nature machine intelligence

Supplementary information

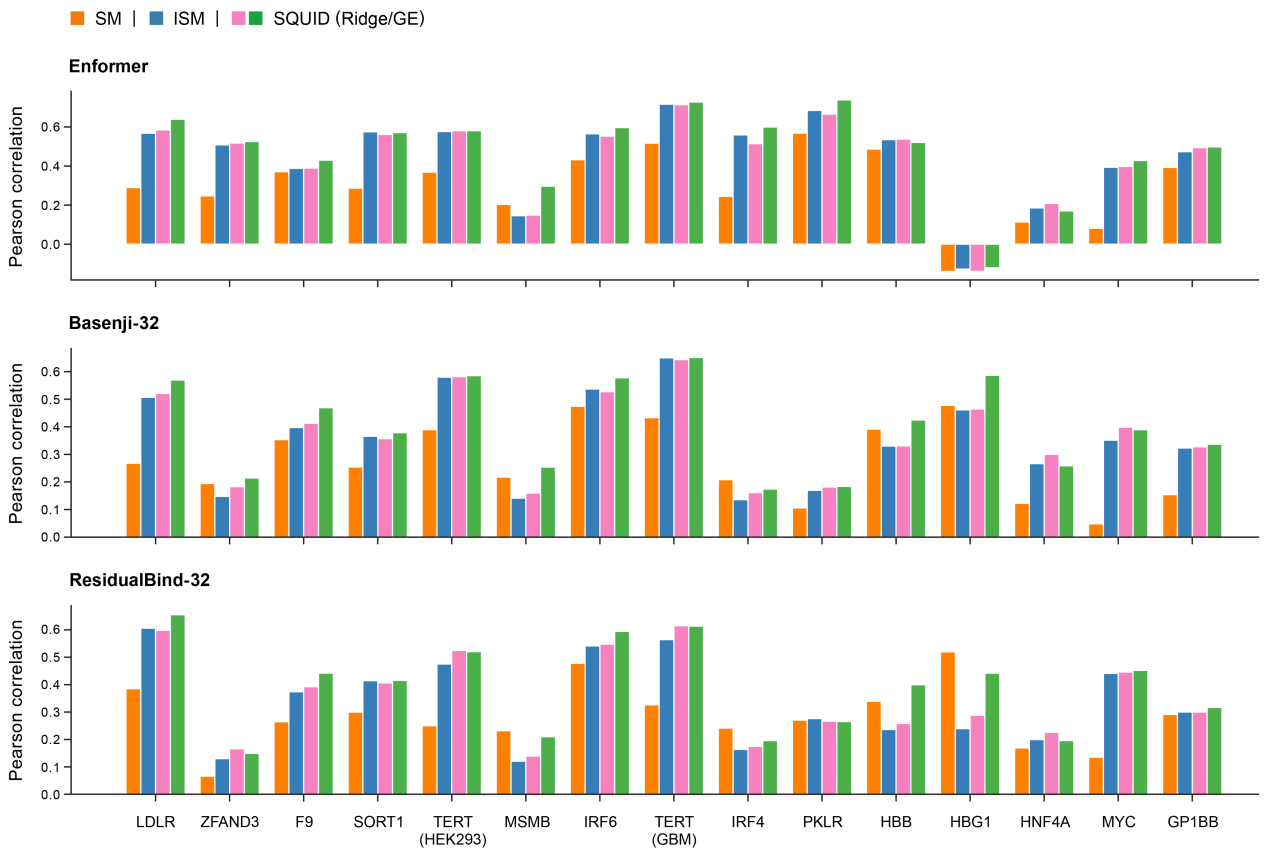
https://doi.org/10.1038/s42256-024-00851-5

Interpreting *cis*-regulatory mechanisms from genomic deep neural networks using surrogate models

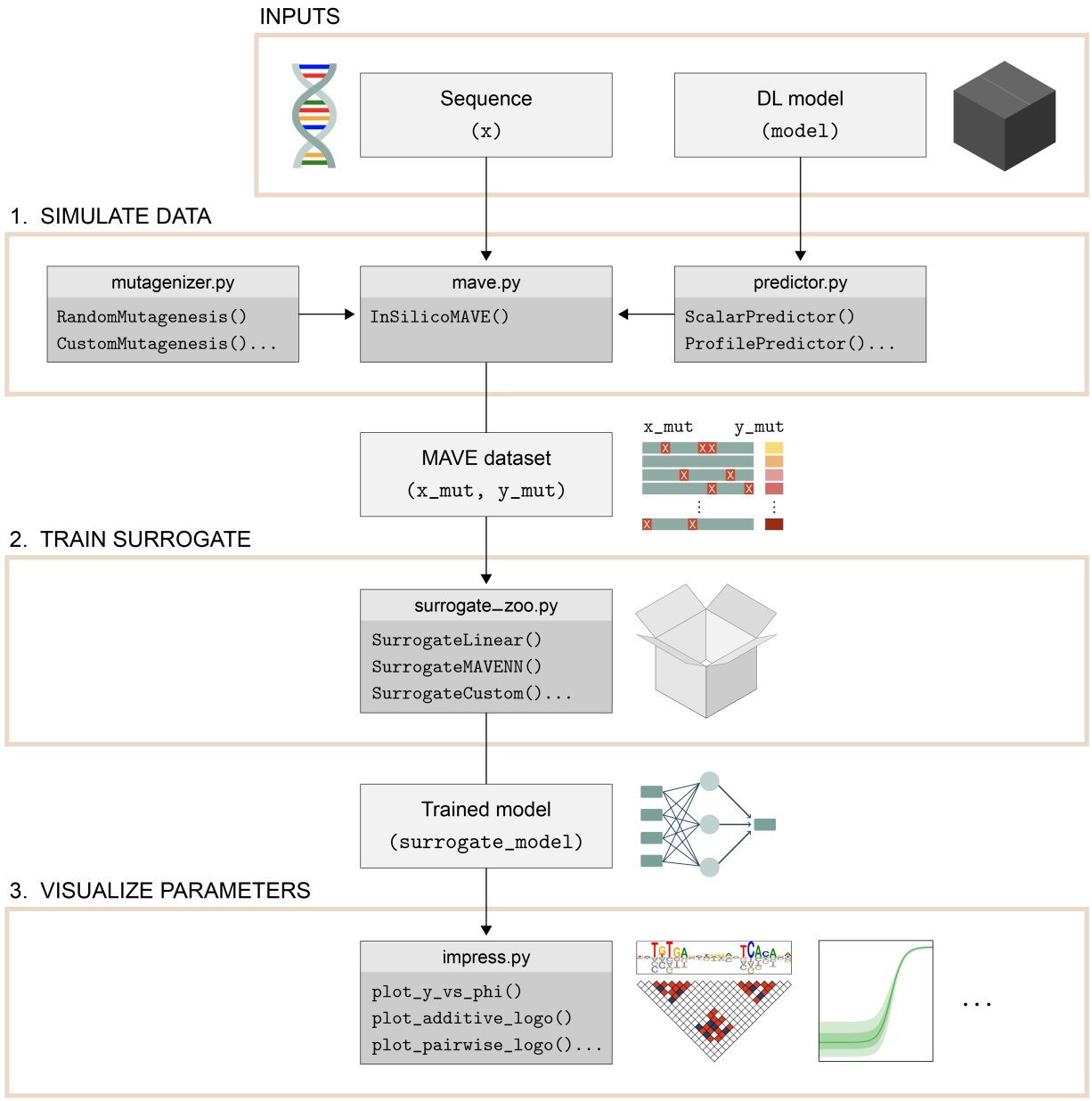
In the format provided by the authors and unedited

TF name	consensus sequence	DNN name	DNN task	PWM ID (database)
AP-1	TGAGTCA	ResBind-32	PC-3	MA0476.1 (Jaspar)
AP-1 / AP-1	TGANTCA···TGANTCA	ResBind-32	PC-3	n/a
IRF1	TGAAAC	ResBind-32	GM12878	n/a
IRF1-long	AANTGAAAC	ResBind-32	GM12878	MA0050.1 (Jaspar)
SPI1	GGAAGT	ResBind-32	GM12878	n/a
FEV	CCGGAA	ResBind-32	HCT116	n/a
AP-1	TGACTCA	DeepSTARR	Dev	MA0476.1 (Jaspar)
Dref	TATCGATA	DeepSTARR	Hk	M00230 (Homer)
Ohler1	AGTGTGACC	DeepSTARR	Hk	M00232 (Homer)
Ohler5	CAGCTG	DeepSTARR	Hk	n/a
Oct4	TTTGCAT	BPNet	Oct4	n/a
Sox2	GAACAATAG	BPNet	Sox2	H12CORE.0.P.B (Hocomoco)
Klf4	GGGTGTGGC	BPNet	Klf4	n/a
Nanog	AGCCATCAA	BPNet	Nanog	H12CORE.1.P.B (Hocomoco)
Nanog / Sox2	AGCCATCAA···GAACAATAG	BPNet	Nanog	n/a

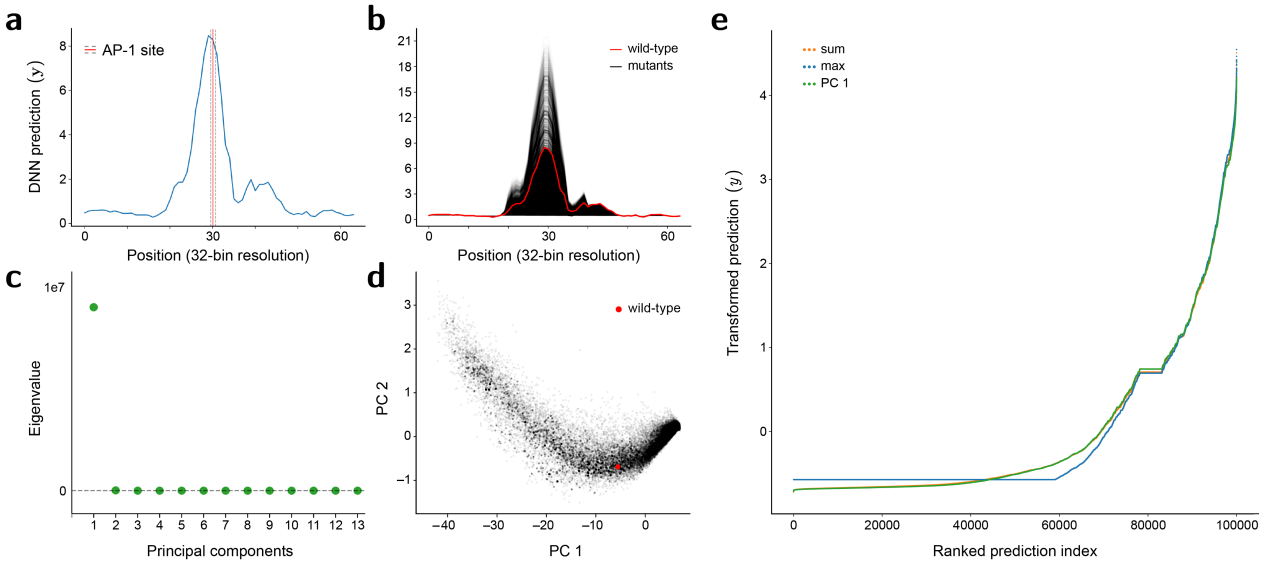
Supplementary Table 1. TFs analyzed in our study. Shown for each TF is the consensus sequence used, the DNN that models the TF, the DNN prediction task to which attribution methods were applied, and PWMs used to investigate weak binding sites. TF, transcription factor; DNN, deep neural network; PWM, position weight matrix.



Supplementary Figure 1. Performance of attribution methods at predicting variant effects at individual loci. Pearson correlation scores for each of the 15 disease-associated loci assayed in CAGI5, computed for the attribution methods and DNNs listed in Table 1. DNN, deep neural network.



Supplementary Figure 2. SQUID workflow. Flowchart representing a typical DNN interpretation analysis pipeline using SQUID. DNN, deep neural network.



Supplementary Figure 3. Dimensionality reduction of DNN predictions using PCA. **a**, Example ATAC-seq profile predicted by ResidualBind-32 for a representative sequence of interest containing a putative AP-1 binding site. The profile shown was cropped to a region spanning the putative site and 30 nt of flanking DNA on either side. **b**, Profiles computed for sequences in the *in silico* MAVE dataset generated by SQUID when analyzing the sequence of interest from panel **a**. **c**, Ranked eigenvalues from a PCA analysis of the profiles in panel **b**. **d**, Projection of profiles onto the first two principal components. **e**, Scalar predictions y for three projection methods: PCA, sum, and max. PCA projections were computed by projecting profiles onto the first principal component. Sum projections were computed by summing the entries in each profile. Max projections were computed by taking the maximum entry in each profile. To aid comparisons between different projection methods, the y values for each method were centered about zero and rescaled to have unit standard deviation. The flat region observed near ranked prediction index 80,000 results from sequences in the *in silico* MAVE library that have no mutations. PCA, principal component analysis.