



Background

Sequence-based deep neural networks (DNNs) for functional genomics predict molecular phenotypes associated with transcriptional regulation from primary sequences. Although DNNs learn predictive functions, it remains difficult to uncover their learned features, such as binding motifs and regulatory grammars.



Attribution methods provide a feature importance score for each nucleotide in the given sequence, with a direct interpretation as single-nucleotide variant effects on model predictions.



Attribution methods operate over a local neighborhood of the DNN-learned function to provide *post hoc* interpretation of DNN predictions for a specific sequence. One **limitation** is that different attribution methods provide different interpretations based on how they characterize the local function.



Schematic of DNN-learned function over local region of sequence space, depicting the **neighborhood sizes** considered by different attribution-based approaches.

CAGI5 challenge measures the effects of single nucleotide variants (SNVs) in massively parallel reporter assays (MPRA) For each of the 15 300–600-bp enhancer/promoter sequences:

Kircher *et al.*, *Nat Commun* (2019)

WT	Reference	ATTCGCTATGCAACT
Ľ	ATTCGCTATGCAACT $\rightarrow 0.9$ ATTCGCCATGCAACT $\rightarrow 0.1$ $\log_2 \Delta$	A C G T
мит		Activity map

The varient effect of each SNV is measured to form an **activity map**, similar to the attribution maps produced via ISM.

5

SQUID: Surrogate Quantitative Interpretability for Deepnets Domain-specific surrogate models SQUID approximates user-defined regions of sequence space with flexible surrogate models that have mechanistically-interpretable parameters. Key features of MAVE-NN: 1. Genotype-phenotype maps (additive, pairwise, higher-order) 2. Global epistasis (GE) nonlinearity in silico Genomic deep Model outputs Surrogate model 3. Complex noise (heteroscedastic) (\mathbf{x}) sequence MAVE data neural network GE nonlinearity & noise model TCASA 2 $\phi = f(\mathbf{x}, \mathbf{\theta})$ G-P map $oldsymbol{ heta}$ additive (2)(3)visualize **—** $\hat{\mathbf{y}}$ simulate train $\boldsymbol{\theta}$ pairwise latent - 68% PI surrogate parameters phenotype 95% PI Generate *in silico* multiplex assays of variant effect (MAVE) data by mutagenizing a sequence pip install squid-nn and using the DNN as a scoring function (i.e., functional readout) 2. Fit the data with mechanistically-interpretable surrogate model designed to fit the MAVE data https://squid-nn.readthedocs.io 3. Visualize parameters for interpretation of *cis*-regulatory mechanisms DNN prediction Tareen et al., Genome Biol (2022) SQUID yields more consistent motif representations across genomic loci Visualizing motif consistency 3 High-error binding motif Avg. of 50 binding motifs **Experiment 1**: Compare the consistency of binding motifs identified by different attribution methods. Low-error binding motif <u>acaggaagtgagtcagcatacyag</u> consensus site GEASCIER STRAT ← AGATGAGTCACGTT → ← GTGTGAGTCACACG → ___(2)___→ 1 find $-3 \rightarrow$ → () ATGACTCAT ALGACTCAT ← ATGTGAGTCAAGGC → compute compute ← GATTGAGTCATATA → attribution attribution – error · genome ← TTATGAGTCAGTCA → errors (θ_1) maps TGAGTC ← CAGTGAGTCAGGAT → sequence $\frac{1}{m}\sum \theta_i$ Method context flank flank flank flank core flank flank core core ***DeepSTARR ResidualBind BPNet — GE nonlinearity (y)95% PI *** 200 -ISM ••• test data tio Saliency 150 -wildtype 4 DeepSHAP DeepLIFT <u>ධි</u> 100 DNN ■ SQUID (GE) SPI1 IRF1 Ohler1 AP-1 AP-1 Oct4 Sox2 Nanog Dref SQUID Toneyan et al., Nat Mach Intell (2022) SQUID additive effect (ϕ) de Almedia et al., Nat Genetics (2022) Avsec et al., Nat Genetics (2021) (Ridge) SQUID extensions provide different insights for interpreting DNNs Global DNN interpretations **Experiment 2**: Implement pairwise-interaction surrogate models to quantify the effects of pairs of binding sites. N N N N N N N N N N N C A A T C A A N N N N N N N N N N N N N N N G A A C <mark>N N N</mark> A G N N N N N N N N N N M — → random DNA TGA STRATC DNA Sox2 ___9 nt -Nanog Sox2 chr8: 47652876intra-site intra-site interactions (A) interactions (B) 47652920 PW max ▶ ADD max ▶ ← <u>AP-1</u> <u>AP-1</u> → inter-site interactions (AB) PW min ▶ Inter-site distance (nt) Averaged additive/pairwise effects Conclusions SQUID improves zero-shot variant effect predictions 6 **Experiment 3:** Closer to ground truth, we next evaluated the performance of attribution-based methods to recapitulate experimental SQUID leverages domain knowledge on how to characterize a SNV measurements made in the **CAGI5 challenge** dataset. regulatory genomic locus to flexibly interpret genomic DNNs SQUID identifies motifs more consistently across genomic loci, and yields improved variant effect predictions compared to Statistical significance existing attribution methods GE vs. ISM | GE vs. Ridge aliency SQUID provides different mechanistic Avsec e insights by swapping surrogate models Metho (e.g., additive, pairwise) ** *** Toneyan *e* • In silico MAVEs can be designed in Mach Inte different ways to address different

A surrogate modeling framework for interpreting deep neural networks in functional genomics Evan E Seitz¹, David M McCandlish¹, Justin B Kinney^{1*}, Peter K Koo^{1*}

¹Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Rd,

Cold Spring Harbor, 11724, NY, USA.





*



	DNN		Average Pearson correlation				
	Architecture	Task	Saliency	ISM	SQUID (Ridge)	SQUID (GE)	GE vs. Sa
et al., Nat ds (2021)	Enformer	DNASE	0.2977	0.4498	0.4486	0.4801	***
et al., Nat	Basenji-32	ATAC-seq	0.2727	0.3575	0.3700	0.4036	***
ell (2022)	ResBind-32	ATAC-seq	0.2846	0.3388	0.3567	0.3912	**











biological questions (e.g., local, global)

