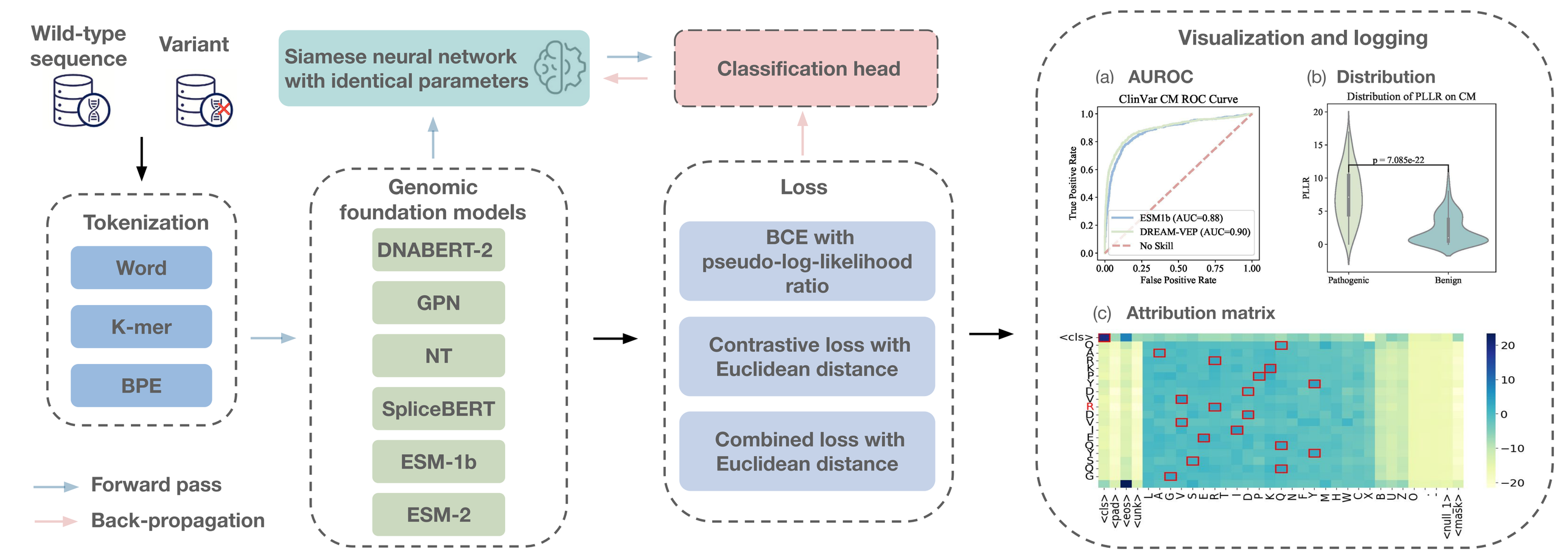
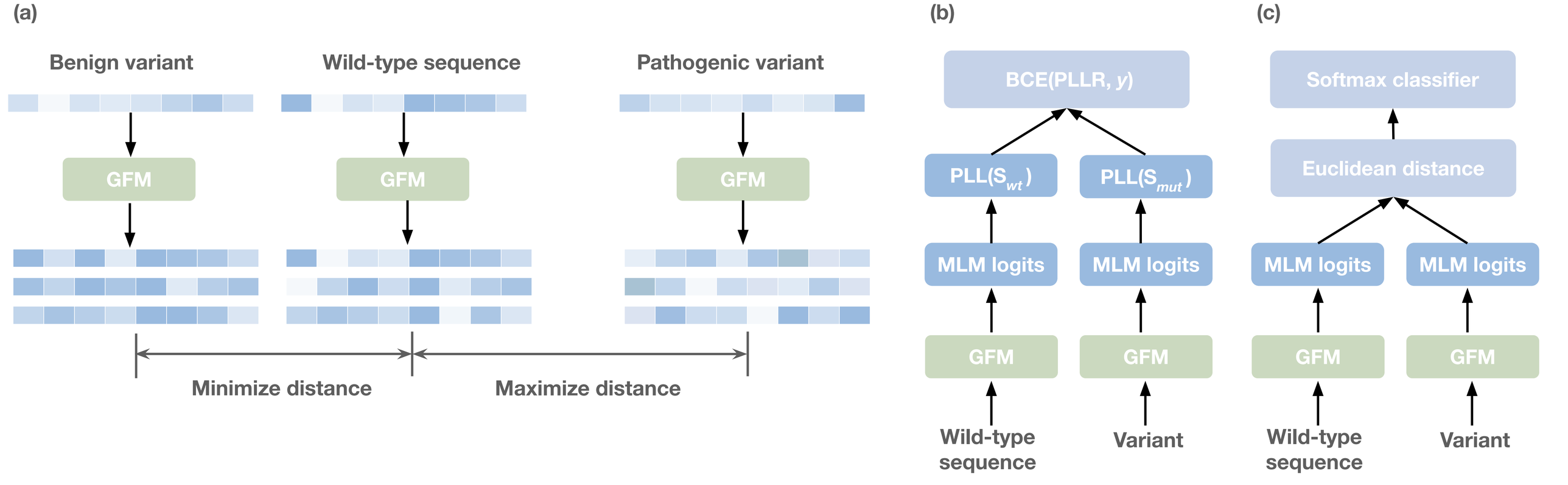
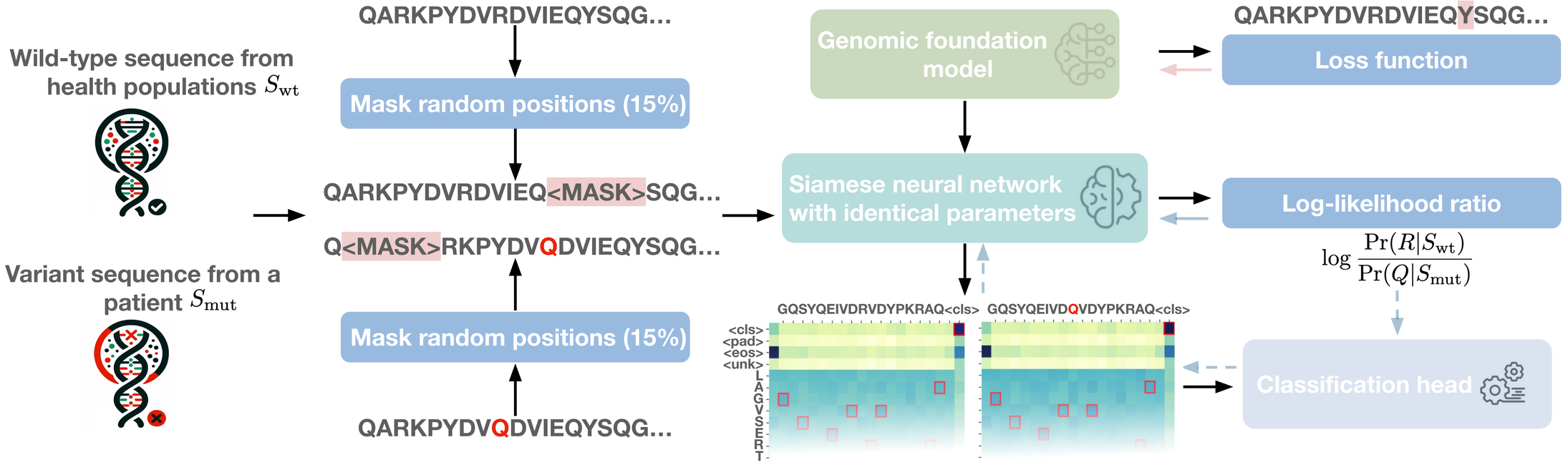
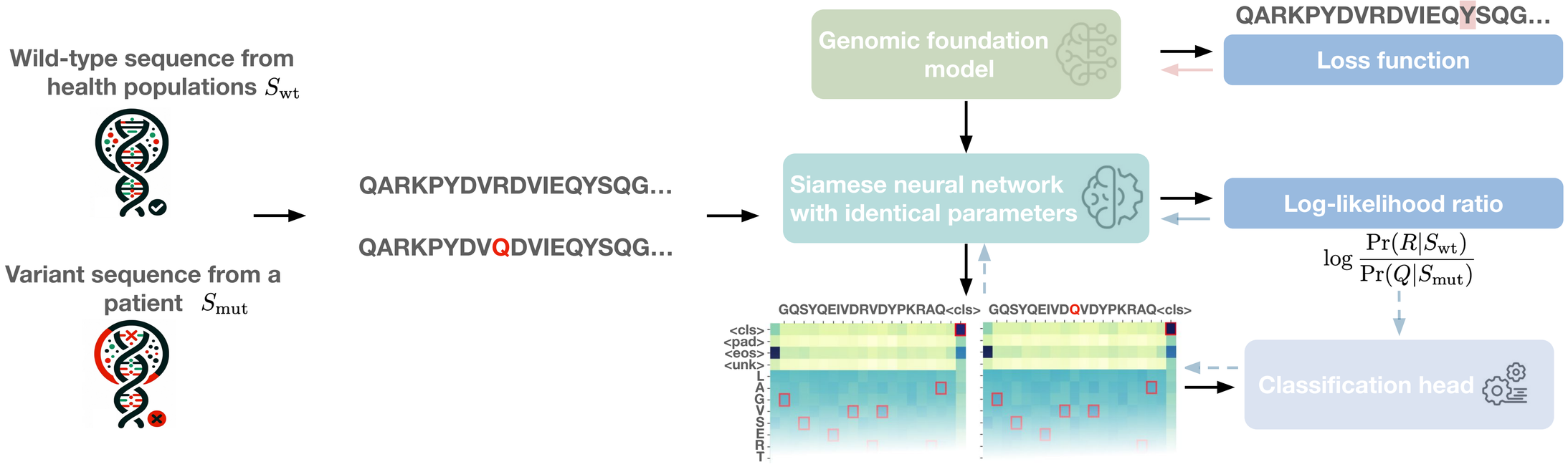
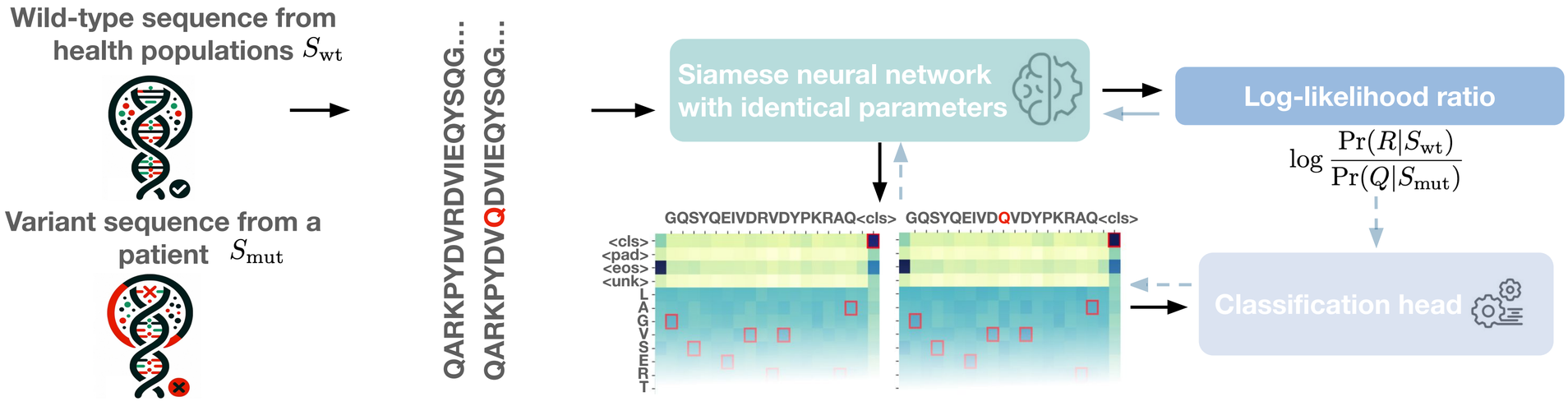
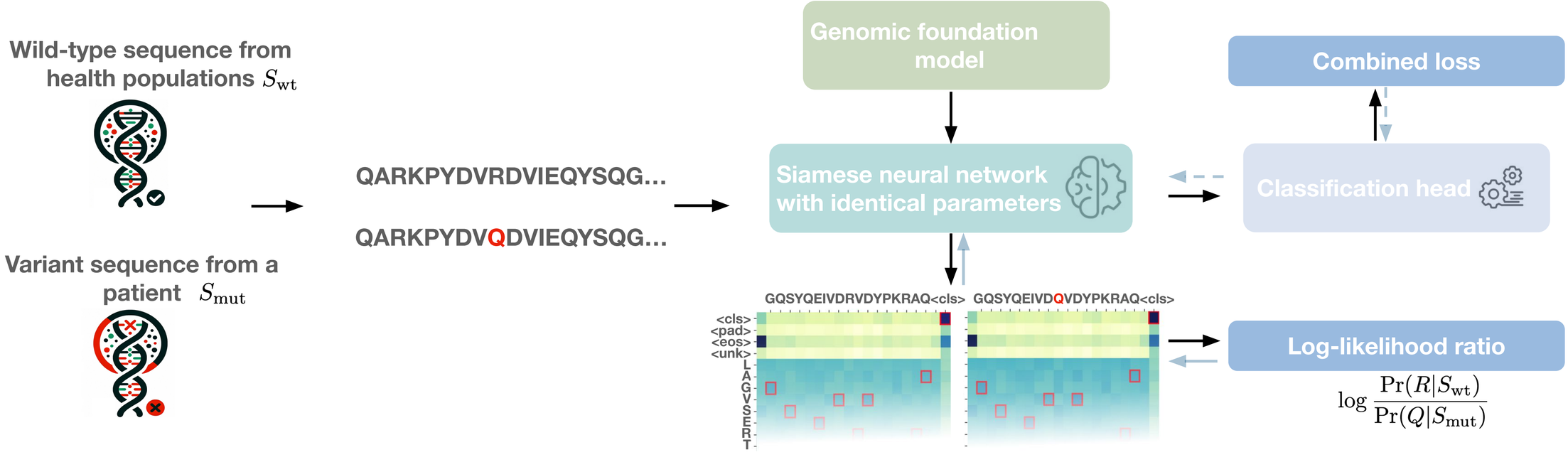

Wild-type sequence
Variant
Siamese neural network with identical parameters
Classification head
Tokenization
Word
K-mer
BPE
Genomic foundation models
DNABERT-2
GPN
NT
SpliceBERT
ESM-1b
ESM-2
Loss
BCE with pseudo-log-likelihood ratio
Contrastive loss with Euclidean distance
Combined loss with Euclidean distance
Visualization and logging
(a) AUROC
ClinVar CM ROC Curve
True Positive Rate
False Positive Rate
ESM1b (AUC=0.88)
DREAM-VEP (AUC=0.90)
No Skill
(b) Distribution
Distribution of PLLR on CM
PLLR
p = 7.085e-22
Pathogenic
Benign
(c) Attribution matrix
Forward pass
Back-propagation

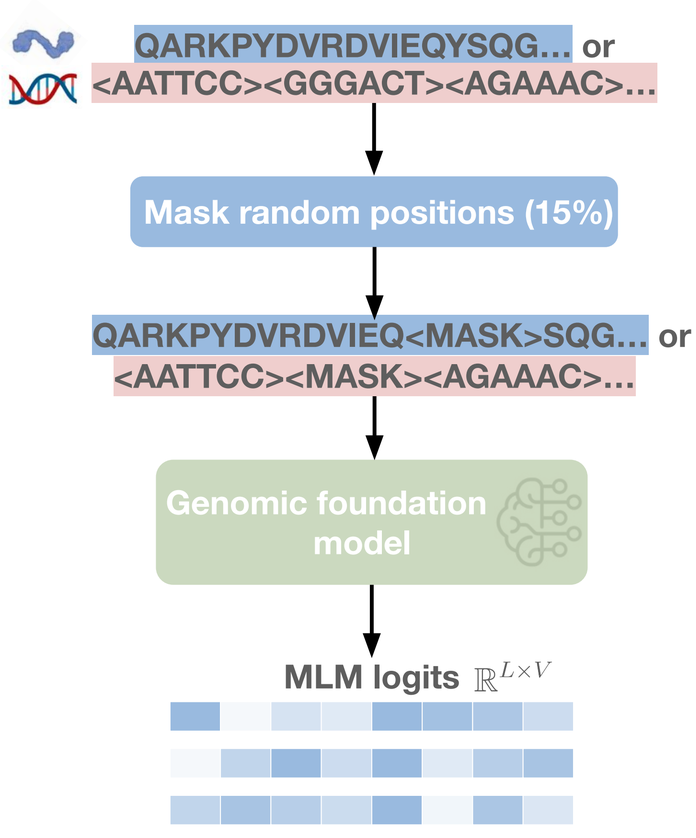

(a)
Benign variant
Wild-type sequence
Pathogenic variant
GFM
GFM
GFM
Minimize distance
Maximize distance
(b)
BCE(PLLR, y)
PLL(S_wt)
PLL(S_mut)
MLM logits
MLM logits
GFM
GFM
Wild-type sequence
Variant
(c)
Softmax classifier
Euclidean distance
MLM logits
MLM logits
GFM
GFM
Wild-type sequence
Variant


Wild-type sequence from health populations S_{wt}
Variant sequence from a patient S_{mut}
QARKPYDVRDVIEQYSQG...
Mask random positions (15%)
QARKPYDVRDVIEQ<MASK>SQG...
Q<MASK>RKPYDVQDVIEQYSQG...
Mask random positions (15%)
QARKPYDVQDVIEQYSQG...
Genomic foundation model
Siamese neural network with identical parameters
GQSYQEIVDRVDYPKRAQ<cls>
GQSYQEIVDQVDYPKRAQ<cls>
<cls>
<pad>
<eos>
<unk>
L A G V S E R T
QARKPYDVRDVIEQYSQG...
Loss function
Log-likelihood ratio
$\log \frac{\Pr(R|S_{\mathrm{wt}})}{\Pr(Q|S_{\mathrm{mut}})}$
Classification head


Wild-type sequence from health populations S_{wt}
Variant sequence from a patient S_{mut}
QARKPYDVRDVIEQYSQG...
QARKPYDVQDVIEQYSQG...
Genomic foundation model
Siamese neural network with identical parameters
QARKPYDVRDVIEQYSQG...
Loss function
Log-likelihood ratio
$\log \frac{\Pr(R|S_{\mathrm{wt}})}{\Pr(Q|S_{\mathrm{mut}})}$
Classification head
GQSYQEIVDRVDYPKRAQ<cls>
GQSYQEIVDQVDYPKRAQ<cls>
<cls>
<pad>
<eos>
<unk>
L
A
G
V
S
E
R
T


Wild-type sequence from health populations S_{wt}
Variant sequence from a patient S_{mut}
QARKPYDVRDVIEQYSQG...
QARKPYDVQDVIEQYSQG...
Siamese neural network with identical parameters
Log-likelihood ratio
$\log \frac{\Pr(R|S_{\mathrm{wt}})}{\Pr(Q|S_{\mathrm{mut}})}$
GQSYQEIVDRVDYPKRAQ<cls>
GQSYQEIVDQVDYPKRAQ<cls>
<cls>
<pad>
<eos>
<unk>
L
A
G
V
S
E
R
T
Classification head


Wild-type sequence from health populations S_{wt}
Variant sequence from a patient S_{mut}
QARKPYDVRDVIEQYSQG...
QARKPYDVQDVIEQYSQG...
Genomic foundation model
Siamese neural network with identical parameters
Combined loss
Classification head
GQSYQEIVDRVDYPKRAQ<cls>
GQSYQEIVDQVDYPKRAQ<cls>
<cls>
<pad>
<eos>
<unk>
L A G V S E R T
Log-likelihood ratio
$\log \frac{\Pr(R|S_{\mathrm{wt}})}{\Pr(Q|S_{\mathrm{mut}})}$


QARKPYDVRDVIEQYSQG… or
<AATTCC><GGGACT><AGAAAC>…
Mask random positions (15%)
QARKPYDVRDVIEQ<MASK>SQG… or
<AATTCC><MASK><AGAAAC>…
Genomic foundation model
MLM logits $\mathbb{R}^{L\times V}$