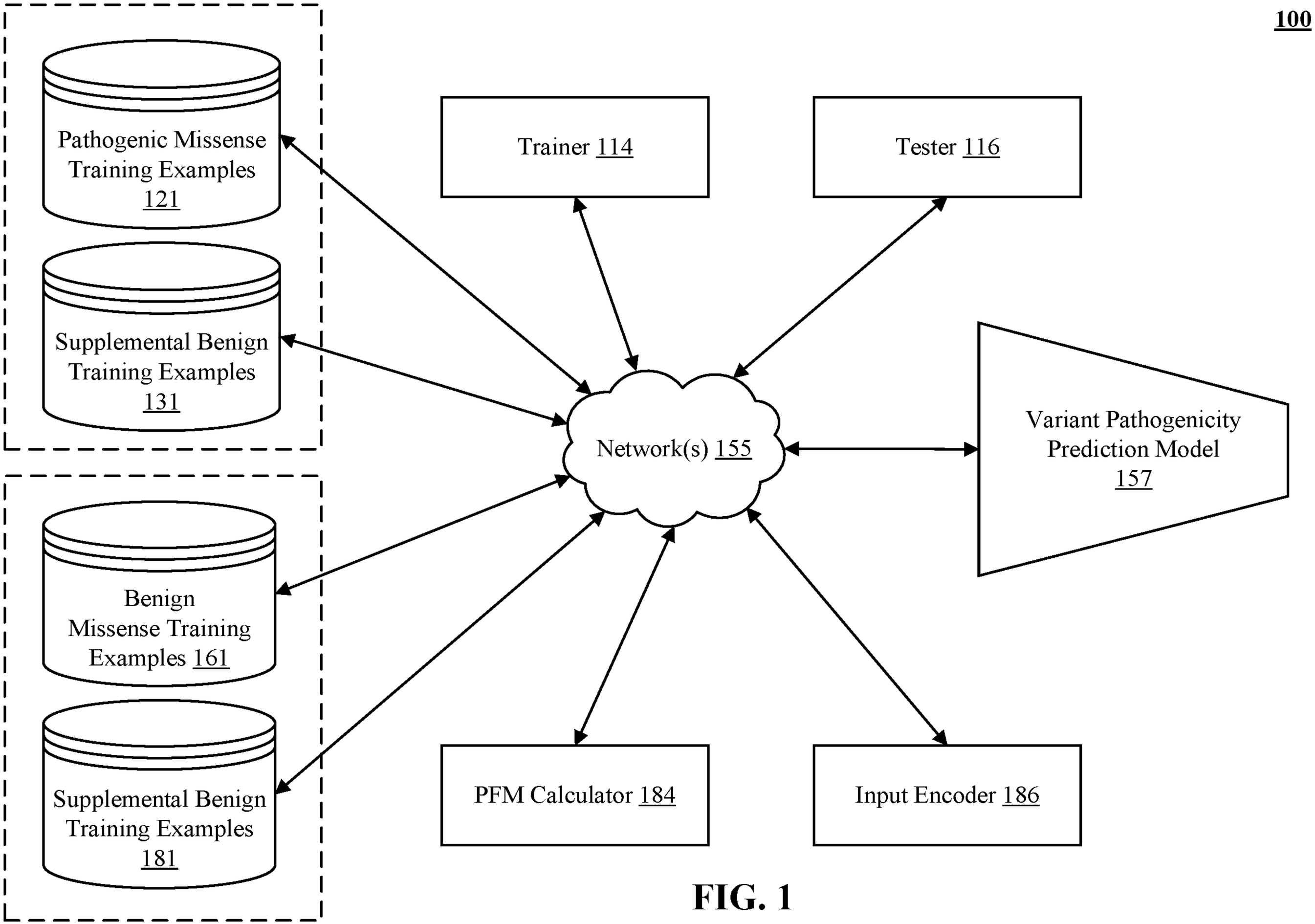

100
Pathogenic Missense Training Examples 121
Supplemental Benign Training Examples 131
Trainer 114
Tester 116
Network(s) 155
Variant Pathogenicity Prediction Model 157
Benign Missense Training Examples 161
Supplemental Benign Training Examples 181
PFM Calculator 184
Input Encoder 186

FIG. 1

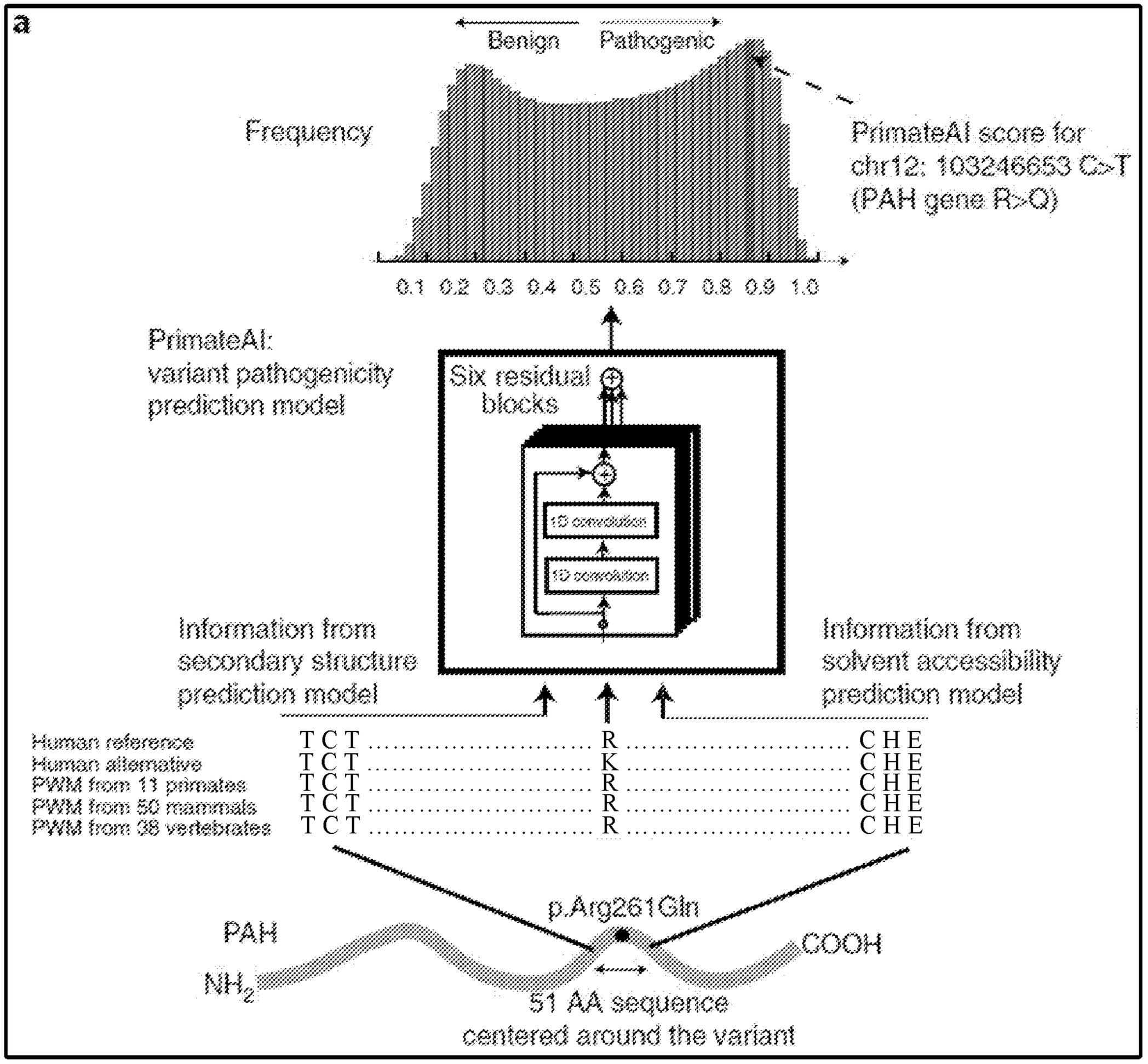


FIG. 2

PAH – phenylalanine hydroxylase
NH_2 – amino-terminus
COOH – carboxyl-terminus

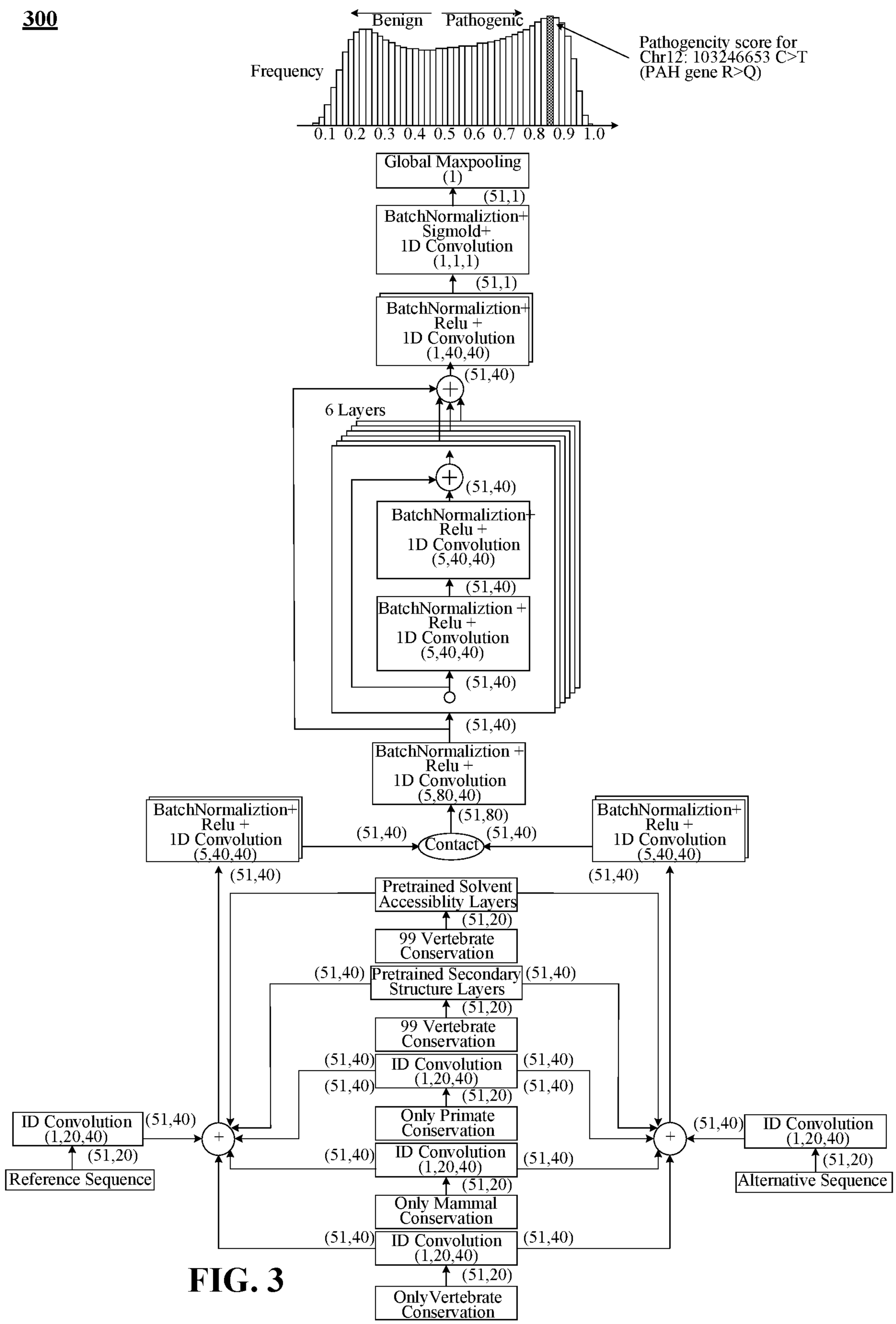

300
Benign
Pathogenic
Frequency
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Pathogencity score for Chr12: 103246653 C>T (PAH gene R>Q)
Global Maxpooling (1)
(51,1)
BatchNormaliztion+ Sigmold+ 1D Convolution (1,1,1)
(51,1)
BatchNormaliztion+ Relu + 1D Convolution (1,40,40)
(51,40)
6 Layers
(51,40)
BatchNormaliztion+ Relu + 1D Convolution (5,40,40)
(51,40)
BatchNormaliztion + Relu + 1D Convolution (5,40,40)
(51,40)
(51,40)
BatchNormaliztion + Relu + 1D Convolution (5,80,40)
(51,80)
BatchNormaliztion+ Relu + 1D Convolution (5,40,40)
(51,40)
Contact
(51,40)
BatchNormaliztion+ Relu + 1D Convolution (5,40,40)
(51,40)
(51,40)
Pretrained Solvent Accessiblity Layers
(51,20)
99 Vertebrate Conservation
(51,40)
Pretrained Secondary Structure Layers
(51,40)
(51,20)
99 Vertebrate Conservation
(51,40)
ID Convolution (1,20,40)
(51,40)
(51,40)
(51,40)
(51,20)
Only Primate Conservation
ID Convolution (1,20,40) (51,40)
(51,40)
ID Convolution (1,20,40)
(51,40)
(51,20)
Reference Sequence
(51,40)
ID Convolution (1,20,40)
(51,20)
Alternative Sequence
(51,20)
Only Mammal Conservation
(51,40)
ID Convolution (1,20,40)
(51,40)
(51,20)
OnlyVertebrate Conservation

FIG. 3

400

Convolutional Neural Network

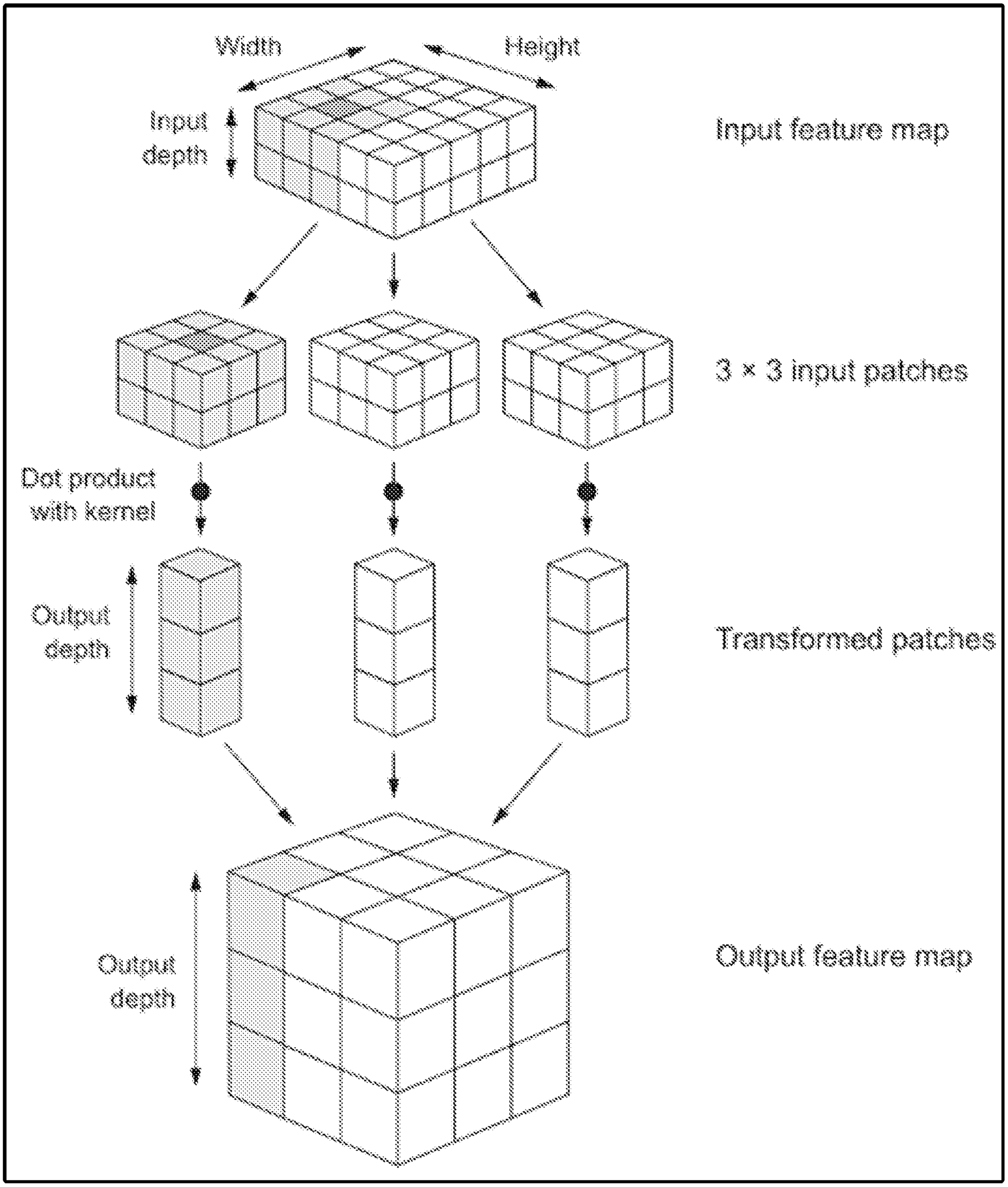


FIG. 4

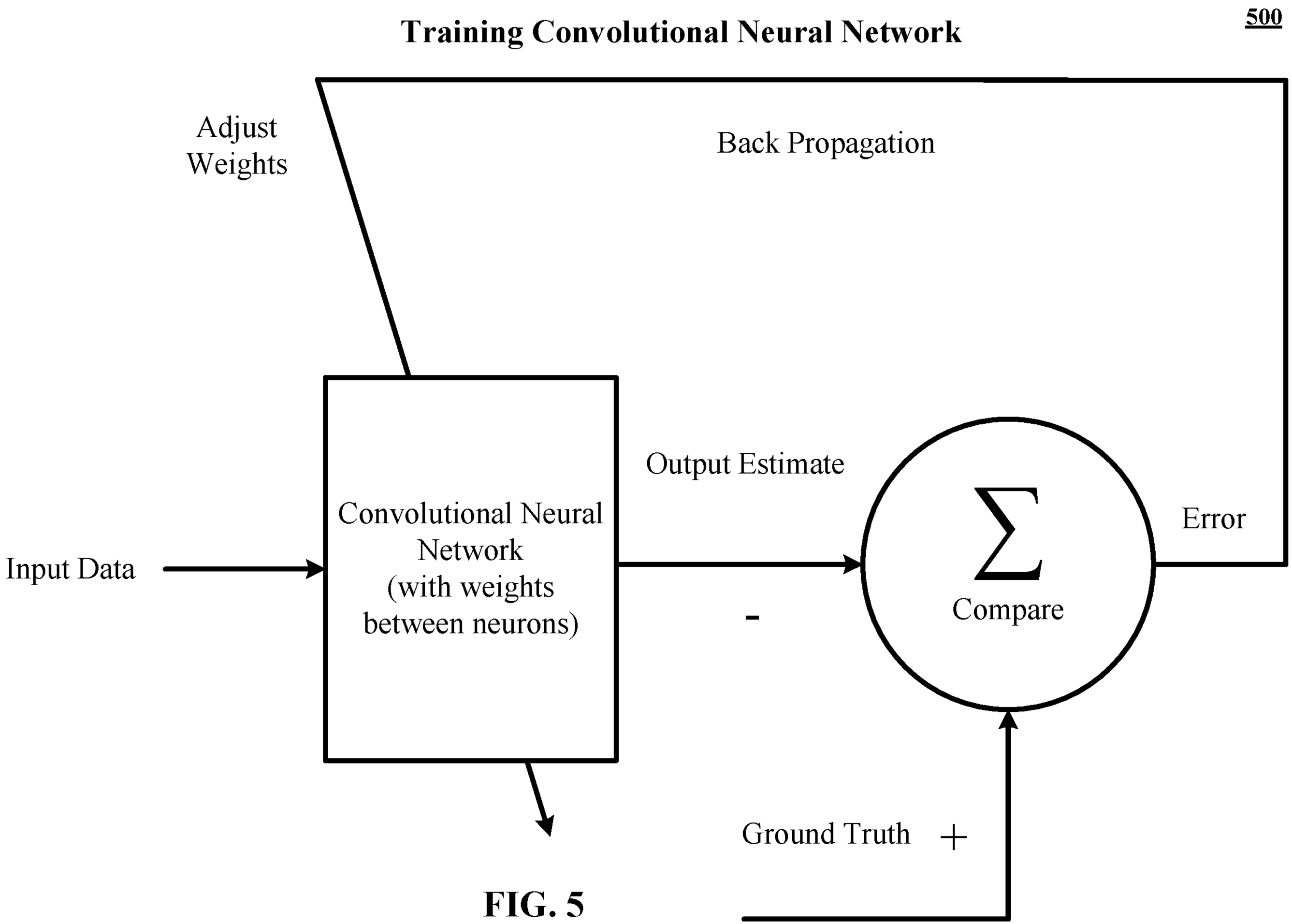

Training Convolutional Neural Network
Adjust Weights
Back Propagation
Input Data
Convolutional Neural Network (with weights between neurons)
Output Estimate
-
Σ
Compare
Error
Ground Truth
+

FIG. 5

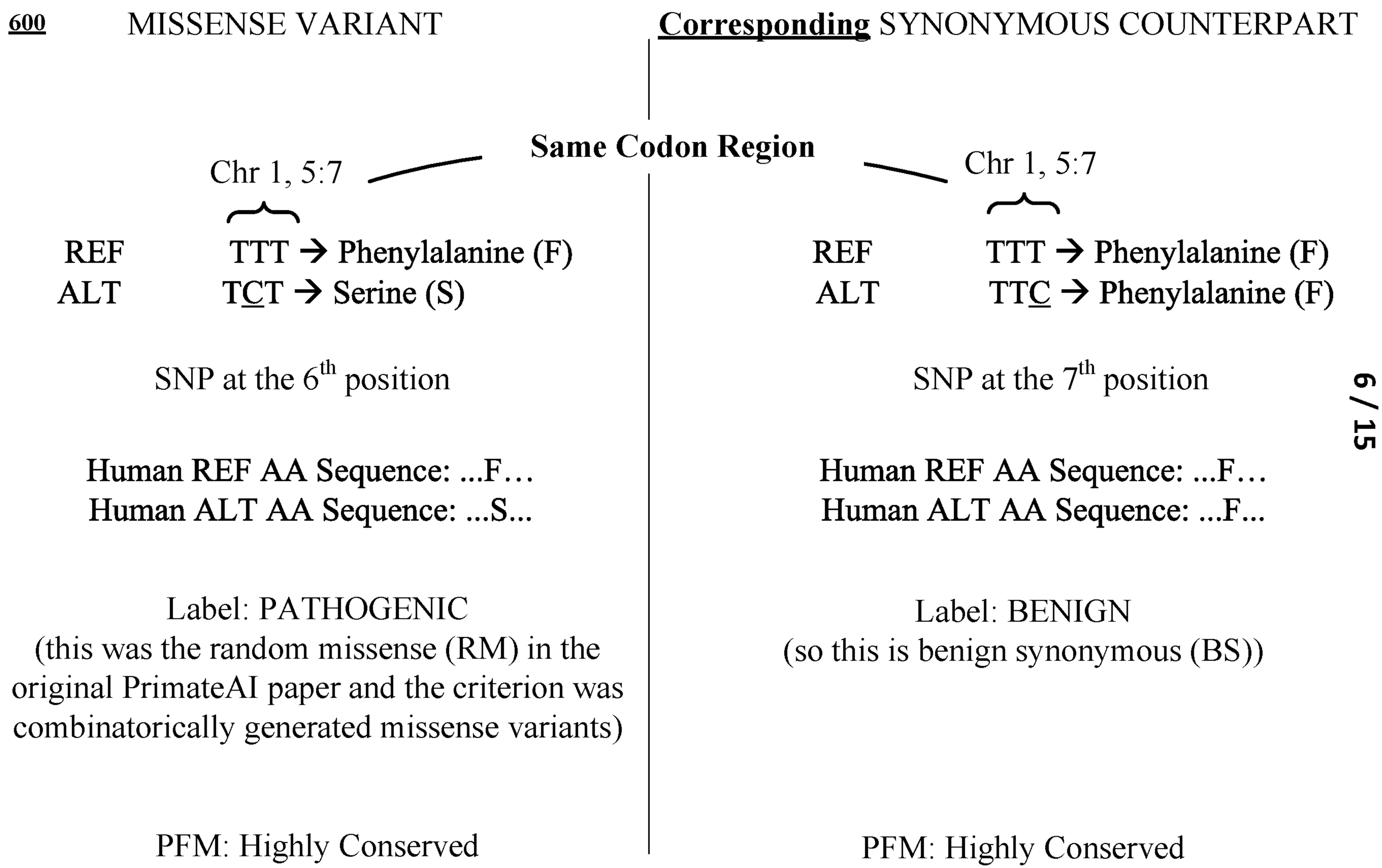

600
MISSENSE VARIANT
Corresponding SYNONYMOUS COUNTERPART
Same Codon Region
Chr 1, 5:7
REF TTT → Phenylalanine (F)
ALT TCT → Serine (S)
SNP at the 6th position
Human REF AA Sequence: ...F…
Human ALT AA Sequence: ...S...
Label: PATHOGENIC
(this was the random missense (RM) in the original PrimateAI paper and the criterion was combinatorically generated missense variants)
PFM: Highly Conserved
Chr 1, 5:7
REF TTT → Phenylalanine (F)
ALT TTC → Phenylalanine (F)
SNP at the 7th position
Human REF AA Sequence: ...F…
Human ALT AA Sequence: ...F...
Label: BENIGN
(so this is benign synonymous (BS))
PFM: Highly Conserved
(but now the model will see as benign label)

FIG. 6

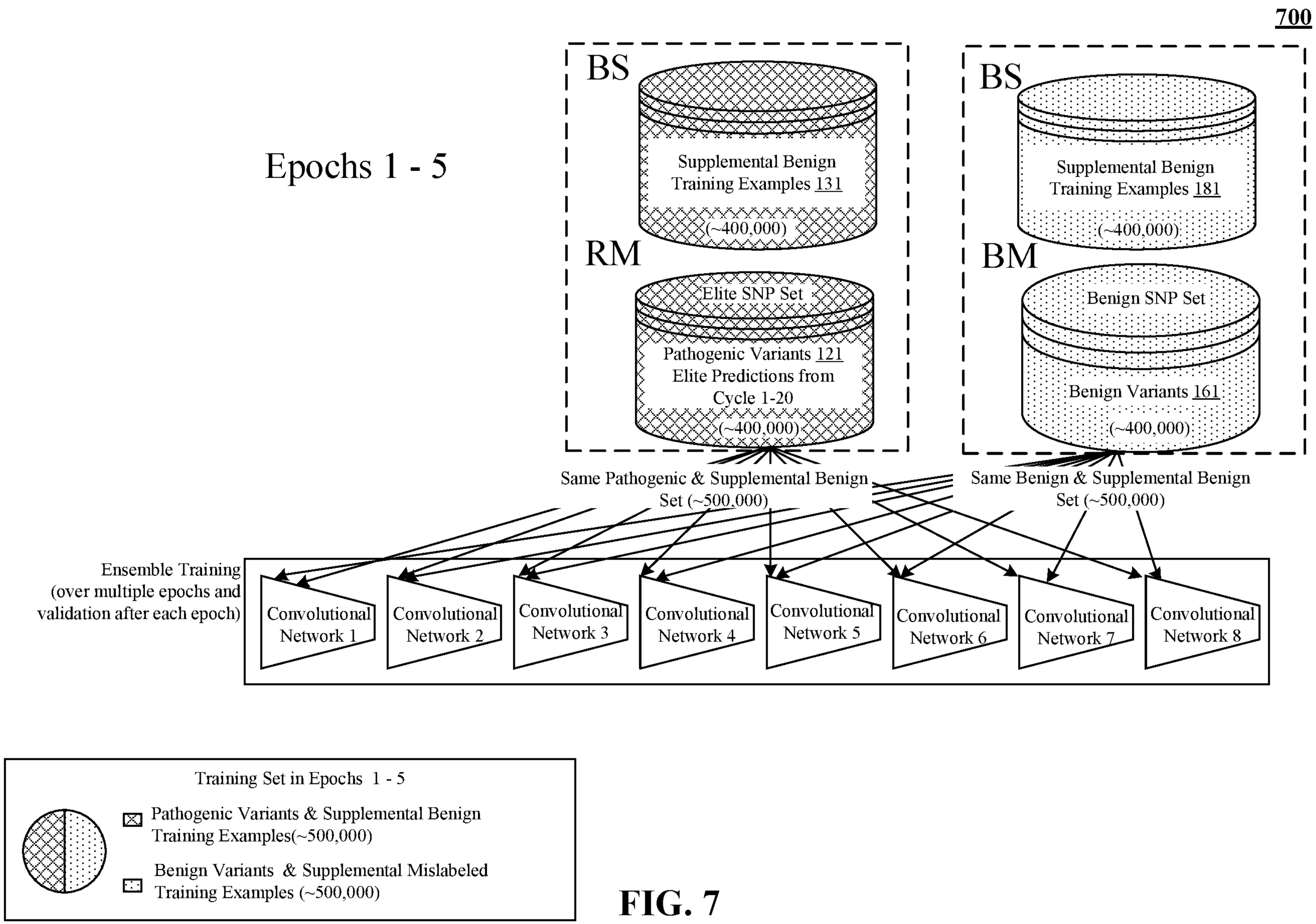

700
Epochs 1 - 5
BS
Supplemental Benign Training Examples 131
(~400,000)
RM
Elite SNP Set
Pathogenic Variants 121
Elite Predictions from Cycle 1-20
(~400,000)
BS
Supplemental Benign Training Examples 181
(~400,000)
BM
Benign SNP Set
Benign Variants 161
(~400,000)
Same Pathogenic & Supplemental Benign Set (~500,000)
Same Benign & Supplemental Benign Set (~500,000)
Ensemble Training (over multiple epochs and validation after each epoch)
Convolutional Network 1
Convolutional Network 2
Convolutional Network 3
Convolutional Network 4
Convolutional Network 5
Convolutional Network 6
Convolutional Network 7
Convolutional Network 8
Training Set in Epochs 1 - 5
Pathogenic Variants & Supplemental Benign Training Examples(~500,000)
Benign Variants & Supplemental Mislabeled Training Examples (~500,000)

FIG. 7

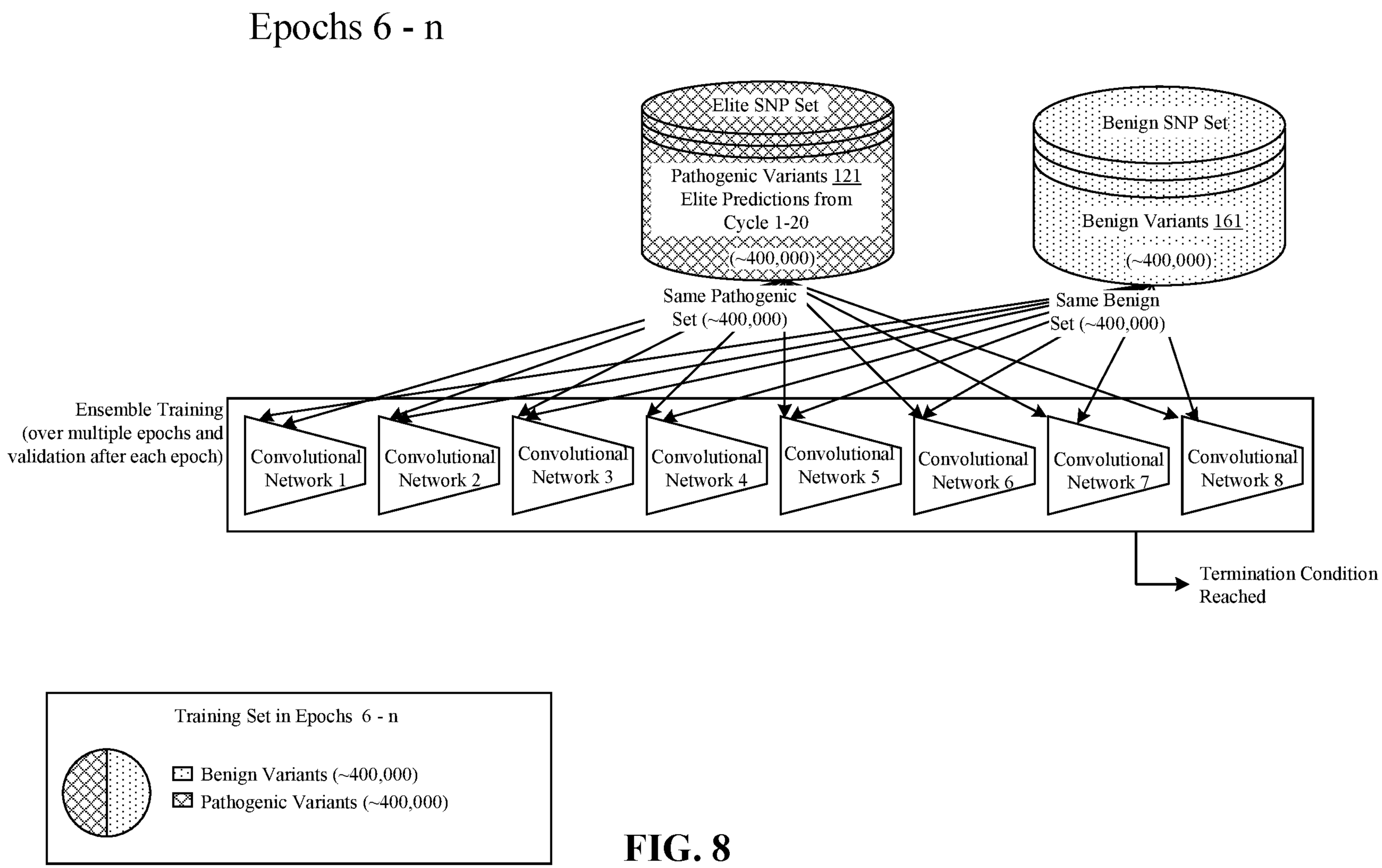

Epochs 6 - n
Elite SNP Set
Pathogenic Variants 121
Elite Predictions from
Cycle 1-20
(~400,000)
Benign SNP Set
Benign Variants 161
(~400,000)
Same Pathogenic
Set (~400,000)
Same Benign
Set (~400,000)
Ensemble Training
(over multiple epochs and
validation after each epoch)
Convolutional Network 1
Convolutional Network 2
Convolutional Network 3
Convolutional Network 4
Convolutional Network 5
Convolutional Network 6
Convolutional Network 7
Convolutional Network 8
Termination Condition
Reached
Training Set in Epochs 6 - n
Benign Variants (~400,000)
Pathogenic Variants (~400,000)

FIG. 8

900

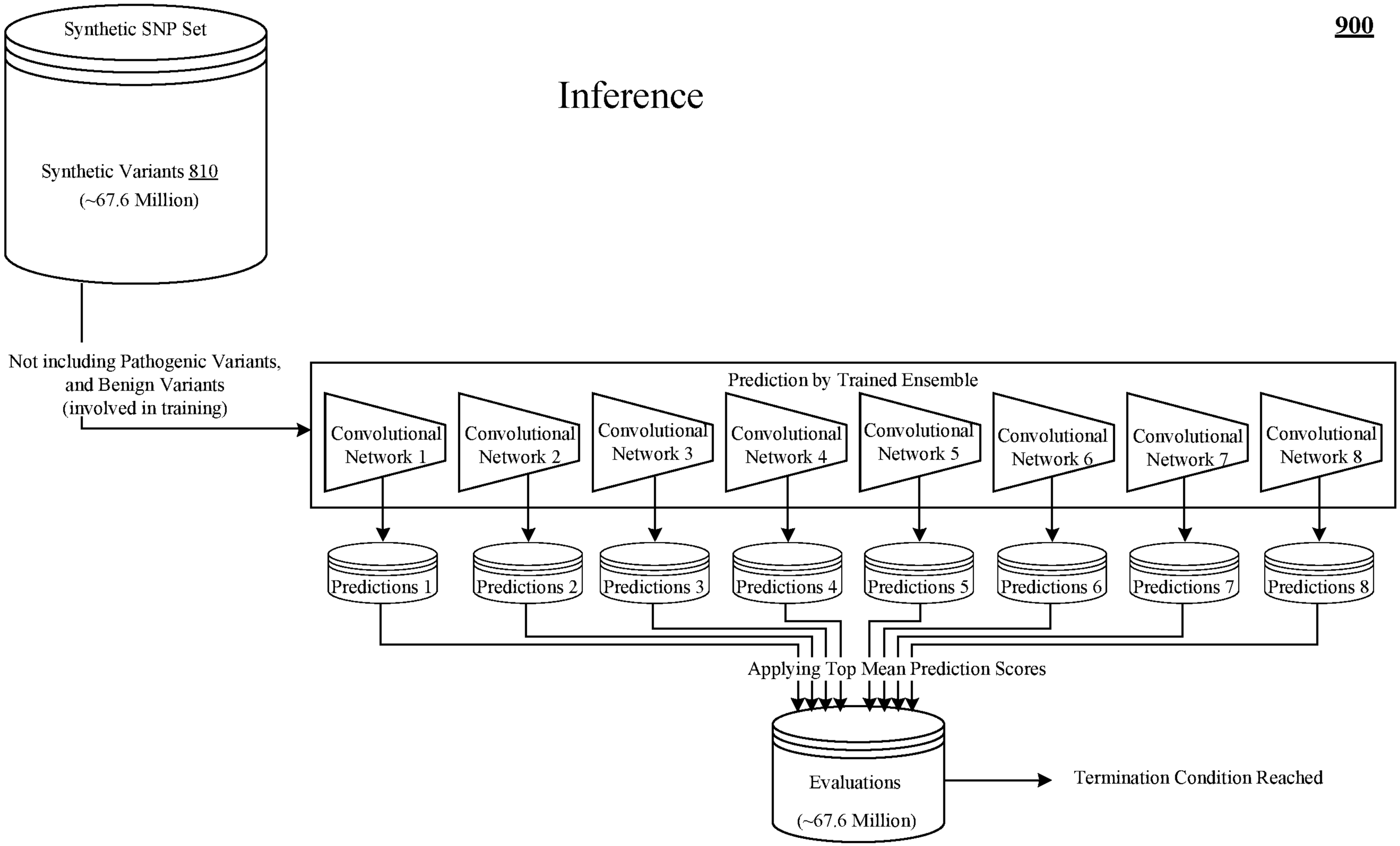
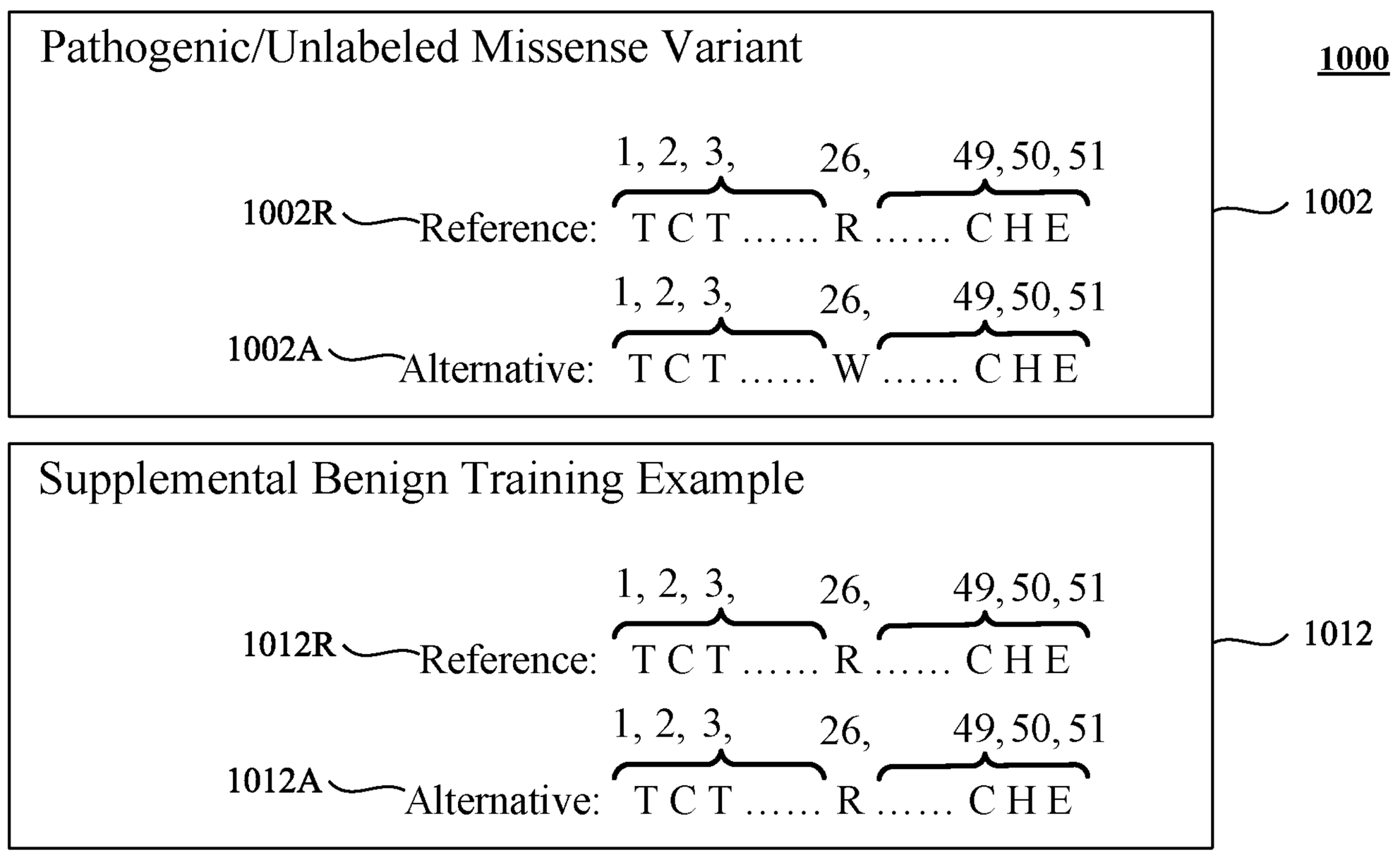

Inference
Synthetic SNP Set
Synthetic Variants 810
(~67.6 Million)
Not including Pathogenic Variants, and Benign Variants (involved in training)
Prediction by Trained Ensemble
Convolutional Network 1
Convolutional Network 2
Convolutional Network 3
Convolutional Network 4
Convolutional Network 5
Convolutional Network 6
Convolutional Network 7
Convolutional Network 8
Predictions 1
Predictions 2
Predictions 3
Predictions 4
Predictions 5
Predictions 6
Predictions 7
Predictions 8
Applying Top Mean Prediction Scores
Evaluations
(~67.6 Million)
Termination Condition Reached

FIG. 9


Pathogenic/Unlabeled Missense Variant
1000
1, 2, 3, 26, 49, 50, 51
1002R
Reference: T C T …… R …… C H E
1002
1, 2, 3, 26, 49, 50, 51
1002A
Alternative: T C T …… W …… C H E
Supplemental Benign Training Example
1, 2, 3, 26, 49, 50, 51
1012R
Reference: T C T …… R …… C H E
1012
1, 2, 3, 26, 49, 50, 51
1012A
Alternative: T C T …… R …… C H E

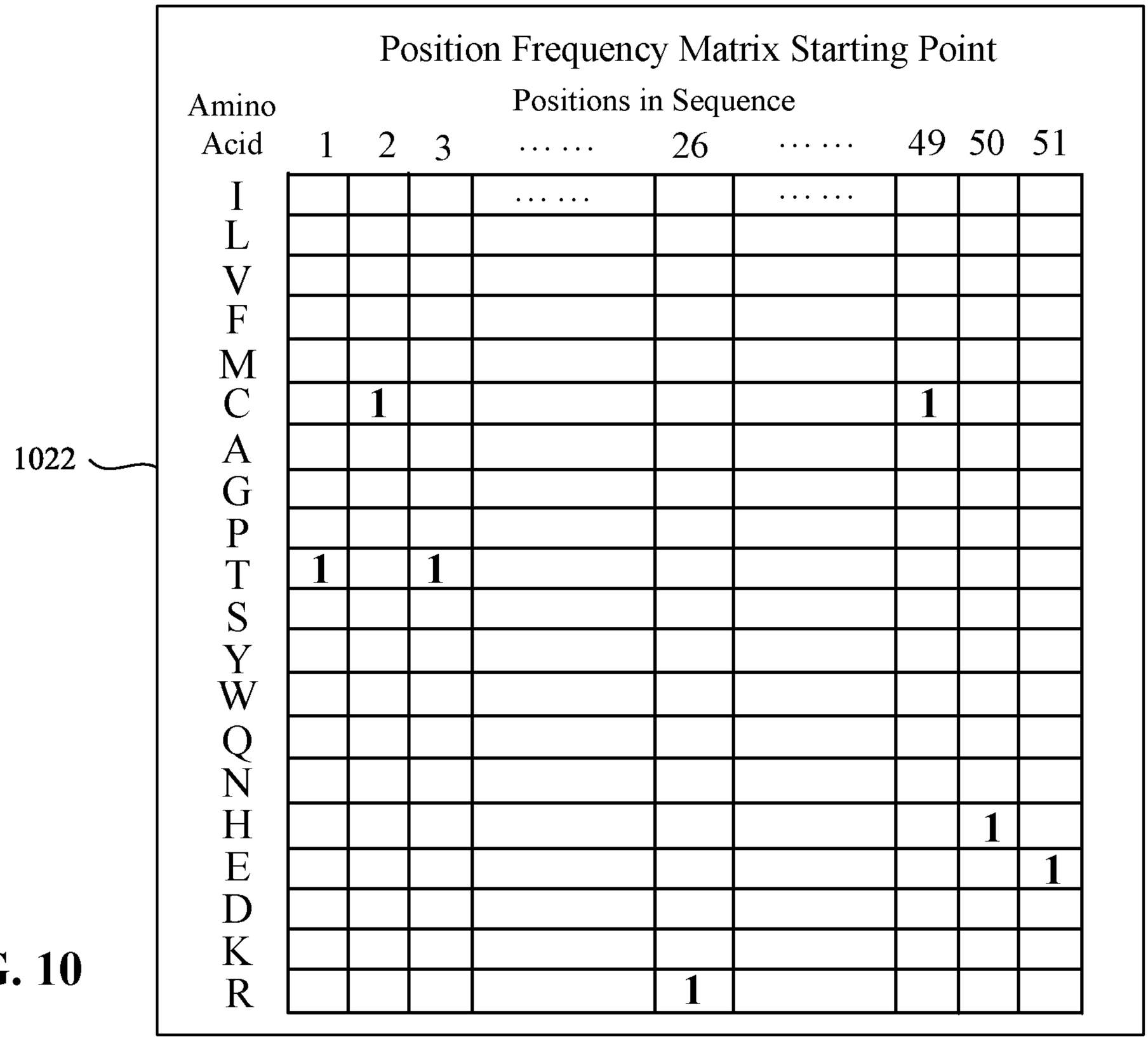

Position Frequency Matrix Starting Point
Amino Acid
Positions in Sequence
1 2 3 …… 26 …… 49 50 51
I
L
V
F
M
C
A
G
P
T
S
Y
W
Q
N
H
E
D
K
R
1022

FIG. 10

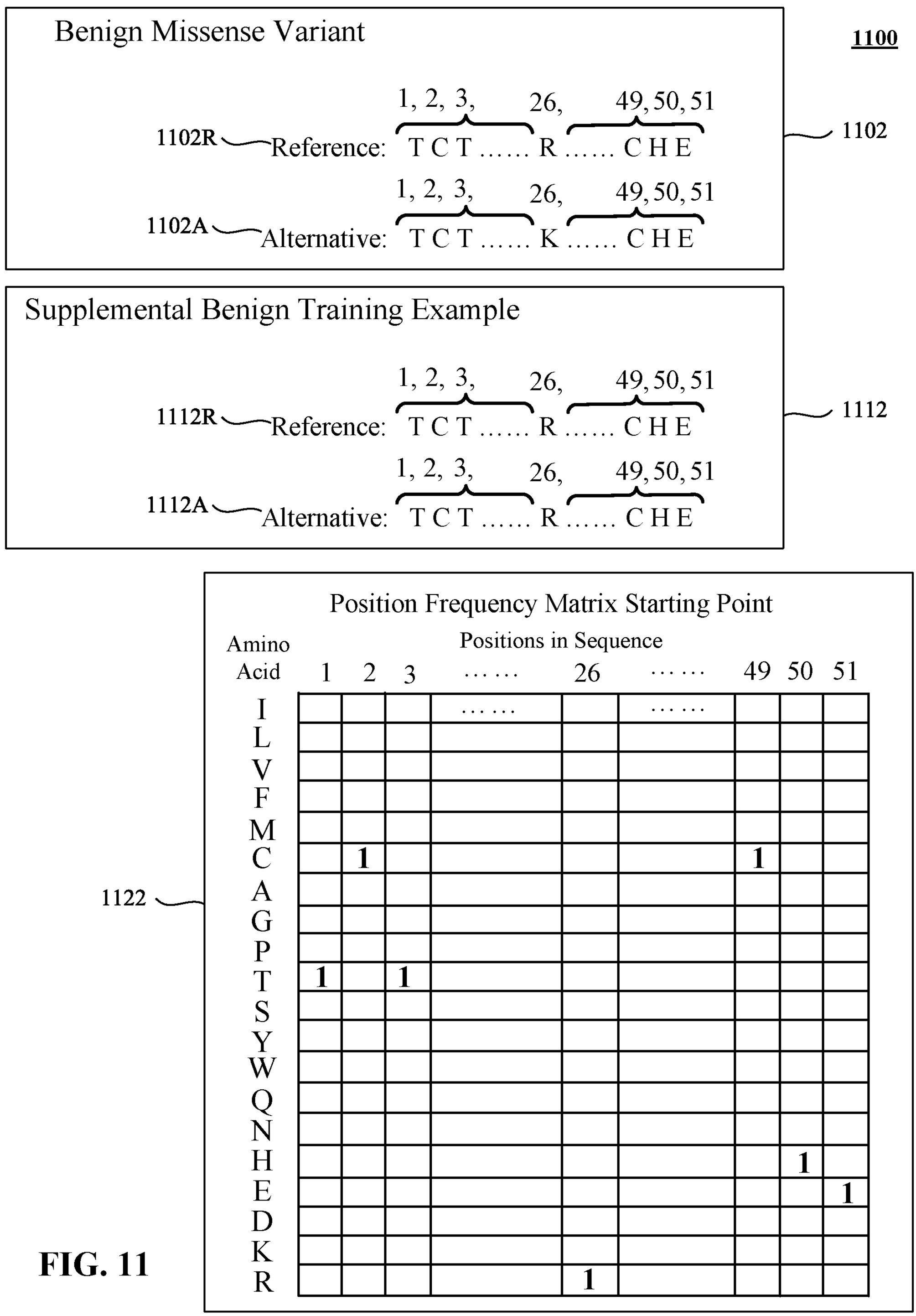

Benign Missense Variant
1100
1, 2, 3,
26,
49, 50, 51
1102R
Reference:
T C T …… R …… C H E
1102
1, 2, 3,
26,
49, 50, 51
1102A
Alternative:
T C T …… K …… C H E
Supplemental Benign Training Example
1, 2, 3,
26,
49, 50, 51
1112R
Reference:
T C T …… R …… C H E
1112
1, 2, 3,
26,
49, 50, 51
1112A
Alternative:
T C T …… R …… C H E
Position Frequency Matrix Starting Point
Amino Acid
Positions in Sequence
1 2 3 …… 26 …… 49 50 51
1122
I
L
V
F
M
C 1 1
A
G
P
T 1 1
S
Y
W
Q
N
H 1
E 1
D
K
R 1

FIG. 11

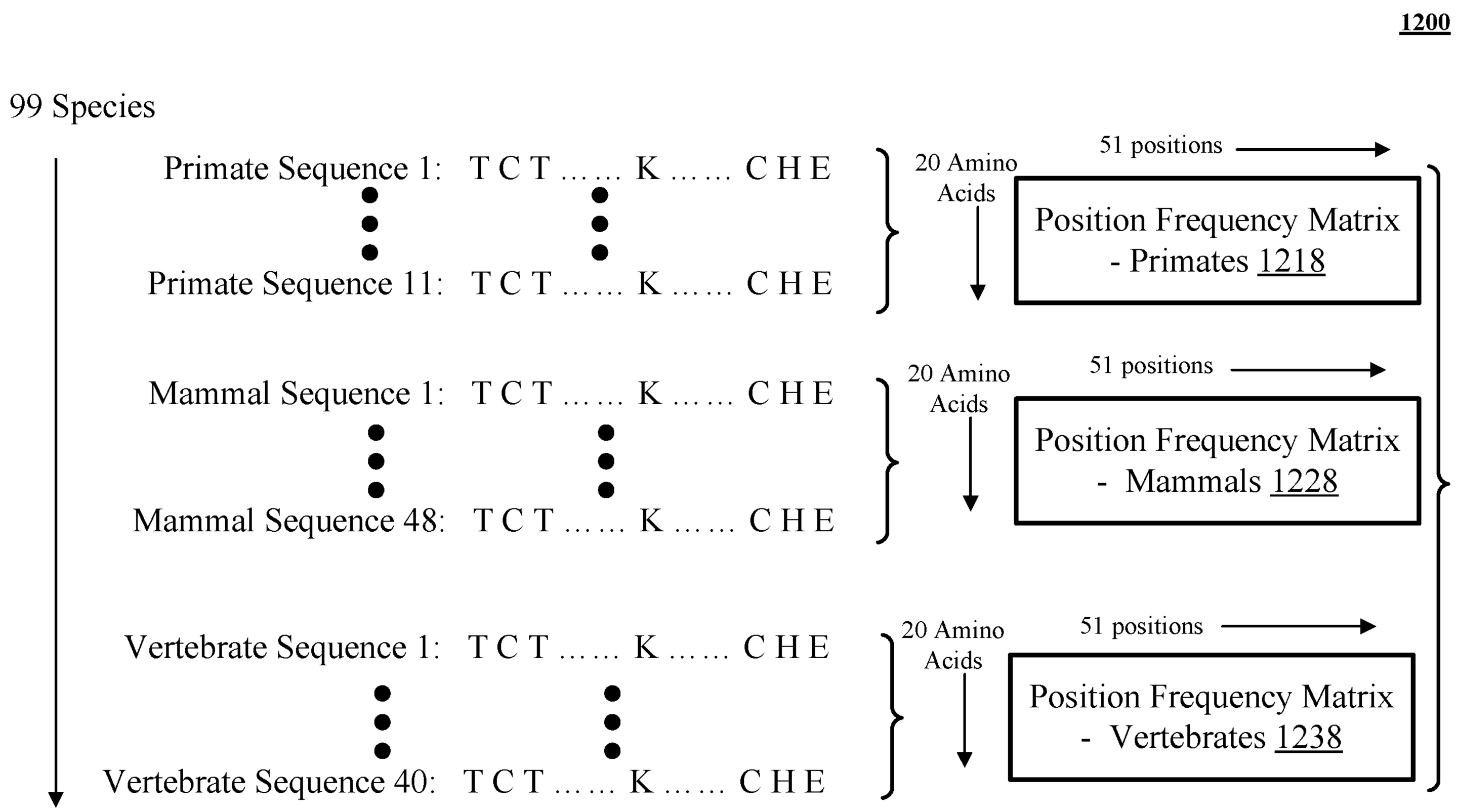

1200
99 Species
Primate Sequence 1: T C T …… K …… C H E
Primate Sequence 11: T C T …… K …… C H E
20 Amino Acids
51 positions
Position Frequency Matrix - Primates 1218
Mammal Sequence 1: T C T …… K …… C H E
Mammal Sequence 48: T C T …… K …… C H E
20 Amino Acids
51 positions
Position Frequency Matrix - Mammals 1228
Vertebrate Sequence 1: T C T …… K …… C H E
Vertebrate Sequence 40: T C T …… K …… C H E
20 Amino Acids
51 positions
Position Frequency Matrix - Vertebrates 1238

FIG. 12

Supplemental Benign Training Example

1300

Human Reference: $\overbrace{\text{T C T} \ldots\ldots}^{1,\,2,\,3,}$ $\overset{26,}{\text{R}}$ $\overbrace{\ldots\ldots\ \text{C H E}}^{49,\,50,\,51}$

Human Alternative: $\overbrace{\text{T C T} \ldots\ldots}^{1,\,2,\,3,}$ $\overset{26,}{\text{R}}$ $\overbrace{\ldots\ldots\ \text{C H E}}^{49,\,50,\,51}$

One Hot Encoded Human Reference

Amino Acid Encoding Row	Positions in Sequence 1	2	3	……	26	……	49	50	51
I	0	0	0	……	0	……	0	0	0
L	0	0	0		0		0	0	0
V	0	0	0		0		0	0	0
F	0	0	0		0		0	0	0
M	0	0	0		0		0	0	0
C	0	1	0		0		1	0	0
A	0	0	0		0		0	0	0
G	0	0	0		0		0	0	0
P	0	0	0		0		0	0	0
T	1	0	1		0		0	0	0
S	0	0	0		0		0	0	0
Y	0	0	0		0		0	0	0
W	0	0	0		1		0	0	0
Q	0	0	0		0		0	0	0
N	0	0	0		0		0	0	0
H	0	0	0		0		0	1	0
E	0	0	0		0		0	0	1
D	0	0	0		0		0	0	0
K	0	0	0		0		0	0	0
R	0	0	0	……	1	……	0	0	0

One Hot Encoded Human Alternate

Amino Acid Encoding Row	Positions in Sequence 1	2	3	……	26	……	49	50	51
I	0	0	0	……	0	……	0	0	0
L	0	0	0		0		0	0	0
V	0	0	0		0		0	0	0
F	0	0	0		0		0	0	0
M	0	0	0		0		0	0	0
C	0	1	0		0		1	0	0
A	0	0	0		0		0	0	0
G	0	0	0		0		0	0	0
P	0	0	0		0		0	0	0
T	1	0	1		0		0	0	0
S	0	0	0		0		0	0	0
Y	0	0	0		0		0	0	0
W	0	0	0		1		0	0	0
Q	0	0	0		0		0	0	0
N	0	0	0		0		0	0	0
H	0	0	0		0		0	1	0
E	0	0	0		0		0	0	1
D	0	0	0		0		0	0	0
K	0	0	0		0		0	0	0
R	0	0	0	……	1	……	0	0	0

FIG. 13

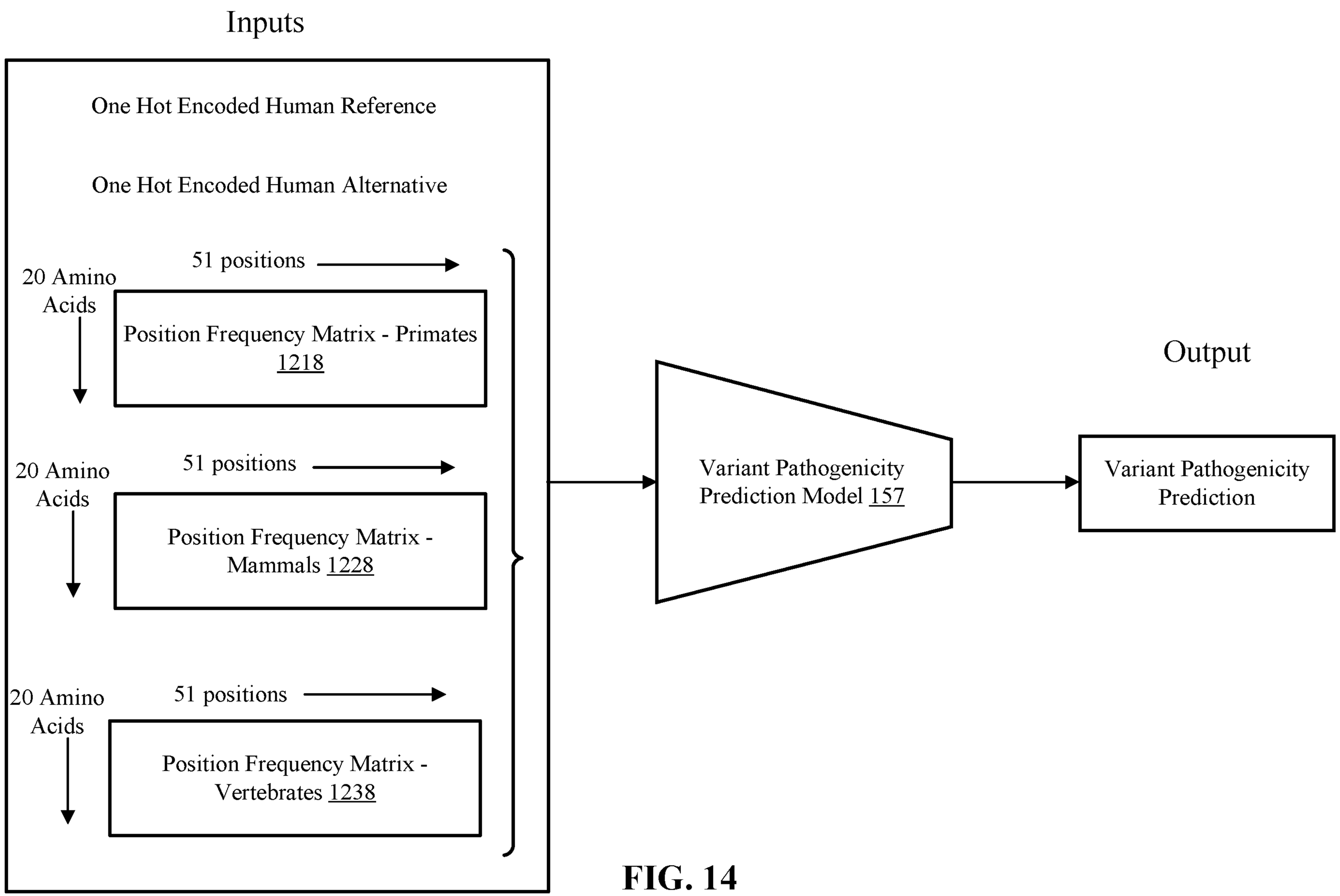

Inputs
One Hot Encoded Human Reference
One Hot Encoded Human Alternative
20 Amino Acids
51 positions
Position Frequency Matrix - Primates 1218
20 Amino Acids
51 positions
Position Frequency Matrix - Mammals 1228
20 Amino Acids
51 positions
Position Frequency Matrix - Vertebrates 1238
Variant Pathogenicity Prediction Model 157
Output
Variant Pathogenicity Prediction

FIG. 14

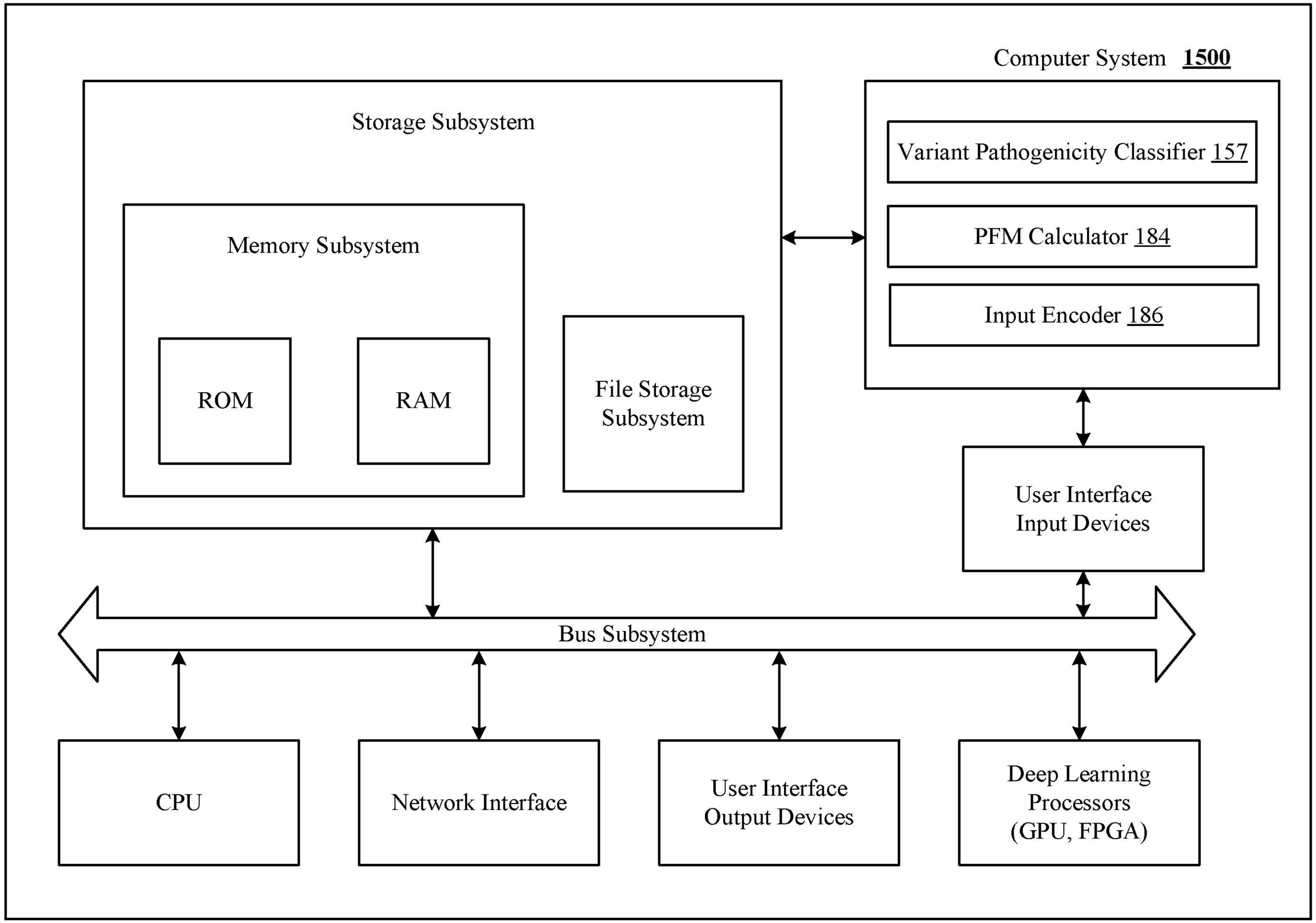

Computer System 1500
Storage Subsystem
Memory Subsystem
ROM
RAM
File Storage Subsystem
Variant Pathogenicity Classifier 157
PFM Calculator 184
Input Encoder 186
User Interface Input Devices
Bus Subsystem
CPU
Network Interface
User Interface Output Devices
Deep Learning Processors (GPU, FPGA)

FIG. 15