General Artificial Intelligence (1)

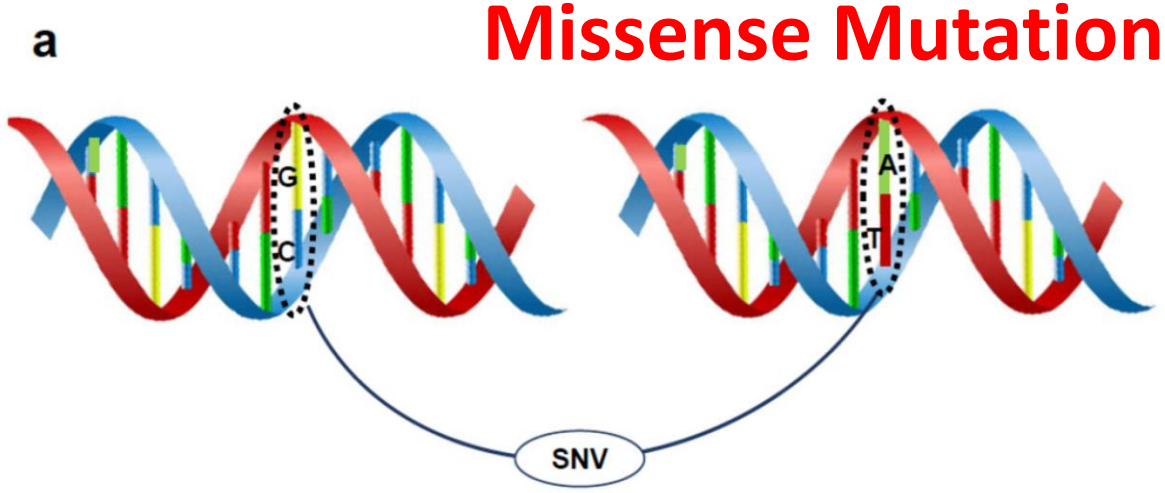
SAIR 2-05 What is next after AlphaMissense? Identifying pathogenic non-coding variants and incorporating association and causal analysis into genomic variation foundation models

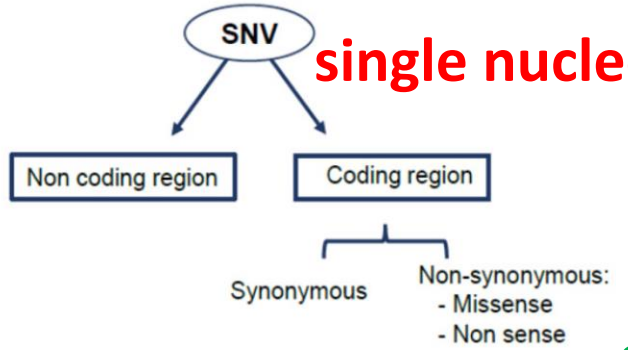
Momiao Xiong, Schicheng Guo
Society of Artificial Intelligence Research

Outlines

- Identify Pathogenic Missense Mutations and Estimate Their Effects
- Identify Pathogenic Non-coding Variants and Estimate Their Effects
- Incorporate Association and Causal Analysis into Fundation Models

Missense Mutation

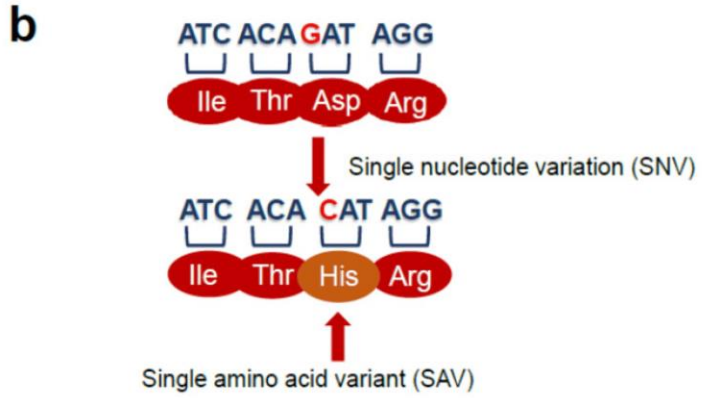


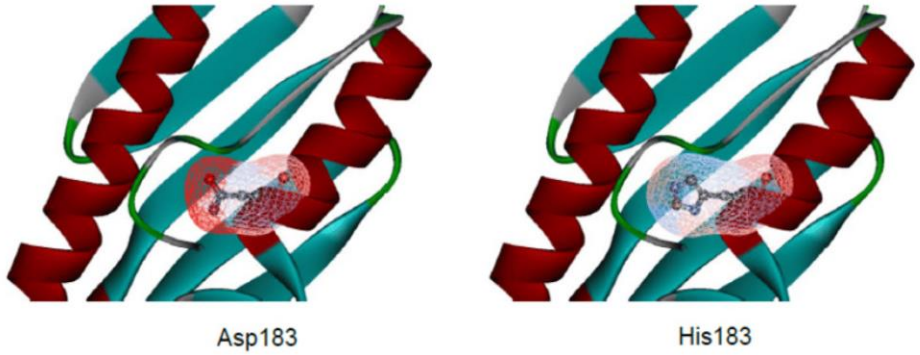


single nucleotide variations (SNVs)

either in coding or
in non-coding regions

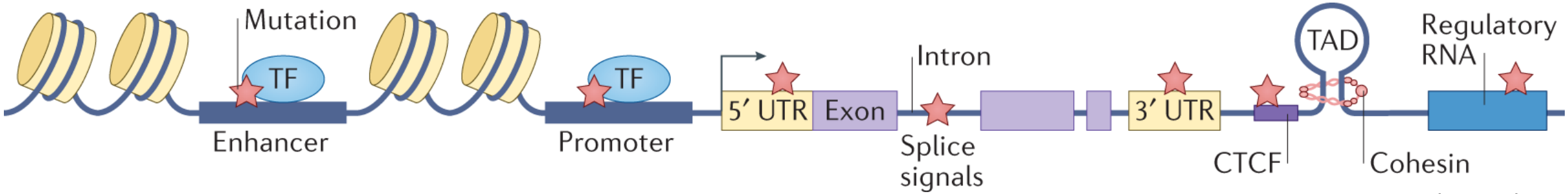
copy number, insertions, deletions, duplications, and rearrangements



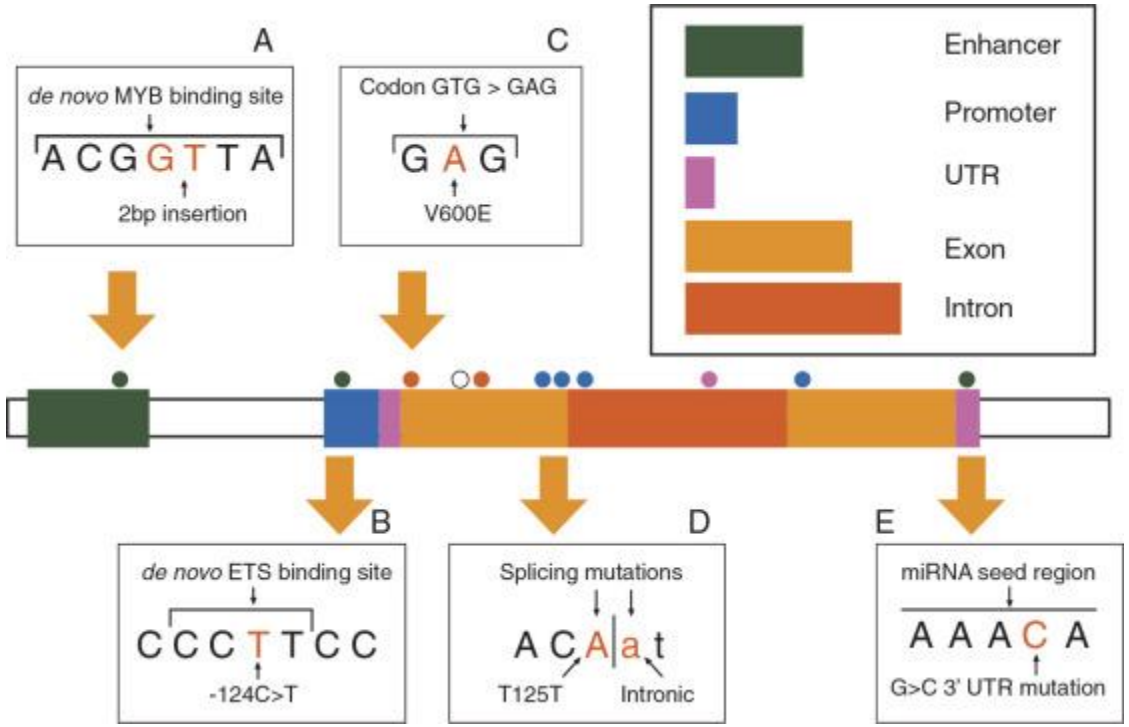


(**a**) Single nucleotide variations (SNVs) can occur in the coding or in the non-coding region. SNVs in the coding region can be synonymous if no amino acid changes are produced, non-synonymous if the single nucleotide substitution induces changes in the protein sequence. Usually, two types of non-synonymous changes can be described: missense mutation, that produces an amino acid change in the protein (SAV) and nonsense mutation which produces a truncated or a longer protein. (**b**) A single nucleotide substitution can lead to a single amino acid change generating a protein variant with structural and/or functional alterations as shown in the substitution of the residue Asp183 with a His in the human frataxin protein.

Non-coding Mutations



Non-coding driver mutations in human cancer. Nature Reviews Cancer volume 21, pages500–509 (2021)



Beyond the exome: the role of non-coding somatic mutations in cancer.

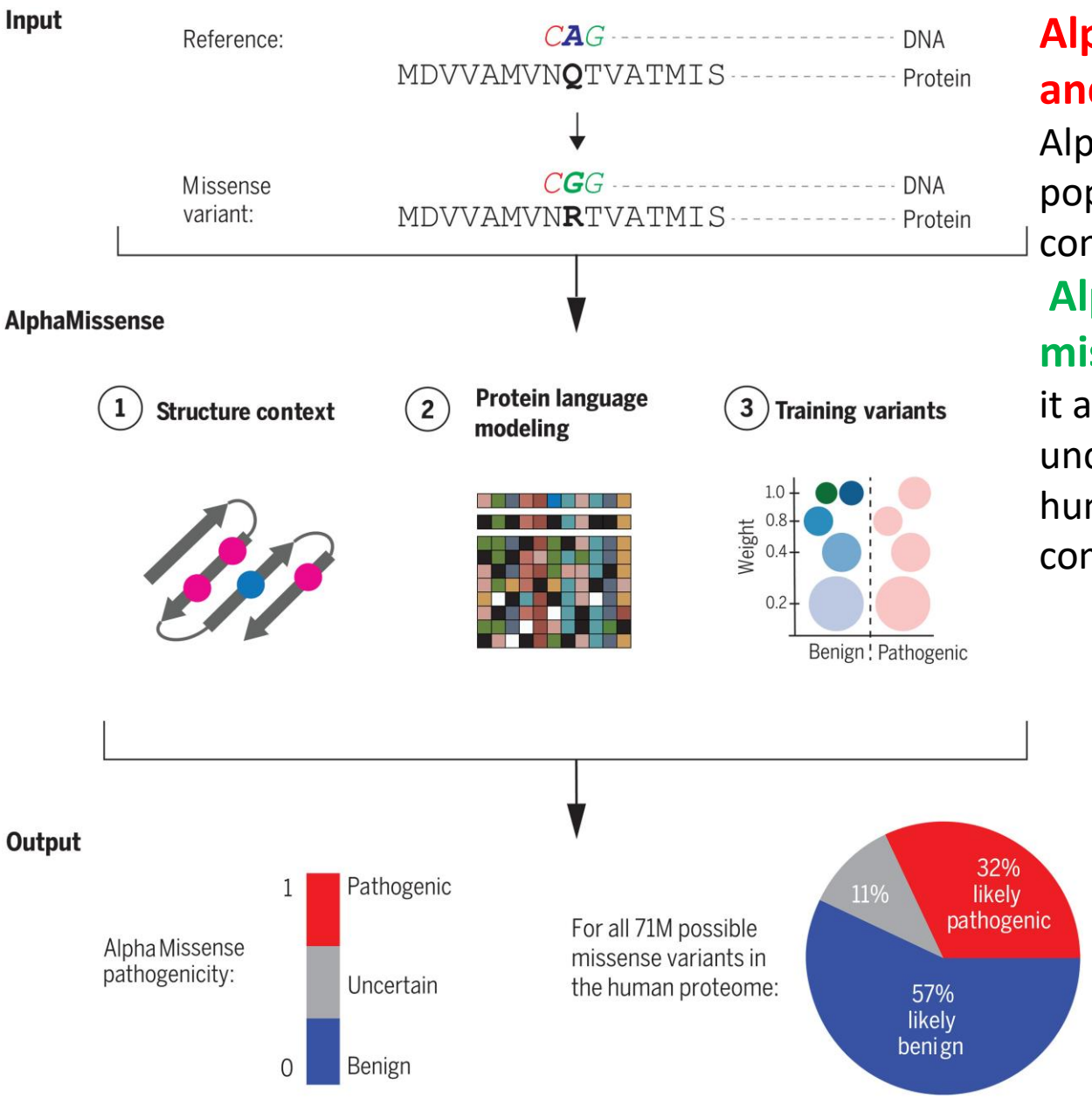
Annals of Oncology .Volume 27, Pages 240-248

AlphaMissense

Accurate proteome-wide missense variant effect prediction with AlphaMissense

JUN CHENG et al.

SCIENCE, 19 Sep 2023, Vol 381, Issue 6664, DOI: 10.1126/science.adg7492

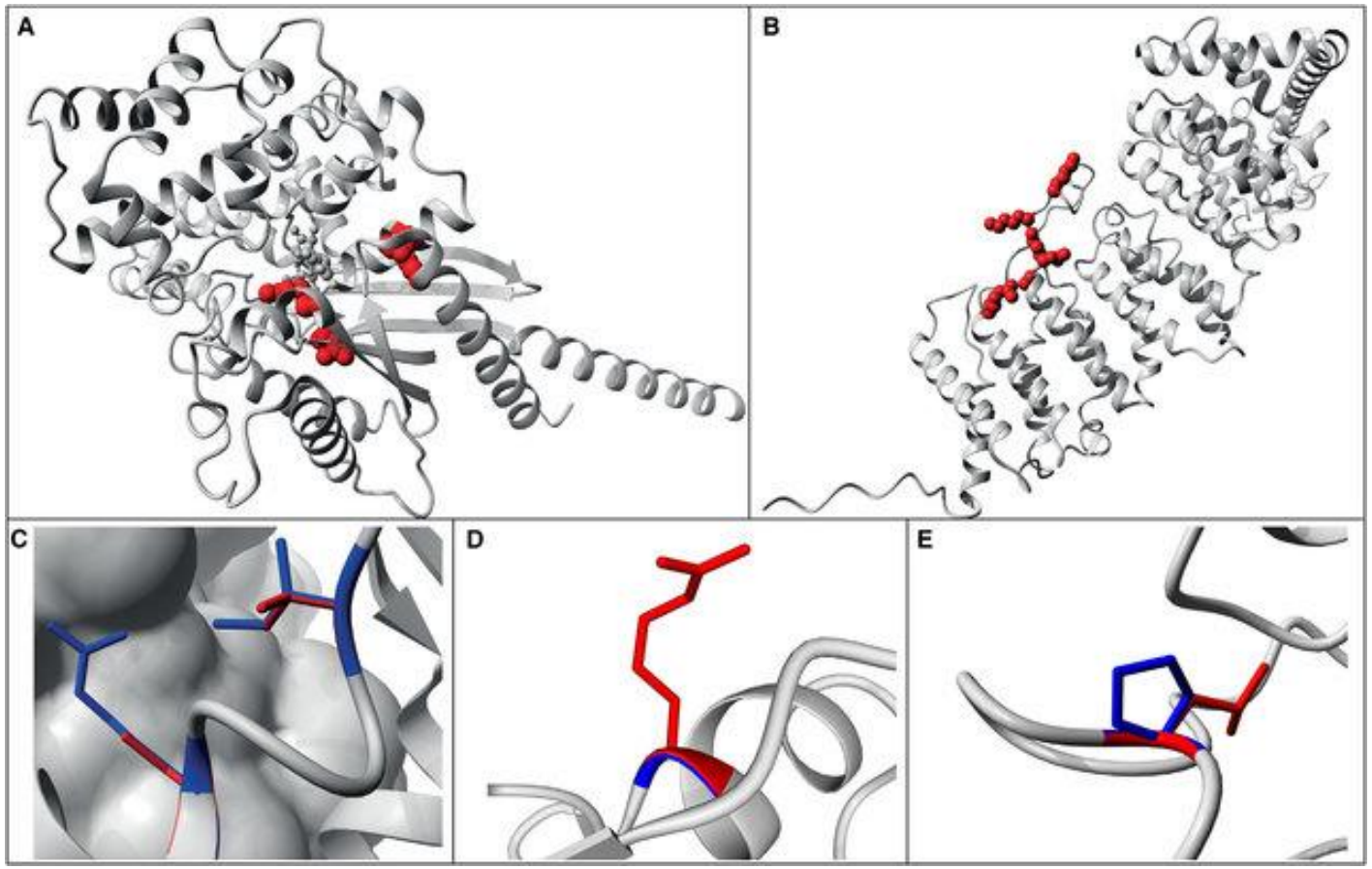


AlphaMissense takes as input a missense variant and predicts its pathogenicity. We fine-tuned AlphaFold on human and primate variant population frequency data and calibrated the confidence on known disease variants.

AlphaMissense predicts the probability of a missense variant being pathogenic and classifies it as either likely benign, likely pathogenic, or uncertain. We provide predictions for all possible human missense variants as a resource for the community

89% of all 71 million possible missense variants as either likely pathogenic or likely benign.

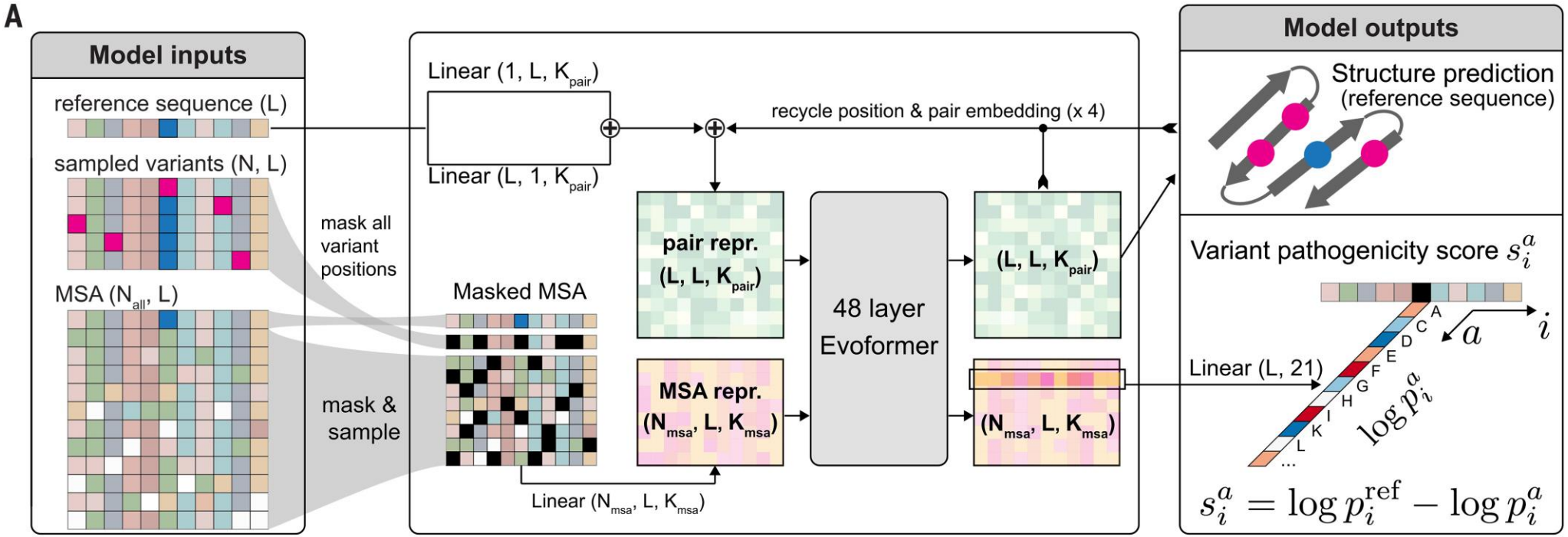
By contrast, only **0.1% have been confirmed by human experts.**

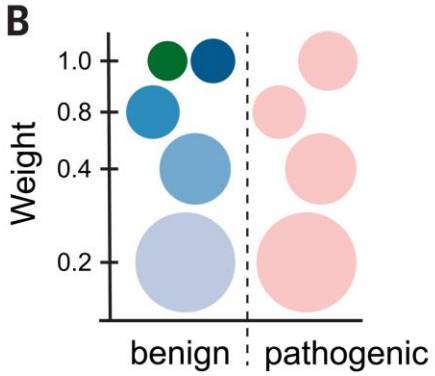


Examples of Modeling of Missense Mutations on 3D Protein Structures **Wild-type** residues are marked in **blue**; de novo **mutations** are indicated as **red** globes or lines

AlphaMissense Procedures

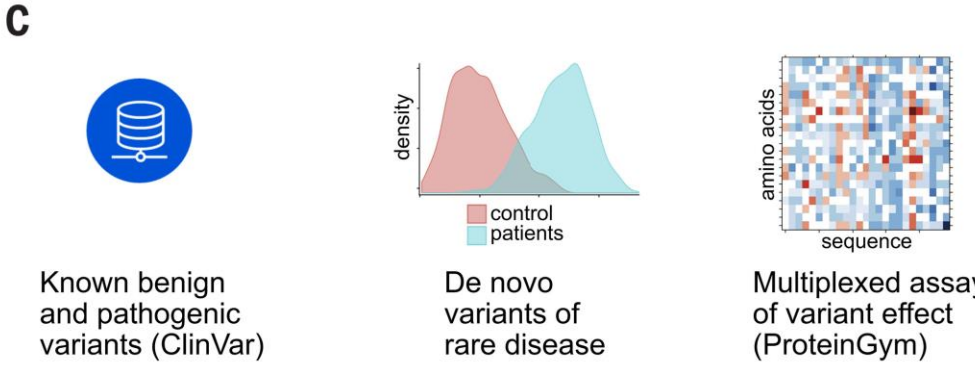
- **Goal: AlphaMissense takes as input an amino acid sequence and predicts the pathogenicity of all possible single amino acid changes at a given position in the sequence. AlphaMissense is trained in two stages.**
- **In the first stage, the network is trained like AF to perform single-chain structure prediction (AF pretraining) along with protein language modeling by predicting the identity of the amino acids masked at random positions in the MSA.**
  After pretraining, the masked language modeling head can already be used for variant effect prediction by computing the loglikelihood ratio between the reference and alternative amino acid probabilities.
- **In the second stage (Fig. 1A), the model is fine-tuned on human proteins with an additional variant pathogenicity classification objective defined for a variant sequence presented in the second row of the MSA (Fig. 1A). For the training set, we assign benign labels to variants frequently observed in the human and primate populations, and pathogenic labels to variants absent from human and primate populations, as is done in Primate AI (12) (Fig. 1B; see methods). We stop training the model once it starts to overfit on the validation set (2526 ClinVar variants with an equal number of pathogenic and benign variants per gene; see methods).**





Primates:
- MAF > 1e-3, total 9.7e4

Humans:
- MAF > 2e-4, total 1.2e5
- 1e-5 < MAF < 2e-4, total 1.1e6
- sample from unobserved, match counts of benign

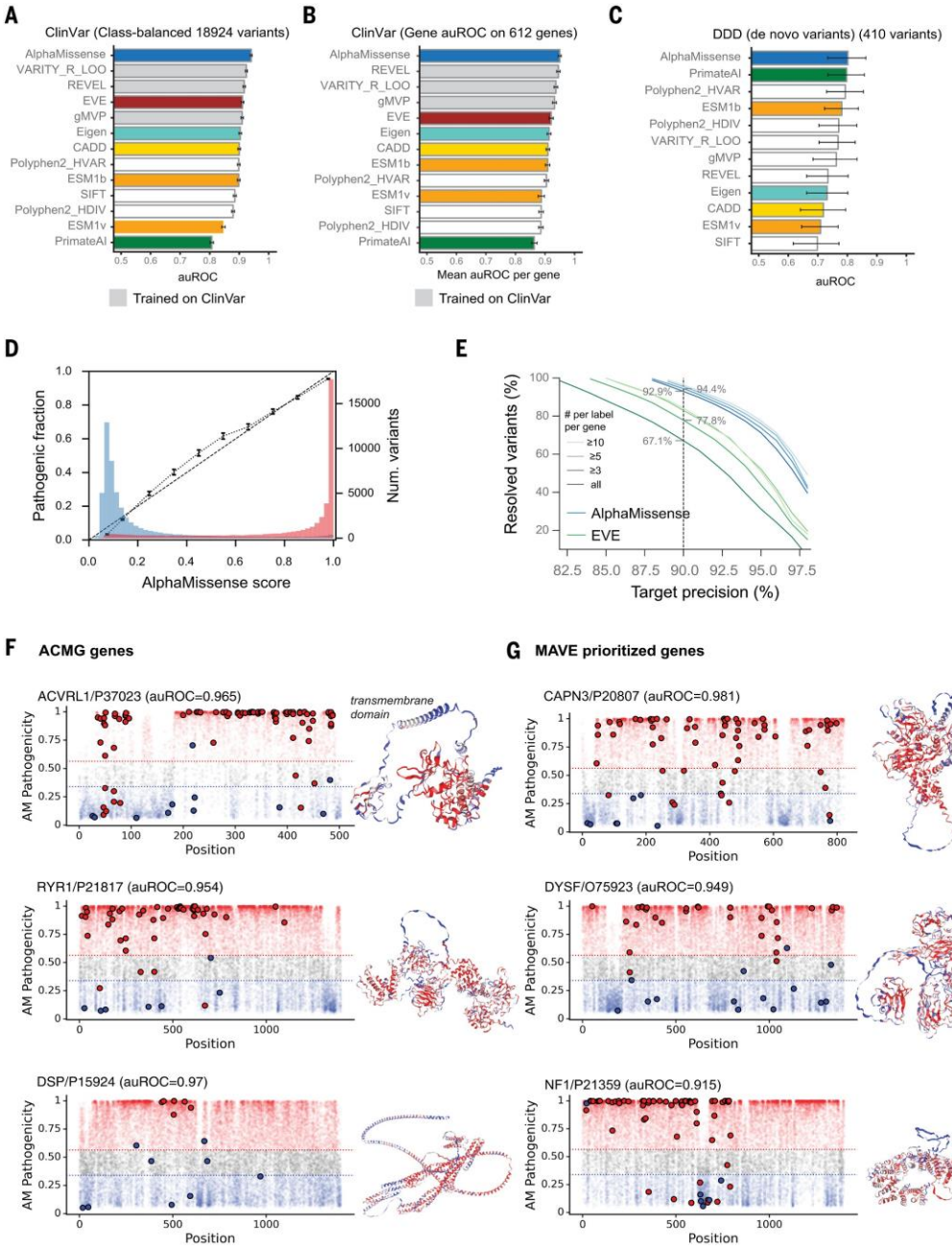


Known benign and pathogenic variants (ClinVar)

De novo variants of rare disease

Multiplexed assay of variant effect (ProteinGym)

Overview of AlphaMissense. (A) AlphaMissense architecture. The model inputs consist of the reference protein sequence [cropped to length (L) = 256 residues], a set of variants sampled from the training set for the same sequence (up to N = 50 variants), and multiple sequence alignments (MSAs, up to Nall = 2048). Inference is performed for one variant at a time (N = 1). The reference sequence is repeated in the second row of the MSA with all sampled variant positions masked (see methods). As in AlphaFold, the model constructs the pair representation (i.e., encodes information about two-way interactions between residues) from the reference sequence (embedding size Kpair), and the MSA representation from the masked MSA (embedding size Kmsa). The MSA and pair representations are processed by a stack of Evoformer layers with recycling. Finally, the model predicts the structure of the reference sequence and the pathogenicity score S_i^a for the variant, which is derived from the masked residue prediction head as the log-likelihood difference between residue a relative to the reference residue at position i (see methods). (B) The pathogenicity score is fine-tuned as a binary classification of variants as benign (observed or frequent missense variants in human or primate populations) or pathogenic (unobserved human missense variants). We split the benign variants into clusters by their minor allele frequency (MAF) and introduce weights in the loss function that reduce the contribution of rare variants. For each observed variant in the benign set, we sample a missense variant from the pathogenic set and assign it the same loss weight as for the benign variant (see methods). (C) We evaluated AlphaMissense on a diverse set of benchmark datasets, including annotated missense variants in ClinVar (30), de novo disease variants (54), and MAVE data collected in ProteinGym (19).



Performance of AlphaMissense on clinically curated classification benchmarks.

Benchmarks are evaluated by area under the receiver operator curve (auROC). Error bars show the 95% confidence interval of 1000 bootstrap resamples (see methods). A few manually chosen methods are colored to illustrate the relative position on different benchmarks

(A) Performance on classification of ClinVar variants (9462 pathogenic and 9462 benign variants from 999 proteins)

(D)The histograms show the distribution of scores among **pathogenic (red)** and **benign (blue) variants**

(E) Precision is defined as the fraction of true predictions in both pathogenic and benign class prediction. The resolved fractions are computed with ClinVar test set variants from proteins scored by EVE (dark lines, all)

(F) Variants predicted as likely **pathogenic** are shown in **red**, variant predicted as likely **benign** are shown in **blue**, and ambiguous variants are shown in gray

Ablating components of AlphaMissense reveals key drivers of performance

- **three types of components:**

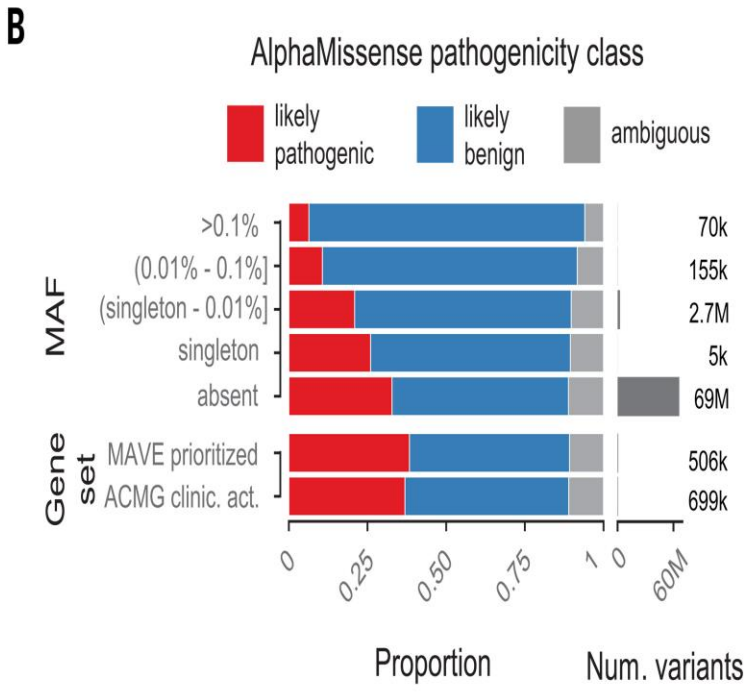
  structure prediction, variant sampling, and training data.

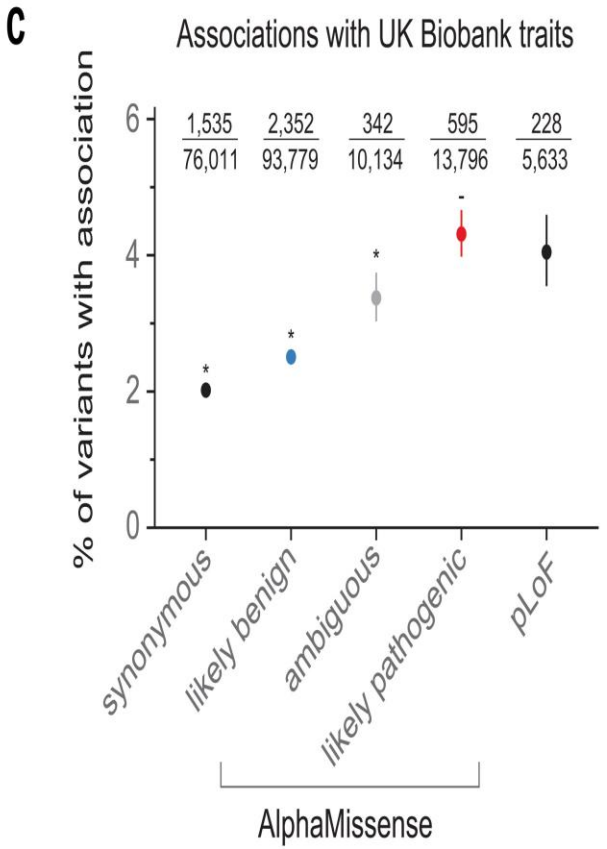
- **Overall, these results emphasize the importance of both training stages: pretraining on a large database of structures and fine-tuning directly for the target application**

AlphaMissense predictions as a community resource

A

CHROM	POS	REF	ALT	genome	uniprot_id	transcript_id	protein_variant	am_pathogenicity	am_class
chr7	44149825	T	C	hg38	P35557	ENST00000403799	D205G	0.999516	likely_pathogenic
...	...	...	...	...	...	...	...	...	...





- 71 million missense variant predictions.
- The second resource is gene-level AlphaMissense pathogenicity predictions, defined as the average pathogenicity over all possible missense variants in a gene
- The third is the expanded dataset of all 216 million possible single amino acid substitutions across the 19,233 canonical human proteins.
- Finally, we provide predictions for all possible missense variants and amino acid substitutions across 60,000 alternative transcript isoforms for future research and evaluation of isoform-specific effects.

An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics

Albi Celaj et al. September 13, 2023. bioRxiv preprint doi:
https://doi.org/10.1101/2023.09.20.558508
Deep Genomics, Brendan J. Frey

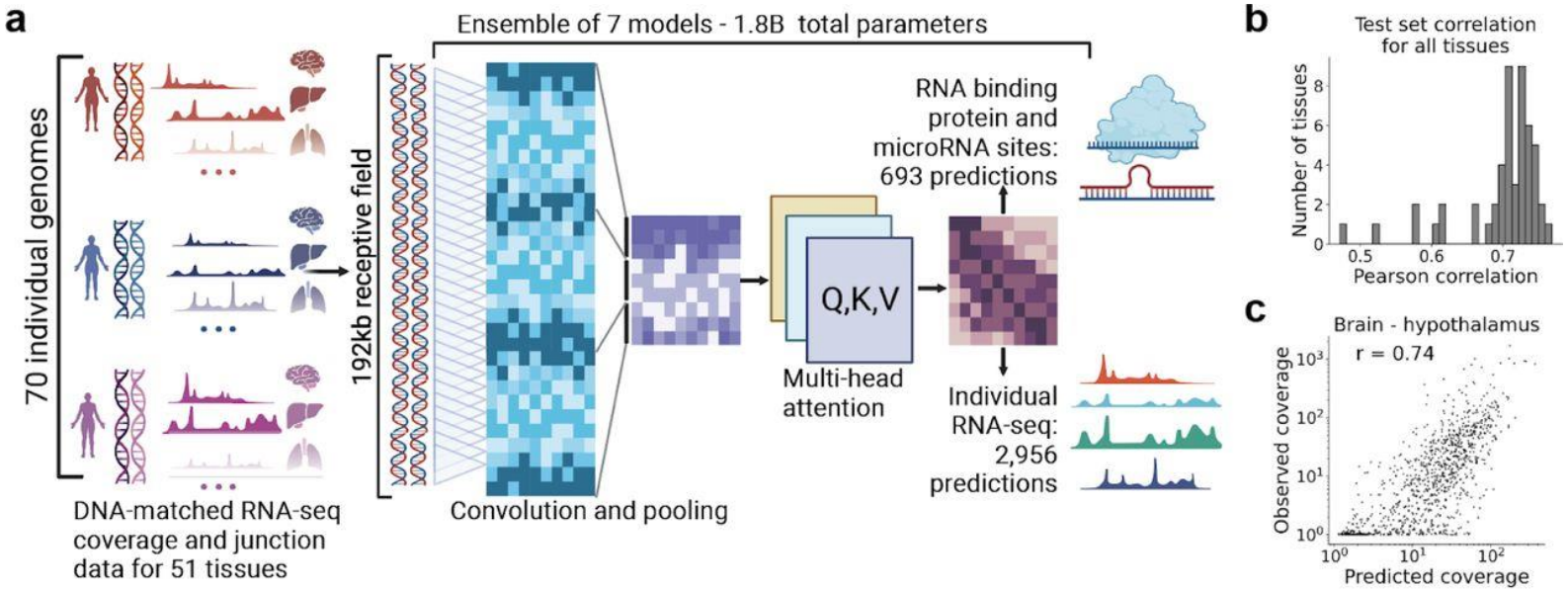
从英文翻译而来-Brendan John Frey FRSC是一位出生于加拿大的企业家，工程师和科学家。他是Deep Genomics的创始人兼首席执行官，Vector人工智能研究所的联合创始人以及多伦多大学工程与医学教授。维基百科（英文)
查看原文说明
学术顾问：杰弗里·辛顿
图书：Graphical Models for Machine Learning and Digital Communication
教育背景：多伦多大学

a foundation model for RNA biology, “BigRNA”

- **trained on thousands of genome-matched datasets to predict**
  tissue-specific RNA expression,
  splicing, microRNA sites, and
  RNA binding protein specificity
  from DNA sequence.
- **BigRNA can identify pathogenic non-coding variant effects across diverse mechanisms**
- **BigRNA accurately predicted the effects of steric blocking oligonucleotides (SBOs) on increasing the expression**

- **Building foundation models**
  that can predict gene expression from DNA sequence

- BigRNA learns from paired genotype and 128bp resolution RNA expression
- downstream tasks
  predicting RNA-binding protein (RBP) specificity and
  microRNA binding sites.
- BigRNA directly models RNA-seq data, not overall expression level



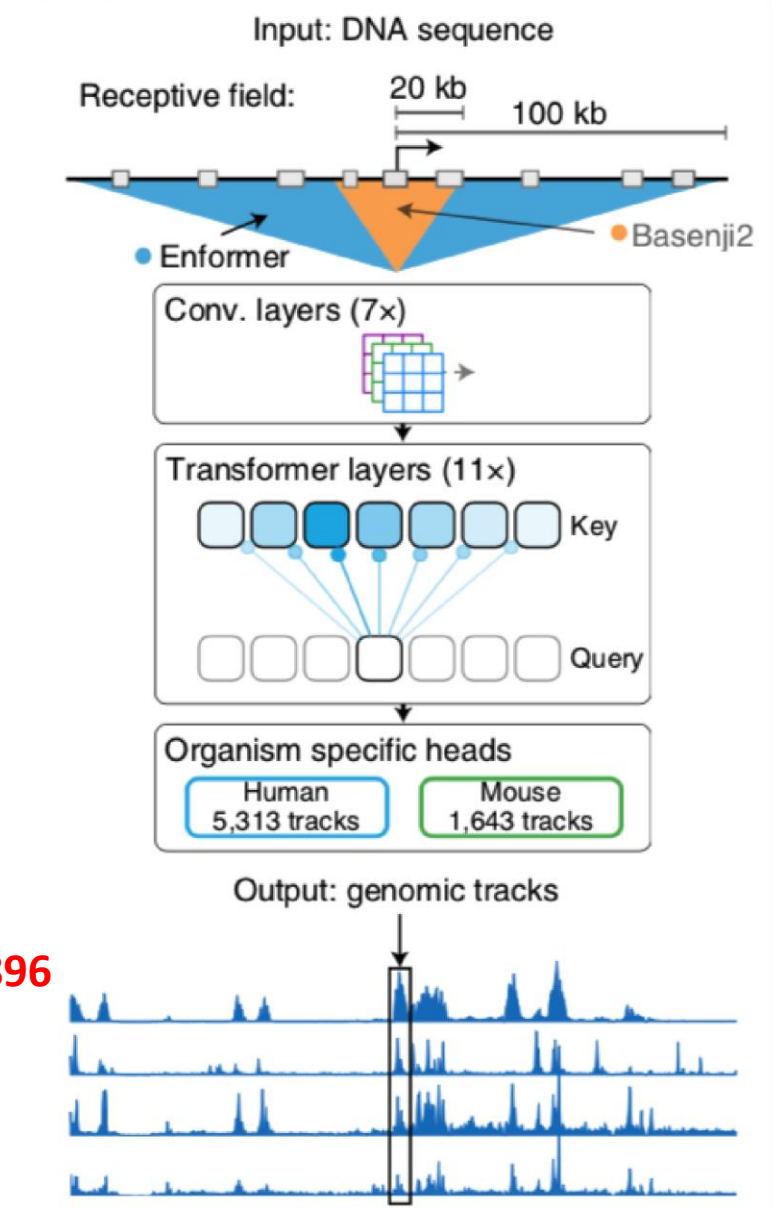
RNA-seq model training

downloaded and aligned RNA-seq data from the GTEx consortium 5 V6 release, processing all available data from the set of 70 individuals

Each RNA-seq sample was processe into two data tracks: Coverage and junction, where the junction track contains a subset of read counts at splice junctions.

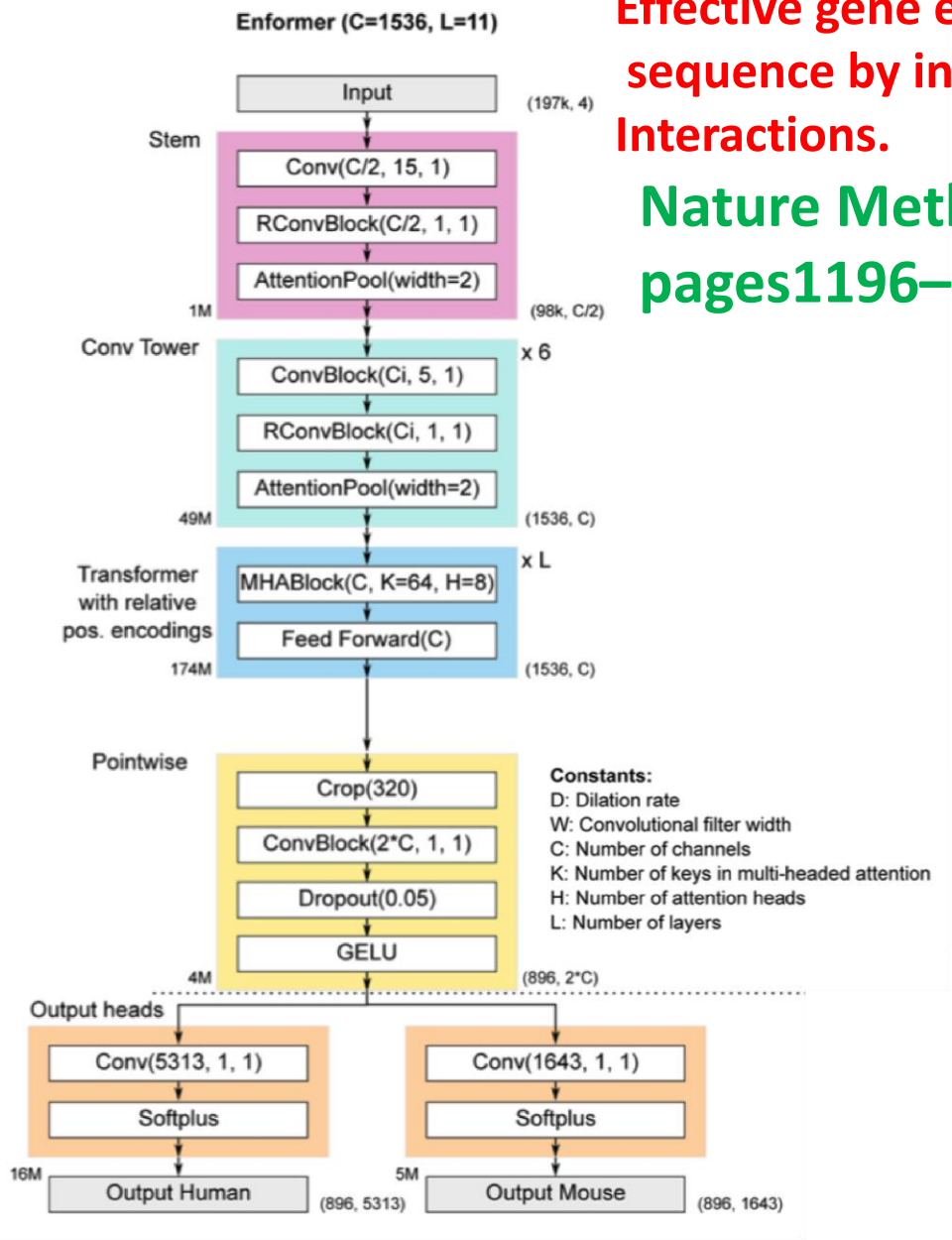
we applied 128bp-window average-pooling on coverage tracks And 128bp-window sum-pooling on junction tracks

196608bp DNA sequence



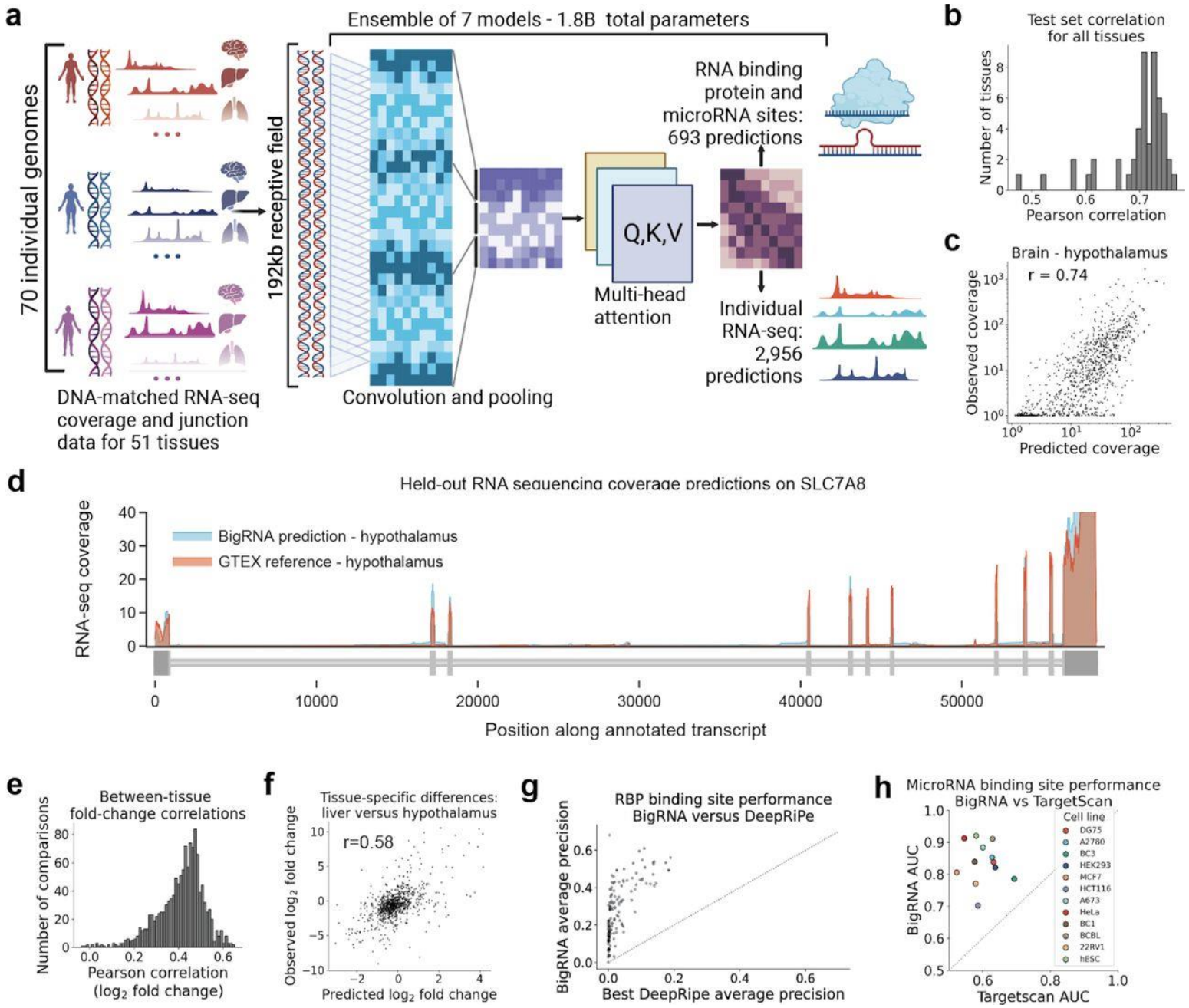
896

896 × 128 = 114688bp



Effective gene expression prediction from sequence by integrating long-range Interactions.

Nature Methods volume 18, pages1196–1203 (2021)



a. BigRNA was trained on the genomes of **70 individuals, to predict a total of 2,956 RNA-seq datasets over 51 tissues**, plus 693 datasets corresponding to RNA binding protein and microRNA sites.

b. Distribution of correlations between **predicted and measured RNA-seq coverage in exonic regions** for genes held-out during training (averaged across individuals)

c. Correlation between predicted and measured RNA-seq coverage for the hypothalamus samples

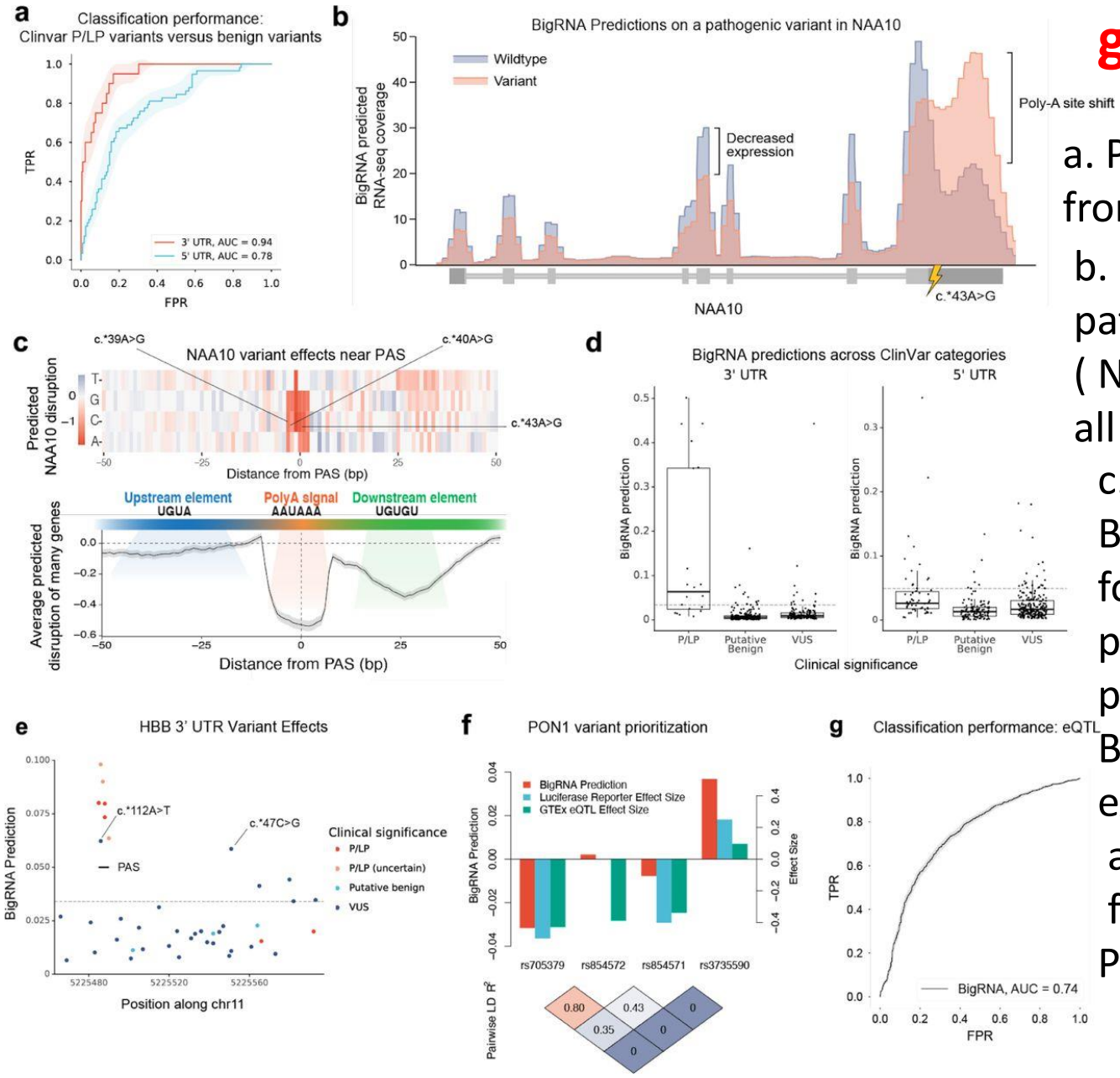
d. Predicted versus measured coverage for *SLC7A8* , averaged Across hypothalamus samples for all individuals

g. Comparison of BigRNA and a previously published method, DeepRiPe, for predicting the binding sites of 98 RNA Binding proteins across 2 cell lines.

While some accurate methods exist for predicting the pathogenic impact of rare missense variants , non-coding variants, such as those located within the 3' and 5' untranslated regions (UTRs) of genes, remain difficult to interpret.

To address this gap, authors evaluate BigRNA's ability to predict the impact of a curated set of pathogenic or likely pathogenic (P/LP) UTR variants from ClinVar

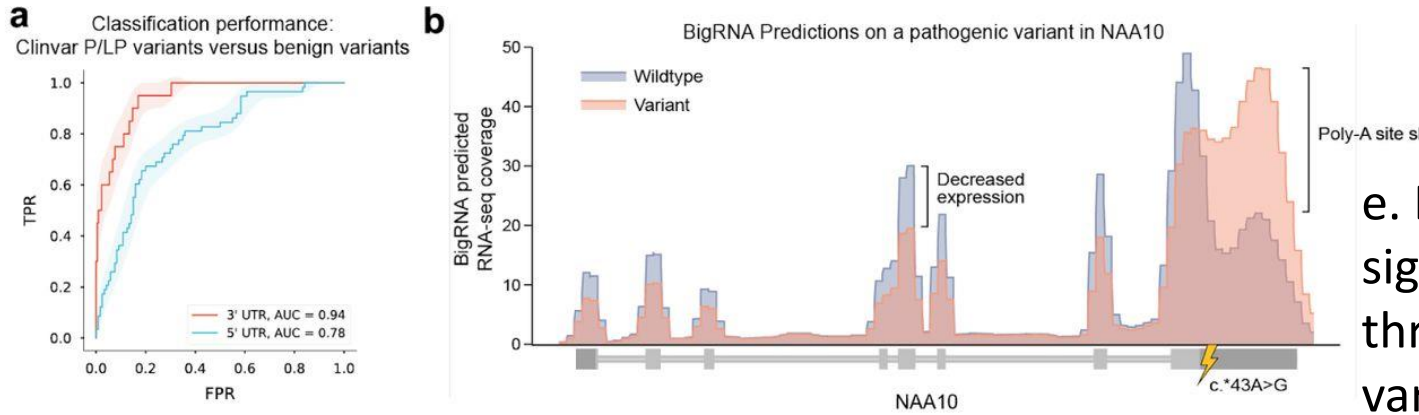
Predicting the effects of variants on gene expression

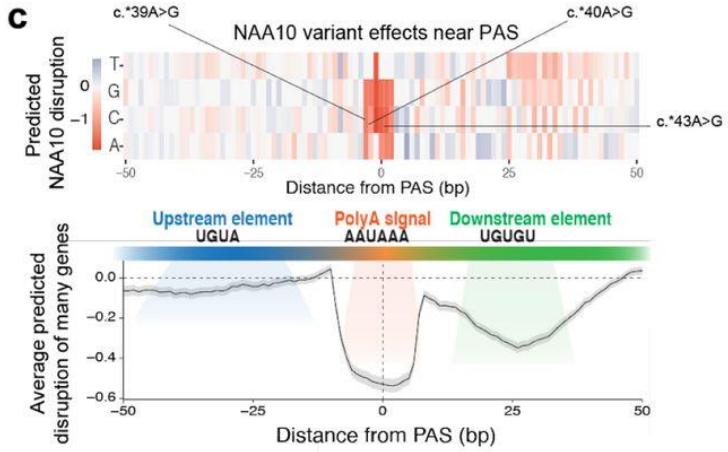


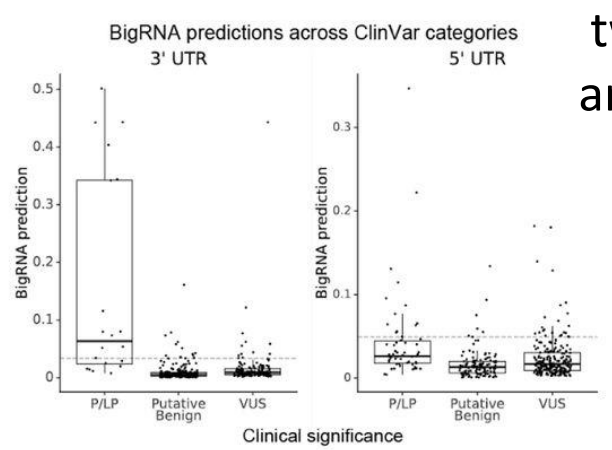
a. Performance of BigRNA on classifying P/LP variants from putative benign variants in the 3’ UTR and 5’ UTR

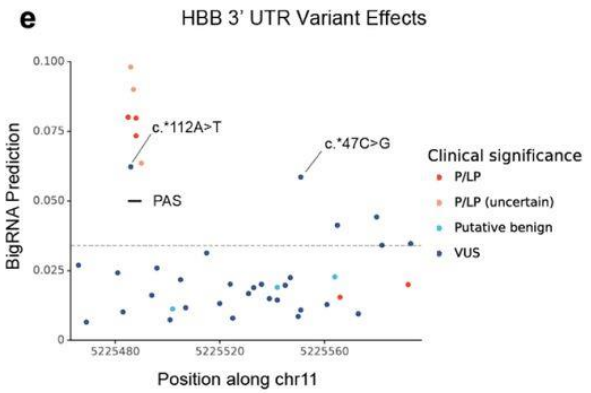
b. RNA-seq coverage predictions for the effects of a pathogenic variant in the 3’ UTR of NAA10 (NM_003491.4), averaged across all individuals and all tissue types

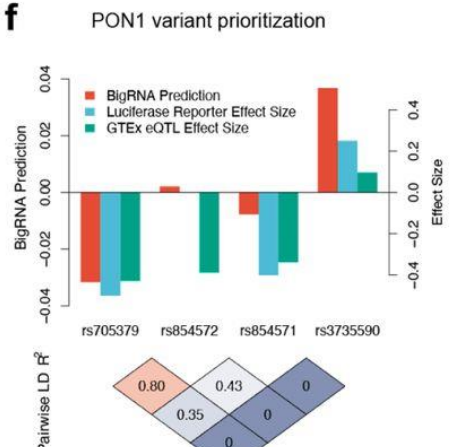
c. Top:
BigRNA predictions showing the change in expression for all possible point mutations around the polyadenylation site (PAS) of NAA10 . Three variants previously identified as impacting the PAS are labeled
Bottom: Relationship between the change in expression predicted by BigRNAfrom ablating regions around the PAS relative to the distance from the PAS for 200 human poly(A) signal sequences selected fror PolyASite 2.0.

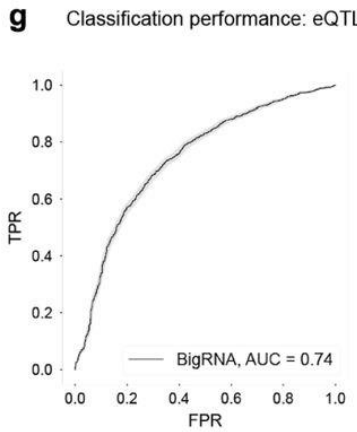












e. BigRNA predictions for variants of varying clinical significance in HBB . The dashed line represents the threshold of classifying P/LP from putative benign variants at a 5% FPR in the 3’ UTR (y = 0.0341). The two highest scoring VUS variants in this gene are annotated.

f. Top: Comparing BigRNA predicted effects to GTEx eQTL effect size and results of a luciferase reporter assay for four variants suspected to impact PON1 expression.

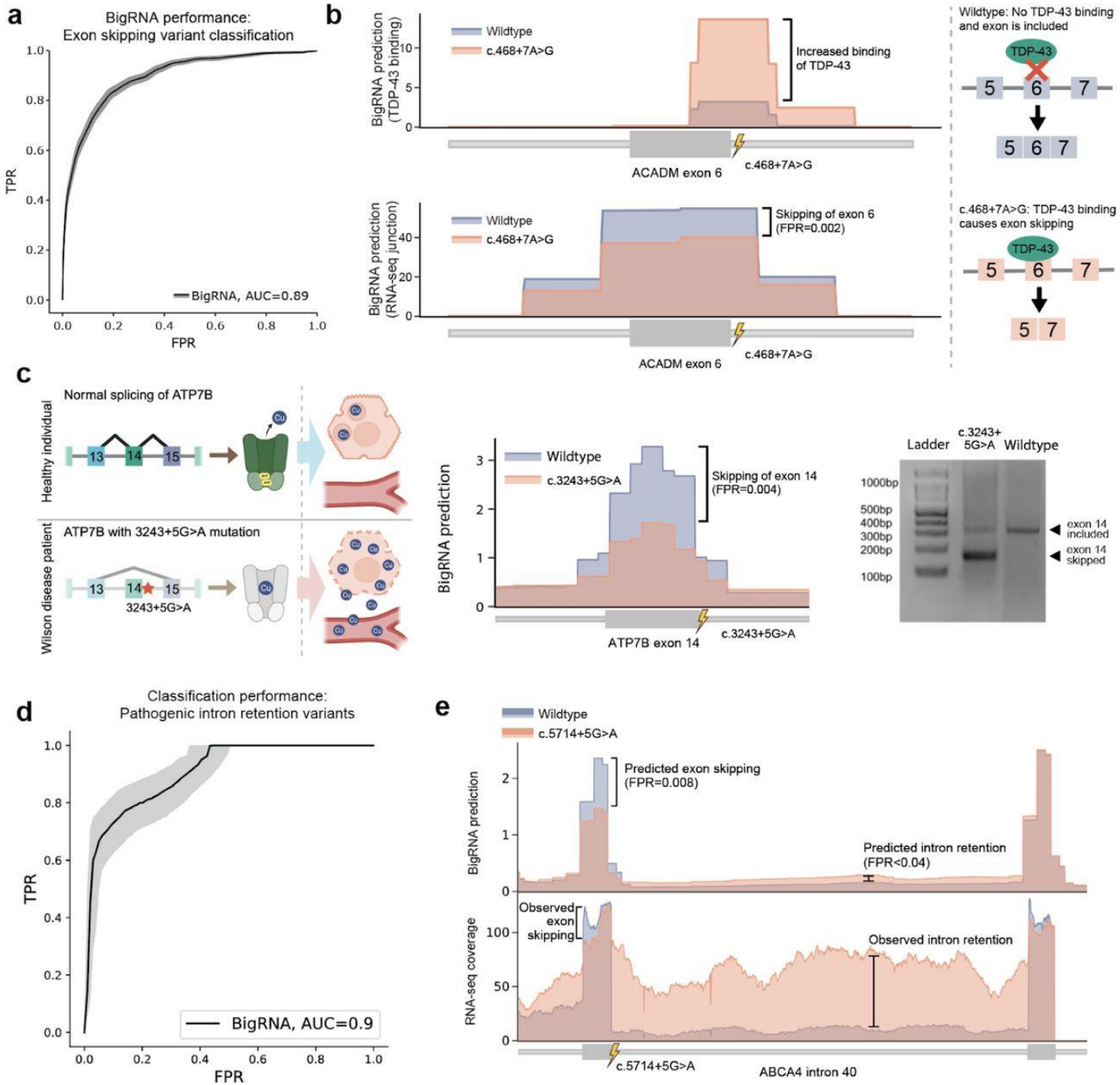
Bottom:

Estimated linkage disequilibrium between variants.

g . Performance of BigRNA at distinguishing fine-mapped expression quantitative trait loci (eQTLs) from controls matched by effector gene (eGene), distance to the transcription start site of the eGene, and minor allele frequency

Predicting the effects of variants on splicing and intron retention

An important subset of pathogenic variants affect splicing, such as those which cause skipping of an exon. These variants often occur in coding regions, and may be incorrectly classified as benign mutations based on their amino acid substitutions, despite their pathogenic splicing effects.



a. BigRNA performance on classifying exonic variants that result in exon skipping by at least 50%, from exonic variants that do not cause skipping, both obtained from MaPSy.

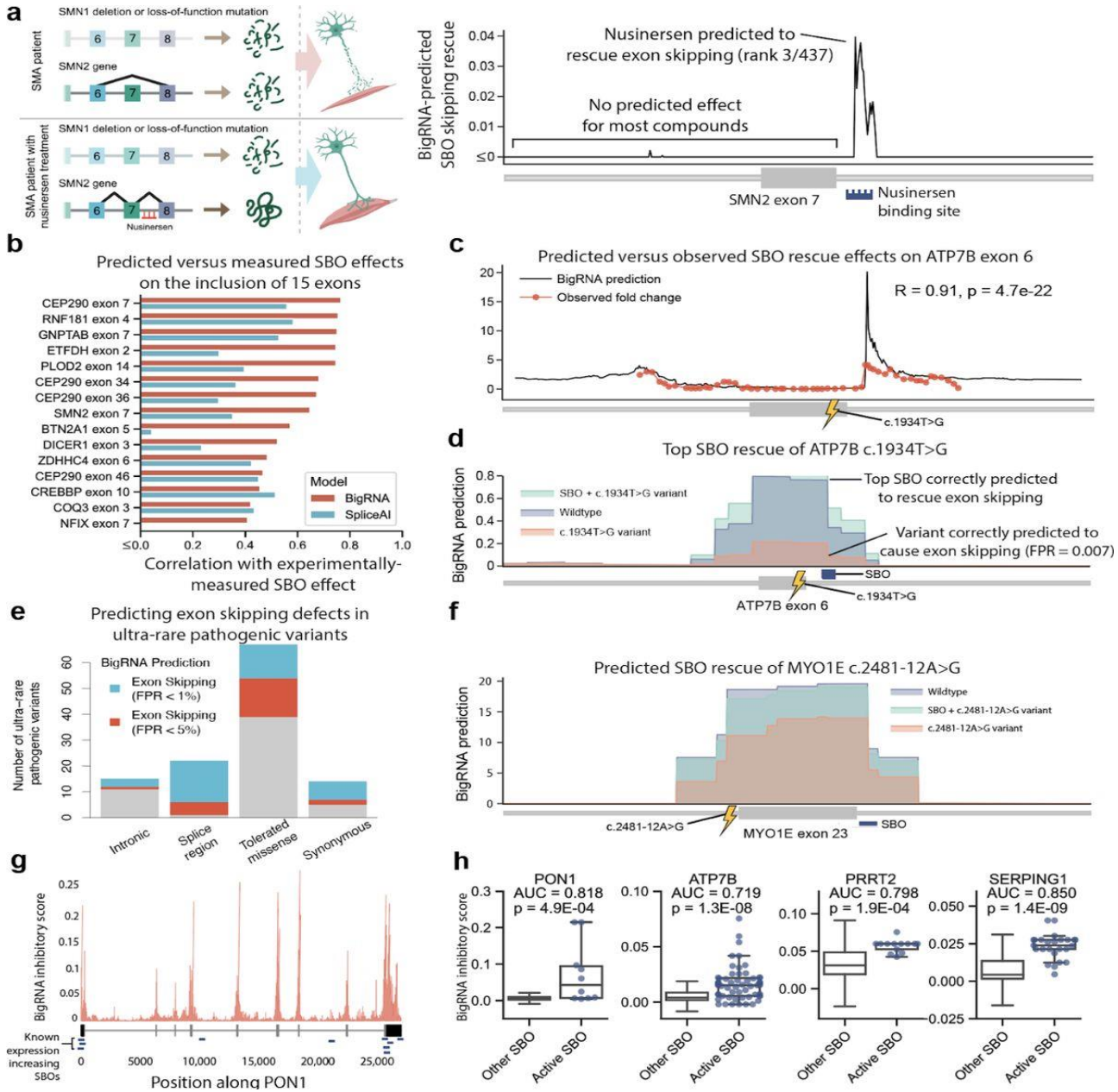
b. BigRNA predicts that the c.468+7A>G variant will result in increased TDP-43 binding and skipping of ACADM exon 6.

c. The ATP7B VUS c.3243+5G>A is predicted by BigRNA to cause in-frame skipping of exon 14. This results in reduced levels of functional ATP7B protein, leading to copper buildup in the cell. Right: An RT-PCR in HepG2 cells edited to be homozygous for 3243+5G>A confirms the expected fragment from exon skipping.

d. BigRNA performance on classifying variants that cause intron retention (n = 25) from a set of matched variants that do not impact splicing (n = 63).

f. Top: BigRNA coverage predictions of the c.5714+5G>A variant in ABCA4 . Bottom: RNA-seq of wildtype WERI cells and WERI cells edited to be homozygous for the variant confir both exon skipping and intron retention effects.

Designing splice-switching and expression-increase molecules

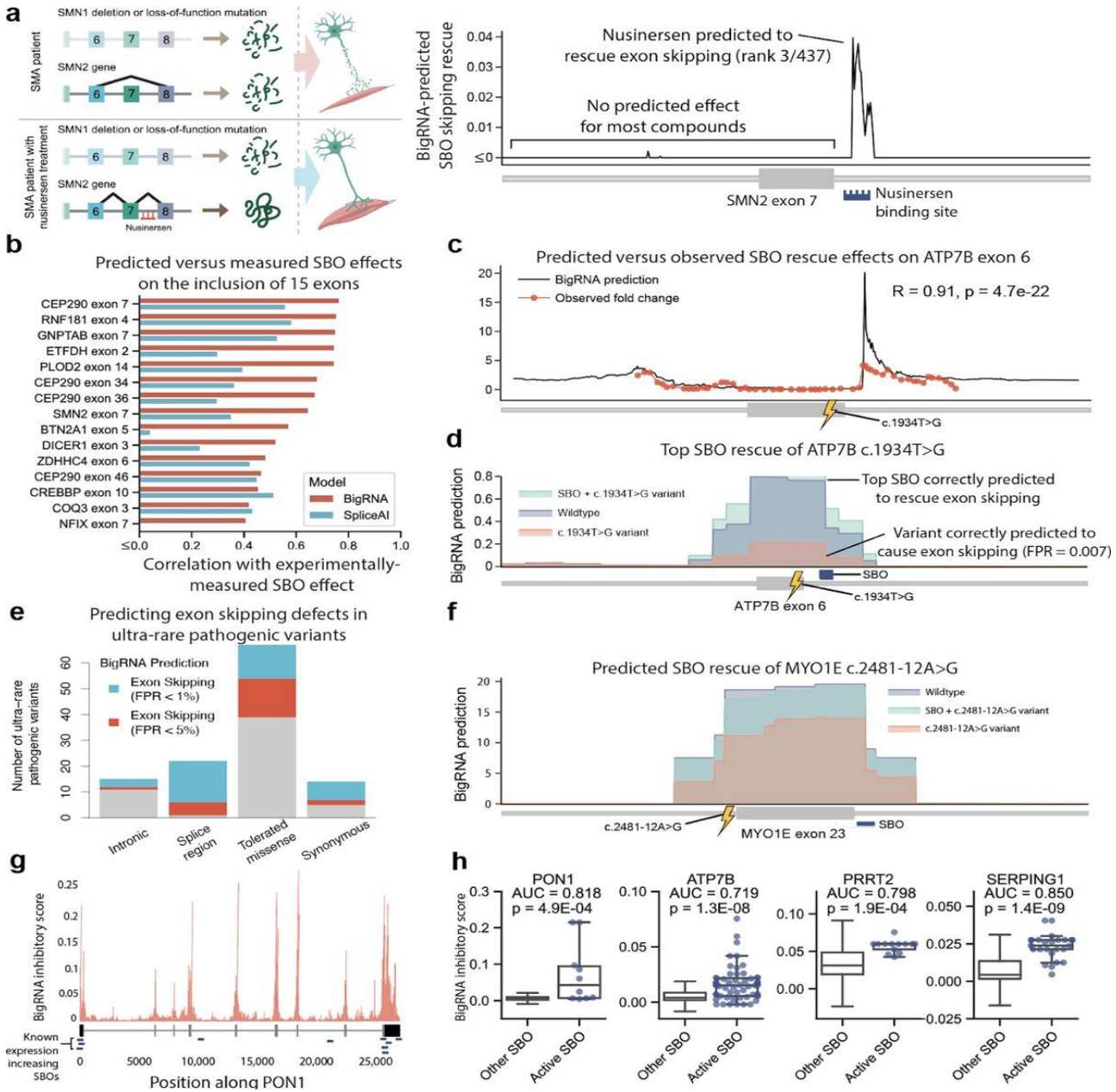


a. Mechanism of action of the splice-switching oligonucleotide Nusinersen, an approved treatment for spinal muscular atrophy (SMA). BigRNA predictions are shown for the exon-restoring effects of all 18-mer SBOs within 200 bp of SMN2 exon 7. The blue bar shows the position of Nusinersen. Predictions were truncated at zero for the plot

b. Spearman correlation between experimentally observed exon-inclusion levels and predictions generated by BigRNA and SpliceAI . A negative correlation for NFIX exon 7 versus SpliceAI (r=-0.13) was truncated to zero

c. BigRNA predictions of SBO effects on ATP7B exon 6 inclusion. 55 SBOs were screened by qPCR to measure total ATP7B expression relative to control (fold change), and the Spearman correlation was computed between the BigRNA predictions and observed fold changes.

d. BigRNA predictions for wildtype, Met645Arg (c.1934T>G) variant, and Met645Arg variant with treatment (lead SBO targeting ATP7B exon 6). The junction count tracks pertaining to individual samples of the liver tissue are average for plotting



e. Proportion of ultra-rare pathogenic variants associated with AR disorders with BigRNA exon skipping predictions above the 1% and 5% FPR thresholds. Intronic (>8bp from splice site), splice region (<8bp from splice site excluding the core dinucleotides), tolerated missense (SIFT score > 0.05) and synonymous variants are shown.

f. BigRNA predictions for wildtype, c.2481-12A>G variant and the variant with treatment (lead SBO targeting MYO1E exon 23).

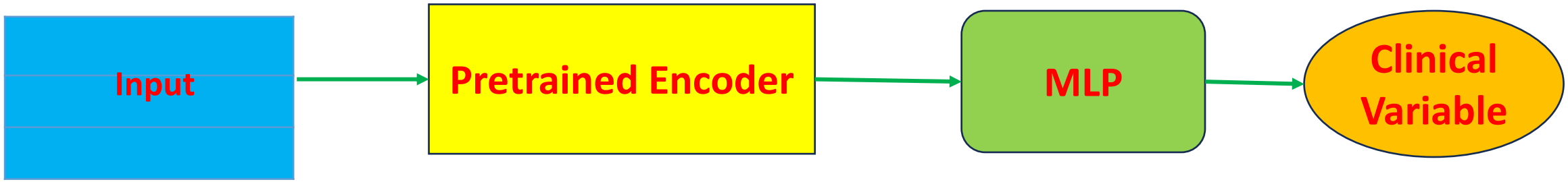
g. BigRNA predicts expression increase SBOs in PON1 . BigRNA inhibitory scores are plotted by region of the gene. The transcript structure is shown under the scores, and the locations of the 10 dose-response hits are shown with blue bars. The distribution of BigRNA inhibitory scores for the 10 dose-response hits is significantly different from the distribution for other length-matched SBOs targeting PON1

h. BigRNA scores of screening hits compared to background of all possible SBOs of same length for PON1, ATP7B , PRRT2 , and SERPING1

Next is Association and Causation Analysis

Tools

- Hypothesis Testing in AI
- Causal Inference


Input
Pretrained Encoder
MLP
Clinical Variable

Cases – Control Studies

- **Null Hypothesis:**

H_0: There is no difference in fitness between cases and controls. $\boldsymbol{H_0: V^A = V^C}$

H_a: Presence of difference in fitness between cases and controls. $\boldsymbol{H_a: V^A \neq V^C}$

- **Notations and Fitness in Cases and Controls.**

n_A: Number of cases
n_C: Number of controls
$l(x_n^i)$: fitness of the individual n in cases at the i^{th} position in a gene with a genotype $x_{j_n}^i$.
$l(y_n^i)$: fitness of the individual n in controls at the i^{th} position in a gene with a genotype $y_{j_n}^i$.

- **Association Tests**

Single Marker

$$T_s = \frac{n_A n_C}{n_A + n_C} \frac{\left(\bar{l}_A - \bar{l}_C\right)^2}{\hat{\sigma}^2} \qquad \bar{l}_A \sim N\left(V^A, \frac{1}{n_A}\hat{\sigma}^2\right), \bar{l}_C \sim N\left(V^C, \frac{1}{n_C}\hat{\sigma}^2\right) \qquad \textbf{(3)}$$

Distribution

Under the null hypothesis $T_s \sim \chi^2_{(1)}$

Define

Embedding vector of genotype of individual n_A in cases

Embedding vector of genotype of individual n_C in Controls

Semantic Embedding and Mutation Effect

- **Test Statistic**

  Single Marker

- **Null Hypothesis** $H_0: \mu_A = \mu_C$

H_0: There is no difference in embedding of genotype in position i between cases and controls

H_a: Presence of difference in embedding of genotype in position i between cases and controls

Embedding vector of genotypes of individual n in controls

$$T_s = \left(\bar{Z}_A^i - \bar{Z}_C^i\right)^T \hat{\Lambda}^{-1} \left(\bar{Z}_A^i - \bar{Z}_C^i\right)$$

Embedding vector of genotype of individual n in cases

$$\bar{Z}_A^i \sim N\left(\mu_A, \frac{1}{n_A}\Sigma\right), \bar{Z}_C^i \sim N\left(\mu_C, \frac{1}{n_C}\Sigma\right)$$

Under the null hypothesis, $T_s \sim \chi^2_{(H)}$

Genomic Regions

$$Var(\bar{Z}_A - \bar{Z}_C) = \Lambda = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\Sigma$$

$$\hat{\Lambda} = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\mathrm{S},$$

$$\mathrm{S} = \frac{1}{n_A + n_C - 2}\left[\sum_{n=1}^{n_A}(\bar{Z}_{An} - \bar{Z}_A)(\bar{Z}_{An} - \bar{Z}_A)^T + \sum_{n=1}^{n_C}(\bar{Z}_{Cn} - \bar{Z}_C)(\bar{Z}_{Cn} - \bar{Z}_C)^T\right]$$

- **Null Hypothesis**

H_0: There is no difference in the total embedding of the genotype in a genomic region between cases and controls.

H_a: Presence of difference in the total embeddings of the genotypes in a genomic region between cases and controls.

$$\Sigma_A = Cov(Z^i_{An}, Z^i_{An}), \Sigma_C = Cov(Z^i_{Cn}, Z^i_{Cn})$$

Define test statistics

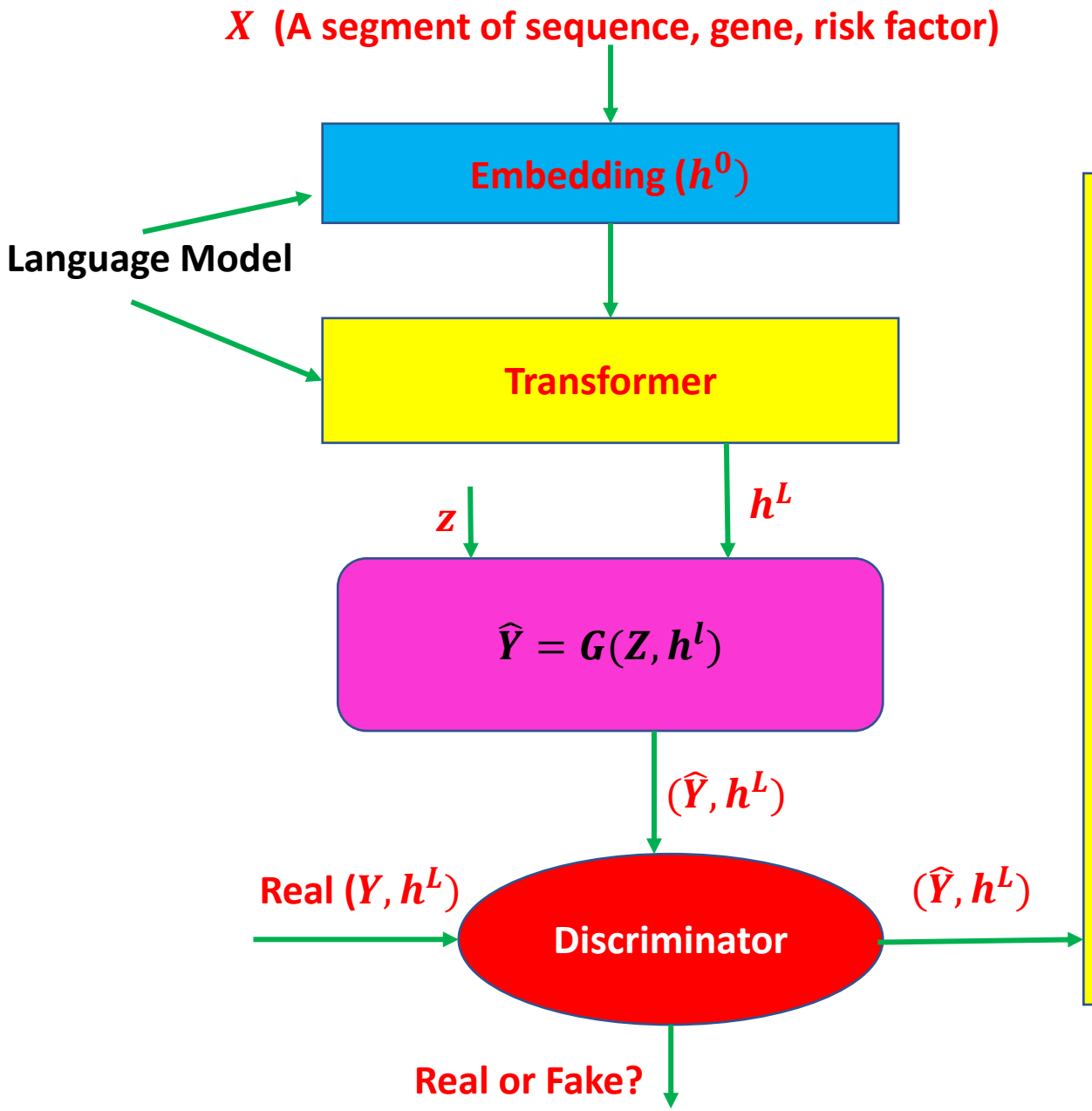
$$\boldsymbol{T_m = (\bar{Z}_A - \bar{Z}_C)^T \hat{\Lambda}^{-1}(\bar{Z}_A - \bar{Z}_C)}$$

$$\bar{Z}_A \sim N\left(\mu_A, \frac{1}{n_A}\Sigma_A\right), \bar{Z}_C \sim N(\mu_C, \frac{1}{n_C}\Sigma_C)$$

Under the null hypothesis, $T_m \sim \chi^2_{(H)}$

Causation Test

Pair-wise causal analysis (GAN)



Two Sample Test

$$D_{X\to Y} = \{h_i^L, \hat{Y}_i = G(Z_i, h_i^L), i = 1, \dots, n\}$$

$$D_t = \{h_i^L, Y_i, i = 1, \dots, n\}$$

$$D = \{[(\hat{Y}_1, 1), \dots, (\hat{Y}_n, 1)] \cup [(Y_1, 0), \dots, (Y_n, 0)]\}$$

$$= \{(z_1, l_1), \dots, (z_{2n}, l_{2n})\}$$

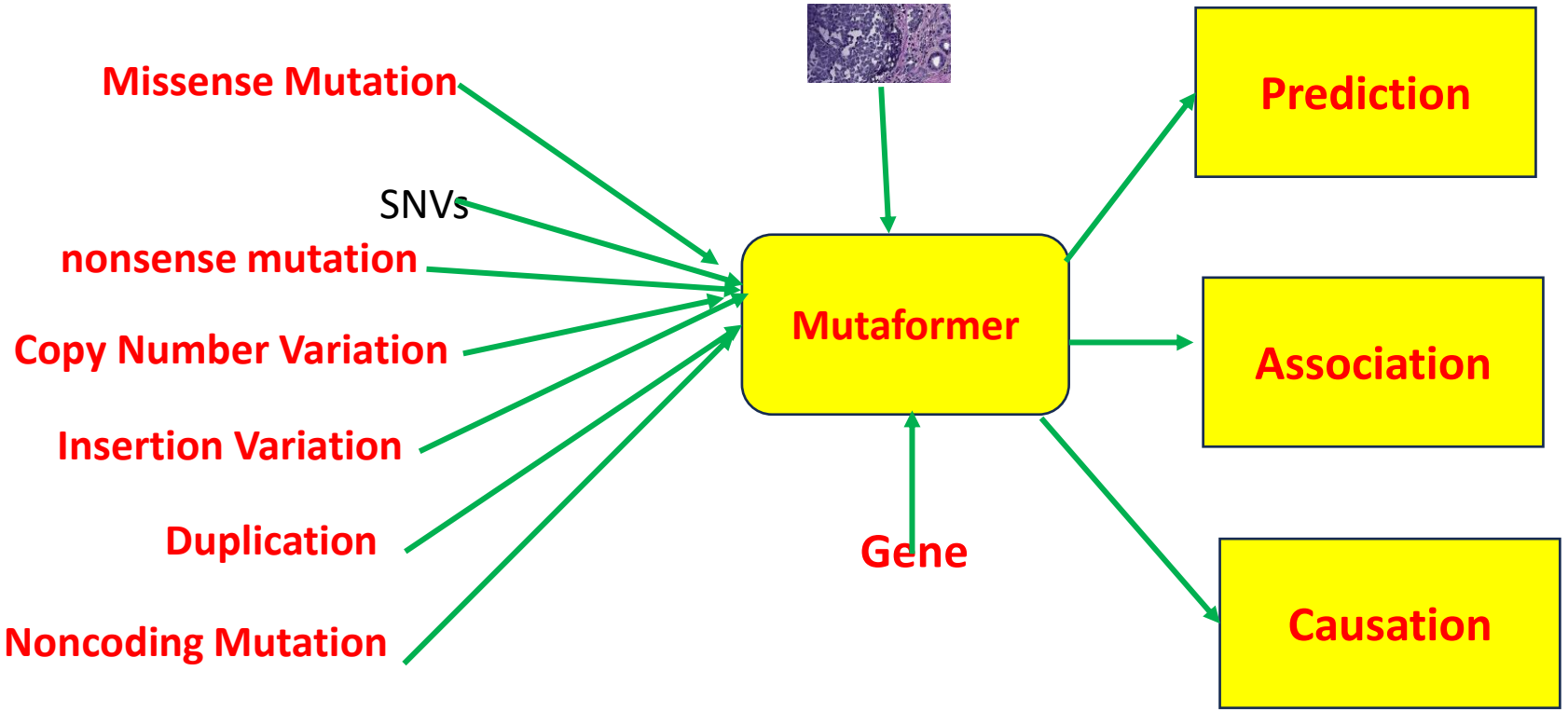
$$l = \begin{cases} 1 & \hat{Y}_i \\ 0 & Y_i \end{cases}$$

$$T_{C(X\to Y)} = \frac{1}{n_{te}} \sum_{(h_i, l_i)\in D_{te}} w_i$$

$$w_i = I[I\left(f(h_i) > \frac{1}{2}\right) = l_i]$$

$T_{C(X\to Y)} \sim 0.5, X \to Y$
X causes Y.

Generalist Model for mutation, gene expression and Image



Apply Causal Methods to Foundation Model