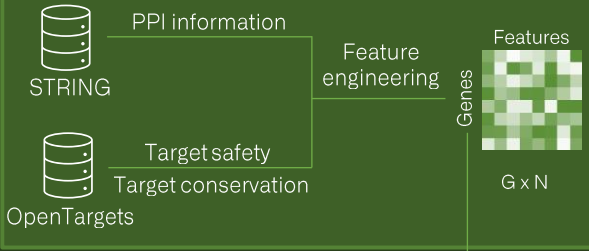ELGH exome data

Gene name
Optional: cell type

Variant type
Variant sequence
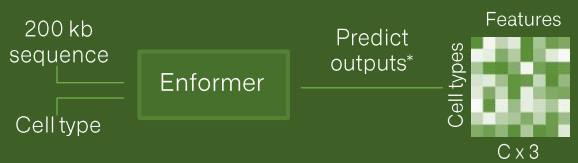
Gene characterization

Cell type-agnostic

STRING — PPI information

Feature engineering

Target safety
Target conservation

OpenTargets

Features

Genes

G x N

for gene in genes:

Cell type-specific

200 kb sequence

Cell type

Enformer

Predict outputs*

Features
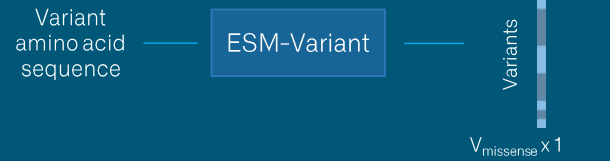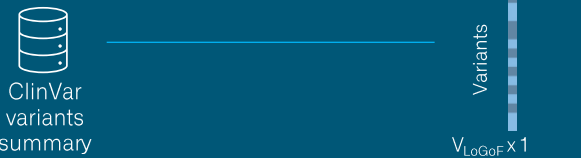
Cell types

C x 3

*Enformer predicts gene expression (CAGE), chromatin accessibility (DNAse), and histone modifications (ChIP-seq).
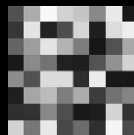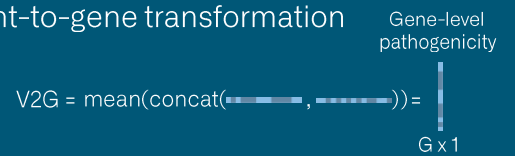
Pathogenicity inference

if variant_type == 'missense':

Variant amino acid sequence — ESM-Variant

Pathogenicity classification

Variants

$V_{missense} \times 1$

else:

ClinVar variants summary

Pathogenicity classification

Variants

$V_{LoGoF} \times 1$

Variant-to-gene transformation

$V2G = mean(concat(\text{━━━━} , \text{┅┅┅┅})) =$

Gene-level pathogenicity

G x 1

Final feature matrix

C x (N + 4)

Machine learning pipeline