Supplement to "Multi-scale deep tensor factorization learns a latent representation of the human epigenome "

Jacob Schreiber[1], Timothy Durham[2], Jeffrey Bilmes[1,3], and William Stafford Noble[1,2]

[1]Paul G. Allen School of Computer Science, University of Washington, Seattle, USA
[2]Department of Genome Sciences, University of Washington, Seattle, USA
[3]Department of Electrical Engineering, University of Washington, Seattle, USA

July 8, 2018

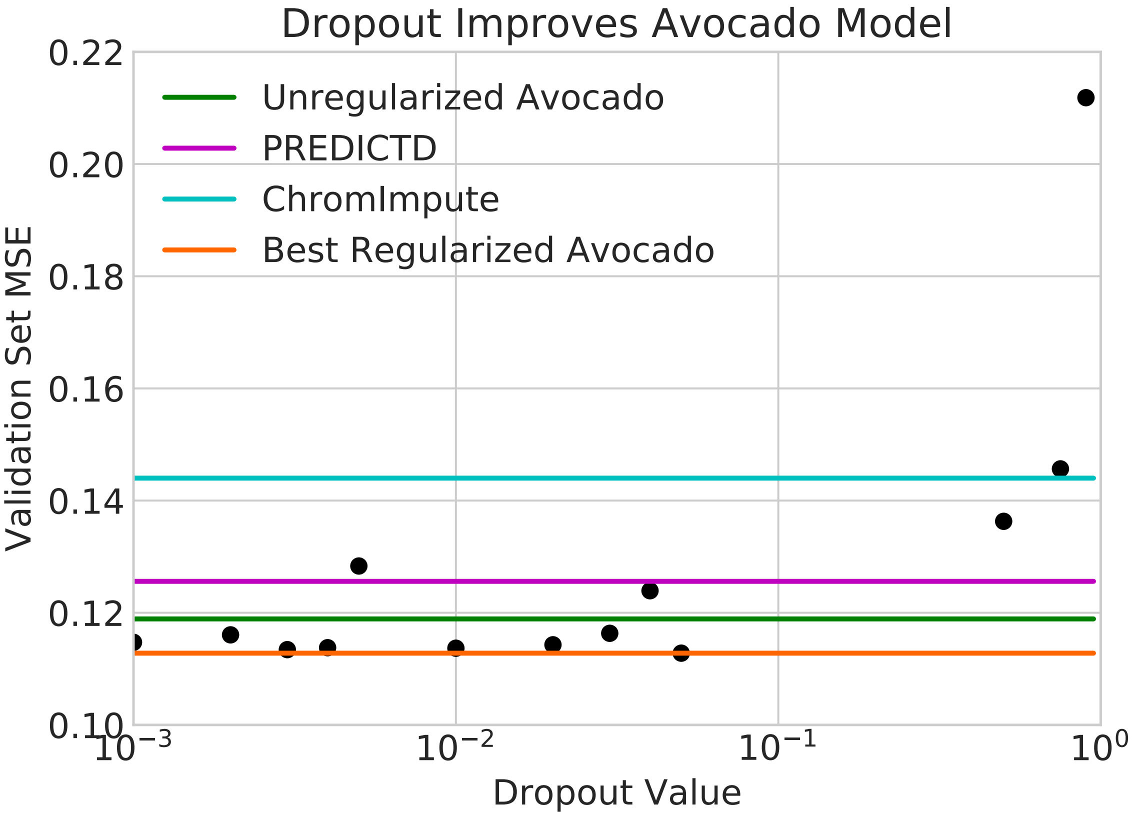

Figure S1: **Dropout improves the validation set performance of Avocado.** Each point corresponds to the performance of an Avocado model trained with a given dropout probability in the two hidden layers. The best performing model (in orange) outperforms not only the unregularized model (in green) but further improves over PREDICTD (in magenta) and ChromImpute (in cyan).

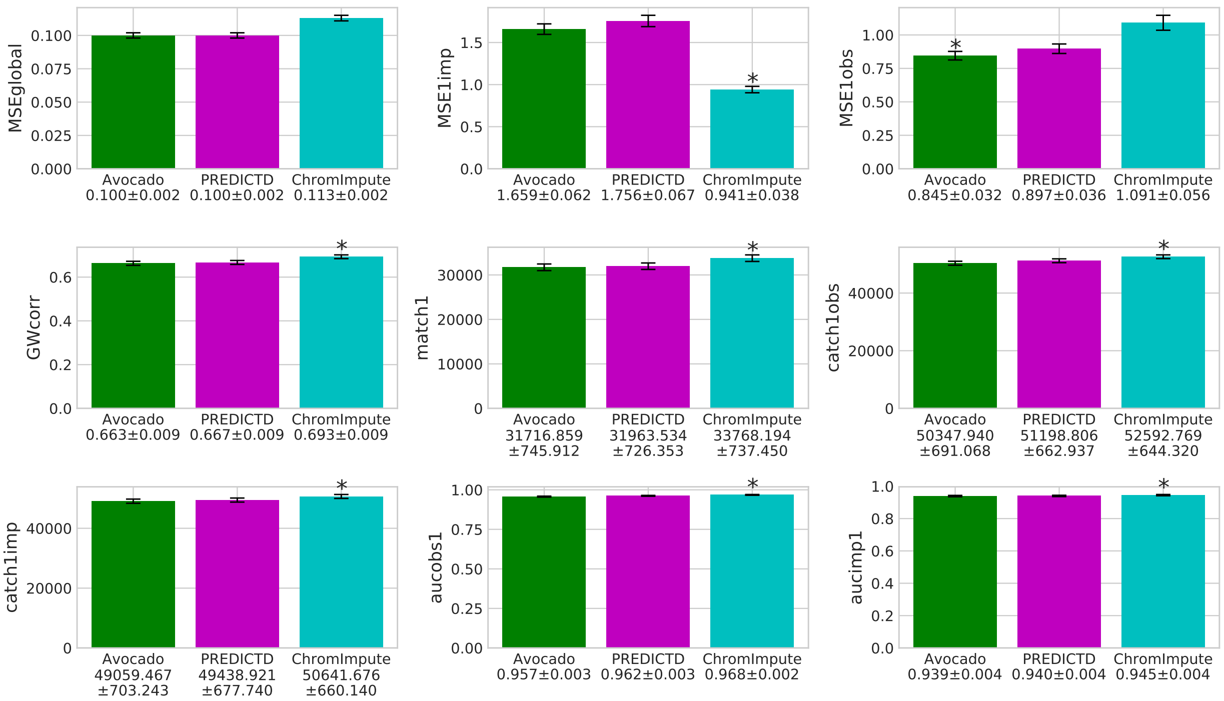

Figure S2: **Nine performance measures evaluated across the full genome for each imputation approach**. Each panel plots the value of a specified performance measure (y-axis), averaged across all 1,014 tracks. The performance measures correspond to those proposed by either Durham et al. or Ernst and Kellis. Briefly, MSEglobal is the MSE across the full span of the genome, MSE1imp is the MSE in the top 1% of genomic positions as ranked by the observed signal value, MSE1imp is the MSE in the top 1% of as ranked by the imputed signal value for each approach separately, GWcorr is the Pearson correlation across the full span of the genome, match1 is the number of genomic positions in the top 1% as ranked by observed signal value that are also in the top 1% as ranked by imputed signal value, catch1obs is the number of genomic positions in the top 1% as ranked by observed signal that are in the top 5% of genomic positions as ranked by imputed signal value, catch1imp is as catch1obs but reversed, aucobs1 is the area under the receiver operator characteristics curve (AUROC) when using the imputed signal to recover the top 1% as ranked by observed signal value, and aucimp1 is as aucobs1 but reversed. 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.

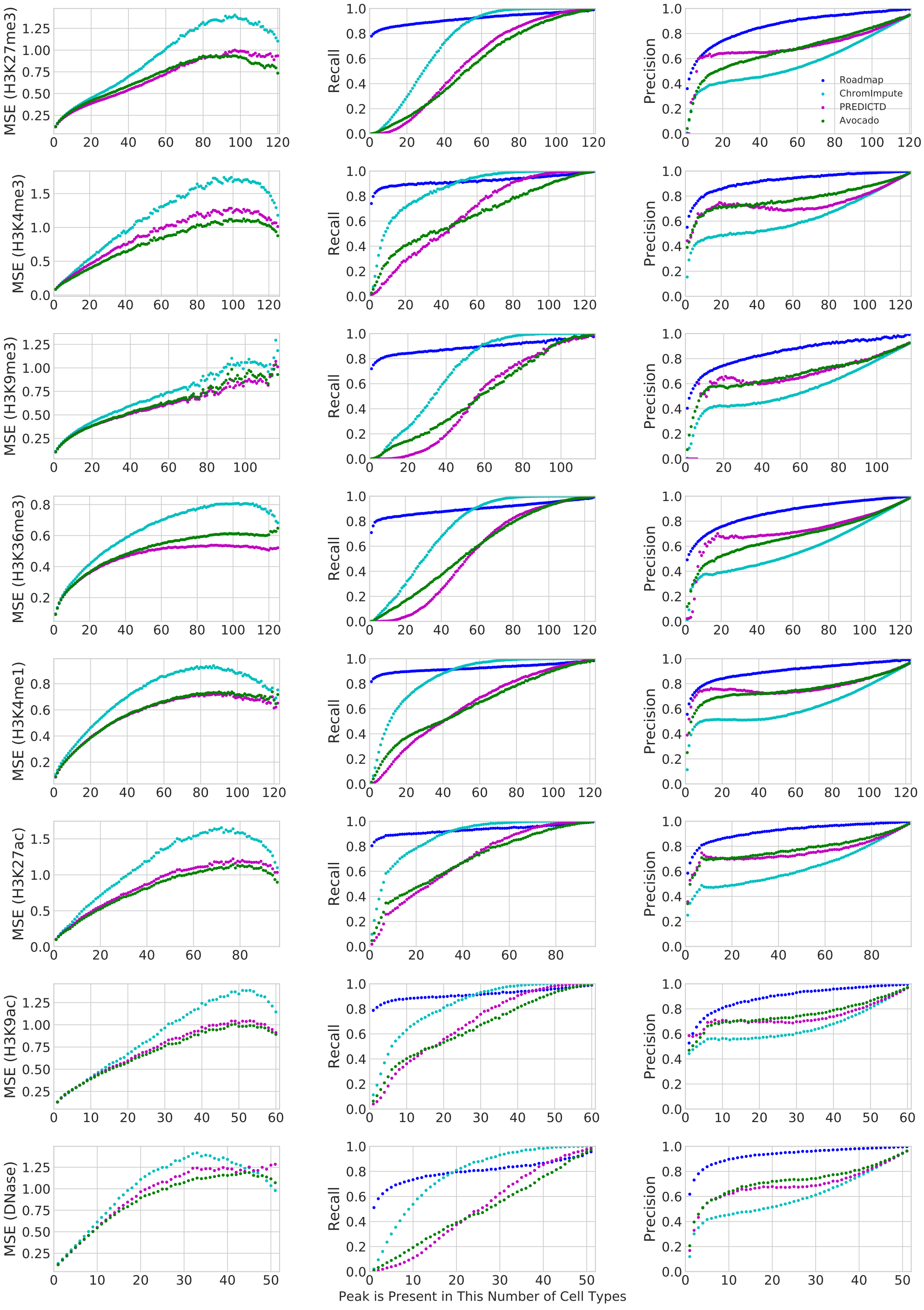

Figure S3: **Ability to recover cell type-specific peaks.** Each panel plots, for a given assay type, the MSE (left column), recall (middle column) or precision (right column) as a function of the number of cell types in which a given peak occurs. Only the 12 assays that have been performed in more than 10 cell types are shown.

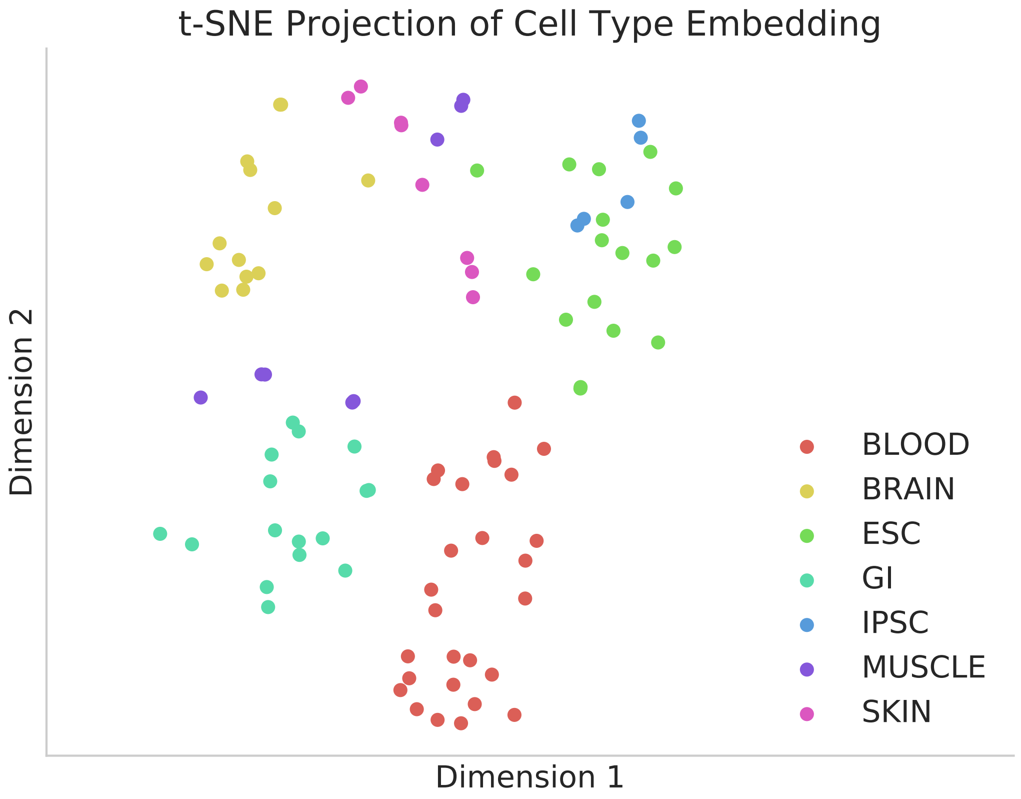

Figure S4: **Avocado latent cell type representation.** The figure shows a t-SNE projection [1] into two dimensions of the learned cell type embedding. Each point corresponds to a cell type and is colored by the Roadmap anatomy label. Cell types that belonged to anatomies with fewer than five cell types are not displayed.

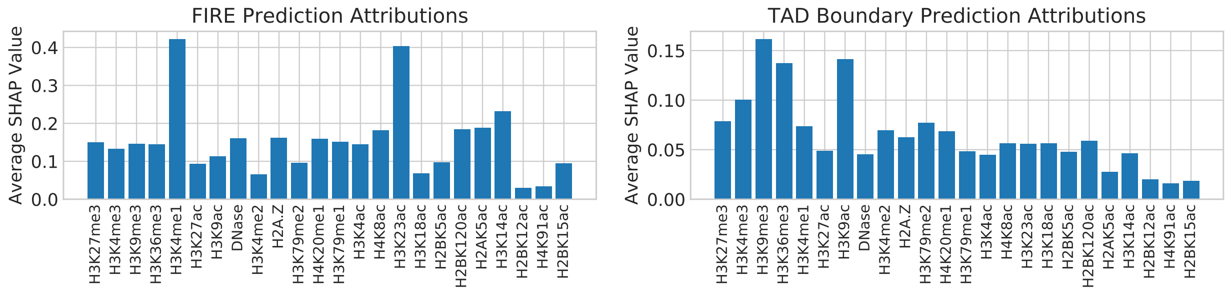

Figure S5: **Feature attribution performed on chromatin architecture prediction models.** (a) Feature attributions for the seven histone modifications in each of the seven cell types for the prediction of FIREs, as determined by SHAP. These attributions are calculated by retraining a gradient boosting classifier on all samples for a cell type, calculating the SHAP values for each sample, and returning the mean average SHAP value across for each histone modification. (b) Similar to (a), but for TAD boundary prediction.

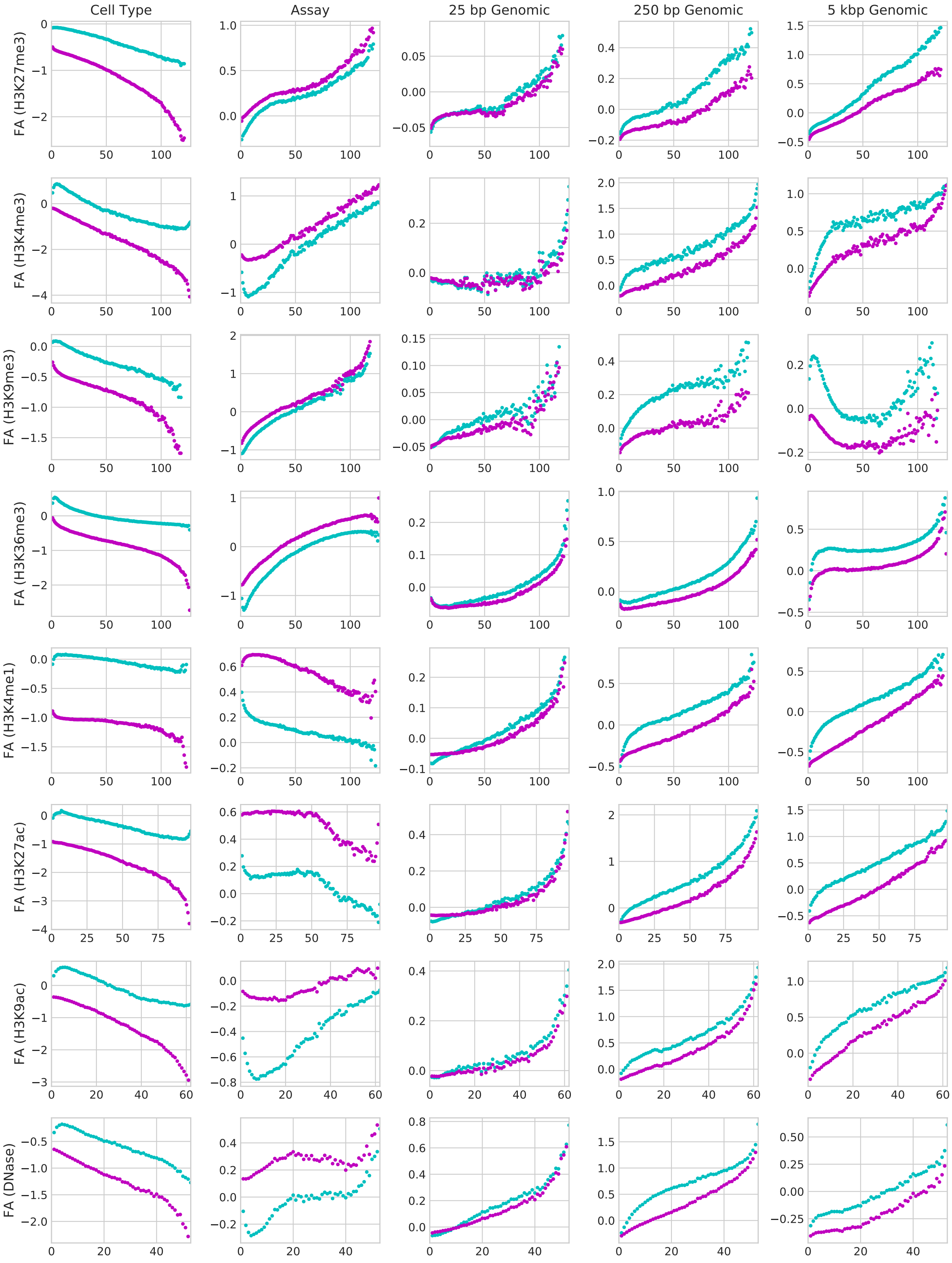

Figure S6: **Feature attribution performed on the Avocado model.** Feature attribution was performed for each position in chromosome 20 across all 1,014 experiments. The results were then aggregated in a manner similar to the analysis of cell-type specific imputations. Instead of calculating the MSE, 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).

Supplementary Note 1: 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 ∈ (16, 32, 64, 128, 256), assay factors ∈ (16, 32, 64, 128, 256), 25 bp resolution genome factors ∈ (5, 10, 15, 20, 25), 250 bp resolution genome factors ∈ (10, 20, 30, 40, 50), 5 kbp resolution genome factors ∈ (15, 30, 45, 60, 75), number of layers in the neural network ∈ (0, 1, 2, 3, 4), and number of neurons in the neural network ∈ (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. In this grid we trained 1,000 models out of a possible ∼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 (Supplementary Figure S7). 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 (Supplementary Figure S8). 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 (Supplementary Figure S9). 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.

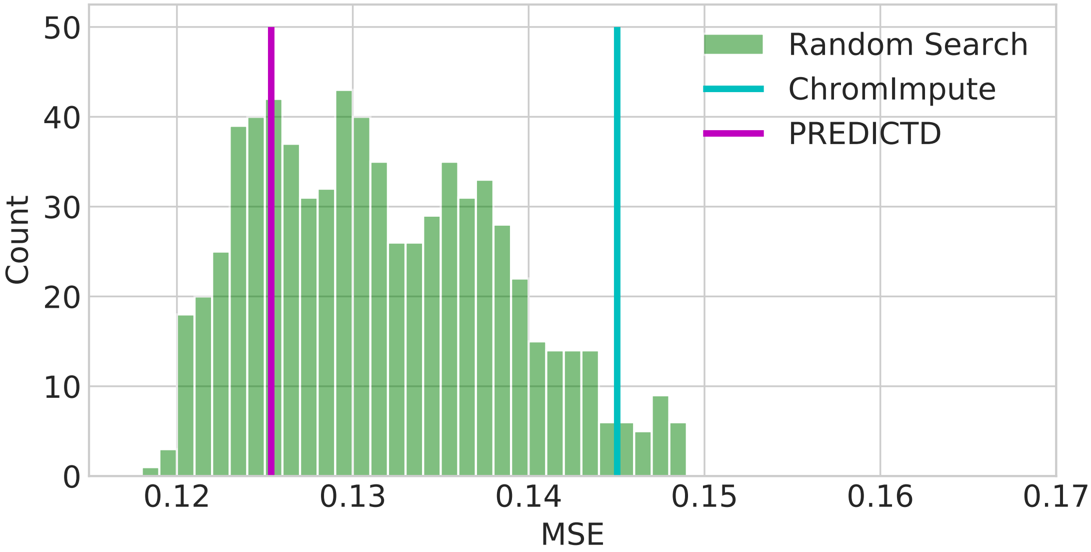

Figure S7: **Random search results on ENCODE pilot regions.** The figure plots a histogram of Avocado validation set MSE values across each hyperparameter setting. For reference, MSE values on the same data set for ChromImpute and PREDICTD are depicted as vertical lines.

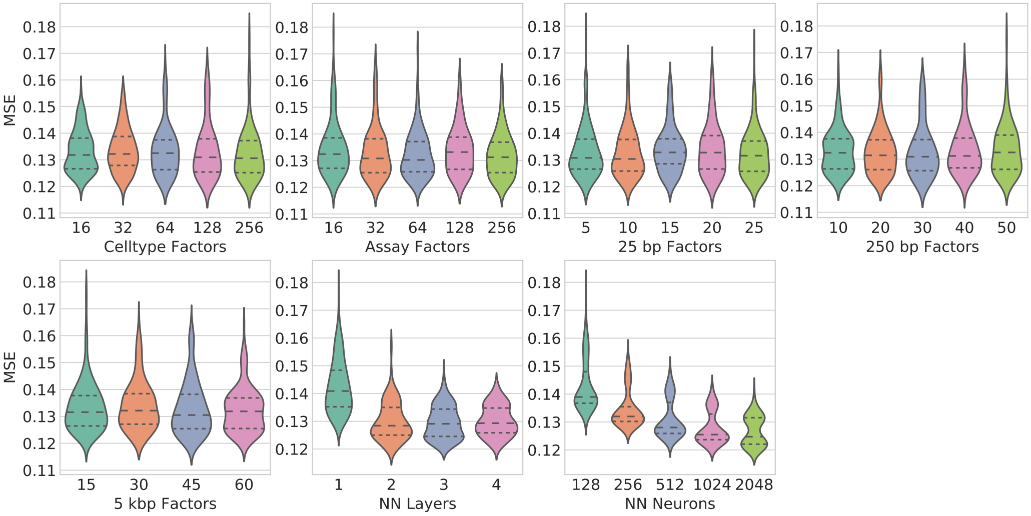

Figure S8: **The performance of the Avocado models learned during random search when stratified by values for each hyperparameter individually.** Each panel shows results for all models that had at least one hidden layer in the neural network. The median is indicated in each violin plot with the longer 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.

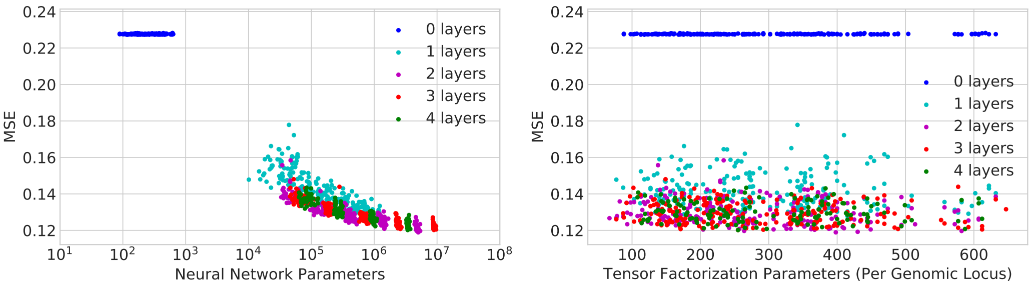

Figure S9: **The number of parameters in each model considered as a part of the random search procedure compared to validation set performance for both the neural network and the tensor factorization aspects.** Left: The trend appears to be that the greater the number of parameters, the better the performance 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.

Supplementary Note 2: 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 [2, 3]. 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 (Supplementary Figure S10a). 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 [4]. 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 (Supplementary Fig S10a/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 (Supplementary Figure S10c), 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 (Supplementary Figure S10d).

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 (Supp. Figure S11). 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 (Supplementary Table S1). We also verified that Avocado does a better job than the other two imputation methods at capturing a lack of correlation between unrelated marks (Supplementary Figure S11), such as the repressive mark H3K27me3 and enhancer-associated mark H3K4me1 (Supplementary Table S1).

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. (Supplementary Figure S11).

	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 S1: 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.

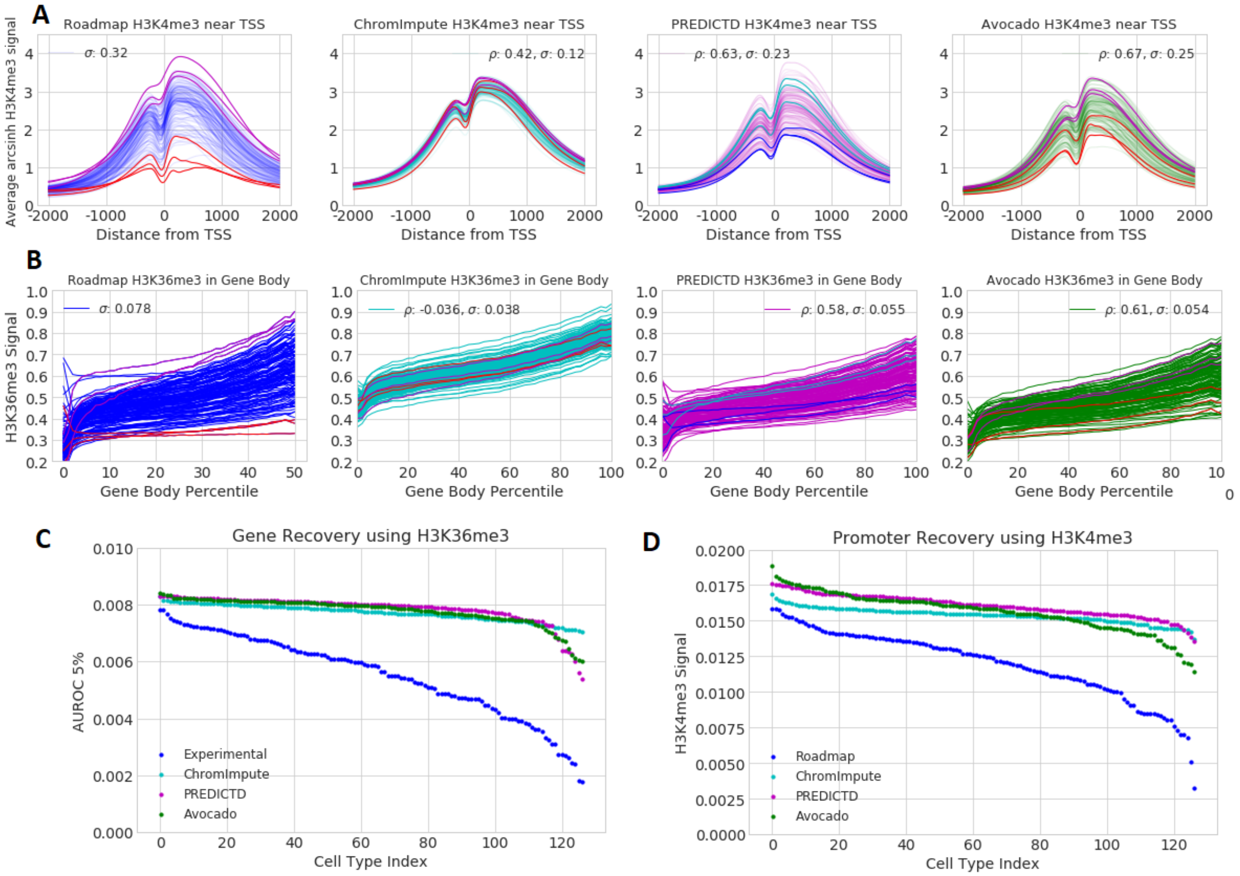

Figure S10: **Aggregate measures of H3K4me3 and H3K36me3 in ChIP-seq experiments and across imputation methods.** (a) H3K4me3 signal in TSSs. Each line displays the average H3K4me3 signal across all TSSs in chromosomes 1-22 for a single cell type after accounting for strand orientation of the gene. The variance of the signal across all cell types at each position is calculated and then averaged (σ). The area under each line is used to define a ranking, and the spearman correlation (ρ) is calculated between each of the three imputation approaches and the ChIP-seq data. To show how the imputation methods alter ordering, the three cell lines with the highest and lowest ChIP-seq signal using this method are colored magenta and red, respectively. In the PREDICTD panels the three with the highest signal are colored cyan and the three with the lowest signal are colored blue. (b) H3K36me3 signal in gene bodies, taking into account the strand orientation of the gene. Measurements are calculated in the same manner as H3K4me3. (c) The GeneRecov performance measure for each cell type. This performance measure quantifies how well H3K36me3 localizes in gene bodies across cell types. It 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. This performance measure quantifies how well H3K4me3 localizes in promoter regions across cell types. It is the area under the ROC curve at 5% FPR when using H3K4me3 to predict promoters across chromosomes 1 through 22.

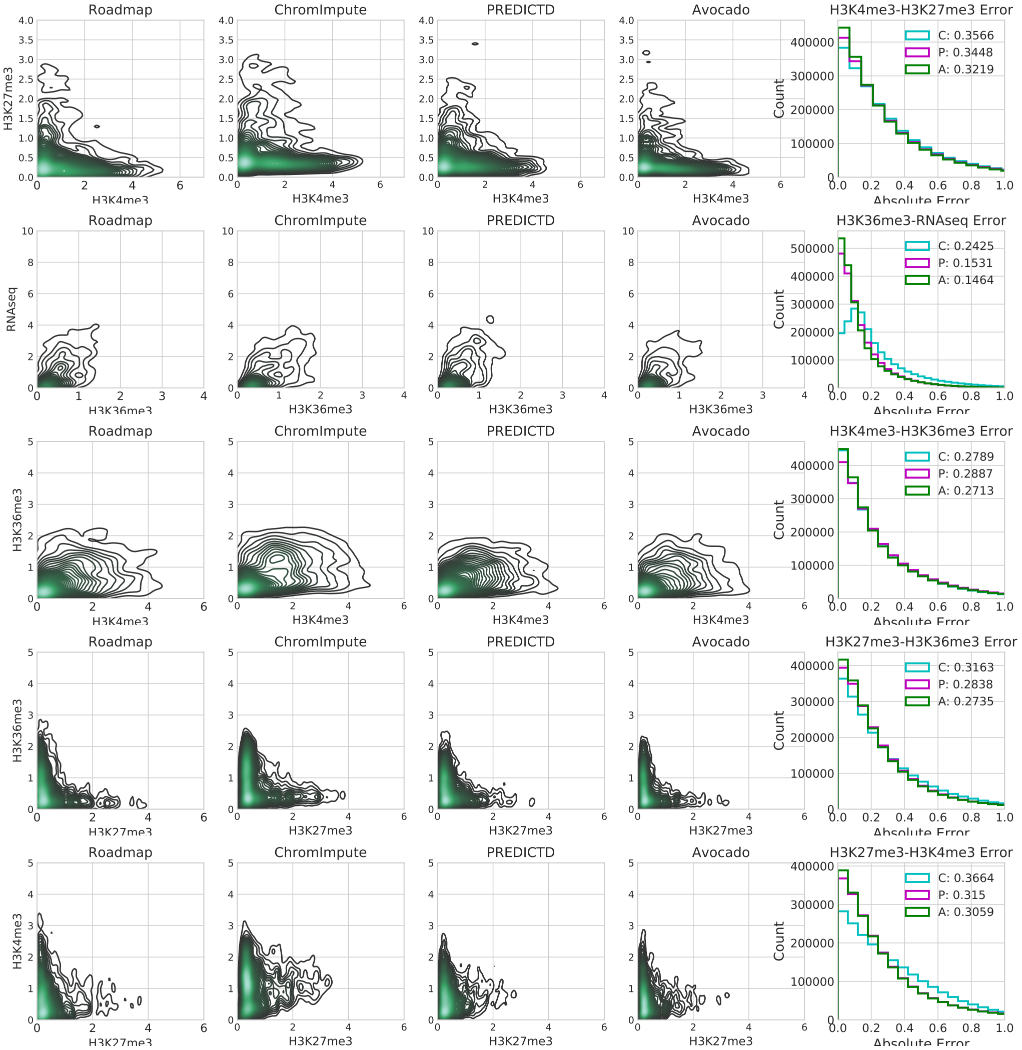

Figure S11: **The relationships between pairs of histone modifications.** These panels show, going from left to right, the signal values in the Roadmap compendium, the imputed signal values from ChromImpute, imputed signal values from PREDICTD, the imputed signal values from Avocado, and the distribution 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 then considered relationships across both histone marks and genomic loci, focusing on the relationship between marks in the promoter and the gene body. Specifically, 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 (Supplementary Table S1). 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 (Supplementary Figure S11). This trend exists to a much lesser extent in the Avocado model, but is not supported by the ChIP-seq signal.

References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

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

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

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