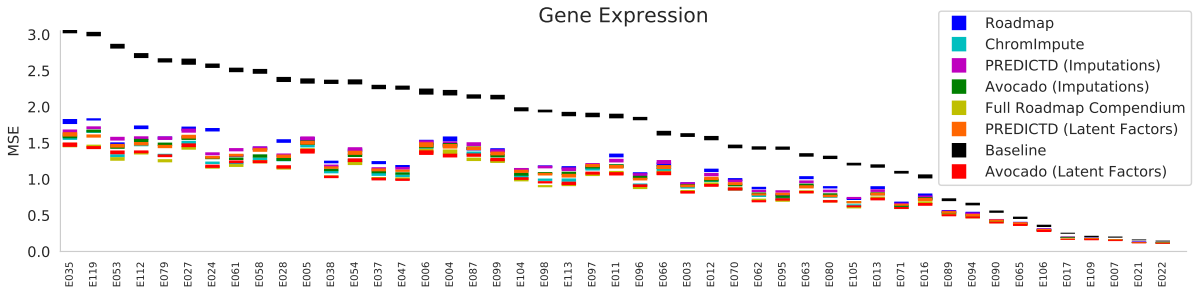



Figure B.7: **Performance of machine learning models trained using various feature sets at regressing gene expression values.** This plot shows the performance of models trained in the same manner as those in Fig. 5a except that the models are trained on the regression task of predicting gene expression values directly. Accordingly, the models are evaluated using mean squared error rather than average precision.

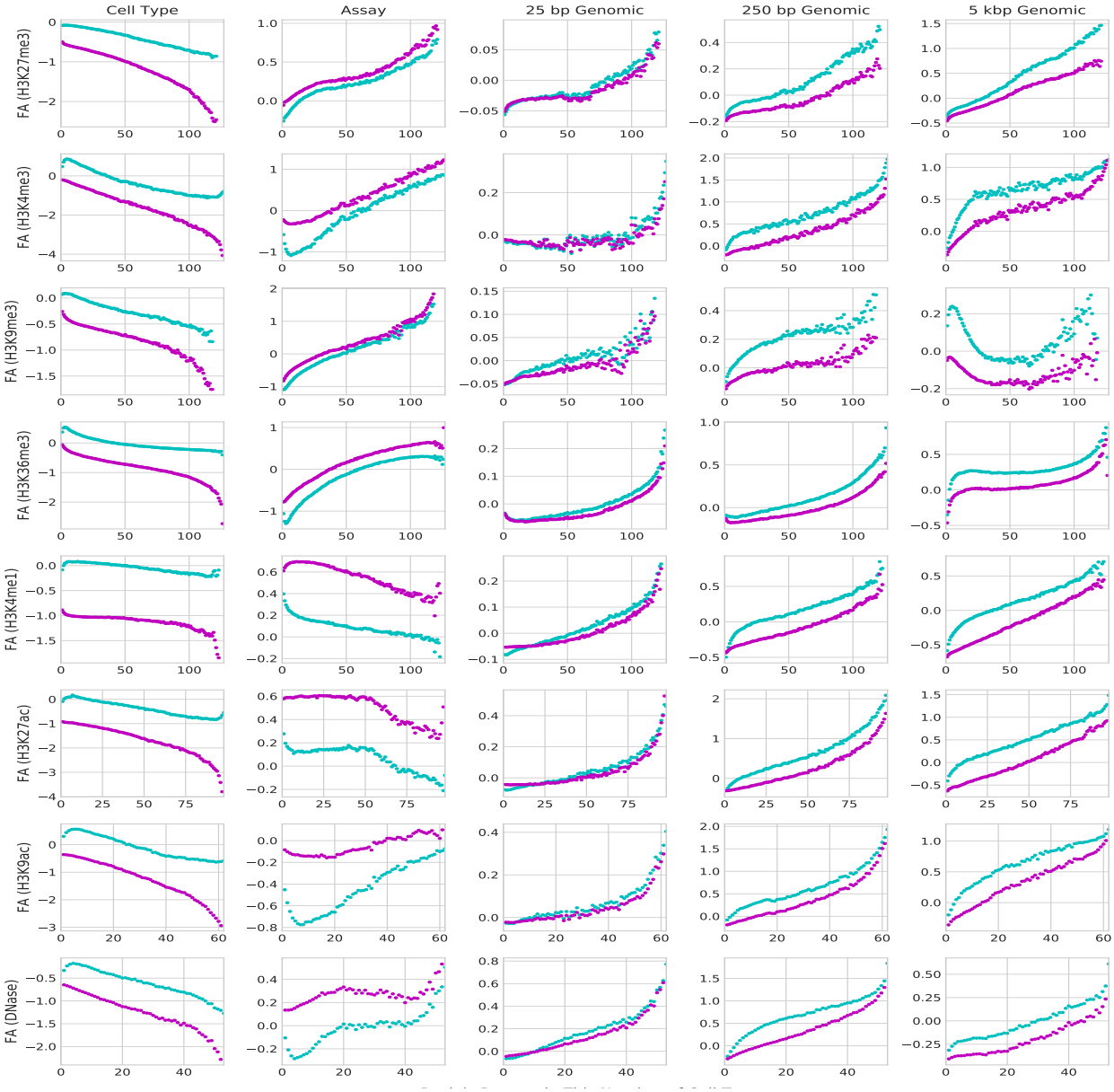



Figure B.8: **Feature attribution performed on the Avocado model.** Feature attribution was performed for each position in chromosome 20 across all 1,014 experiments. The results were then aggregated in a manner similar to the analysis of cell-type specific imputations. Instead of calculating the MSE, precision, and recall, instead only the average attribution value is calculated. However, this is done for each of the five model components (the columns). Additionally, the average attribution value is calculated both for those cell types where a peak is exhibited (cyan) and those cell types where a peak is not exhibited (magenta).

# Appendix C

# AVOCADO'S IMPUTED TRACKS ARE CONSISTENT WITH KNOWN BIOLOGY

To better understand the relative behavior of the three imputation methods, we evaluated the imputed measurements for specific histone marks based on their enrichment in functional elements. In particular, H3K4me3 is known to form peaks within transcription start sites (TSSs) and H3K36me3 is known to localize within transcribed genes [61, 8]. We began by extracting the values of H3K4me3 from all TSSs and H3K36me3 from all gene bodies for each cell type. We note that the average H3K4me3 profile across TSSs forms a distinctive bimodal peak (Additional file 4: Fig. C.1a). Previously, Ernst and Kellis showed that imputed versions of these histone marks exhibit significantly less variation across cell types than the same signal from ChIP-seq tracks, a trend that is also exhibited by PREDICTD and Avocado [14]. An open question is whether this observed reduced variance corresponds to reduction in noise or reduction in true variation among cell types.

To address this question, we first test whether the observed reduction in variation preserves cellular variation by calculating the rank correlation across cell types between imputed signal and ChIP-seq signal according to the area under each cell types' average mark profile (Additional file 4: Fig C.1a/b). This analysis shows that Avocado preserves the ordering of cell types the best in both H3K4me3 and H3K36me3, while still reducing the variation of the signal. In contrast, while ChromImpute reduces the variation across cell types the most, there is almost no correlation of this measurement between the ChromImpute-imputed H3K36me3 signal and the ChIP-seq measurements. We next test whether cellular variation is maintained by re-implementing the PromRecov and GeneRecov performance measures

proposed by Ernst and Kellis that measure how well these two marks localize within their respective regions. All three imputation strategies show similar localization of H3K36me3 in gene bodies (Additional file 4: Fig. C.1c), but Avocado shows the highest localization of H3K4me3 in promoter regions in 23 cell types, and a higher localization than ChromImpute in 87 cell types (Additional file 4: Fig. C.1d).

To expand on this investigation, we then looked at each techniques' ability to reconstruct relationships among multiple histone marks at the same locus in the genome. We began by looking at the signal values of repressive mark H3K27me3 and the activating mark H3K4me3 in promoter regions, because the two marks tend not to co-localize in differentiated cells (Additional file 4: Fig. C.2). To quantitatively evaluate this relationship, we calculate the difference between H3K4me3 and H3K27me3 across all 127 cell lines for all promoter regions and calculate the mean absolute error (MAE) between the ChIP-seq signal and the corresponding imputed tracks. This performance measure measures how well the imputation strategies are able to preserve the difference between the two marks. We find that Avocado achieves a lower MAE at reconstructing this relationship than either other method (Additional file 4: Table F.1). We also verified that Avocado does a better job than the other two imputation methods at capturing a lack of correlation between unrelated marks (Additional file 4: Fig. C.2), such as the repressive mark H3K27me3 and enhancer-associated mark H3K4me1 (Additional file 4: Table F.1).

We then consider how well the methods can reconstruct the relationship between H3K36me3, a mark typically associated with active gene transcription, and RNA-seq measurements in gene bodies. We restricted our comparison to 47 cell types in which RNA-seq measurements were available from the Roadmap consortium. In this analysis, Avocado captures the relationship the best, and ChromImpute the worst. (Additional file 4: Fig. C.2).

We then considered relationships across both histone marks and genomic loci, focusing on the relationship between marks in the promoter and the gene body. Specifically,

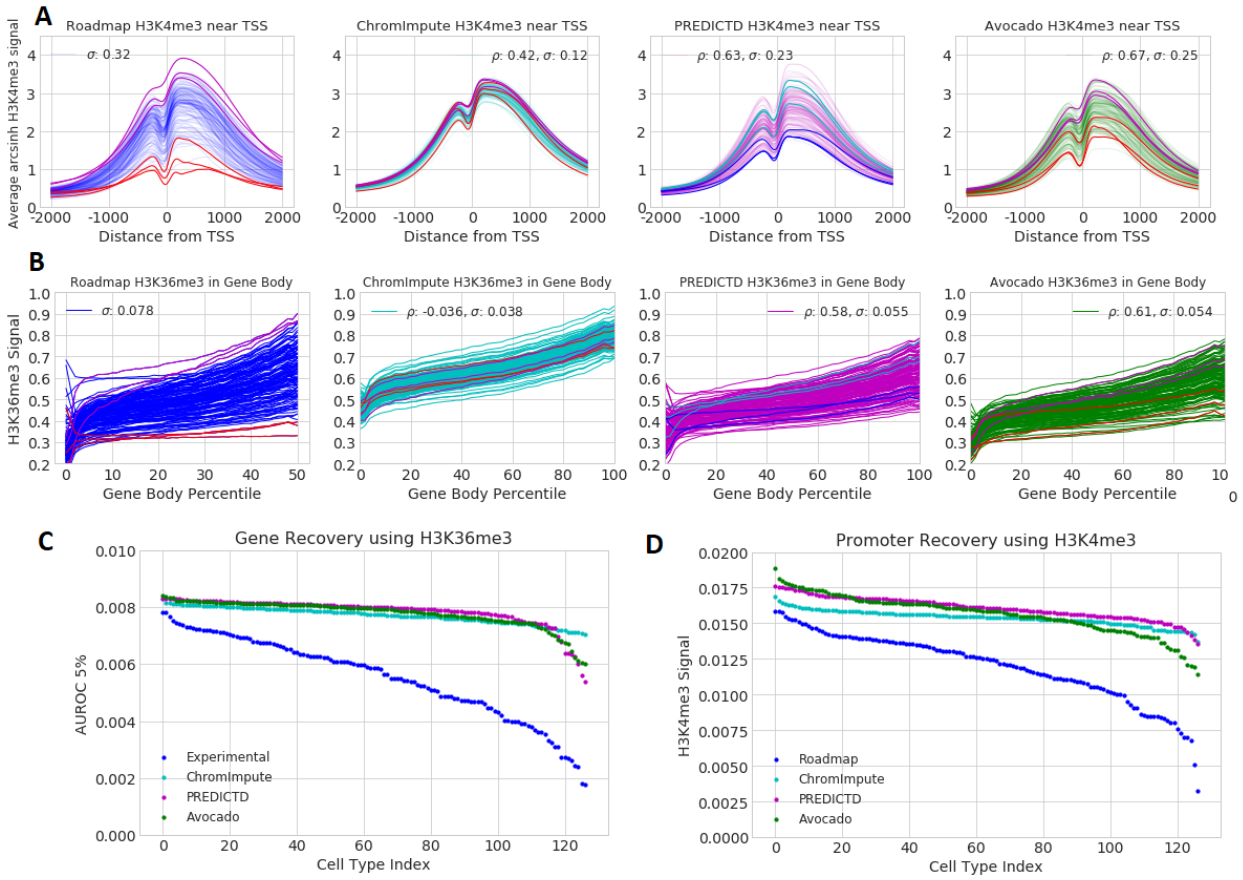



Figure C.1: **Aggregate measures of H3K4me3 and H3K36me3 in ChIP-seq experiments and across imputation methods.** (a) Each line displays the average H3K4me3 signal across all TSSs in chromosomes 1-22 for a single cell type after accounting for strand orientation of the gene. The variance of the signal across all cell types at each position is calculated and then averaged ($\sigma$). The area under each line is used to define a ranking, and the spearman correlation ($\rho$) is calculated between each of the three imputation approaches and the ChIP-seq data. (b) The same as (a)a except for H3K36me3 signal in gene bodies. (c) The GeneRecov performance measure for each cell type, which is the area under the ROC curve at 5% FPR when using H3K36me3 to predict gene bodies across chromosomes 1 through 22. (d) The PromRecov performance measure for each cell type, which is the area under the ROC curve at 5% FPR when using H3K4me3 to predict promoters across chromosomes 1 through 22.

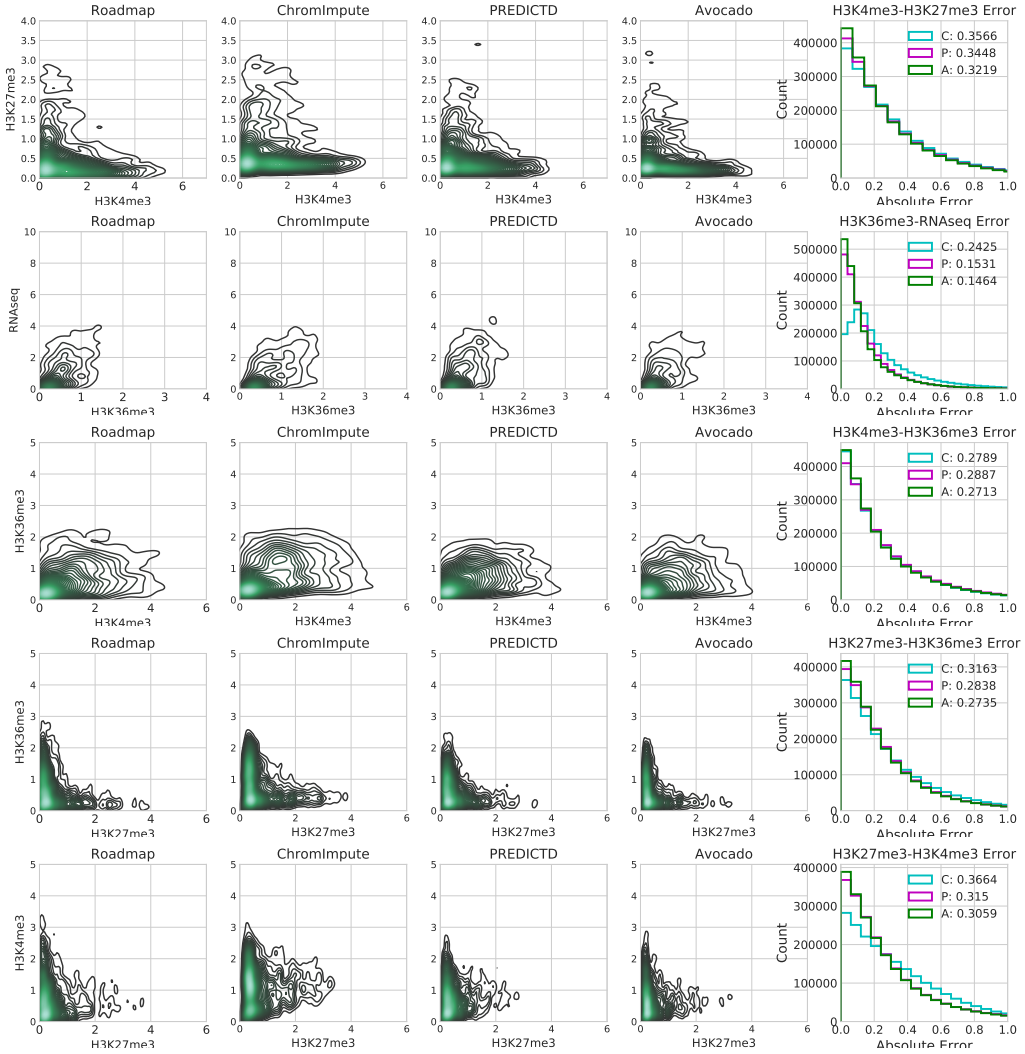



Figure C.2: **The relationships between pairs of histone modifications.** These panels show, going from left to right, the signal values in the Roadmap compendium, the imputed signal values from ChromImpute, imputed signal values from PREDICTD, the imputed signal values from Avocado, and the distribution of the absolute error in reconstructing the relationship. In the rightmost panels the legend denotes ChromImpute as C, PREDICTD as P, and Avocado as A. Because each plot contains over 2 million samples, the contour plots are generated on a randomly selected one thousandth of the data, though the error histogram is generated from the full set of samples.

we consider the relationship between H3K4me3 in the promoter region with H3K36me3 in the gene body, because an enrichment of the activating mark should lead to higher levels of the transcription-associated mark. Likewise, we would expect that an enrichment in H3K27me3 in the promoter region should lead to a depletion of H3K36me3 in the gene body. A priori, we expect that ChromImpute and Avocado would do particularly well at reconstructing these interactions because they both take as input information from many nearby genomic loci, whereas PREDICTD treats each genomic position independently. However, we find that while PREDICTD does the worst at reconstructing the relationship between H3K4me3 and H3K36me3, ChromImpute performs much worse at connecting H3K27me3 and H3k36me3 (Additional file 4: Table F.1). Interestingly, despite ChromImpute having an overall negative correlation between H3K27me3 and H3K36me3, as ChromImpute's imputed value of H3K27me3 increases so too does the minimum value of H3K36me3 (Additional file 4: Fig. C.2). This trend exists to a much lesser extent in the Avocado model, but is not supported by the ChIP-seq signal.

| | ChromImpute | PREDICTD | Avocado |
|---|---|---|---|
| **H3K4me4 - H3K27me3** | 0.3566 | 0.3448 | **0.3219** |
| **H3K4me1 - H3K27me3** | 0.3664 | 0.3150 | **0.3059** |
| **H3K36me3 - RNAseq** | 0.2425 | 0.1531 | **0.1464** |
| **H3K4me3 - H3K36me3** | 0.2789 | 0.2887 | **0.2713** |
| **H3K27me3 - H3K36me3** | 0.3163 | 0.2838 | **0.2735** |

Table C.1: Evaluation of ChromImpute, PREDICTD, and Avocado at reconstructing relationships between different histone marks across the genome according to the mean absolute error. The best result is in boldface for each comparison.

# Appendix D

# INSPECTION OF AVOCADO'S LEARNED EMBEDDINGS

We inspected the three clusters in Fig. 3a to better understand the types of loci that each cluster contains. Specifically, we wanted to understand whether the loci comprising the "mixed" cluster exhibited low average signal across all cell types because the loci were active in a very cell type-specific manner, or simply always demonstrated low signal. We began with the epigenomic signal ±2kbp around each locus in each cell type. This data was then divided into four sets: (1) H3K4me3 in promoters, (2) H3K27ac in promoters, (3) H3K4me3 in enhancers, and (4) H3K27ac in enhancers. We averaged each signal across all cell types but partitioned the loci based on which cluster from Fig. 3a they were a part of (Fig. 3b). We will refer to the epigenomic signal ±2kbp around a locus in a particular cell type as a "profile", so that it is not confused with the term "locus". Thus, each locus has one profile per cell type.

Next, we applied k-means clustering to each of these four sets separately to split the profiles into "high-signal" profiles and "low-signal" profiles. We adopt this terminology rather than the more traditional term "cluster" so as to not confuse these with the three clusters from Fig. 3a. As expected, the average high-signal profile shows patterns commonly seen with active functional elements, whereas the average low-signal profile shows almost no signal (Additional file 5: Fig. D.1). Furthermore, the average high-signal profile looks consistent across all three clusters, giving initial evidence that the mixed cluster is not made up exclusively of low-signal profiles.

Lastly, we adopted a more comprehensive view of the signal partitions by examining the number of high-signal profiles per locus. For each set, we examined the partition that each

locus was assigned to in each cell type (i.e. each profile). We then summed the number of cell types that exhibited high-signal profiles per locus (Additional file 5: Fig. D.2). We found that, although the mixed cluster appeared to be made up predominately of loci that exhibit low signal in all cell types, there are indeed many loci that exhibit high signal in a very cell type-specific manner. It is likely that, at loci that exhibit lower signal in all cell types, a weaker regulatory signal is sufficient for regulatory function. These observations explain why a model like Avocado, which is trained using the signal strength directly, groups these loci together, separate from either the promoter or the enhancer cluster.

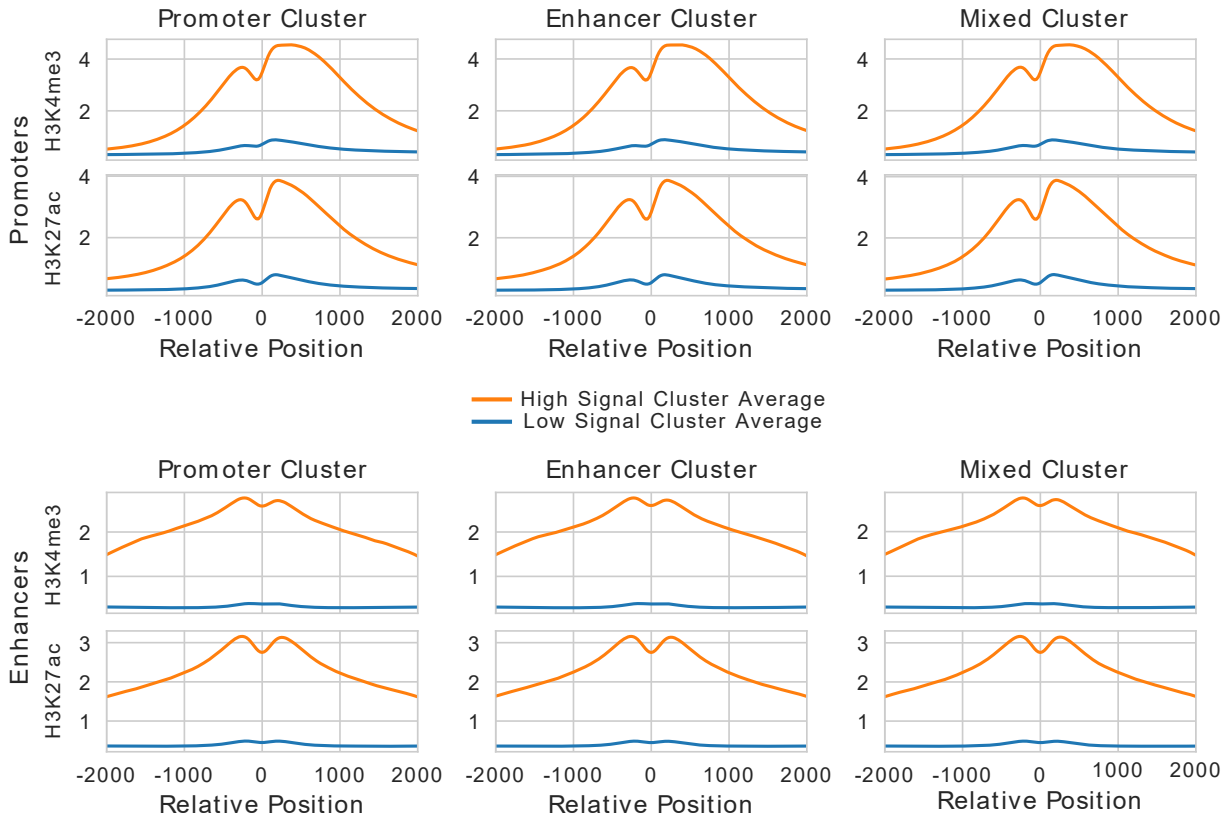



Figure D.1: **Average epigenomic profiles of clustered loci.** The average epigenomic activity of loci clustered into a "high" signal cluster (orange) and a "low" signal cluster (blue). The average profile for these clusters is shown for each of the three clusters (columns) and four sets (rows)

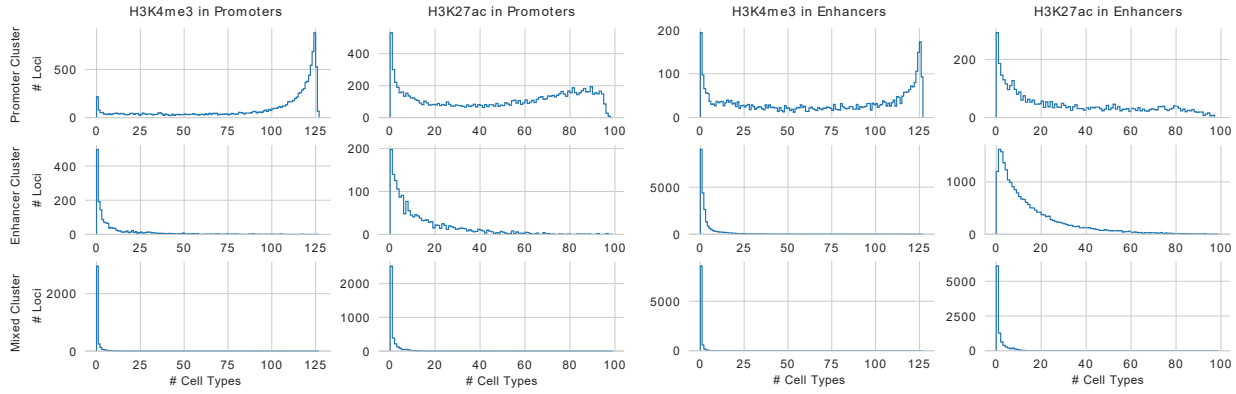


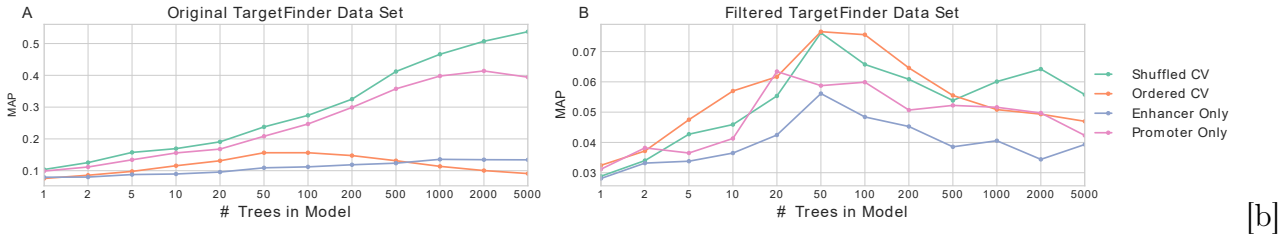

Figure D.2: **Cell type specificity of profile signals** Each panel shows a distribution of the number of cell types that each profile exhibits high signal. These profiles come from each of the four sets (columns) and are partitioned according to the three original clusters (rows).

# Appendix E

# PROMOTER-ENHANCER INTERACTION DATA SET

The set of promoter-enhancer interactions used to evaluate the TargetFinder model [16] has been recently shown to contain biases related to the pairwise nature of the task [37]. The bias arises for two reasons. First, the data set includes features derived from the window between the promoter and the enhancer, and these features are highly correlated between examples whose windows overlap. This correlation leads to a leakage of information when regions of the genome are in windows of examples in both the training and the test set. Fortunately, this issue can be easily corrected by simply removing the problematic features. The second issue is that when the data set was constructed, an equal number of positive and negative interactions were sampled at each genomic distance. Consequently, many promoters occur repeatedly and only in the context of a negative interaction. When promoter-enhancer pairs are randomly assigned to both the training and test sets, as is the case with the TargetFinder model, then a sufficiently complicated model can simply memorize these repeated promoters as never interacting. These issues are described more thoroughly by Xi and Beer [37].

To construct a data set without these biases, we choose the simple approach of filtering out interactions such that each promoter occurs only once in each cell type. While we do not also enforce that enhancers can only occur once, we greedily select pairs where the enhancer has not yet been part of an example. This approach yields a data set with 27,048 interactions across all four cell types in chromosomes 1 through 22, where each interaction corresponds to a unique promoter in its cell type. Among these interactions, nearly all (26,707) have unique enhancers as well; 158 enhancers are seen twice, 7 are seen 3 times, and 1 is seen four




[b]

Figure E.1: **Model performance on the original and filtered TargetFinder data sets.** (A) The performance of gradient boosting classifiers on the TargetFinder data set split by randomly assigning interactions to folds (cyan) or ordering interactions by genomic coordinate and then splitting into consecutive blocks (orange). Further, when randomly assigning interactions to folds, the performance is shown when using only features from the enhancer (blue) and when using features only from the promoter (pink). (B) Similar to (A), but on the new filtered data set.

times. After this filtering step, IMR90 has 4,702 pairs, of which 82 are positive interactions; GM12878 has 7,881 pairs, of which 181 are positive interactions; HeLa-S3 has 7,060 pairs, of which 121 are positive interactions; and K562 has 7,405 pairs, of which 145 are positive interactions. Promoters and enhancers were defined by Whalen *et al.* to be those identified using combined Segway and ChromHMM annotations for the respective cell types. Further, promoters were then filtered to be only those in GENCODEv19 that were actively transcribed (mean FPKM $> 0.3$ and IDR $< 0.1$ using corresponding ENCODE RNA-seq data for each cell type). Thus, both the positive and negative sets for our predictive tasks were defined on active regulatory elements.

We verify that the source of bias has been removed using the same techniques used by Xi and Beer [37]. First, we plot the performance of gradient boosting models with an increasing number of trees evaluated using five-fold cross-validation with examples randomly assigned to folds. We observe a steadily increasing performance on the original data set, similar to

that reported by Whalen et al. [16], but not on the new data set (Additional file 6: Fig. E.1, orange). Next, we sort examples based on their genomic coordinates and assigned samples to folds based on this ordering. We observe the same diminished performance on the original data set in comparison to random splitting that was observed by Xi and Beer (blue), but similar performance on the new data set compared to random splitting. Lastly, to confirm that this issue is related to memorizing which promoters never interact, we train models using features from only the promoter (green) or the enhancer (brown). We observe similar performance in the original data set when using all features or only using features derived from the promoter region. However, we do not observe this trend in the new data set. Taken together, these results confirm that the new data set does not exhibit the same issue as the original data set used by Whalen et al [16].

# Appendix F

# INITIAL TRAINING STEP

We tested whether the performance of Avocado was sensitive to the set of regions used in the initial training step when the assay embeddings, cell type embeddings, and neural network weights are learned. Our primary model uses the ENCODE Pilot Regions for this step. In this supplementary analysis, we trained five additional models using signal from contiguous blocks of the same size as the ENCODE Pilot Regions extracted from the centers of chromosomes 1 through 5. Then, for each of the five models, we froze the assay embeddings, cell type embeddings, and the neural network weights, and we fit the genome factors for chromosome 16. These models were each trained using experiments from four of the five folds from the five-fold cross-validation in both of these steps and then evaluated based on their ability to impute the remaining fifth fold of experiments in chromosome 16. We found that the model trained using the ENCODE Pilot Regions were similar to those trained using the contiguous blocks of the genome.

| Step 1 Trained On | Step 2 Trained On | Test Set MSE |
|---:|:---:|:---:|
| ENCODE Pilot Regions | chr16 | **0.0733** |
| chr1 | chr16 | **0.0755** |
| chr2 | chr16 | **0.0689** |
| chr3 | chr16 | **0.0700** |
| chr4 | chr16 | **0.0711** |
| chr5 | chr16 | **0.0770** |

Table F.1: Performance of six models when evaluated using the same region—chromosome 16—but trained using different regions for the initial training step.

# Appendix G

# CHAPTER 2 SUPPLEMENT

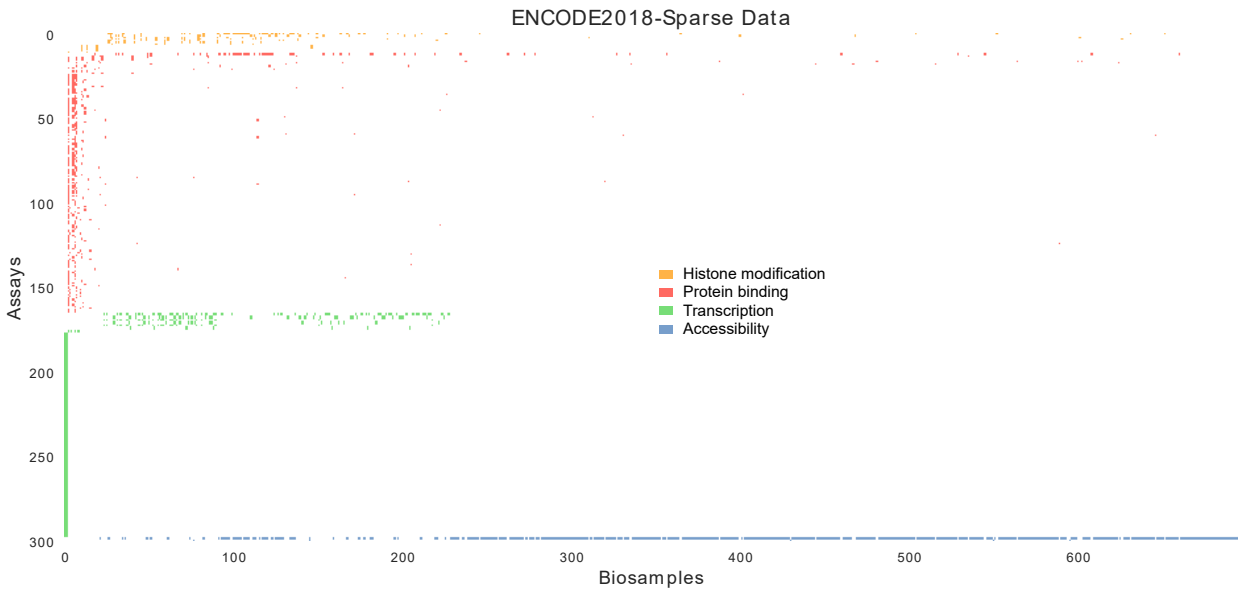



Figure G.1: **The ENCODE2018-Sparse data matrix.** The ENCODE2018-Sparse data matrix includes all assays that were performed in fewer than 5 biosamples, and all biosamples that were characterized by fewer than 5 assays. Experiments that have been performed are displayed as colored rectangles, and experiments that have not been performed are displayed as white. The color corresponds to the general type of assay, with blue indicating chromatin accessibility, orange indicating histone modification, red indicating protein binding, and green indicating transcription. This figure displays all biosamples and the top 300 assays ranked number of biosamples that they were performed in.

| Biosample | iPSC | PC-3 | liver | liver | liver | liver | liver | liver | liver |
|---|---|---|---|---|---|---|---|---|---|
| Assay | CTCF | CTCF | EGR1 | FOXA1 | GABPA | JUND | MAX | REST | TAF1 |
| Method | | | | | | | | | |
| Yuanfang Guan | 0.655 | 0.564 | 0.433 | 0.341 | 0.355 | 0.535 | 0.473 | 0.386 | 0.320 |
| dxquang | **0.811** | 0.717 | 0.315 | 0.440 | 0.340 | 0.286 | 0.394 | 0.384 | 0.323 |
| autosome.ru | 0.709 | 0.458 | 0.364 | 0.323 | 0.360 | 0.441 | 0.434 | 0.353 | 0.261 |
| J-TEAM | 0.754 | 0.688 | 0.379 | **0.484** | 0.334 | 0.450 | 0.444 | 0.271 | 0.337 |
| Avocado | 0.665 | **0.724** | **0.542** | 0.401 | **0.431** | **0.630** | **0.570** | **0.513** | **0.425** |
| Similar Biosample | — | — | 0.410 | 0.437 | 0.257 | 0.581 | 0.500 | 0.457 | — |
| Same Biosample | 0.671 | 0.818 | 0.645 | 0.691 | 0.580 | 0.716 | 0.619 | 0.617 | 0.561 |
| Average Activity | 0.530 | 0.664 | 0.321 | 0.380 | 0.287 | 0.273 | 0.421 | 0.215 | 0.256 |

Table G.1: **Comparison of methods on ENCODE-DREAM challenge test set.** The equal precision-recall (EPR) computed across nine epigenomic experiments in the ENCODE-DREAM challenge test set in chromosome 21. For each track, the score for the best-performing predictive model is in boldface.

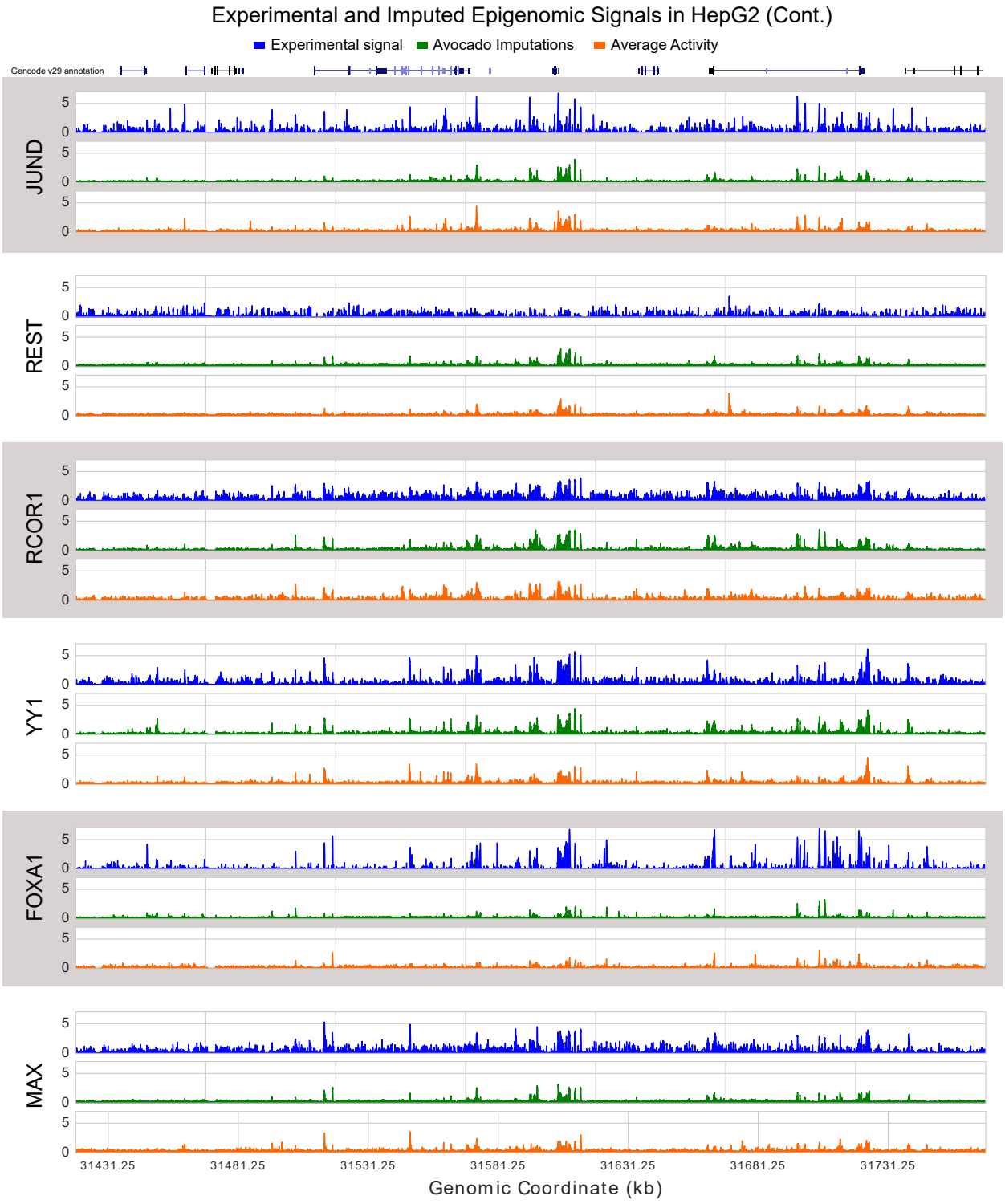



Figure G.2: **Imputations of various transcription factors.** This figure extends Fig. 2a by showing the experimental signal (in blue), Avocado imputations (in green), and average activity baseline (in orange), for six additional transcription factors at the same locus.

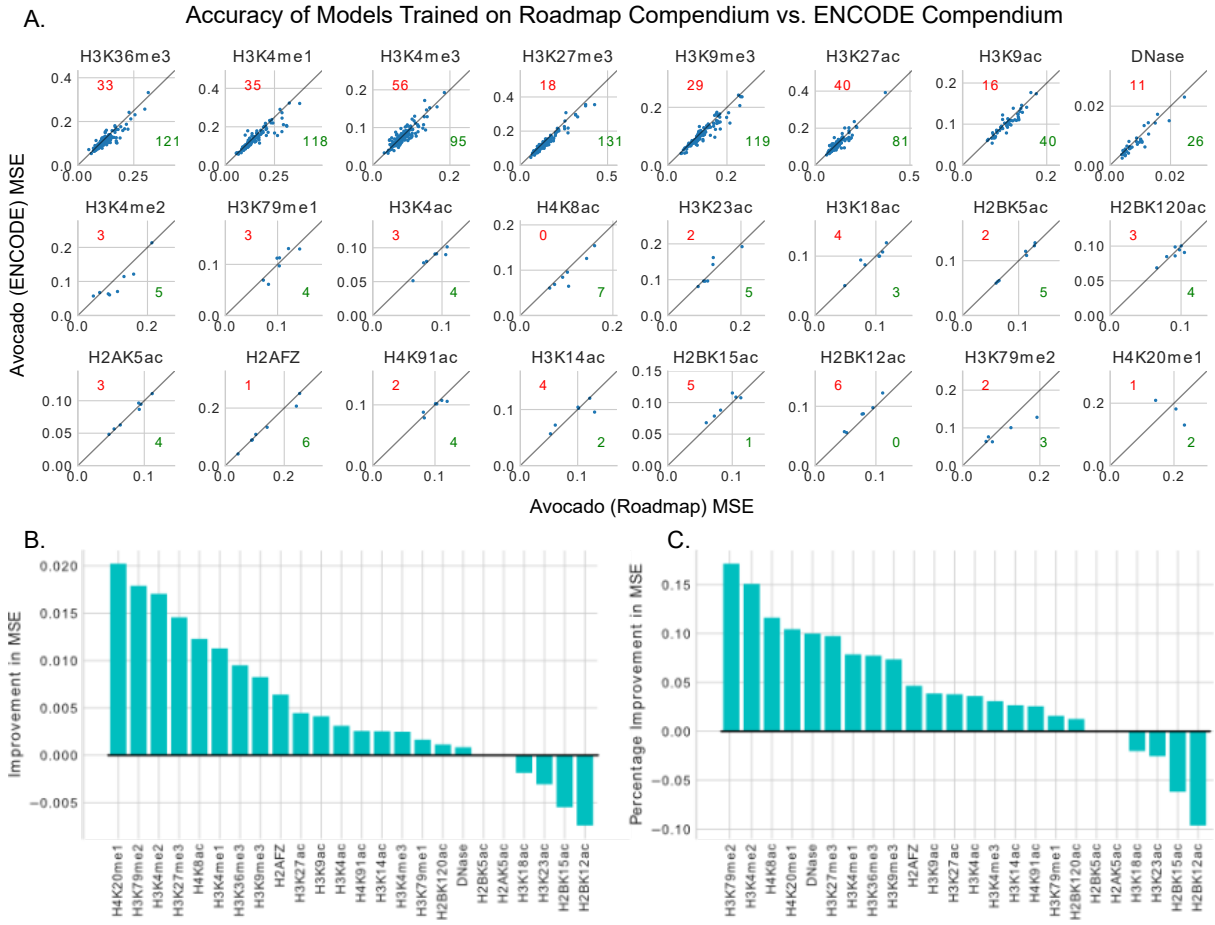



Figure G.3: **Accuracies of models trained on either the Roadmap compendium or the ENCODE2018-Core data.** (A) Each panel depicts the error of models trained on either the ENCODE2018-Core dataset (Avocado (ENCODE)), or those tracks from the ENCODE2018-Core dataset that were provided by the Roadmap Epigenomics Consortium (Avocado (Roadmap)), when imputing the tracks contained in the latter. Each dot corresponds to MSE on a single track, and each panel corrsponds to all tracks from that assay. Dots below the diagonal line indicate that the model trained on the ENCODE2018-Core dataset outperformed the model trained on the Roadmap dataset, with the number in green specifying the number of such tracks, and dots above the line indicate the reverse, specified by the red number. (B) The improvement in performance when using a model trained on the full ENCODE2018-Core dataset versus one trained on only the Roadmap tracks. (C) Similar to (B), except the percentage improvement.

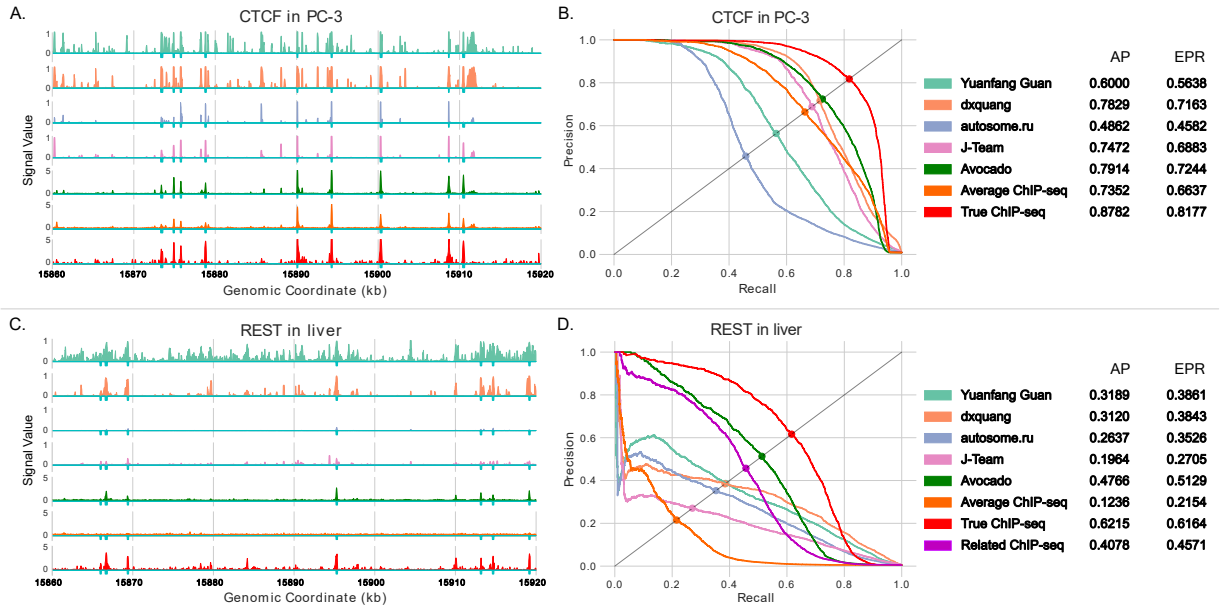



Figure G.4: **Avocado imputes transcription factors correctly.** (A) Example predictions from a region of chromosome 21 for the top four ENCODE-DREAM participants, Avocado, and experimental ChIP-seq data measuring CTCF binding in PC-3. Cyan ticks at the bottom of the tracks indicate peak calls. (B) A precision-recall curve showing the performance of the four participants and Avocado in chromosome 21. As additional baselines, the experimental ChIP-seq signal (red) and the average signal across Avocado's training set (orange) were included in the comparison. For each approach, the average precision (AP) and the equal-precision-recall (EPR) are reported, and the position on the curve where the EPR lies is marked as a dot. (C) Similar to (A), except for REST binding in a liver biosample. (D) Similar to (B), except for REST binding in a liver biosample. The experimental signal from a different liver biosample is used as a further baseline (magenta).

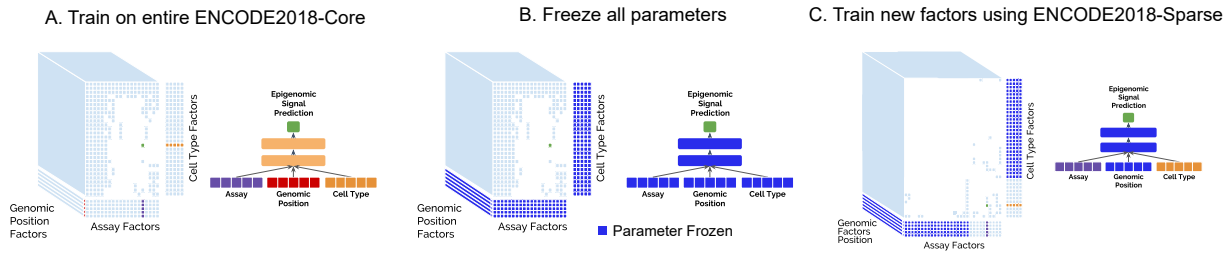



Figure G.5: **Transfer learning methodology.** A schematic of the three step process to train Avocado on the ENCODE2018-Sparse dataset. (A) Train Avocado on the entire ENCODE2018-Core dataset as normal. (B) Freeze the weights of both the neural network and the factors. (C) Train only the factor values for new biosamples and assays that are being added to the model.

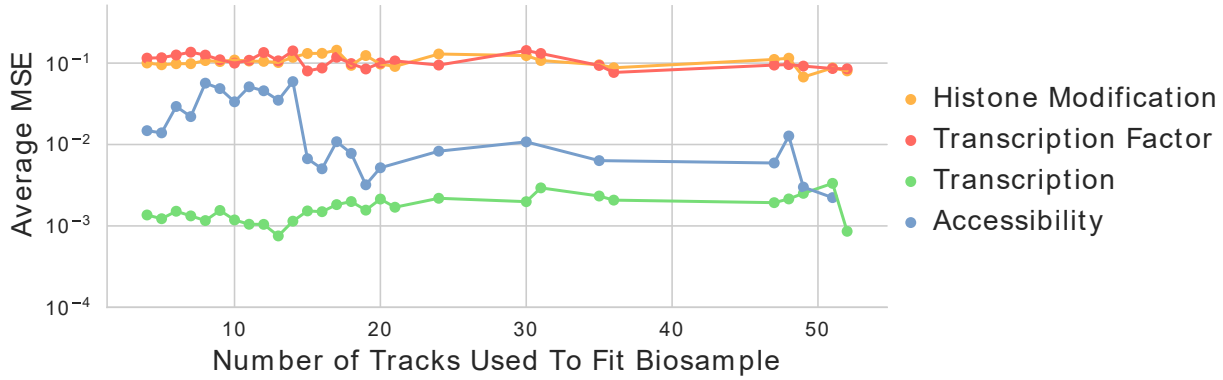



Figure G.6: **Trends in imputation performance by number of assays per biosample.** The MSE of each of the 3,814 experiments in the ENCODE2018-Core data set averaged across both the number of assays used to fit the biosample factors of the experiment and the form of biological activity.

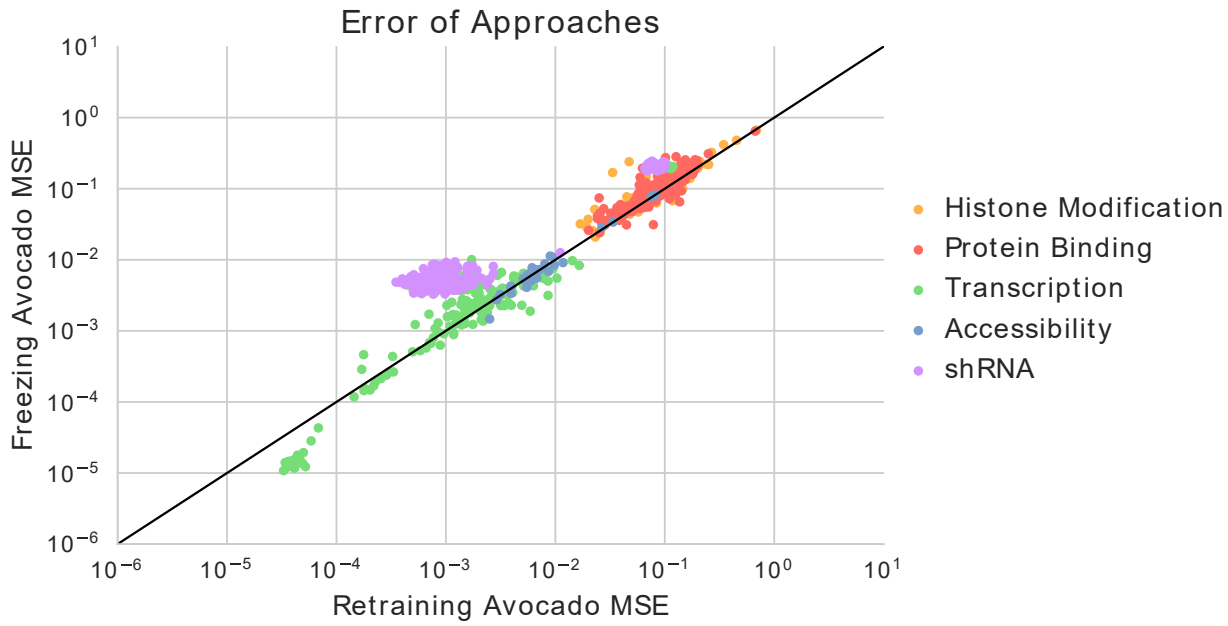



Figure G.7: **Error of two methods for incorporating new experiments.** The MSE from each of 965 tracks of experimental data from the test set of ENCODE2018-Sparse from either retraining Avocado to include new experiments (x-axis) or freezing parameters from a pre-trained model and only training new biosample and assay factors (y-axis). The experiments are colored according to their type of biological activity.

# Appendix H

# FOLLOW-UP ON EXPERIMENTS IN WHICH AVOCADO PERFORMS POORLY

We investigated further the 13 experiments for which Avocado underperforms the average activity predictor. This set is enriched for measurements of transcription: 10 of the 13 experiments (77%) measure gene transcription, such as CAGE, RAMPAGE, microRNA-seq, polyA-depleted RNA-seq, and small RNA-seq. The remaining three assays for which Avocado does not outperform the average activity predictor according to mseGlobal are H3K9me2, EP300, and ATAC-seq. Further investigation on the ENCODE portal showed all H3K9me2 experiments had audit warnings and that only one of the experiments, in iPSC cells, had a fraction of reads in peaks (FRiP) score above the general quality control theshold of 1% used for ChIP-seq experiments [91]. While standards for ATAC-seq experiments have been released, the quality metrics associated with the experiments we used had not yet been released on the ENCODE portal, and so we were unable to verify their quality.

We then investigated those assays that Avocado underperformed the average activity baseline on other performance measures. First, we notice that Avocado imputed transcription poorly across all measures. On all measures except mse1imp, at least 9 of the underperforming assays related to measurements of transcription. Second, we notice that H3K9me2 and ATAC-seq are poor performers across all metrics as well. The consistent poor performance of these 11 assays may give a more pessimistic view of Avocado's performance in general.

We then evaluated assay performance across different performance measures. We noticed that Avocado only underperforms the average activity baseline on only five of the problematic transcription assays. This suggests that Avocado may have a higher precision than recall

when it comes to predicting exon-specific activity. However, one weakness of mse1imp and mse1obs is that the percentage used to approximate peak coverage, 1%, may be appropriate for histone marks, but is not as specific to areas of transcription.

# Appendix I

# FURTHER ANALYSES OF ENCODE CHALLENGE RESULTS

Comparing the performance of Avocado to the ENCODE Transcription Factor Binding Prediction Challenge participants is challenging for several reasons relating to differences in evaluation setting and model inputs. Accordingly, we performed several follow-up experiments to better understand the effect that these differences may have had on performance.

We began by characterizing the effect on predictive power produced by training models on the same loci that predictions were being made for. Described using the evaluation settings from Schreiber *et al.* [69], our original comparisons evaluated Avocado in the "cross-cell type" setting because the model had been trained and evaluated on the same chromosome, but we evaluated the challenge participants in the "hybrid" setting because their models had been trained and evaluated on different chromosomes. This does not mean that Avocado was evaluated on the training set, but rather that Avocado was evaluated on the ability to predict held-out tracks at the same loci that it was trained on, i.e., chromosome 21. When we evaluated all models using the cross-cell type setting by using chromosome 17, which was a part of the ENCODE challenge training set, we observed similar trends as in the original evaluation setting (Additional file 3: Table 1). This suggests that the evaluation setting was not a major confounder of performance.

Next, we investigated the extent to which Avocado leveraged the tracks of epigenomic data that were not available to the participants. This analysis involved training Avocado models in three settings. The first (denoted Avo0 in Additional file 3: Table 2) was to train Avocado on all tracks in the ENCODE2018-Core data set except for those in the challenge test set. This is in contrast to the evaluations presented in the main text, which are done

on imputations made as a part of five-fold cross-validation. Because the model in this first setting was trained using more tracks than the models trained as a part of five-fold cross-validation, the resulting imputations should serve as an upper bound of performance for Avocado using the ENCODE2018-Core data set. The second setting (denoted Avo1) involved training Avocado using only DNase-seq and RNA-seq from the biosamples in the challenge, as well as the transcription factor binding tracks present in the training and validation sets of the challenge. In this setting, the model would have strictly less information than the challenge participants, who also had access to nucleotide sequence. The final evaluation setting was similar to the first setting, except that all tracks from biosamples in the challenge test set that were not DNase-seq and RNA-seq were also removed. This setting evaluates the ability of Avocado to leverage the ENCODE compendium to make imputations in biosamples while still using the same epigenomic input that the participants had.

Unfortunately, while it was simple to use the DNase-seq and RNA-seq experiments in our data sets for two biosamples (PC-3 and iPSC), there were several reasons why it was difficult to find corresponding experiments for the tracks denoted as "liver." The first difficulty is that the challenge test tracks actually come from two different liver biosamples: liver male adult 32 years (J099) and liver female child 4 years (J468), and Avocado treats these as distinct biosamples. The effect that including related data may have had was controlled for with the inclusion of the "similar biosample" row in Table 1. To further complicate matters, neither DNase-seq nor RNA-seq experiments had been performed in either of these liver biosamples. In the challenge, the RNA-seq track originates from a third, embryonic, biosample—liver female embryo 20 weeks and male embryo 22 weeks (J325)—and the DNase-seq track (https://www.encodeproject.org/files/ENCFF530SFF/) comes from a fourth biosample—right lobe of liver female adult 53 years (J288). To further complicate the comparison, the DNase experiment was revoked after the challenge and subsequently replaced with a higher quality experiment before we assembled the ENCODE2018-Core data set.

We addressed these difficulties in two ways. The first (denoted Avo2) was to simply remove J099, J468, and J325 from the model, and to use the RNA-seq and DNase-seq experiments from J288 (a biosample only present in the ENCODE2018-Sparse data set) as our new "liver" biosample. This approach ensured that there were matching DNase-seq and RNA-seq experiments from the same biosample, but had the drawback that neither experiment had been provided to the challenge participants nor were matched with the labels. The second approach (denoted Avo3) was to train two models, one to impute the tracks from J099 and one made to impute the tracks from J468. In each case, we fit the biosample that we are making predictions for using the DNase-seq and RNA-seq tracks from J288, and we fit the other biosample using include all of its assays. This evaluation setting has the benefit of measuring the effect that simply including data from a related liver biosample during training would have on model performance.

We observed the expected results in the first two settings (Avo0 and Avo1, Additional file 3: Table 2). In the first setting, the model either outperformed or exhibited comparable performance to the original Avocado model on each of the challenge test set tracks. In the second setting, the model performed very poorly on all tracks from the liver biosample, potentially due to the issues indicated above, and also performed worse than the original Avocado model at predicting CTCF in PC-3. Interestingly, we observed similar performance at predicting CTCF in iPSC as the first setting, despite having far fewer tracks as input. These results suggest that Avocado does indeed leverage the diversity of signals in the ENCODE compendium to make accurate predictions.

The third setting, where only DNase-seq and RNA-seq were used to fit the test set biosamples, generally showed similar results to the second setting. Specifically, both Avo2 and Avo3 underperformed the challenge participants on each of the tracks that were from liver. The mismatches in the DNase-seq and RNA-seq experiments denoted as liver are likely one reason for this poor performance, but it was difficult for us to assess whether

| Biosample<br>Assay<br>Method | iPSC<br>CTCF | PC-3<br>CTCF | liver<br>EGR1 | liver<br>FOXA1 | liver<br>GABPA | liver<br>JUND | liver<br>MAX | liver<br>REST | liver<br>TAF1 |
|---|---|---|---|---|---|---|---|---|---|
| Yuanfang Guan | 0.742 | 0.627 | 0.455 | 0.358 | 0.520 | 0.570 | 0.520 | 0.427 | 0.368 |
| dxquang | **0.857** | 0.800 | 0.358 | **0.507** | 0.470 | 0.283 | 0.407 | 0.396 | 0.355 |
| autosome.ru | 0.764 | 0.515 | 0.387 | 0.310 | 0.486 | 0.428 | 0.454 | 0.364 | 0.300 |
| J-TEAM | 0.812 | 0.767 | 0.421 | 0.480 | 0.465 | 0.441 | 0.426 | 0.266 | 0.346 |
| Avocado | 0.758 | **0.856** | **0.571** | 0.376 | **0.542** | **0.692** | **0.676** | **0.585** | **0.542** |
| Similar Signal | 0.731 | 0.685 | 0.427 | 0.417 | 0.293 | 0.557 | 0.571 | 0.494 | 0.217 |
| Same Signal | 0.768 | 0.924 | 0.706 | 0.740 | 0.696 | 0.763 | 0.734 | 0.718 | 0.647 |
| Average Signal | 0.634 | 0.796 | 0.435 | 0.335 | 0.384 | 0.364 | 0.437 | 0.386 | 0.363 |

Table I.1: **Comparison of methods on chromosome 17 of the ENCODE-DREAM challenge test set.** The average precision computed across nine epigenomic experiments in the ENCODE-DREAM challenge test set in chromosome 17, which is one of the training set chromosomes. For each track, the score for the best-performing predictive model is in boldface.

they were the sole reason. Another potential reason for the poor performance was that the original Avocado model required the transcription factor binding signal from the related liver biosamples for reasonable performance, even if the model wasn't memorizing this signal. However, because Avo3—a model exposed to the transcription factor binding signal from the related liver biosample—still performed poorly, it seemed unlikely that this was a major reason. Interestingly, Avo2 achieved the highest performance at predicting CTCF of any Avocado model. The improvement in performance over Avo1 at predicting CTCF suggests that Avo2 was leveraging the epigenomic signal in the ENCODE compendium to make good predictions and that it is not necessary to have many assays for each biosample that one would want to make predictions in.

| Biosample | iPSC | PC-3 | liver | liver | liver | liver | liver | liver | liver |
|---|---|---|---|---|---|---|---|---|---|
| Assay | CTCF | CTCF | EGR1 | FOXA1 | GABPA | JUND | MAX | REST | TAF1 |
| Method | | | | | | | | | |
| Yuanfang Guan | 0.729 | 0.600 | 0.397 | 0.282 | 0.353 | 0.533 | 0.441 | 0.318 | 0.281 |
| dxquang | **0.866** | 0.783 | 0.274 | 0.399 | 0.347 | 0.260 | 0.330 | 0.311 | 0.264 |
| autosome.ru | 0.778 | 0.486 | 0.331 | 0.243 | 0.342 | 0.416 | 0.384 | 0.263 | 0.221 |
| J-TEAM | 0.812 | 0.747 | 0.363 | **0.462** | 0.344 | 0.415 | 0.377 | 0.196 | 0.272 |
| Avocado | 0.723 | 0.791 | 0.530 | 0.354 | **0.396** | **0.660** | **0.574** | 0.477 | 0.384 |
| Avo0 | 0.733 | 0.779 | **0.582** | 0.430 | 0.381 | 0.650 | 0.550 | **0.534** | **0.397** |
| Avo1 | 0.735 | 0.640 | 0.010 | 0.192 | 0.199 | 0.145 | 0.179 | 0.078 | 0.124 |
| Avo2 | 0.788 | **0.797** | 0.105 | 0.088 | 0.242 | 0.117 | 0.112 | 0.145 | 0.112 |
| Avo3 | 0.783 | 0.764 | 0.115 | 0.019 | 0.200 | 0.100 | 0.110 | 0.108 | 0.139 |
| Similar Signal | 0.627 | 0.570 | 0.363 | 0.389 | 0.226 | 0.568 | 0.446 | 0.408 | 0.096 |
| Same Signal | 0.741 | 0.878 | 0.648 | 0.716 | 0.573 | 0.731 | 0.622 | 0.622 | 0.556 |
| Average Signal | 0.574 | 0.736 | 0.324 | 0.299 | 0.253 | 0.375 | 0.336 | 0.327 | 0.197 |

Table I.2: **Comparison of alternate Avocado methods on ENCODE-DREAM challenge test set.** The average precision for four alternate Avocado models (Avo0-Avo3) computed across nine epigenomic experiments in the ENCODE-DREAM challenge test set in chromosome 21. The numbers from Table 1 are also shown for comparison. For each track, the score for the best-performing predictive model is in boldface.